UNIVERSITY OF BAMBERG

Selection of models and variables: an automated Bayesian view for handling missing values

A thesis submitted in fulfillment of the requirements for the degree of

DOCTOR RERUM POLITICARUM (Dr. rer. pol.)

to the Faculty for Social Sciences, Economics, and Business Administration at the University of Bamberg

MICHAEL DAVID BERGRAB,
MASTER OF SCIENCE IN SURVEY-STATISTIK

Bamberg 2026

Diese Arbeit hat der Fakultät Sozial- und Wirtschaftswissenschaften der Otto-Friedrich-Universität Bamberg als Dissertation vorgelegen.

Erstgutachter: Prof. Dr. Christian Aßmann
Zweitgutachter: Prof. David Kaplan, PhD
Drittgutachterin: Prof. Dr. Anne Leucht

Tag der Disputation: 14.01.2026

URN: urn:nbn:de:bvb:473-irb-113953x
DOI: https://doi.org/10.20378/irb-113953

Contents

List of Figures

List of Tables

List of Abbreviations

AIC	**A**kaike **I**nformation **C**riterion
BART	**B**ayesian **A**dditive **R**egression **T**rees
BIC	**B**ayesian **I**nformation **C**riterion
CART	**C**lassification **A**nd sequential **R**egression **T**rees
CDF	**C**umulative **D**istribution **F**unction
DAG	**D**irected **A**cyclic **G**raph
EM	**E**xpectation **M**aximization
FCD	**F**ull **C**onditional **D**istribution
GLM	**G**eneralized **L**inear **M**odel
LDA	**L**inear **D**iscriminant **A**nalysis
LRT	**L**ikelihood **R**atio **T**est
MAR	**M**issing **A**t **R**andom
MCAR	**M**issing **C**omplete **A**t **R**andom
MCMC	**M**arkov **C**hain **M**onte **C**arlo
MI	**M**ultiple **I**mputation
MICE	**M**ultiple **I**mputation via **C**hained **E**quations
ML	**M**aximum **L**ikelihood
MLE	**M**aximum **L**ikelihood **E**stimation
MNAR	**M**issing **N**ot **A**t **R**andom
MSE	**M**ean **S**quared **E**rror
NEPS	**N**ational **E**ducational **P**anel **S**tudy
RMSE	**R**oot **M**ean **S**quared **E**rror
ROC	**R**eceiver **O**perating **C**haracteristic

Chapter 1

Introduction

1.1 Bayesian inquiry and missing data: bridging Popperian gaps

In the epistemological landscape, Karl Popper's theory of falsifiability serves as a benchmark that encourages scientists to develop theories that are empirically testable and, above all, falsifiable (Popper, 2002). Popper's philosophy stands in contrast to inductivism and emphasizes that no amount of positive instances can definitively prove a scientific theory. Rather, the strength of a theory lies in the fact that it can withstand rigorous attempts at falsification. This sets a high standard for scientific theories and encourages a climate of constant refinement and improvement. In the empirical realm, the problem of missing data proves to be a formidable challenge, adding complexity to the structure of statistical inference (Little & Rubin, 2002). Tension arises when the requirements of falsifiability meet the practical reality of incomplete observational data. Popper's principles demand precision, but incomplete datasets are often elusive, leading to an existential dilemma for empirical investigations (Godfrey-Smith, 2008). This discrepancy drives to Bayesian methodology - an approach that not only accounts for the uncertainty caused by missing data, but understands it as an inherent part of the scientific narrative (Gelman et al., 2023). In the Bayesian view, probability is a measure of belief or uncertainty, recognizing that even in the absence of complete information, hypotheses can be assigned a reasonable degree of belief.

This view is consistent with Popper's emphasis on empirical falsifiability, since Bayesian methods allow to update beliefs in response to new evidence and thereby refine one's understanding of the empirical world (Jaynes, 2010). In the area of statistical modeling, the challenges of Popper's falsifiability manifest themselves in the problem of model and variable selection. Popper's criterion requires that a theory or model be specific enough to make falsifiable predictions, but not so specific that it cannot be tested empirically. The tension between model complexity and simplicity reflects the delicate balance that statisticians strive for when selecting variables and formulating models. Uncertainty is a central concept in statistical modeling as well as in machine learning, as practitioners often work with incomplete information or models that simplify reality. Popper (2002) asserts that empirical falsifiability is crucial, meaning a model must make predictions that can be tested against empirical evidence. This presents a challenge where the model must be specific enough to make falsifiable predictions but flexible enough to accommodate real-world variability. Bayesian methods provide a formal framework for incorporating uncertainty into model

parameters and model selection. By updating beliefs in response to new data, Bayesian approaches directly address the inherent uncertainty associated with the model's structure. This contrasts with traditional frequentist approaches, which often focus on point estimates and hypothesis testing without explicitly accounting for model uncertainty. Bayesian model averaging and selection methods, among others, offer a principled way to resolve this dilemma by incorporating uncertainty about model structure into the inference process (Hoeting et al., 1999).

The uncertainty in modeling arises from the fact that processes take place in a framework, e.g., in nature, whose input x and output y are known, observable and visible, and can also be made measurable, but the underlying process that governs the relationship between them is hidden. This black box represents the unknown or unobservable mechanism that connects the inputs to the outputs. However, in many frameworks, the underlying black box that controls the process is always closed and the greater the degree of closure, the more significant the uncertainty, not only in the measurement itself but also in the models that are used to interpret it. Beyond the observable data and the hidden mechanisms connecting inputs and outputs, a key methodological step in both Bayesian and classical statistics is the formulation of distributional assumptions. These assumptions are not mere technicalities but essential epistemological commitments that allow us to bridge the gap between theory and data. They guide the specification of the likelihood, define priors, and ultimately shape the inference process. In this sense, distributional assumptions represent an attempt to give structure to uncertainty and facilitate the extraction of knowledge from incomplete or noisy data. Hence, Breiman (2001) emphasized two perspectives in modeling:

- **Prediction**: The goal here is to make accurate predictions based on the relationship between x and y, without necessarily needing to explain the internal workings of the black box. This approach is commonly seen in machine learning, where predictive accuracy is prioritized over interpretability.
- **Information**: The alternative approach focuses on understanding and explaining the relationship between input, output, and the black box. Here, the goal is to derive insights about the underlying process that generates the data, rather than purely focusing on predictive performance.

In consequence, the dataset D represents the observable measurements of inputs and outputs. In statistical terms, the data provides evidence from which the parameters θ that define the model can be inferred or estimated. The parameter θ represents a statistical quantity that summarizes the relationship between the inputs and outputs of a given model. It can be conceived of as the coefficients of a regression model, the weights in a neural network, or other statistical parameters that encapsulate the relationship between the inputs and outputs. In Bayesian methodology, the parameter θ is treated as a random variable with its own distribution, reflecting the uncertainty about its true value (Jaynes, 2010). The process of inference is concerned with the estimation of the parameter vector, represented by θ, using the dataset D in a manner that is as accurate as possible, while accounting for the inherent uncertainty associated with both the data and the model itself.

The process of inference and parameter estimation is inherently more challenging when dealing with missing values in data, as the additional uncertainty introduced by incomplete data makes it more difficult to make accurate inferences and estimate the relevant parameters or prediction (Schafer, 1997; Schafer & Graham, 2002). However, missing values can arise for various reasons, such as technical issues in data collection or non-responses in surveys. The presence of missing values leads to gaps in the observed data, which complicates the task of estimating the parameters accurately. The various types of missing data require different approaches to their resolution, with the specific method depending on the nature of the missingness. There are a number of techniques that may be employed to address the issue of incomplete or absent data, see among other Gelman et al. (2023) and Little and Rubin (2002). Imputation is a strategy whereby missing values are replaced with estimated values. This may entail the application of relatively straightforward methods, such as the replacement of missing values with the observed mean, median, or mode. However, more complex techniques, such as multiple imputation, involve the generation of several plausible imputed datasets and the averaging of these results in order to account for the inherent uncertainty associated with incomplete data. Alternatively, predictive models, including those based on regression analysis, can be utilized to predict and fill in missing data points based on observed variables. An alternative approach is to utilize likelihood-based methodologies that operate directly with the available data, see among other Enders (2022). These methods estimate the model parameters using only the complete cases, without attempting to fill in missing values. This approach is frequently applicable when the data exhibits missing at random (MAR) characteristics and adjustments are made to account for the missingness mechanism. In Bayesian methodology, missing values are treated as additional unknowns, and the inference process integrates over the missing data using a probability model. This allows Bayesian approaches to account for both the uncertainty in the missing values and the uncertainty in the parameter estimates. Techniques such Markov chain Monte Carlo (MCMC), see Geweke (1989), or especially Gibbs sampling, see Gelfand and Smith (1990) and Geman and Geman (1984), are commonly employed to sample from the joint distribution of the parameters and missing data. The impact of estimating parameters θ with missing data on the subsequent analysis is contingent upon the approach employed. Following Little and Rubin (2002), in the frequentist framework, the presence of missing data reduces the effective sample size, which may result in increased variability or bias in parameter estimates if not handled appropriately. Furthermore, the reliability of standard errors and confidence intervals may be compromised. In contrast, the Bayesian framework explicitly models the uncertainty introduced by missing data, and the posterior distribution of θ reflects both the uncertainty in the parameters and the missing data. Consequently, Bayesian methods often provide a more nuanced approach to missing data, particularly when the missingness is non-trivial. The process of inference with missing data follows a similar path to that of inference with complete data; however, additional steps are required to manage the missingness. In the case of complete data, the objective is to estimate the parameters θ using the full dataset, frequently through techniques such as Maximum Likelihood Estimation (MLE) or Bayesian inference. However, when missing values are present, the model must incorporate the uncertainty derived from the

missingness in addition to estimating the relationships between x and y. This renders the estimation of θ more intricate, as it necessitates either the imputation of missing values or the adjustment of the model to account for the missing data. In conclusion, the presence of missing values introduces a layer of complexity into the inference process, as they introduce uncertainty and the potential for bias in parameter estimates. The use of methods such as imputation, likelihood-based approaches, or Bayesian inference allows statisticians and machine learning practitioners to obtain more accurate and reliable estimates of θ. It is therefore crucial for researchers to properly handle missing data in order to ensure that the conclusions drawn from the analysis remain valid, even when faced with incomplete information, which is a common challenge in real-world data analysis.

The uncertainty is also reflected in the trade-off between variance and bias, which thus becomes a central consideration in statistical modeling in the sense of Popper's falsifiability. From an epistemological standpoint, distributional assumptions can be viewed as structured hypotheses about the world. In the spirit of Popper, they are potentially falsifiable through empirical evidence. Yet, in Bayesian terms, they are also vehicles of subjective belief that evolve with data. This dual role—as falsifiable assumptions and as carriers of belief—embodies the dynamic interplay between empirical contradiction and probabilistic reasoning. Hence, not only should a model fit the observed data well, but it should also be generalizable to new, unseen data. The trade-off is to balance the bias introduced by the simplification of the model against the variance introduced by the complexity of the model. Bayesian methods, with their ability to seamlessly integrate prior information, help to mitigate this trade-off by allowing for flexible model structures while accounting for uncertainty in parameter estimates. Addressing the intricacies of Bayesian estimation with missing values goes beyond conventional statistical paradigms (Gelman et al., 2023). This includes not only the technical challenges of imputation and estimation, but also broader epistemological considerations that reconcile the demands of empirical rigor with the inherent uncertainty of incomplete datasets, and thus advance scientific understanding in a way that is consistent with Popperian principles as well as the probabilistic foundations of Bayesian inference and the nuanced subtleties of statistical model selection. Here, a synthesis of Popper's falsifiability, Bayesian principles, and statistical modeling challenges is formed to shape a deeper understanding of the empirical scenery, highlighting the iterative nature of scientific inquiry and the dynamic interplay between empirical evidence, falsification, and probabilistic reasoning.[1]

Ultimately, the conscious formulation of distributional assumptions is not only a methodological necessity but also an epistemological stance—transforming vague uncertainties into structured, testable components of scientific inquiry.

[1] Furthermore, proponents of Bayesian epistemology criticize the binary nature of falsifiability and argue for a more probabilistic and incremental approach to scientific knowledge (Howson & Urbach, 2006). In navigating the complexities of modern philosophy, the tension between the falsifiability criterion and the evolving understanding of scientific enquiry emphasizes the dynamic nature of epistemological debates.

1.2 Bayesian estimation: an overview

The core of Bayesian estimation is Bayes' theorem, which provides a framework for updating beliefs about a parameter (or more parameters) based on new data (Box & Tiao, 1992; Gelman et al., 2023; Jackman, 2009). The general form of Bayes' theorem is:

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} = \frac{P(D|\theta) \cdot P(\theta)}{\int P(D|\theta)P(\theta)d\theta} \tag{1.1}$$

with

- $P(\theta|D)$ is the posterior distribution of the parameter(s) θ given the observed data D,
- $P(D|\theta)$ is the likelihood of the data given the parameter(s),
- $P(\theta)$ is the prior distribution of the parameter(s), representing our beliefs about the parameter(s) before observing the data, and
- $P(D)$ is the marginal likelihood, also known as the evidence, which is a normalization constant ensuring that the posterior distribution integrates to 1.

The goal of Bayesian estimation is to update prior beliefs ($P(\theta)$) in light of new data ($P(D|\theta)$) to obtain the posterior distribution ($P(\theta|D)$). θ represents the parameter(s) of the model which could be anything from the mean and variance of a distribution to the coefficients in a regression model. This posterior distribution captures the uncertainty about the parameters given the data which are observed. In Bayesian estimation, the focus is on the joint posterior distribution of all parameters, which can only be identified analytically in simple textbook cases. However, obtaining samples directly from the joint posterior can be computationally challenging, especially in high-dimensional spheres. This problem can be solved by applying Markov Chain Monte Carlo (MCMC) techniques. One such MCMC algorithm is the Gibbs sampler which samples from the posterior distribution iteratively (Gelfand & Smith, 1990; Geman & Geman, 1984). The key insight is the use of the conditional distributions of the individual parameters as a function of the values of the other parameters, which is essentially a cut through the joint posterior distribution. The Gibbs sampler takes advantage of this by iteratively sampling from these conditional distributions.

The Gibbs sampler is shown schematically:

1. **Initialization**: Start with initial values for the parameter(s) θ.
2. **Iteration**: In each iteration of the loop:
    - Update each parameter block by sampling from its conditional distribution given the initial or updated values of all other parameter blocks.
    - Repeat for all parameter blocks.
3. **Repeat**: Repeat the iteration process for a sufficient number of steps until a point where the samples seem to be stabilized or converged.

Supposing for a distinct model M different parameter blocks, $f(\theta_1, \theta_2, \ldots, \theta_p, \ldots, \theta_P|D)$ for $p = 1, \ldots, P$, the initialization step requires starting values for the set $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_P^{(0)})$. For each parameter θ_i in the set of parameters θ, the full conditional distributions $P(\theta_p|\theta_{-p}, D)$ are identified and the specified. This distribution represents the probability distribution of θ_p given the current values of all other parameters θ_{-p} and the data D. If analytical expressions are available, the analytical form of the full conditional distributions is derived using Bayesian inference principles, such as conjugacy properties or conditional independence assumptions. This involves expressing each full conditional distribution in terms of the likelihood function $P(D|\theta)$ and the prior distribution $P(\theta)$. If analytical expressions are not available, numerical methods or simulation-based techniques are used to approximate the full conditional distributions. To sample from the full conditional distributions MCMC methods, including Gibbs sampling, Metropolis-Hastings or slice sampling, can be implemented, which lead to a more manageable and computationally efficient way. For each parameter θ_p, the current value is sampled given the initial or updates values of all other parameters θ_{-p} and the data D. The sampling process is iterated until a stopping criterion (i.e., convergence) and the values of the parameters θ are updated based on the sampled values from their full conditional distributions so that the sampled values are stabilized around the posterior distribution of the parameters given the observed data (Gelman et al., 2023; Robert & Casella, 2004).

For $r = 1, \ldots, R$ iterations the Gibbs sampler draws from the full conditional posterior distributions of the respective parameter blocks:

1. Draw $\theta_1^{(r)}$ from $f(\theta_1|\theta_2^{(r-1)}, \theta_3^{(r-1)}, \ldots, \theta_P^{(r-1)}, D)$
2. Draw $\theta_2^{(r)}$ from $f(\theta_2|\theta_1^{(r)}, \theta_3^{(r-1)}, \ldots, \theta_P^{(r-1)}, D)$

   $\vdots$

P. Draw $\theta_P^{(r)}$ from $f(\theta_P|\theta_1^{(r)}, \theta_2^{(r)}, \ldots, \theta_{P-1}^{(r)}, D)$

For this purpose, the number of iterations R must be chosen large enough so that after an appropriate and sufficient burn-in phase and after checking whether the Gibbs sampler converges to the target density, it can be assumed that the joint posterior distribution of the parameter blocks can be approximated by their sampled empirical distributions. Estimators of interest given as moments can be assessed in terms of their sample counterparts. MCMC methods, such as Gibbs sampling, are powerful tools for Bayesian inference and probabilistic modeling, and a practical strategy for navigating in high-dimensional parameter spaces and updating (prior) beliefs in the Bayesian framework. They provide a flexible framework for estimating complex probability distributions and making inferences about model parameters. However, like any statistical method, there are considerations and regulatory conditions that researchers should be aware of when using MCMC methods in practice, see Roberts and Smith (1994) and Chib (2001).

The full conditional distributions are important in this context because of (i) simplicity of sampling, (ii) convergence and mixing, (iii) computational efficiency, (iv) flexibility in model specification, and (v) implementation of block Gibbs sampling. In many cases the

joint posterior distribution of parameters may be complex and difficult to sample directly (simplicity of sampling). However, the conditional distributions of individual parameters given the values of the others can be simpler and more amenable to sampling. The Gibbs sampler exploits this by sampling from these simpler conditional distributions iteratively. The Gibbs sampler relies on updating one parameter at a time, and the convergence of the algorithm depends on the properties of the full conditional distributions (Convergence and mixing). If the full conditional distributions are easily sampled and mix well, then the Gibbs sampler is likely to converge more efficiently. Sampling from the full conditional distributions can be computationally more efficient than sampling directly from the joint posterior distribution (computational efficiency). This is especially true when the conditional distributions have a simpler form or when analytical solutions are available. When building complex Bayesian models, specifying the joint distribution of all parameters may be impractical (flexibility in model specification). The Gibbs sampler allows to build and estimate models in a modular way. In some cases, it may be possible to group parameters into blocks, and the full conditional distributions can be sampled for each block, making the Gibbs sampler computationally efficient. This is known as block Gibbs sampling.

In summary, full conditional distributions are essential in Bayesian estimation using the Gibbs sampler because they provide a practical and computationally efficient way to update parameters in a sequential manner, leading to convergence to the joint posterior distribution of interest.

1.3 Handling missing values with data augmentation

Incomplete datasets are a very serious challenge in statistics and in other disciplines, which influences the inference negatively. The primary aim of research is to draw conclusions about the population, not on a sample that is not representative of the population. Missing data is common in many datasets and cannot be avoided in data-driven research, even with the best design and data collection plan. The effect of ignoring incomplete data is that it can reduce the sample size, statistical power, and thus information, thereby increasing the standard errors of the estimates and increasing the estimation bias, especially if the estimation bias is high (Little & Rubin, 2002; Rubin, 1976). For an overview in education research, see Peugh and Enders (2004). To address missing data, most researchers use imputation techniques to handle missing values in the data. The advantage of using imputation techniques is that the full sample size is used, making the results less biased and more accurate. The analysis of incomplete datasets is directly related to the problem of how missingness in the data affects inference and prediction on the data sets. In statistical software, such as SPSS, SAS, or STATA, the traditional way of handling missing values is to drop them from the analysis by default. Reviewing the literature on missing data, several methods have been developed to deal with missing data, see among others Cheema (2014), Dong and Peng (2013), Little and Rubin (2002), and Pigott (2001). There is no universally agreed-upon threshold for the proportion of missing data that can be considered unproblematic. Cohen et al. (2013) suggests a range of 5% to 10% as a rough guideline. However, the impact of missing data also depends on

the research field and the specific variables involved. In addition to the missing rate, the strength of the relationship between missing and observed variables is crucial in determining the potential bias (Heymans & Twisk, 2022).

The focus of this thesis is on data augmentation wrapped in Bayesian estimation to complete the observed data with missing data (Tanner & Wong, 1987). As a Markov Chain Monte Carlo (MCMC) approach, data augmentation improves the quality of the data by adding the missing data to the model, which allows for both: handling the missing data and estimating the model parameters from their posterior full conditional distributions. According to Tanner and Wong (1987), the basic idea behind data augmentation is to improve the quality of inference by augmenting the observed data D_{obs} with the missing data D_{mis} so that it can be sampled from the expanded posterior distribution $p(\theta|D = [D_{obs}, D_{mis}])$.

It is also necessary to consider which failure mechanisms are involved in the missing values, as these have an influence on the inference. For an overview, see Little and Rubin (2002) and Xu (2022). The missing data mechanism influences the potential for bias, although the impact of missing values on research conclusions depends on various factors. Beyond the pattern of missingness, the research objectives and also the reporting transparency are among those factors. Three fundamental mechanisms can be distinguished: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR) (Gelman et al., 2023; Little & Rubin, 2002; Rubin, 1976). For the sake of clarity, the variables X and U are distinguished more precisely in the context of the mechanisms of missing values:

- X represents the observed variables or features in the dataset. These are the measurable characteristics or attributes that are available for analysis. For instance, in a dataset measuring the educational progress of individuals, X could include variables such as age, gender, competence test-scores, or school test grades.
- U denotes the unobserved variables or features that are not included in the dataset but may influence the missingness of data. These variables are typically latent or hidden factors that are not directly measured or observed. In the context of the educational outcomes a dataset U could represent factors like motivation, parental engagement, mental health, or teacher effects, which are not captured in the dataset but may affect the likelihood of missing data.

Understanding the interplay between observed and unobserved variables is essential for annotating the missing data mechanisms and devising appropriate strategies for handling missing data in statistical analysis.

- **Missing Completely At Random (MCAR)**: In MCAR scenarios, the probability of missingness remains independent of both observed and unobserved data. Mathematically, this is denoted as $P(Missing|X, U) = P(Missing)$, where missingness is unrelated to any variables. Consequently, the missing data mechanism is entirely stochastic, devoid of any systematic pattern.
- **Missing At Random (MAR)**: Contrary to MCAR, MAR acknowledges a dependence of missingness on observed data while remaining unrelated to unobserved data. This

is expressed as $P(Missing|X,U) = P(Missing|X)$, indicating that missingness can be explained by observed variables. However, once these variables are considered, missingness follows a random pattern, independent of unobserved factors.

- **Missing Not At Random (MNAR)**: MNAR reflects situations where missingness is influenced by unobserved data, persisting even after accounting for observed variables. Formally, $P(Missing|X,U) \neq P(Missing|X)$, signifying a non-random missingness pattern driven by unobserved factors. In such cases, missing data poses a substantial challenge, as its mechanism cannot be fully characterized by observed variables alone.

Understanding these mechanisms is paramount in statistical analysis, guiding the selection of appropriate imputation methods and modeling strategies. Ignoring the underlying missing data mechanism can lead to biased estimates and erroneous conclusions, underscoring the necessity of robust handling techniques in data analysis.

The challenge of handling missing data arises from the uncertainty surrounding the mechanism of missingness, particularly whether it conforms to the MAR assumption or is contingent upon unobserved predictors. While the MAR assumption posits that the probability of missingness is solely influenced by observed variables, it is often impractical to ascertain with certainty. The inherent difficulty lies in the presence of unobservable variables whose influence on missingness cannot be directly measured. Consequently, researchers typically resort to making assumptions or drawing upon external evidence, such as data from surveys with comprehensive follow-up procedures, to gauge the plausibility of the MAR assumption.

In practice, efforts are made to include as many potential predictors as feasible in the analysis to bolster the credibility of the MAR assumption. By incorporating a comprehensive set of predictors, the assumption that missingness is unrelated to unobserved variables gains greater credence. For instance, while it may be ambitious to assert that nonresponse to a particular question, such as earnings, is exclusively determined by observable characteristics like gender, race, and education, this assumption appears more tenable compared to scenarios where the probability of nonresponse remains constant or relies on a single predictor.

1.4 Modeling and estimating a binary dependent variable

In the field of statistical modeling, the selection of an optimal link function to relate predictor variables to a binary outcome is a critical undertaking. This decision is crucial because it profoundly affects the interpretability of the resulting model and the underlying assumptions about the distribution of the data. In order to make an informed decision among the available options, scientists must balance various factors and considerations.

Supposing a response variable y with binary outcomes observed as 1 or 0, and X representing the corresponding covariates, the data D contains both: $D = (y, X)$ where missings in X are included. Y may represent the decision of an individual i, e.g., success or failure of some device, participation on the labor market, transferring to a special school type or not, and X can contain as many variables with any number of characteristics. Supposing N

individuals making a decision, a $N \times P$ matrix describes the corresponding covariates of the binary dependent variable, called X.

$$y_i = \begin{cases} 1 & \text{if} \quad y_i^* = \alpha + X_i\beta + e_i > 0, \\ 0 & \text{if} \quad y_i^* = \alpha + X_i\beta + e_i \leq 0, \end{cases} \tag{1.2}$$

with y_i describing the observed binary outcome for individual i, y_i^* the latent variable representing the unobserved propensity for success (1) or failure (0), α an intercept or threshold, X_i the corresponding covariate matrix for individual i, β the coefficient vector for the covariates, and e_i an error term. The observed binary outcomes are determined by an unobservable latent variable y_i^* and a threshold. The latent variable y_i^* is a linear combination of covariates ($X_i\beta$), an intercept (α), and an error term (e_i). The decision rule is based on whether y_i^* is greater than 0. If y_i^* is positive, the individual is predicted to have a success (1); otherwise, it's predicted as a failure (0).

Following Greene (2003), in the basic framework of a generalized linear model (GLM), the core tenet revolves around specifying a linear predictor that describes the relationship between covariates and the response variable through a chosen link function. This framework assumes a distribution for the response variable, typically from the exponential family, which includes both systematic and random components. GLMs extend the class of linear models by easing certain assumptions. Linear models assume that the response variable (dependent variable y) is normally distributed conditional on the predictors, with a constant variance regardless of the predicted response value. Unlike linear models, GLMs relax the assumptions of normality and constant variance becaues they allow for more flexible modeling of the relationship between predictors and the response variable by allowing for different types of response distributions and modeling the relationship through a link function.

GLMs can handle a wide range of response distributions, including but not limited to normal (for continuous outcomes), binomial (for binary outcomes), or Poisson (for count data) distributions. By allowing different response distributions, GLMs provide greater flexibility in modeling different types of data. In GLMs, the linear predictor is linked to the mean of the response variable by a link function. This link function defines the relationship between the linear predictor and the expected value of the response variable. Commonly used link functions include identity link (for Gaussian distribution), or logit link (for binomial distribution), or log link (for Poisson distribution). The choice of the link function depends on the nature of the response variable and the assumptions about its relationship with the predictors.

The binary choice is influenced by this underlying observed data and the respondent reflects the binary choice of being or not being, e.g., $Y = 1$ for participating in the labor market or not $Y = 0$. The probability for seeing a positive outcome, i.e., $Y_i = 1$ depends on a vector of observed covariates[2], so that

[2]Also known as linear probability model.

$$\pi_i = Pr(Y_i = 1|\boldsymbol{x}) = Pr(Y_i^* > t) \tag{1.3}$$

with the observed outcome Y_i depending on a latency Y_i^* being over a threshold t to link the observed outcome to the linear combination of the covariates X_i and their corresponding coefficients β. The linear dependence can be written as

$$Y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + U_i \tag{1.4}$$

where U_i being an error term, assuming to have a cumulative distribution function (CDF) $F(u)$. The probability π_i of observing an outcome $Y_i = 1$ equals to

$$\pi_i = Pr(Y_i > 0) = Pr(U_i > -\eta_i) = 1 - F(-\eta_i), \tag{1.5}$$

with supposing $\eta_i = \boldsymbol{x}'\boldsymbol{\beta}$. Assuming the distribution of the error term U_i is symmetric around zero it holds that $F(u) = 1 - F(-u)$, it follows that $\pi_i = F(\eta_i)$.

The relationship between the linear predictor η_i and the expected value of the response variable $E(y|\boldsymbol{x})$ is established through a link function $g(\cdot)$, i.e., $E(y|\boldsymbol{x}) = g^{-1}(\boldsymbol{x}, \boldsymbol{\beta})$.

$$Pr(Y_i = 1|\boldsymbol{x}) = g(\boldsymbol{x}, \boldsymbol{\beta}) \tag{1.6}$$
$$Pr(Y_i = 0|\boldsymbol{x}) = 1 - g(\boldsymbol{x}, \boldsymbol{\beta}) \tag{1.7}$$

1.4.1 Statistical models for describing binary dependent variables

Statistical models especially for binary dependent variables are designed to analyze situations where individuals make dichotomous decisions, such as choosing between two alternatives; see among others Greene (2003), Train (2009), and Wooldridge (2010). These models are used to understand the factors influencing binary outcomes and predict the probability of a particular choice. The type of model as mentioned in equation 1.2 is related to the probit model and logistic regression, where the link function transforms the linear combination of covariates into probabilities. The probit model assumes that the unobservable latent variable y_i^* follows a standard normal distribution.

The **logit** link function is derived from the logistic distribution. The logit function is the inverse of the logistic cumulative distribution function (CDF), enabling the transformation of probabilities to the entire real line:

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) \tag{1.8}$$

The logit transformation, characterized by the natural logarithm of the odds ratio, is a cornerstone of logistic regression modeling. Prized for its wide applicability in binary classification tasks, the logit function adeptly maps probabilities from the (0,1) interval to the entire real line. Its versatility and interpretability make it a preferred choice in various empirical contexts, facilitating insightful inferences about binary outcomes.

The **probit** link function is derived from the standard normal distribution. The probit function is the inverse of the standard normal CDF, providing an alternative approach to modeling binary outcomes based on normality assumptions:

$$g(\pi_i) = \Phi^{-1}(\pi_i) \tag{1.9}$$

The probit transformation, which is based on the inverse cumulative distribution function (CDF) of the standard normal distribution, provides an alternative to the logit function. Frequently used in probit regression models, it is assumed that the underlying response variable follows a standard normal distribution. Like the logit function, the probit transformation extends probabilities across the real line, providing nuanced insight into binary outcome modeling.

The **complementary log log** link function, also known as the cloglog transformation, is derived from a transformation of the complement of the probability, offering a unique mapping from probabilities to the real line, i.e., the inverse of the CDF of the extreme value (or log-Weibull) distribution:

$$g(\pi_i) = \log(-\log(1 - \pi_i)) \tag{1.10}$$

Distinguished by the logarithm of the negative logarithm of the complement of the probability, the cloglog transformation represents a unique approach. Widely used in survival analysis and scenarios involving binary outcomes with bounded probabilities, the cloglog function provides a distinctive mapping from the (0,1) interval to the real line. Its idiosyncratic properties and shape make it well suited for modeling challenges where traditional link functions may fall short.

The chosen link function $g(\cdot)$ is used to transform the linear predictor to ensure that the predicted values conform to the distributional assumptions inherent in the response variable. The latent variable in such models can be interpreted as an unobservable utility associated with a particular choice. In essence, the decision process involves a comprehensive evaluation of the characteristics of the data and the analytical objectives at hand to arrive at an informed choice among the logit, probit, and cloglog transformations. A nuanced understanding of the basic framework underlying these transformations is paramount to their judicious application in statistical modeling efforts.

Hence, binary choice models operationalize the concepts from choice theory into a statistical framework. Estimation techniques like MLE are used to estimate parameters and test hypotheses related to the factors influencing choices. Once the link function is established, GLMs can accommodate a wide range of response distributions, such as the binomial distribution for binary outcomes. The likelihood function in a GLM represents the probability of observing the data given the parameters of the model:

$$L_i(\beta) = f(y_i|\beta) \tag{1.11}$$

Maximum Likelihood Estimation (MLE) is employed to estimate the parameters of the

GLM, optimizing the likelihood function to find the values of the parameters that maximize the probability of observing the observed data.

1.4.2 Maximum Likelihood Estimation

The general estimation function for Maximum Likelihood Estimation (MLE) involves optimizing the likelihood function with respect to the parameters of interest. This likelihood function is defined by $L(\theta|X)$, which represents the probability of observing the data $D = (y, X)$ given the parameters θ. Then optimization techniques are used to maximize the log-likelihood function, $l(\theta|X) = \log(L(\theta|X))$, with respect to the parameters θ which simplifies computations and helps avoid numerical underflow compared to the optimization of the likelihood function:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta}\{l(\theta|x)\}, \tag{1.12}$$

where $\hat{\theta}_{\text{MLE}}$ represents the Maximum Likelihood (ML) estimates of the parameters θ. Common optimization methods include gradient-based optimization algorithms (e.g., gradient descent, Newton-Raphson) and numerical optimization routines (e.g., BFGS, Nelder-Mead). Once the maximum of the log-likelihood function is found, the corresponding values of the parameters θ represent the ML estimates of the parameters (Casella & Berger, 2002; Greene, 2003; Mittelhammer, 2013) In practice, the specific form of the likelihood function $L(\theta|x)$ depends on the statistical model being used (e.g., normal distribution, binomial distribution, etc.), and the optimization algorithm used to maximize the log-likelihood function can vary based on the properties of the likelihood function and the computational resources available.

Estimating the parameters (α, β) for equation 1.2 involves maximizing the likelihood function, given the observed data and assuming a distribution for the error term (e_i). As mentioned, common choices for the distribution of e_i include the logistic distribution for logistic regression and the standard normal distribution for the probit model. The probability of the binary outcome y_i is given by the logistic function applied to y_i^*:

$$P(y_i = 1|X_i) = \frac{1}{1 + exp(-(\alpha + X_i\beta + e_i))} \tag{1.13}$$

where $exp(-(\alpha + X_i\beta + e_i))$ is the odds of success.

Hence, the likelihood function for logistic regression is the product of the probabilities of observing the binary responses given the predictors and model parameters:

$$L(\beta_0, \beta_1, \ldots, \beta_P|X, y) = \prod_{i=1}^{n} P(y_i|X_i) \tag{1.14}$$

The log-likelihood function is then:

$$l(\beta_0, \beta_1, \ldots, \beta_P|X, y) = \sum_{i=1}^{n} [y_i \log(P(y_i|X_i)) + (1 - y_i) \log(1 - P(y_i|X_i))] \tag{1.15}$$

In the probit case the probability of the binary outcome y_i is given by the cumulative distribution function of the standard normal distribution applied to y_i^*:

$$P(y_i = 1) = \Phi(\alpha + X_i\beta + e_i) \tag{1.16}$$

with Φ being the standard normal cumulative density function, α being the intercept, X_i being the covariate matrix, β being the coefficient vector, and e_i being the error term. The likelihood function for probit regression is similar to logistic regression, but the probabilities are computed using the standard normal CDF. The ML estimates for the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are obtained by maximizing the log-likelihood function using the same optimization methods as in logistic regression.

Non-linear estimators, such as logistic regression, present complex relationships between independent and dependent variables. The impact of a specific independent variable on the probability of an outcome is not constant but varies based on its value and the values of other variables in the model. Therefore, understanding and interpreting the marginal effects of these variables is crucial. Marginal effects measure the change in the predicted probability of the outcome resulting from a unit change in the independent variable. This can be calculated for specific values of the independent variable, averaged across all observations, or evaluated at the mean values of other variables.

To visualize the different MLE methods, a dataset is generated for $N = 8,000$ respondents by simulating two covariates, X_1 and X_2, from standard normal distributions, while β_0 as a constant is set to .75, β_1 to .5, and β_2 to .8. A linear predictor is computed as a weighted sum of the covariates with β_1 and β_2, an intercept β_0, and with standard normally distributed random noise added to create a latent variable. If the latent variable is greater than zero, the respondent is classified as successful ($y_i = 1$); otherwise, the respondent is classified as unsuccessful ($y_i = 0$), i.e., the data follows a probit link. This deterministic rule creates a threshold-based decision mechanism which is modeled with different frameworks.

First a probit model is implemented that assumes the probability of success is linked to the covariates X_1 and X_2 through the cumulative normal distribution Φ. It models the probability of success as $P(y_i = 1) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$. This model is particularly suited for binary outcomes where an underlying latent variable with a normal distribution determines the observed outcome. The logit model uses the logistic function to relate the covariates X_1 and X_2 to the probability of success ($y_i = 1$). It specifies as: $P(y_i = 1) = \frac{1}{1 + exp\{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2)\}}$. The complementary log-log (cloglog) model is used for binary outcomes with (highly) skewed response probabilities. It links the covariates to the probability of success with the relationship $P(y_i = 1) = 1 - exp(-exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2))$, i.e., an asymmetric transformation of the probability, so that the cloglog model accommodates situations where extreme probabilities (very close to 0 or 1) are more likely. Finally, the linear model assumes a direct linear relationship between the binary outcome success and the covariates X_1 and X_2. It models the outcome as $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, where ϵ represents the error term. While simple, this model is less suitable for binary outcomes because it does not constrain predictions to probabilities within the range of 0 and 1. The results are presented in table 1.1 for the full model, and in table 1.2 for the null models, where only the constant β_0 is included. In both cases, all estimates are highly significant and the

information criteria are quite similar.

In table 1.3 the marginal effects that measure how a small change in an independent variable affects the predicted probability of the outcome variable are presented. For continuous predictors, marginal effects represent the instantaneous rate of change in probability associated with a one-unit increase in the predictor, assuming the change is small. For binary predictors, marginal effects quantify the change in probability when the variable shifts from 0 to 1. These effects are especially useful in nonlinear models, such as logit and probit, where the relationship between predictors and probabilities is not constant but depends on the levels of the predictors and other factors. Thus, the results are quite similar and highly significant above all model types.

1.4.3 Bayesian estimation

Referring to the above, the Bayesian estimation is based on the calculation of the posterior distribution. Albert and Chib (1993) describe at first a data augmentation solution to estimate a probit model. The Gibbs Sampler is a specific Markov Chain Monte Carlo (MCMC) algorithm designed for Bayesian probit regression models. The key idea behind the Gibbs Sampler is to use a data augmentation approach to represent the latent variables in a probit model as additional parameters, transforming the problem into a higher-dimensional space. The Bayesian model is completed by specifying a necessary prior distribution for the coefficients; e.g., a multivariate normal distribution can be used:

$$\beta \sim N(b_0, B_0)$$

with b_0 being the mean and B_0 being the covariance matrix of a normal distribution.

A directed acyclic graph (DAG) helps visualizing the casual relationship of the Bayesian estimation. In the context of statistical modeling and causal inference, DAGs are often used to visually represent causal relationships between variables. In a causal DAG, nodes represent variables (e.g., treatment, outcome, confounders), and directed edges represent causal relationships between them (Cunningham, 2021; Thulasiraman & Swamy, 1992). Figure 1.1 shows a DAG of the probit model mentioned above and illustrates the relationship between the latent variables, the observed variables, and parameters, incorporating prior information on the parameter β. At the core of the DAG is the latent variable y_i^* representing the unobserved outcome, influenced by the observed covariates. The parameter β represents the coefficients associated with the covariates, while priors for β are denoted as b_0 and B_0, representing the prior mean and covariance respectively. Furthermore unobserved variables are represented by circle nodes, observed variables by rectangular nodes, and the dependencies between the variables by arrows. Arrows depict the directional influence, such as from the covariates to the latent variable and from the priors to the parameter β, reflecting how changes in these components affect the likelihood of the latent variable y_i^* and thus of the outcome y. The arrow between the latent variable y_i^* and the observed dependent variable y_i should generally be represented by a directional arrow pointing from the latent variable to the observed variable. This direction symbolizes that the observed dependent

variable is derived from the latent variable meaning that changes in the latent variable result in changes in the observed dependent variable.

In a Bayesian view, each node is a random variable which helps to derive the full conditionals for the Gibbs sampler. Via a Gibbs sampler the parameters are updated iteratively, including the latent variables, by drawing from their full conditional distributions for the parameters, given the observed data and the current values of the other parameters. Hence, the Gibbs sampler described by Albert and Chib (1993) is in particular a simplified sampling process for the latent variables, making it computationally efficient. In literature, alternative Gibbs samplers are presented because of the lack of mixing chains, see among others Imai and van Dyk (2005), Jackman (2009), and van Dyk and Meng (2001).

Let $\eta_i = \alpha + X_i^\top \beta$ and $p_i = \Pr(Y_i = 1 \mid X_i)$ with link $g(p_i) = \eta_i$ given by the logit, probit, or complementary log–log function (see section 1.2). Coefficients are assigned weakly informative normal priors on the link scale with $\alpha \sim \mathcal{N}(0, s_0^2)$ and $\beta \sim \mathcal{N}(0, s^2 I_p)$. In the probit specification, identification is achieved by fixing $\mathrm{Var}(\varepsilon) = 1$. Estimation proceeds with the data augmentation Gibbs sampler following Albert and Chib (1993), which alternates between drawing the latent vector z from univariate normal truncated at zero according to y_i and sampling β from its multivariate normal full conditional with closed-form mean and covariance. In the logit specification, $g(p) = \log \frac{p}{1-p}$, the latent-variable view corresponds to logistic errors with fixed variance $\pi^2/3$, which sets the scale of β. The posterior is non-conjugate under normal priors, so that coefficients are updated with a random-walk Metropolis–Hastings algorithm as implemented in `MCMCpack::MCMClogit`, see Martin et al. (2011). Proposal scales are tuned to achieve stable acceptance and adequate effective sample sizes.

For the complementary log–log (cloglog) link, $g(p) = \log(-\log(1-p))$, the latent-error interpretation corresponds to a type-I extreme-value (Gumbel) distribution. The link is asymmetric – steeper near $p \approx 0$ and flatter near $p \approx 1$ – and is closely related to discrete-time hazard models, which makes it attractive when event probabilities are very small or very large. The posterior is again non-conjugate; estimation employs Hamiltonian Monte Carlo via the No-U-Turn Sampler in `rstanarm::stan_glm`, see Stan Development Team (2025). The gradient-based exploration adapts the step size and mass matrix to the local curvature and is effective for the skewed posterior geometries that can arise with rare events; diagnostics focus on $\widehat{R}$, effective sample sizes, and the absence of divergent transitions (in which case `adapt_delta` can be raised). For the Bayesian OLS (linear probability) model, $Y_i = \eta_i + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, conjugacy yields Gibbs updates for (β, σ^2) under $\beta \sim \mathcal{N}(0, 5^2 I_p)$ and $\sigma^2 \sim \mathrm{InvGamma}(a_0, b_0)$ with diffuse hyperparameters. This model provides a linear benchmark on the probability scale but does not constrain fitted values to $[0, 1]$. All models are estimated under common MCMC settings. For the `MCMCpack` fits, 5,000 iterations are discarded as burn-in and 20,000 iterations are retained with thinning by 10. For the `rstanarm` fits, four chains with 2,000 iterations per chain are run, yielding approximately 1,000 post-warmup draws per chain; the target acceptance `adapt_delta` is increased only if divergent transitions occur. Convergence is assessed by visual inspection of trace plots, autocorrelation functions, and effective sample sizes for the `MCMCpack` chains, and by $\widehat{R}$ values close to 1.00,

large effective sample sizes, and the absence of divergent transitions for the `rstanarm` fits. Posterior summaries are reported as posterior means together with 95% credible intervals.

The null specifications in table 1.5 primarily reflect the sample prevalence of y mapped onto each link scale and therefore exhibit tight credible intervals given the large N. In the full models (Table 1.4), the probit posteriors recover the data-generating parameters closely, with 95% credible intervals excluding zero. Logit coefficients lie on a rescaled metric and are therefore larger in magnitude (approximately consistent with the logistic vs. normal scale), whereas cloglog coefficients reside on an asymmetric link scale but agree in sign and yield comparable predictive probabilities. The Bayesian OLS (linear probability) model reproduces signs and average effects on the identity scale but does not constrain fitted values to $[0, 1]$ and is treated as a linear benchmark. Marginal effects on the probability scale (see table 1.6) are computed as average marginal effects for continuous regressors. Across probit, logit and cloglog they are very similar despite different coefficient scales, with the cloglog effects showing slight asymmetry at extreme probabilities, while OLS marginal effects equal the slope coefficients and typically approximate the average nonlinear effects when predicted probabilities are moderate.

1.5 Bias-variance trade-off

As shown above, in a probit model, a directed acyclic graph (DAG) could represent the hypothesized causal relationships between variables. For example, given a binary outcome variable Y and covariates X, the DAG could represent the causal relationships leading to the outcome Y. If the DAG helps to identify and include all relevant confounding variables in the model, it will help to reduce bias by ensuring that the estimated relationships between variables are closer to the true causal relationships in the underlying system. Otherwise, if the DAG suggests including many unnecessary variables or complex relationships in the model, it may increase the model's sensitivity to variation and thus increase variance. The bias-variance trade-off is a fundamental concept in statistical learning and model selection that describes the trade-off between the bias of a model and its variance. In addition to the analytical or computational issues caused by the (high) dimensionality of a dataset, the selection of appropriate variables reduces the level of complexity. In high-dimensional estimation, many parameters increase the overall error of the estimation (Hastie et al., 2009; James et al., 2013)

In model and variable selection, underfitting, or high bias, occurs when the model is overly simplistic and fails to capture the true underlying relationships present in the data. This often happens when important confounding variables are excluded from the model, akin to omitting crucial nodes in a DAG. Consequently, an underfitted model may exhibit poor predictive performance in machine learning tasks and yield inaccurate estimation results in statistical inference. Conversely, overfitting arises when the model is excessively complex, capturing noise or including unnecessary variables that do not contribute to the underlying relationships. This situation can manifest as a model fitting the training data too closely but failing to generalize well to new data, akin to including excessive variables in a

DAG. In machine learning and statistical inference, overfitting leads to poor generalization performance, while in statistical inference, it can result in high standard errors and unstable estimates due to the model's sensitivity to small fluctuations in the data, ultimately leading to unreliable conclusions.

The bias-variance trade-off is a fundamental concept in both statistics and machine learning, aiming to strike a balance between model complexity and generalization performance. Figure 1.2 illustrates the relationship between total error and the bias-variance trade-off (Pearl, 2016). This trade-off is particularly crucial in models like the probit model, which uses a nonlinear function to predict binary outcomes based on input variables.

In statistics, bias refers to the error introduced by approximating a realworld phenomenon with a simplified model, while variance quantifies the model's sensitivity to variations in the training data (Cunningham, 2021; Gelman et al., 2023; Hastie et al., 2009). In machine learning, bias typically refers to the error due to overly simplistic assumptions in the model, while variance captures the model's tendency to overfit to the training data. In the context of a probit model and DAG, the bias-variance trade-off underscores the importance of careful model selection. A model that is too simplistic (high bias) may fail to capture the underlying relationships, leading to inaccurate predictions or estimates. On the other hand, a model that is too complex (high variance) may capture noise in the training data, resulting in poor generalization to new data or unstable estimates. Therefore, finding the optimal level of model complexity involves balancing bias and variance to minimize overall prediction error on new data. This process requires careful consideration of the trade-offs involved in model selection, ensuring that the chosen model is appropriately suited to the underlying data-generating process represented by the DAG.

Also in machine learning the bias-variance trade-off is a fundamental concept, essential for achieving models that generalize well to unseen data (Bishop, 2009). In machine learning, bias refers to the error introduced by the model's assumptions, often resulting in the model's inability to capture the true underlying patterns in the data. Variance, on the other hand, measures the model's sensitivity to fluctuations in the training data, with high variance leading to overfitting and poor generalization. In the context of machine learning, striking the right balance between bias and variance is crucial for building models that perform well on unseen data. Models with high bias may be too simplistic and fail to capture complex patterns, while models with high variance may be too flexible and capture noise instead of underlying trends. Thus, finding the optimal level of model complexity is paramount, and techniques such as cross-validation, regularization, and ensemble methods are commonly employed to navigate the bias-variance trade-off. For example, in the case of a probit model used in machine learning, the bias-variance trade-off remains central. Choosing an appropriate set of features, regularization parameters, and model architecture can help mitigate bias and variance, ensuring that the model generalizes well to new data while capturing meaningful relationships. In summary, the bias-variance trade-off is a critical consideration in machine learning, guiding model development and selection to achieve models that balance complexity and generalization performance effectively.

1.5.1 Bias and variance in inference

In the context of statistical inference, the bias-variance trade-off manifests in the estimation of model parameters and the subsequent uncertainty associated with these estimates. The goal in statistical inference is to construct an estimator for the unknown parameter θ^* given the observed dataset D (Mittelhammer, 2013).

Bias in inference In statistical inference, bias refers to the systematic error introduced by the estimation procedure. A biased estimator systematically deviates from the true parameter value. For example, if a model consistently underestimates or overestimates the true parameter value, it exhibits bias. A biased estimator may result from simplifying assumptions or model misspecification. The error due to bias is the difference between expected values and optima. Consider the bias of an estimator $\hat{\theta}$ as $\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta} - \theta^* | D]$. The estimator is called unbiased if and only if $\mathbb{E}[\hat{\theta} - \theta^* | D] = 0$, i.e., $\mathbb{E}[\hat{\theta} | D] = \theta^*$ for all values of θ^*. The bias refers to the systematic error introduced by an estimation or testing procedure and can arise due to various factors, such as model mis-specification, improper assumptions, or inadequate sample size. High bias can lead to consistently underestimating or overestimating the true parameter values.

Variance in inference Variance quantifies the variability of an estimator $\hat{\theta}$ across different samples drawn from the same population. High-variance estimators are sensitive to small fluctuations in the data, potentially producing substantially different estimates across samples. This phenomenon often arises due to overfitting, when the model captures noise rather than the underlying signal. Formally, for a univariate estimator, the variance is $\text{Var}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\right]$, and for a multivariate estimator $\hat{\theta} \in \mathbb{R}^p$, it is given by the covariance matrix $\text{Var}(\hat{\theta}) \equiv \text{Cov}[\hat{\theta}]$. Note that the variance measures only the spread of the estimator values around their expected value and does not directly depend on the true parameter value θ^*.

Bias-Variance trade-off in inference Considering that the first and second moment of any distribution just are not necessarily related, the bias and variance are not necessarily related, either. The classical bias-variance trade-off predicts that bias decreases and variance increases with model complexity, leading to a U-shaped risk curve. High bias does not necessarily cause high variance and vice versa, but looking at the squared error, bias and variance of an estimator $\hat{\theta}$ are related as (Hastie et al., 2009):

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \mathbb{E}\left[\|\hat{\theta} - \theta^*\|^2\right] \\
&= \mathbb{E}\left[\|\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta^*\|^2\right] \\
&= \mathbb{E}\left[\|\hat{\theta} - \mathbb{E}[\hat{\theta}]\|^2\right] + \|\mathbb{E}[\hat{\theta}] - \theta^*\|^2 \\
&= \text{Var}(\theta) + \text{Bias}^2(\hat{\theta})
\end{aligned}$$

Implementing techniques to reduce variance often increases bias, and vice versa. The challenge in most model considerations is to minimize the sum of both. When choosing

a statistical model or hypothesis test, there is a trade-off between model complexity and flexibility. A simpler model may introduce bias but have lower variance, while a more complex model may capture nuances but exhibit higher variance. In statistical modeling, selecting a model involves balancing bias and variance. A model that is too simple may have high bias, while a model that is too complex may have high variance. Therefore, the optimal model strikes a balance between the two, and model selection techniques aim to achieve this balance.

Finding the appropriate balance between bias and variance is essential for obtaining reliable estimates and making accurate inferences. Techniques such as regularization, model selection, and cross-validation are commonly employed to manage the bias-variance trade-off in statistical inference. Regularization methods impose a penalty on model complexity, thereby reducing variance while introducing a controlled amount of bias to improve generalization (Hastie et al., 2009). Model selection involves choosing models with the right level of complexity to minimize both bias and variance. Cross-validation assesses the generalization performance of the estimator, providing insights into its bias and variance properties. In summary, the bias-variance trade-off in statistical inference requires careful consideration of the trade-offs between bias and variance to obtain robust and accurate parameter estimates (Freedman, 2009; James et al., 2013).

1.5.2 Bias and variance in prediction

Rather than constructing an estimator $\hat{\theta}$ for an unknown parameter, the goal of a prediction setting, e.g., machine learning, differs from statistical inference by looking for a function $f(x)$ that can predict y given X well with respect to some loss function (Bishop, 2009). Assume $y = f(x) + \epsilon$ with a disturbance term ϵ is distributed with mean $\mathbb{E}[\epsilon] = \mu = 0$ and $\text{Var}[\epsilon] = \eta^2$.[3] The bias-variance trade-off in prediction is a key consideration when selecting and evaluating predictive models, especially in machine learning.

Estimating a model $\hat{f}(x)$ of $f(x)$ for a training set S, such as in most machine learning frameworks, will emulate $f(x)$ well on all unseen datasets. Due to the disturbance term embedded in the training and test dataset, randomness is introduced, so $\hat{f}(x)$ is also random. Considering an unseen example pair (y^*, x^*), the expected squared error is defined as follows (Bishop, 2009; Hastie et al., 2009):

$$\begin{aligned}
\text{MSE}(\hat{f}(x^*)) &= \mathbb{E}\left[\left(y^* - \hat{f}(x^*)\right)^2\right] \\
&= \mathbb{E}\left[\left(f(x^*) + \epsilon - \hat{f}(x^*)\right)^2\right] \\
&= \mathbb{E}\left[(f(x^*) - \hat{f}(x^*))^2 + 2\epsilon(f(x^*) - \hat{f}(x^*)) + \epsilon^2\right] \\
&= \mathbb{E}\left[(f(x^*) - \hat{f}(x^*))^2\right] + \mathbb{E}\left[2\epsilon(f(x^*) - \hat{f}(x^*))\right] + \mathbb{E}\left[\epsilon^2\right].
\end{aligned}$$

[3] Not necessarily Gaussian-normal.

With $\mathbb{E}[\epsilon^2] = \eta^2$ and ϵ being independent of both $f(x^*)$ and $\hat{f}(x^*))$, and $\mathbb{E}[\epsilon] = 0$ the term can be set to zero. So the equation simplifies to:

$$\begin{aligned}
\text{MSE}(\hat{f}(x^*)) &= \mathbb{E}\Big[(f(x^*) - \hat{f}(x^*))^2\Big] + \eta^2 \\
&= \mathbb{E}\Big[(f(x^*) - \mathbb{E}[\hat{f}(x^*)]) + (\mathbb{E}[\hat{f}(x^*)] - \hat{f}(x^*))^2\Big] + \eta^2 \\
&= \mathbb{E}\Big[(f(x^*) - \mathbb{E}[\hat{f}(x^*)])^2 \\
&+ 2(f(x^*) - \mathbb{E}[\hat{f}(x^*)])(\mathbb{E}[\hat{f}(x^*)] - \hat{f}(x^*)) + (\mathbb{E}[\hat{f}(x^*)] - \hat{f}(x^*))^2\Big] + \eta^2 \\
&= \mathbb{E}\Big[(f(x^*) - \mathbb{E}[\hat{f}(x^*)])^2\Big] \\
&+ 2\mathbb{E}\Big[(f(x^*) - \mathbb{E}[\hat{f}(x^*)])(\mathbb{E}[\hat{f}(x^*)] - \hat{f}(x^*))\Big] \\
&+ \mathbb{E}\Big[(\mathbb{E}[\hat{f}(x^*)] - \hat{f}(x^*))^2\Big] \\
&+ \eta^2
\end{aligned}$$

By definition, the $\mathbb{E}[\mathbb{E}[\hat{f}(x^*)] - \hat{f}(x^*)]$ has an expectation of 0 zero because $\mathbb{E}[\hat{f}(x^*)]$ is the expected value of $\hat{f}(x^*)$, so: $\mathbb{E}[\mathbb{E}[\hat{f}(x^*)] - \hat{f}(x^*)] = \mathbb{E}[\mathbb{E}[\hat{f}(x^*)]] - \mathbb{E}[\hat{f}(x^*)] = 0$. Finally simplifying further:

$$\begin{aligned}
\text{MSE}(\hat{f}(x^*)) &= \mathbb{E}\Big[\Big(f(x^*) - \mathbb{E}[f(x^*)]\Big)^2\Big] + \mathbb{E}\Big[\Big(\mathbb{E}[\hat{f}(x^*)] - \hat{f}(x^*)\Big)^2\Big] + \eta^2 \\
&= \underbrace{\Big(f(x^*) - \mathbb{E}[\hat{f}(x^*)]\Big)^2}_{\text{Bias}^2 \text{ of method}} + \underbrace{\text{Var}(\hat{f}(x^*))}_{\text{Variance of method}} + \underbrace{\eta^2}_{\text{Irreducible error}} .
\end{aligned}$$

Irreducible error in prediction Additionally, the irreducible error term η^2, i.e., the noise term, can fundamentally not be reduced by any model, even if the true $f(x)$ is known. The irreducible error term $\mathbb{E}[\epsilon^2]$ represents the inherent variability in the data that cannot be reduced by any model, i.e., it captures the variability in y that cannot be explained by $f(x)$. The irreducible error in prediction is caused by measurement noise (errors in measuring the data D), latent variables (factors influencing y that are not included in D), stochasticity in the measurement system like true randomness in the data generating process in clinical experiments, as well as human behavior (measuring depending on human decisions or preferences).

Bias in prediction The bias is the difference between the expected (or average) prediction of a model and the correct value which is tried to predict. The bias term represents the squared difference between the average prediction of the model $f(x)$ and the true underlying function value $f(x)$.

Variance in prediction The error due to variance is taken as the variability of a model prediction for a given data point. Assuming repeating the model building and predicting multiple times, the variance is the amount for a given point to vary between different realizations of the model. The variance is represented by the expected squared deviation of the model predictions $f(x)$ from their average.

Bias-Variance trade-off in prediction A model that is too simple (high bias) may have a low variance but may not capture the underlying complexity of the data, leading to a high bias term in the MSE. On the other hand, a more complex model (low bias) may capture the data's complexity but might be sensitive to fluctuations in the training data, resulting in a high variance term in the MSE. The goal is to find the right level of model complexity that minimizes the overall MSE, balancing the bias and variance terms. The variance and bias can be controlled and reduced to zero. However, dealing with incomplete models and finite data sets allows for a trade-off between minimizing the bias and minimizing the variance. Such a clean decomposition into bias and variance terms exists only for a few types of models, such as the squared error loss or the k-nearest neighbor algorithm. Increasing k, the number of model parameters and variables used in the model, will increase the variance and decrease the bias, and vice versa by decreasing k. The complexity of a model increases as more and more parameters are added to a model. The bias has a negative first-order condition in response to model complexity, while the variance has a positive one.

Focusing on the total error rather than the decomposition means understanding the bias-variance trade-off as finding a balanced fit so that the model is not (too) under- or over-fitted. Such a sweet spot for a set of models with different numbers of parameters is the level of complexity at which the increase in bias is equal to the decrease in variance. If the model exceeds this sweet spot, the model complexity is increased by a higher variance and a lower bias, which is called overfitting, otherwise underfitting. Practically speaking, there is no analytical way to find this sweet spot.

1.5.3 Bias and variance in a Bayesian view

Overfitting is a property of Maximum Likelihood that does not occur when marginalizing over parameters in a Bayesian setting. Bias-variance decomposition is a valuable tool for understanding the complexities of models, though its practical limitations stem from its reliance on ensembles of datasets. In practice only a single dataset is observed for estimation. Thus, for model evaluation most machine learning algorithm or statistical techniques split the datasets into test and training data or combine them with techniques such as cross-validation, which would reduce over-fitting for a given model complexity.

According to Gelman et al. (2023), a Bayesian approach gives useful insights into over-fitting and the bias can be controlled by setting the prior because bias and variance are still important concepts, but their interpretation and analysis may differ somewhat from the frequentist perspective. In Bayesian statistics, parameters are treated as random variables with probability distributions. The concept of bias is less straightforward than in frequentist statistics because Bayesian methods consider uncertainty in parameter estimates as a fundamental

aspect of the modeling process. As noted-above, the goal of Bayesian statistics is to estimate the posterior distribution of parameters given the observed data which incorporates both prior beliefs (prior distribution) and information from the data (likelihood). Bayesian bias is often considered in the context of prior information. If the prior distribution is misspecified or biased, it can affect the posterior distribution and, consequently, the parameter estimates.

The loss-function approach, also known as decision theory, is a framework used in statistical modeling and machine learning to make decisions under uncertainty (Gelman et al., 2023; Jeffreys, 1961; Smith, 2010). In this approach, decisions are made by minimizing or optimizing a predefined loss function, which quantifies the cost or penalty associated with different actions or outcomes. Bayesian decision theory takes a broader view of bias by considering the loss function associated with decision-making which incorporates both prior and posterior distributions to minimize expected loss, and what might be considered biased from a frequentist perspective may be justified in a Bayesian framework. Instead of thinking of variance as a measure of variability in repeated sampling (as in frequentist statistics), Bayesian statistics considers the variability in parameter estimates within the posterior distribution. The spread or uncertainty in the posterior distribution reflects the model's uncertainty about the true parameter values given the observed data. In Bayesian statistics, a loss function quantifies the penalty associated with making incorrect decisions or predictions. The first step is to define a loss function that quantifies the penalty or cost associated with making incorrect decisions or predictions. Common loss functions include squared error loss for regression tasks, log-loss for classification problems, and others depending on the specific problem and decision context. In Bayesian decision theory, the goal is to minimize the expected loss under the posterior distribution of parameters. This involves integrating the loss function over the posterior distribution to compute the expected loss (Bissiri & Walker, 2019; Huber & Ronchetti, 2011). The expected loss can be expressed as:

$$\text{Expected Loss} = \int L(\theta, \hat{\theta}(x)) \cdot p(\theta|x)\, d\theta \tag{1.17}$$

where $L(\theta, \hat{\theta}(x))$ is the loss function, θ is the parameter of interest, $\hat{\theta}(x)$ is the estimator or decision rule based on the data x, and $p(\theta|x)$ is the posterior distribution of θ given the data. The decision rule, crucial in Bayesian inference, guides decision-making by leveraging the posterior distribution of parameters and the specified loss function. It aims to minimize the expected loss and can be established through diverse criteria, including maximum a posteriori estimation (MAP), Bayesian decision boundaries, or expected utility maximization. Once determined, the decision rule facilitates predictions or decisions based on new data by applying it to the posterior distribution of parameters. This integration of Bayesian inference with the loss function approach enables principled decision-making under uncertainty, effectively balancing the costs and benefits associated with different actions or outcomes. In Bayesian estimation, leveraging loss functions allows the derivation of point estimates (e.g., maximum a posteriori estimation) or interval estimates (e.g., credible intervals) that minimize the expected loss, enhancing the robustness of inference.

In frequentist statistics, loss functions are used to evaluate the performance of estimators

or classifiers (Casella & Berger, 2002; Hastie et al., 2009). Common loss functions include mean squared error, negative log-likelihood, and classification error. In the frequentist setting, estimation is often framed as empirical risk minimization, where the goal is to minimize the average loss over the training data. Maximum Likelihood Estimation is a common frequentist approach where the parameters of a model are chosen to maximize the likelihood function, which is equivalent to minimizing the negative log-likelihood loss. In summary, while the fundamental concepts of loss functions are similar between Bayesian and frequentist frameworks, their implementation and interpretation can differ. Bayesian decision theory aims to minimize the expected loss under the posterior distribution, while frequentist methods focus on minimizing empirical risk over the training data. Both approaches have their advantages and are commonly used in different statistical applications.

The loss-function approach for the above mentioned probit model can be formulated in both contexts. In the Bayesian context, a probit regression model is specified where the probability of the outcome y_i for the i-th observation is modeled as the cumulative distribution function of a standard normal distribution evaluated at the linear combination of predictor variables x_i. While prior distributions are specified for the regression coefficients β, the posterior is obtained given the data and the prior. The loss-function, in the Bayesian context, could be the posterior predictive loss, which is the expected loss under the posterior distribution of β. Common choices include the 0-1 loss function for classification tasks, where the loss is 0 if the predicted class matches the true class and 1 otherwise.

In the frequentist context, the regression coefficients β is estimated using MLE, i.e., the likelihood function is maximized or equivalently the negative log-likelihood function is minimized (Berger, 2006b). The loss function in the frequentist context is typically the negative log-likelihood function, which quantifies the discrepancy between the observed data and the model predictions For probit regression, the negative log-likelihood function is derived from the standard normal cumulative distribution function. For optimization algorithms such as gradient descent or Newton-Raphson are used to minimize the negative log-likelihood function and estimate the parameters β. After parameter estimation, the model's performance is evaluated using metrics such as accuracy, precision, recall, or the area under the ROC curve. In summary, in both the Bayesian and frequentist contexts, the loss function plays a central role in model estimation and evaluation. While the formulation of the loss function may differ, the ultimate goal is to minimize the expected loss or maximize the likelihood of the observed data to obtain accurate and reliable predictions. The loss function and the bias-variance trade-off are interconnected concepts in the realm of statistical modeling and machine learning, as both play crucial roles in model development, evaluation, and decision-making. A loss function quantifies the penalty or cost associated with the model's predictions. It measures the discrepancy between the predicted outcomes and the true outcomes. The choice of loss function depends on the nature of the problem and the goals of the analysis. Different loss functions are appropriate for different types of tasks, such as regression, classification, or ranking. Examples of loss functions include mean squared error (MSE), cross-entropy loss, hinge loss, and absolute error.

In summary, the loss function and the bias-variance trade-off are intertwined aspects

of model development and evaluation. The choice of loss function affects the model's performance and can influence the trade-off between bias and variance. Understanding this relationship is essential for building predictive models that achieve a suitable balance between accuracy and generalization.

Furthermore, in Bayesian statistics, model averaging is a common approach to account for model uncertainty. Instead of selecting a single best model, Bayesian model averaging considers a weighted average of predictions from multiple models, with weights determined by the posterior model probabilities. In addition, variable selection and model selection are also used in Bayesian statistics where the focus is on characterizing uncertainty through probability distributions rather than providing point estimates and standard errors. Bayesian methods provide a coherent framework for incorporating prior information, accounting for uncertainty, and making decisions based on posterior distributions. The concepts of bias and variance are often interpreted in the context of this broader probabilistic framework.

The handling of missing values before or during the estimation can increase the bias and variance. Bayesian imputation methods can also introduce bias if model assumptions are violated or if the model is misspecified (Gelman et al., 2023; Little & Rubin, 2002). However, the Bayesian framework allows for explicit modeling of uncertainty and priors, potentially mitigating bias. Bayesian methods naturally provide estimates of uncertainty, and the variability of imputed values is explicitly represented in the posterior distributions. This can provide a more comprehensive view of the uncertainty associated with imputed values. Bayesian methods naturally incorporate uncertainty into the imputation process by treating missing data as parameters with associated probability distributions. Multiple imputations are often generated to account for this uncertainty. Bayesian models can explicitly model the missing data mechanism, allowing joint modeling of observed and missing data. This can help reduce bias by accounting for the relationships between variables. Finally, both Bayesian and frequentist approaches involve choices about model complexity. A more complex imputation model may reduce bias, but may also increase variance. Therefore, both frameworks benefit from sensitivity analyses that assess the impact of different imputation strategies on bias and variance. Bayesian methods allow for the incorporation of prior information, which can affect both bias and variance. Careful consideration of priors is critical to balancing bias and variance. In both perspectives, handling missing data involves navigating the bias-variance trade-off by choosing appropriate imputation methods and modeling strategies. Bayesian methods, with their inherent probabilistic nature, provide a framework for explicitly addressing uncertainty in imputation and capturing the complexity of the missing data mechanism. Frequentist methods, while different in their approach, also seek to balance bias and variance when dealing with missing data.

1.6 Navigating complexity: the imperative of model and variable selection

In the realm of inference, the process of model and variable selection assumes paramount importance, particularly in the pursuit of rigorous inferential insights from data. This

critical undertaking involves judiciously identifying the subset of relevant variables within a potentially expansive covariate space and crafting a model that best annotates the underlying relationships. As mentioned above, a fundamental consideration in this endeavor is the delicate balance between model complexity and interpretability, which lies at the heart of statistical inference. Central to the task of model and variable selection is the principle of parsimony, wherein the goal is to construct a model that achieves maximal explanatory power with minimal complexity. This principle aligns closely with the overarching objective of inference, which seeks to uncover the underlying structure of the data while avoiding spurious relationships and overfitting (Gelman et al., 2023; James et al., 2013).

In the pursuit of parsimony, various methodological approaches present themselves as viable strategies. Shrinkage techniques, such as Ridge regression and Lasso regularization, offer effective means of tempering model complexity by imposing penalties on the magnitude of regression coefficients (Zou & Hastie, 2005). These methods not only facilitate variable selection but also enhance the stability and interpretability of the estimated coefficients. Principal component analysis (PCA) represents another powerful tool for dimensionality reduction, wherein the original variables are transformed into a smaller set of orthogonal components capturing the most substantial variation in the data (Jolliffe, 2004). By condensing the information content of the covariates into a reduced-dimensional space, PCA enables the identification of key underlying patterns while mitigating the risk of multicollinearity and overfitting. Subset selection methods, including forward, backward, and stepwise selection, offer yet another avenue for model refinement. These iterative procedures systematically evaluate different combinations of variables to identify the subset that optimally balances model fit and complexity. While computationally intensive, subset selection methods provide valuable insights into the relative importance of individual covariates and facilitate the construction of parsimonious models amenable to interpretation (Hocking, 1976). The strategic selection of models and variables in inference constitutes a foundational aspect of inferential research. By embracing the principles of parsimony and judiciously navigating the trade-offs between bias and variance, researchers can construct models that not only yield meaningful insights but also engender confidence in the validity and generalizability of their findings.

In addition to its significance in inference, model and variable selection also play a crucial role in predictive modeling and machine learning (James et al., 2013). The quality of predictions hinges on the ability of the model to generalize well to unseen data, a task that demands careful consideration of model complexity and variable relevance. Overly complex models risk overfitting to the idiosyncrasies of the training data, resulting in poor performance on new observations. Conversely, overly simplistic models may fail to capture the underlying patterns in the data, leading to suboptimal predictive accuracy. By selecting an appropriate subset of variables and tuning the model complexity, practitioners can strike a balance between bias and variance, thereby maximizing predictive performance. Techniques such as cross-validation and regularization further aid in the selection of models that generalize well to new data, ensuring robust and reliable predictions in real-world applications.

Returning to Popper's approach, it can be summarized that Popper's philosophy of

science offers valuable insights into the process of model and variable selection in both inference and machine learning. Popper emphasized the importance of falsifiability – the notion that scientific theories should be formulated in a way that allows for empirical testing and potential refutation. Applied to model and variable selection, this principle suggests that models should be constructed with the explicit aim of being subjected to empirical scrutiny, allowing for the rejection or modification of hypotheses based on observed data. In the context of inference, this entails selecting models and variables that can be rigorously tested against empirical evidence, ensuring that conclusions drawn from statistical analyses are grounded in empirical reality. Similarly, in machine learning, Popper's ideas advocate for the development of models that are not only predictive but also interpretable and falsifiable. By embracing falsifiability as a guiding principle, practitioners can foster a culture of critical inquiry and empirical validation, leading to more robust and reliable scientific conclusions in both inference and machine learning contexts.

1.7 Outline

This thesis is organized as follows: In chapter 2 a submitted article is presented which discusses that the Bayesian view on variable selection while handling missing values. This article is co-authored with Christian Aßmann. Chapter 3 presents additional information on variable selection for missing values. Next, the Bayesian model selection approach is presented in chapter 4, also in the context of missing values in the covariates. Finally, a conclusion (chapter 5) summarizes the overall results.

1.8 Appendix section 1

1.8.1 Figures

FIGURE 1.1: Bayesian probit model: Directed acyclic graph (DAG)

This directed acyclic graph (DAG) highlights the role of the probit link function in transforming the latent variable y_i^* into a probability of the observed outcome y_i. Observed variables (X_i, y_i) are represented by squares, while latent variables (y_i^*, β) are represented by circles. Arrows indicate the direction of influence.

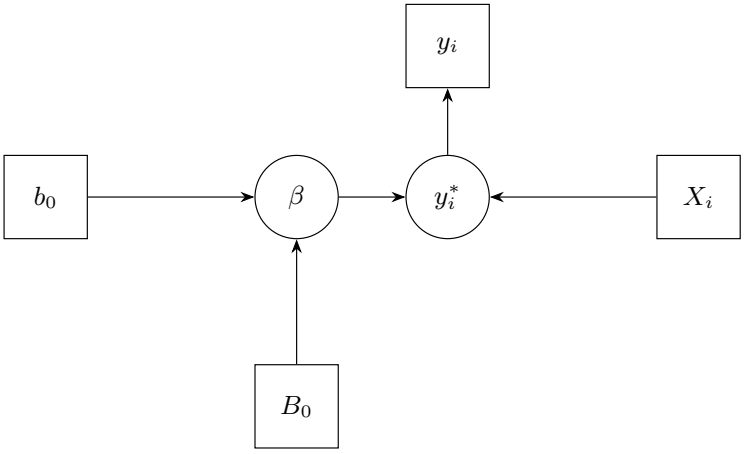
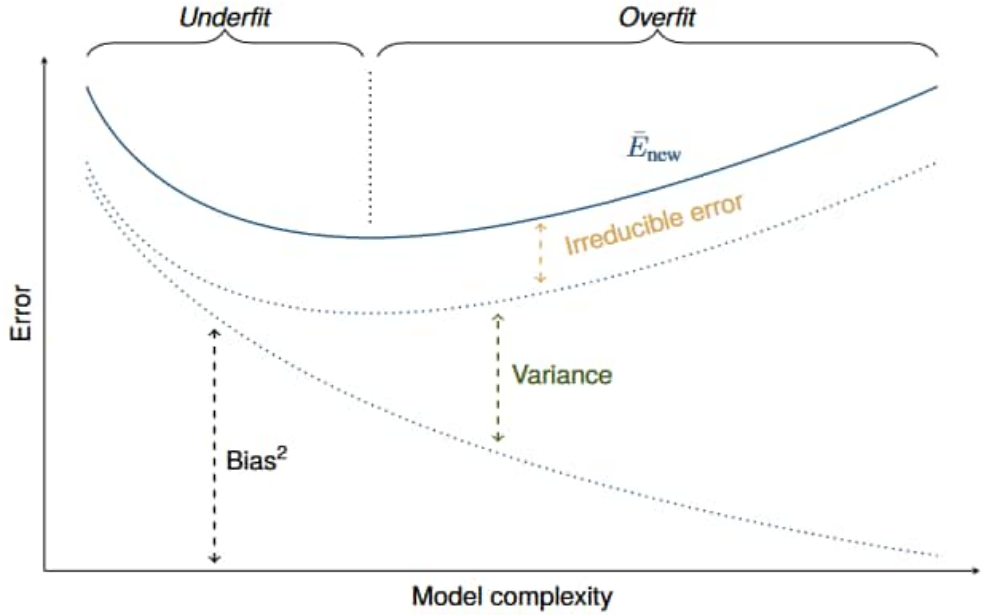


FIGURE 1.2: Model complexity vs. error: Bias-variance trade-off

This graph illustrates the bias-variance trade-off in machine learning. As model complexity increases, bias decreases (the model fits the training data better), but variance increases (the model becomes more sensitive to noise in the training data). The goal is to find the optimal balance between bias and variance. The bias-variance trade-off highlights the challenge of building models that generalize well to unseen data. The irreducible error represents the inherent noise in the data that cannot be reduced by any model.



1.8.2 Tables

TABLE 1.1: Comparison of full models for binary choices (ML results)

The ML estimates are presented for the logistic model, the probit model, the complementary log-log (cloglog) model and for the linear regression (OLS) model with 95% confidence intervals in square brackets.

	Dependent variable:			
	success			
	logit	*probit*	*cloglog*	*OLS*
X1	.873***	.508***	.487***	.129***
	[.808, .938]	[.471, .545]	[.452, .522]	[.121, .137]
X2	1.334***	.772***	.726***	.195***
	[1.260, 1.408]	[.731, .813]	[.685, .767]	[.187, .203]
Constant	1.263***	.737***	.310***	.706***
	[1.198, 1.328]	[.702, .772]	[.279, .341]	[.698, .714]
Observations	8,000	8,000	8,000	8,000
Log likelihood	−3,593.354	−3,590.658	−3,615.357	−3,821.023
AIC	7,192.708	7,187.316	7,236.714	7,650.046

Note: *p<0.1; **p<0.05; ***p<0.01

TABLE 1.2: Comparison of null models for binary choices (ML results)

The ML estimates are presented for the logistic null model, the probit null model, the complementary log-log (cloglog) null model and for the linear regression (OLS) null model with 95% confidence intervals in square brackets.

	Dependent variable:			
	success			
	logit	*probit*	*cloglog*	*OLS*
Constant	.877***	.542***	.203***	.706***
	[.828, .926]	[.513, .571]	[.176, .230]	[.696, .716]
Observations	8,000	8,000	8,000	8,000
Log likelihood	−4,844.679	−4,844.679	−4,844.679	−5,061.253
AIC	9,691.359	9,691.359	9,691.359	10,126.510

Note: *p<0.1; **p<0.05; ***p<0.01

TABLE 1.3: Average marginal effect for binary choice models (ML results)

For the full models of logit, probit, cloglog, and OLS the average marignal effects with 95% confidence intervals in square brackets.

	Dependent variable:			
	success			
	logit	*probit*	*cloglog*	*OLS*
X1	.128***	.128***	.125***	.129***
	[.120, .136]	[.120, .136]	[.117, .133]	[.121, .137]
X2	.195***	.195***	.187***	.195***
	[.187, .203]	[.187, .203]	[.179, .195]	[.187, .203]

Note: *p<0.1; **p<0.05; ***p<0.01

TABLE 1.4: Comparison of full models for binary choices (Bayesian results)

The Bayesian estimates are presented for the logistic model, the probit model, the complementary log-log (cloglog) model and for the linear regression (OLS) model with 95% credible intervals in square brackets.

	Dependent variable:			
	success			
	logit	*probit*	*cloglog*	*OLS*
X1	.874	.508	.487	.128
	[.809, .939]	[.472, .544]	[.453, .523]	[.120, .137]
X2	1.333	.772	.726	.195
	[1.256, 1.1403]	[.734, .812]	[.698, .765]	[.187, .204]
Constant	1.263	.737	.310	.706
	[1.199, 1.327]	[.700, .770]	[.279, .341]	[.697, .714]
Observations	8,000	8,000	8,000	8,000
Log likelihood	−3,593.355	−3,590.660	−3,615.357	−3,821.025
AIC	7,192.709	7,187.320	7,236.715	7,650.050

TABLE 1.5: Comparison of null models for binary choices (Bayesian results)

The Bayesian estimates are presented for the logistic null model, the probit null model, the complementary log-log (cloglog) null model and for the linear regression (OLS) null model with 95% credible intervals in square brackets.

	Dependent variable:			
	success			
	logit	*probit*	*cloglog*	*OLS*
Constant	.877	.542	.202	.706
	[.830, .927]	[.514, .570]	[.174, .230]	[.696, .716]
Observations	8,000	8,000	8,000	8,000
Log likelihood	−4,844.679	−4,844.680	−4,844.679	−5,061.253
AIC	9,691.359	9,691.359	9,691.360	10,126.510

TABLE 1.6: Average marginal effect for binary choice models (Bayesian results)

For the full models of logit, probit, cloglog, and OLS the average marignal effects with 95% credible intervals in square brackets.

	Dependent variable:			
	success			
	logit	*probit*	*cloglog*	*OLS*
X1	.128	.128	.125	.128
	[.120, .136]	[.120, .136]	[.118, .133]	[.120, .136]
X2	.195	.194	.187	.195
	[.187, .202]	[.187, .202]	[.180, .193]	[.187, .204]

Chapter 2

Automated Bayesian variable selection methods for binary regression models with missing covariate data

Note

This chapter has been published in the journal "AStA Wirtschafts- und Sozialstatistisches Archiv" co-authored by Christian Aßmann, see (Bergrab & Aßmann, 2024).

2.1 Introduction

In regression modeling a long-standing problem has been to select an appropriate model in terms of the considered set of conditioning variables. The selection of appropriate variables is always related to the associated selection of models, where various approaches arising in the domain of statistical or machine learning algorithms are discussed in the literature to solve this task. While model selection is in principle straightforward in terms of a decision theoretic approach typically pursued in the context of model averaging, see Hansen (2007), implementation of such a model selection strategy considering all possible model setups is often impossible given available computing capacities. Since the seminal paper of Schwarz (1978) providing a benchmark criterion for model complexity obeying Occam's razor and allowing for model comparison on a common scale, many papers have addressed variable selection in frequentist and Bayesian model setups, see among others Bottolo and Richardson (2010), Ishwaran and Rao (2005), O'Hara and Sillanpää (2009), Raftery (1995), and Tibshirani (1996). Clyde and George (2004) depict variable selection problems become a special part of model selection corresponding every subset of covariates to a distinct model, and finally every selection problem is a description of uncertainty to the data.[1] To avoid computationally difficulties when considering all possible models, selection methods help to pick up relevant

[1] Model selection is further often related to prediction and estimation performance, which in the sense of model fitness is used as a model selection criterion. In the context of variable selection, the prediction and estimation performance relate to the ability to identify the relevant set of covariates and the implied regression relation correctly with higher prediction and estimating performance indicating higher quality of the applied approach. A side aspect of quality relates to the conceptual stringency and interpretability of the considered approach.

variables and reduce the amount of variables that become part of the modeling process.[2] Further, formalized model selection strategies guard against ad-hoc multiple testing approaches invalidating the use of p-values and informal model assessment potentially results in incorrect inference due to strong multicollinearity among the set of variables. This points to the choices with regard to the set of variables considered within model selection. Next to the set of actually observed variables, say X, any combination of the variables in terms of higher order moments and cross products, or functional transformation thereof could be considered as well to handle nonlinear relationships. This increases also for moderate P the number of variables to be considered. While Hansen (2007) and Frühwirth-Schnatter (2010), as typical for the applied literature, consider complete data scenarios, model selection strategies at least in the field of surveyed and administrative data should also address potentially incomplete data, i.e., the estimation strategy requires handling of missing data.

Typical formalized approaches to variable selection discussed in the statistical and machine learning literature are shrinkage estimators, with Lasso, Ridge regression, and Elastic net as the prominent variants. Next to established procedures such as stepwise selection, see Marill and Green (1963), also spike-and-slab prior formulations have been suggested, see Ročková and George (2018) and O'Hara and Sillanpää (2009) for an overview. As pointed out by Korobilis and Shimizu (2022) shrinkage estimators can be well aligned to the Bayesian estimation paradigm, where the different penalization terms correspond to assumed prior distributions, whereas the considered loss functions correspond to likelihood functions. However, the Bayesian estimation approach can be readily extended to handle missing values via the device of data augmentation, see Tanner and Wong (1987). Data augmentation in combination with Markov chain Monte Carlo methodology (MCMC) allows for derivation of estimators via sample averages. Further, the more complex the assumed likelihood structures are, the more compelling MCMC approaches to provide estimators may become.[3] When considering binary data or hierarchical model structures, the involved loss and likelihood functions, serving as optimization criteria in the statistical and machine learning context, become more complex although in principle straightforward to handle via MCMC techniques, see among others Aßmann and Boysen-Hogrefe (2011).

In this article, we hence illustrate how model selection for binary regression models can be performed simultaneously to handling missing values in the considered set of covariate data in a Bayesian framework. Comparison is provided regarding alternative statistical and machine learning approaches arising in the context of shrinkage estimation, such as Lasso, Ridge regression, and Elastic net regression for binary dependent variables. We provide the corresponding Bayesian approach based on a MCMC implementation for the handling of missing values accomplished in conjunction with estimation and variable selection and review the close relationships between shrinkage estimators. The described approach uses

[2]Once a statistical modeling has been agreed upon, model selection problems can also be classified in terms of the number of variables (P) and number of observations (N). Thereby, the case with $N >> P$ typically arises for survey data, whereas the case with $P >> N$ can be found for medical data.

[3]With more complex model structure also the burden of involved numerical optimization may rise. In combination with efforts to handle missing values via multiple imputation, see Rubin (1984), the computational costs increased considerably.

classification and regression trees to approximate the full conditional distribution of missing values. The holistic Bayesian approach allows for the incorporation of prior uncertainty and the flexibility to consider any function of observed or augmented data within the set of conditioning variables.

We assess the quality of different variable selection approaches when missing values occur, where the considered shrinkage estimation approaches are combined with multiple imputation, via a simulation study and an empirical illustration. As indicators of quality, we use indicators assessing the prediction performance regarding variable selection. The results suggest that for simple setups in terms of numbers of variables, missing mechanism, and dependency structures, all considered approaches perform well with regard to model selection. The more complex the setups becomes, the less reliable standard model selection approaches work, where especially inference is more accurate for the suggested Bayesian approach based on spike-and-slab priors and simultaneous consideration of missing values. The empirical application illustrates that the considered Bayesian approach is well suited to provide weights that can be used in subsequent analyses. A main finding of the paper is hence that the quality of variable selection approaches remains high in principle even in the context of incomplete data situations. The paper also documents the required computational resources to apply estimation and model selection of this kind. Finally, the Bayesian approach to handling missing values and variable selection is a powerful and flexible framework that offers several advantages over established methods. By providing a coherent and unified framework, the Bayesian approach enables the comparison of different models on a common scale and the incorporation of prior knowledge and expert opinion. Additionally, standardizing the variables and ranking them based on the effect sizes can provide a simple and intuitive way to interpret the results. However, it is important to consider other concepts of variable importance when interpreting the results to gain a more comprehensive understanding of the relationships between the variables and the response variable.

The paper is organized as follows. Section 2.2 reviews the relationship between shrinkage estimators and Bayesian estimation and provides the suggested Bayesian approach towards model selection in the presence of missing covariate data in terms of a spike-and-slab approach for binary regression models. Also, the details with regard to posterior sampling and corresponding inference are provided. Section 2.3 subsumes the variable selection methods in statistical learning handling missing values in general and section 2.4 adds quality assessments of variable selection. Section 2.5 presents results for simulated datasets and Section 2.6 provides the empirical illustration. Section 2.7 concludes.

2.2 Bayesian estimation for binary regression models with variable selection and handling of missing values

A Bayesian estimation approach in general can be motivated in terms of a decision theoretic approach using a loss function to assess the difference between parameter of interest taking value θ and the corresponding estimator θ^*. A loss function $L(\theta, \theta^*)$ is defined as a mapping of the estimators θ^* from the set of possible estimators and each of the parameter values θ

within the parameter space in the real line. The optimal estimator $\tilde{\theta}^*$ in terms of minimal expected posterior loss is then defined as

$$\tilde{\theta}^* = \arg\min \int_\theta L(\theta, \theta^*) f(\theta|D) d\theta,$$

where $f(\theta|D)$ denotes the posterior density of all parameters of interest θ. The resulting minimization problem is similar to optimization problems arising in the context of penalized estimation if the involved loss function is defined to imply the mode of the posterior distribution and if the target function in penalized estimation (possibly in log scale) is similar to the structure of the posterior distribution proportional to the product of likelihood and prior distribution.[4]

This framing in terms of a decision theoretic approach points out that also all model selection decisions are an integral part of the estimation process with various model selection approaches being discussed within the literature. Standard estimation procedures are typically conditional on one specific statistical model and the dataset under consideration, where properties of the considered data need to be reflected within the statistical model. These properties refer to the scale type of variables within the data, the dimensionality of the dataset, and the completeness. Whereas the scale of the variables is reflected in the considered statistical model, strategies to handle the dimension or incompleteness of the data are typically not an integral part of the estimation routine, but considered in a sequential manner. The straightforward strategy to address all issues simultaneously provided by complete enumeration of all possible models (including all possible subsets of variables and incomplete data constellations) is often hindered by the tremendous computational efforts involved and the intractability of the incomplete data likelihood functions. The computational intensity is one of the main reasons why, from a historic viewpoint, strategies such as stepwise variable selection, both forward and backward, have been discussed in the context of complete data early.[5] Thus, the model selection process involves in complete data situations at most the evaluation of $P(P+1)/2$ model specifications instead of 2^P model specifications to find a maximum or minimum of the underlying break-up criterion. Thereby, the number of models considered within the selection is dramatically reduced although at the cost of path dependency.[6]

For illustration of the mechanisms involved in complete model enumeration, consider a set of models $\{\mathcal{M}_m\}_{m=1}^M$. Given data D, prediction or inference can be based on the asymptotic distribution or the posterior distribution of a parameter estimator for θ being

[4] A direct correspondence is given when the prior is directly comparable in structure to the penalization term and the function assessing model fitness in the penalized context reflects the properties of the negative (logarithmic) likelihood function.

[5] For completeness, note that stepwise selection backwards starts with the most general still manageable model specification involving all P variables and selects from P candidate models, where these P models each leave one variable out. If the predefined model selection criterion detects better fit among the candidate models, the candidate model with the largest increase in model fit is chosen, and the process is repeated until no further increase in model fit is detected. Stepwise selection - forwards or backwards - add or removes just one variable per step that changes the criterion most.

[6] Note that a similar strategy is also inherent to other approaches like classification and regression trees, see Breiman et al. (1984), where splitting points are defined in terms of single variables only.

model specific, i.e., $f(\theta|D, \mathcal{M}_m)$ being specific for the considered models $m = 1, \ldots, M$ with $2^P \leq M$ in case model selection coincide with selecting the appropriate subset of covariate variables, or M even larger than 2^P when alternative model frameworks or missing data patterns are considered additionally.[7] A common pitfall is relying on a single model, especially when multiple models are equally likely but provide differing predictions or inferences.

Bayesian model averaging provides a formal mechanism to aggregate estimation results from different models arising when performing a complete enumeration. Aggregated prediction or inference can be obtained via

$$f(\theta|D) = \sum_{m=1}^{M} f(\theta|D, \mathcal{M}_m) f(\mathcal{M}_m|D),$$

with

$$f(\mathcal{M}_m|D) = \frac{f(D|\mathcal{M}_m) f(\mathcal{M}_m)}{\sum_{m'=1}^{M} f(D|\mathcal{M}_{m'}) f(\mathcal{M}_{m'})}, \quad m = 1 \ldots, M$$

and $f(\mathcal{M}_m|D)$ denoting the posterior and $f(\mathcal{M}_m)$ the prior model probability, whereas $f(D|\mathcal{M}_m)$ denotes the marginal model specific likelihood implied via

$$f(\theta|D, \mathcal{M}_m) = \frac{\mathcal{L}(D|\theta, \mathcal{M}_m) f(\theta|\mathcal{M}_m)}{f(D|\mathcal{M}_m)} = \frac{\mathcal{L}(D|\theta, \mathcal{M}_m) f(\theta|\mathcal{M}_m)}{\int \mathcal{L}(D|\theta, \mathcal{M}_m) f(\theta|\mathcal{M}_m) d\theta}, \quad m = 1, \ldots, M.$$

$\mathcal{L}(D|\theta, \mathcal{M}_m)$ and $f(\theta|\mathcal{M}_m)$ thereby denote the model specific likelihood and prior distributions, respectively. In case the model specification addresses missing values, the likelihood

When a specific model, say $\mathcal{M}_{m'}$, has by far the highest probability aggregated prediction and inference resembles the inference and prediction conditional on model $\mathcal{M}_{m'}$. In case the posterior model probabilities of different models are similar, the aggregated and conditional predictions and inferences will also be similar.

This scheme allows for aggregating inference and prediction from a set of considered models, whether nested or non-nested. However, operationalization and implementation of this scheme require access to the likelihood function as well as tractability of the integration involved to derive the marginal model likelihood. Further, the scheme may be criticized to depend on prior assumption, although the set of considered models may include different prior settings as well for a given likelihood specification. As the computational efforts can become easily prohibitively large, this strategy is applied in the literature in case only relatively small sets of alternative models are considered, see Aßmann and Boysen-Hogrefe (2011), Aßmann (2012), and Frühwirth-Schnatter and Kaufmann (2008), where the computational issues are tractable.

Given the tremendous efforts possibly involved in this general strategy, alternative strategies are discussed in the literature providing model selection based on adaptive strategies. These alternative strategies may consider a restricted class of statistical models only or use alternative model criteria for selection and aggregation purposes which involve tractable

[7] When making prediction or inference, it's essential to recognize that the both are based on a specific model and may be optimal only within the context of the considered models.

computational efforts. In particular, model selection comes down to variable selection, also referred to as feature selection in the literature, when the statistical model is restricted to take the form of a regression model with the conditional expectation of the dependent variable taking the form $X\beta$, where β is a $P \times 1$ vector of parameters. In general, looking on the 2^P possible regression models in total requires to calculate 2^P different comparative measurements, e.g., the Bayes-factor, which leads to model-averaging to get a posterior distribution that takes into account the uncertainty about all M models which requires computing the posterior distribution over the parameters of interest in each model $\mathcal{M}_m$ (Clyde & George, 2004). Additional computation is required for the posterior distribution over all such models. For linear regression models with complete covariate data Hansen (2007) discusses model averaging allowing as well as for model selection. Otherwise, in a Bayesian view ignoring the model or parameter by setting the prior to zero violates Cromwells's rule (Jackman, 2009). Further, for a restricted model class, the model selection issue can be tackled as well by means of shrinkage estimation often also labeled as penalized estimation approaches. In a model-averaging perspective we are interested in all posterior model probabilities for models in which the regression parameters are unequal to zero which leads to variable selection (George, 2000).

In the following, we will consider binary regression models with missing data in covariate variables and discuss shrinkage estimators and Bayesian variable selection approaches to handle the model selection issue arising in form of selecting the most appropriate subset of conditioning variables. The discussion will also point out how these approaches relate to the general strategy. The considered model framework can be described as follows. Let $D = \{y, X\}$ comprise a $N \times 1$ vector $y = (y_1, \ldots, y_N)'$ of binary dependent variables and a $N \times P$ matrix of covariate data $X = (X_1', \ldots, X_N')'$ not including a constant. We will introduce the probit specification for the binary regression model in the following as the involved MCMC sampling scheme is more tractable compared to a corresponding Logit specification. Hence, the binary regression model with probit link is given as

$$y_i = \begin{cases} 1 & \text{if} \quad y_i^* = \alpha + X_i\beta + e_i > 0, \\ 0 & \text{if} \quad y_i^* = \alpha + X_i\beta + e_i \leq 0, \end{cases}$$

where $e = (e_1, \ldots, e_N)'$ is a $N \times 1$ vector of independent standard normally distributed error terms and $y^* = (y_1^*, \ldots, y_N^*)'$ a vector of latent variables. The corresponding likelihood and augmented likelihood functions take the forms

$$\mathcal{L}(y|X, \beta, \alpha) = \prod_{i=1}^{N} \Phi((2y_i - 1)(\alpha + X_i\beta)) \tag{2.1}$$

and

$$\mathcal{L}(y, y^*|X, \beta, \alpha) = \prod_{i=1}^{N} \phi(y_i^* - (\alpha + X_i\beta))\{y_i\mathcal{I}(y_i^* > 0) + (1 - y_i)\mathcal{I}(y_i^* < 0)\}, \tag{2.2}$$

where $\Phi(\cdot)$, $\phi(\cdot)$, and $\mathcal{I}(\cdot)$ denote the cumulative density of the standard normal distribution, the density of the standard normal distribution, and the indicator function, respectively.

The consideration of the augmented likelihood function is based on Albert and Chib (1993) as it simplifies the implementation of a MCMC sampling scheme for estimation. Contextualizing this for variable selection within binary regression models, a representation for all possible model specifications is required. In general, the model setup is implied via the likelihood and the prior distribution which yields to approximate the posterior distribution as proportion of the likelihood and the prior distribution. To describe differences between Bayesian estimators and shrinkage or Maximum Likelihood estimators in general it may be helpful to recall that Bayesian estimators can be formulated as a decision theoretic problem aiming at minimizing Bayes risk. This risk is associated with an appropriate loss function, see Mood et al. (1974). Depending on the loss function, the posterior mean or median are appropriate Bayesian estimators. In this sense, shrinkage estimators may be interpreted as posterior mode estimators, although as pointed out by Gneiting (2011) it might be hard to reconcile this kind of estimator with the decision theoretic duality of loss functions and estimators in case of interaction between penalization and cross-validation. In this sense, and as the posterior distribution is hardly ever accessible by analytical means, Bayesian estimators are typically derived as sample means, where samples from the assumed posterior distribution are obtained using Markov chain Monte Carlo (MCMC) methods. Different MCMC techniques for Bayesian variable and model selection are developed varying the prior distribution or the mechanism in the MCMC sampler, for details see Yang et al. (2005).

To arrive a general model specification encompassing all possible models, the binary regression setup with $\beta = (\beta_1, \ldots, \beta_P)'$ described above can be extended as follows. Following Lee et al. (2003), we use a $P \times 1$ indicator vector γ, where each single γ_j with $j = 1, \ldots, P$ is defined by

$$\gamma_j = \begin{cases} 1, & \text{if variable } X_j \text{ is considered corresponding to } \beta_j \neq 0, \\ 0, & \text{if variable } X_j \text{ is not considered corresponding to } \beta_j = 0. \end{cases} \tag{2.3}$$

Taking γ as a condition into account, the model described in Equations (2.1) and (2.2) would become

$$\mathcal{L}(y|X, \beta, \alpha, \gamma) = \prod_{i=1}^{N} \Phi(2y_i - 1)(\alpha + X_i \text{diag}(\gamma)\beta)$$

and

$$\mathcal{L}(y, y^*|X, \beta, \alpha, \gamma) = \prod_{i=1}^{N} \phi(y_i^* - (\alpha + X_i \text{diag}(\gamma)\beta))\{y_i \mathcal{I}(y_i^* > 0) + (1 - y_i)\mathcal{I}(y_i^* < 0)\}, \tag{2.4}$$

with the diag() operator stacking the indicated vector on the main diagonal of corresponding square matrix. To complete the model setup, priors for α, β, and γ need to be specified when variable selection is considered for binary regression models. The implied posterior can be

described via

$$p(\beta, \alpha, \gamma | y, X) \propto \mathcal{L}(y | X, \beta, \alpha, \gamma) f(\beta, \alpha, \gamma). \tag{2.5}$$

Thereby, all quantitative continuous covariate variables in the dataset are considered to be standardized via a z-transformation.[8] In case an intercept is considered in the model specifications, the intercept is then the common parameter for all possible model and represents the overall mean of the model. Following George and McCulloch (1993), Lamnisos et al. (2009), and Lee et al. (2003) we use a normal prior for α with expected value α_0 and variance h, i.e.,

$$f(\beta, \alpha, \gamma) = f(\alpha) f(\beta, \gamma) = \phi(\alpha | \alpha_0, h) f(\beta, \gamma), \tag{2.6}$$

with $\phi(\cdot|\cdot, \cdot)$ denoting the normal distribution with indicated expectation and variance parameter. Typically, h is set as a large value corresponding to an uninformative prior setting with regard to the intercept (Lamnisos et al., 2009). Table 2.1 outlines the details with regard to hyperparameters of all prior distributions.

The prior setting for β and γ is based on George and McCulloch (1993) assuming an independent marginal conditional setup

$$f(\beta, \gamma) = \prod_{j=1}^{P} f(\beta_j, | \gamma_j) f(\gamma_j). \tag{2.7}$$

The functional forms follow spike-and-slab priors first suggested by Mitchell and Beauchamp (1988) for Bayesian variable selection for normal linear regression models. The according mixture distribution for the coefficients is given as

$$f(\beta_j | \gamma_j) = (1 - \gamma_j) f_1 + \gamma_j f_2, \tag{2.8}$$

where f_1 and f_2 are placeholders for any appropriate continuous or discrete probability density function. Given this mixture form, f_1 is used to steer coefficients to zero (spike), e.g., f_1 assigns a unit point mass at $\beta_j = 0$ and f_2 allows for non-zero coefficients (slab), which can be an absolutely continuous density otherwise such as uniform or normal. Hence, this setup directly incorporates selective shrinkage, i.e., the separating effect in the coefficients caused by the spike-and-slab so that most coefficients are peaked at zero and significant coefficients are set different from zero. A mixture of two normal distributions with different variances are widely used implying

$$f(\beta_j | \gamma_j) = (1 - \gamma_j) \phi(\beta_j | \beta_0, \tau_1^2 \sigma_\beta^2) + \gamma_j \phi(\beta_j | \beta_0, \tau_2^2 \sigma_\beta^2), \tag{2.9}$$

[8] The situation with incomplete covariate data requires that each imputed dataset is subject to a specific z-transformation or that within each iteration of the MCMC algorithm a z-transformation is performed as explained below. In addition, while the use of standardized covariate data is an implicit requirement in the context of shrinkage estimation as a single parameter steers the shrinkage, the Bayesian formulation in form of a variable specific prior allows for more flexibility. Most important, while for continuous quantitative variables standardization is possible using a z-transformation or other stabilizing transformations such as singular value decomposition, standardization is less straightforward implemented for categorical variables.

where typically $\tau_2 >> \tau_1$. Note that different approaches for spike-and-slab prior setups can be found, such as Laplace priors (Tibshirani, 1996) or Horseshoe priors (Carvalho et al., 2010) or setting f_1 to unit mass at zero (George & McCulloch, 1997). Hence, if β_j is found to differ substantially from zero, it will be assigned in the model (slab) or otherwise will be skipped out of the model (spike). Note that the different prior setups correspond to different setups of the penalization function in the context of shrinkage estimators.

George and McCulloch (1993) introduced the Stochastic Search Variable Selection (SSVS) method, which is a Bayesian approach for variable selection in linear regression models. SSVS uses a mixture of two normal distributions as a prior for the regression coefficients, where one distribution has a very small variance and the other has a large variance. This prior encourages shrinkage of the coefficients towards zero, and the stochastic search algorithm explores the model space to identify the most important variables. For the prior for γ specifies the model space we follow George and McCulloch (1993, 1997) and Lee et al. (2003) and consider a Bernoulli prior framework given as

$$f(\gamma) = \prod_{j=1}^{P} w_j^{\gamma_j} (1 - w_j)^{1-\gamma_j} \tag{2.10}$$

with $w_j \in (0, 1)$ governing the probability that the j-th column of X is considered within the regression. It is also common to set $w_j = w$ for $j = 1, \ldots, P$, thereby assuming homogeneity of the inclusion probabilities.

In this case, the prior distribution of γ is binomial and the a priori expected number of selected variables of X can be modeled in terms of w. A fixed value for w can be set if there is consolidated knowledge. If $P >> N$, small values of w are chosen, to bound the number of variables in the model. Hence, the prior penalizes larger models by setting w to a small percentage.[9] Otherwise, a maximum for the model size $P_{\max}$ can be set as in Dobra (2009).[10] In the following, we use the binomial prior on γ with homogeneous inclusion probabilities.

The model setup so far describes the situation with completely observed data and is, as discussed in the literature, accessible to posterior sampling, see also Albert (1992). Data augmentation can also be used to handle missing values. To perform Bayesian inference, an MCMC sampling scheme, see among others Aßmann and Preising (2020), Gelfand and Smith (1990), and Geman and Geman (1984), is implemented to generate a sample from the posterior distributions of interest. Handling of latent structures and missing values is conceptually straightforward in the Bayesian context via the device of data augmentation

[9] In the case of a huge more variables than individuals, a small w selects only a few variables. e.g., a dataset with $10,000$ variables and $1,000$ individuals a value of $w = 0.001$ means that only 10 variables a expected to be selected into the model.

[10] An alternative way is to specify a hierarchical prior distribution for w. Thereby, uncertainty of w can be modeled by implementing a prior for w following a distinct distribution, e.g., a Beta distribution with $\mathcal{B}$ is a Beta function and $w \sim \mathcal{B}(\delta_1, \delta_2)$ (Kohn et al., 2001) so that

$$f(w) = \frac{w^{\delta_1 - 1}(1 - w)^{\delta_2 - 1}}{\mathcal{B}(\delta_1, \delta_2)}$$

with $\mathcal{B}(\cdot, \cdot)$ denoting the Beta function. The prior belief of the model size, i.e., the number if included variables can be parameterized with both $\delta_1 > 0$ and $\delta_2 > 0$.

since the full conditional distributions of missing values can be added as outlined in Aßmann et al. (2023). The prior is thereby also augmented where we opt for a prior for the missing values proportional to the distribution of observed values, see also below. By augmenting the parameter vector with the missing values which are then in turn available as conditions for all other full conditional distributions of interest. In the context of a binary probit models, data augmentation involves augmenting the dependent variable y by drawing a new value y^* from a conditional distribution depending on the other current model parameters, thus the observed binary data is augmented with the latent continuous variable. Data augmentation can be operationalized via including a set of appropriately specified full conditional distribution for the missing values within the MCMC sampling scheme. First the latent variable is drawn or the missing variable are handled, then the model parameters are updated by using the current augmented data. Finally, it allows to estimate the model parameters more accurately and flexibly, and can be used in a variety of applications.[11] This framework has been widely used in various applications, including missing data imputation.

The quantities of interest are hence $y^*, \beta, \alpha, \gamma$, and X_{mis}, where X_{mis} denotes the missing values of the covariate data X with $X = (X_{\text{obs}}, X_{\text{mis}})$. The corresponding posterior of interest results from Equations (2.4) in combination with the assumed operationalizations of Equation (2.6), see also Table 2.1. The prior for the missing values X_{mis} is discussed when providing the assumed full conditional distribution. Starting point for sampling and inference is hence the augmented posterior distribution, see also Aßmann et al. (2023), given as

$$p(\beta, \alpha, \gamma, y^*, X_{\text{mis}}|y, X_{\text{obs}}) \propto \mathcal{L}(y, y^*|X, \beta, \alpha, \gamma) f(\alpha) f(\beta, \gamma) f(X_{\text{mis}}|X_{\text{obs}}).$$

Thus, we draw the posterior values iteratively for $m = 1, \ldots, M$ from the respective full conditional distributions of the considered parameter blocks $y^*, \beta, \gamma, X_{\text{mis}}$. After setting appropriate starting values for $\beta, \gamma, X_{\text{mis}}$, the set of full conditional distributions in the Gibbs sampler is set up as follows. Figure 2.1 shows the schematic progression of the Gibbs sampler with the setting of the start values and the sequential structure of the full conditional distributions.

$f(y^*|\cdot)$ The full conditional distributions of the latent variables y^* corresponds to a product of truncated normal distributions, since the single elements y_i^*, $i = 1, \ldots, N$, are mutually independent. Sampling for each element is hence performed from a truncated normal distribution with moments $\mu_{y_i^*} = \alpha + X_{i,\gamma}\beta_\gamma$ and variance equal to one with the truncated sphere ranging from $-\infty$ to 0 in case $y_i = 0$ and in case $y_i = 1$ ranging from 0 to ∞.

[11] Depending on the scale of the variables under consideration, both parametric and non-parametric models may be appropriate to specify the full conditional distribution of the variables showing missing values. Following Burgette and Reiter (2010) and Doove et al. (2014), we use classification and regression trees (CART) as discussed by Breiman et al. (1984) to approximate the full conditional distributions. This offers a flexible yet computationally feasibly way to model missing values. As previously stated by Aßmann and Preising (2020), data augmentation is employed directly within the MCMC sampler. During the individual draws from the full conditional distributions of an MCMC run, not only are the estimates calculated and a variable selection performed, but also the missing values are imputed. The imputation is conducted using the approach proposed by Tanner and Wong (1987), where a general framework for data augmentation is proposed, which involves introducing latent variables to the model to simplify the estimation of parameters.

$f(\alpha, \beta|\cdot)$ Following Albert and Chib (1993), the full conditional distribution follows in principle the standard Bayesian linear regression given the latent continuous variable y^*. Consideration of the underlying continuous spike-and-slab prior the full conditional distribution has the form of a multivariate normal with variance and expectation given as

$$V_{\alpha,\beta} = (D^{-1} + \tilde{X}'\tilde{X})^{-1} \quad \text{and} \quad m_{\alpha,\beta} = V(D^{-1}(\alpha_0, \beta_0')' + \tilde{X}'y^*)$$

respectively, where D is a diagonal matrix with $D = \text{diag}(h, (\iota_P - \gamma)\tau_1 + \gamma\tau_2)$ with $\iota_\cdot$ denoting a vector of ones with indicated size and $\tilde{X} = (\iota_N,\ X)$, see also Biswas et al. (2022) for a discussion of this full conditional distribution.[12]

$f(\gamma|\cdot)$ The full conditional distribution for γ is implied via the assumed prior structure and corresponds to P independent Bernoulli distributions as the single elements γ_j, $j = 1, \ldots, P$, are mutually independent. The corresponding implied probabilities are given as

$$p_j = \frac{w\tau_2^{-1} \exp\left\{-\frac{\beta_j^2}{2\tau_2^2}\right\}}{w\tau_2^{-1} \exp\left\{-\frac{\beta_j^2}{2\tau_2^2}\right\} + (1-w)\tau_1^{-1} \exp\left\{-\frac{\beta_j^2}{2\tau_1^2}\right\}}, \quad j = 1, \ldots, P,$$

with the hyperparameters $0 < w < 1$ and $\tau_2^2 >> \tau_1^2 > 0$, see Biswas et al. (2022) for discussion and Table (2.1) for chosen values.

$f(X_{\textbf{mis}}|\cdot)$ Values of X_{mis} are sampled sequentially for each column vector of X, i.e., $X = (X^{(1)}, \ldots, X^{(P)})$, based on the non-parametric approximation suggested in the form of classification and sequential regression trees (CART), see Burgette and Reiter (2010). Let $X_{\text{com}}^{(k)} = (X_{\text{obs}}^{(k)}, X_{\text{mis}}^{(k)})$, $k = 1 \ldots, P$, denote the completed variables, and $X_{\text{com}}^{(\setminus k)}$, $k = 1, \ldots, P$, denotes the completed matrix of variables except column k. It is a major advantage of the data augmentation approach that the latent variables possibly serving as kinds of sufficient statistics can be used for the approximation of the full conditional distribution of missing values. In a first step, a decision tree is built for $X_{\text{com}}^{(k)}$ conditional on the corresponding values of all remaining variables $X_{\text{com}}^{(\setminus k)}$ as well as conditional on y and y^* serving as a kind of sufficient statistic for y.[13] In each iteration the covariates are standardized in the spike-and-slab approach as well as in the imputation for the variable selection. To incorporate a prior uncertainty on the hyperparameters of the sequential partitioning regression trees, we build trees with a randomly varying minimum number of elements within nodes. Every missing observation can then be assigned to a node and thus a grouping of observations implied by the binary partition in terms of the

[12] Drawing directly from the distribution is very time intensive if $P >> N$. Following Biswas et al. (2022) the direct drawing routine requires computational cost of $\mathcal{O}(p^3)$ and thus can be modified based on the Woodbury matrix identity summarized by Bhattacharya et al. (2016), which requires still cost of $\mathcal{O}(N^2 p)$ but instead of other approaches converges to the posterior distributions.

[13] Note that this specification is also used when performing imputation in the context of alternative approaches serving as comparative benchmarks.

conditioning variables. The values within each node provide access to an empirical distribution function serving as an approximation to the full conditional distribution of a missing value and thus as the key element for running the data generating mechanism for missing values. Thereby, this modeling approach is in line with prior distributions of missing values proportional to observed data densities. Draws from the empirical distribution function within a node correspond to draws from the full conditional distributions of missing values, where sampling is performed via the Bayesian bootstrap to account for the estimation uncertainty of the full conditional distribution, see Rubin (1981). The considered approach further offers the flexibility to consider any function of observed or augmented data within the set of conditioning variables as well. We incorporate the implementation available within the rpart package available for R (R Core Team, 2020), see Therneau and Atkinson (2018) for further details. The sampled X_{mis} values allow to refer to an updated completed matrix of covariates in all other steps of the MCMC algorithm including a renewed standardization of the covariate data.

Given the MCMC output, estimators are readily defined via corresponding sample moments, either arithmetic means or medians. Whether arithmetic means or medians are reported depends on the loss function involved in defining a Bayesian point estimators based on a Risk function, see Mood et al. (1974). Arithmetic means correspond to quadratic Bayesian loss, whereas medians correspond to absolute Bayesian loss. Furthermore, the model structure implies that estimators conditional on $\gamma_j = 1, j = 1, \ldots, P$, may be considered as well as unconditional estimators. Whereas unconditional estimators can be obtained via using the complete MCMC output, conditional estimators correspond to discarding those draws for which the sample elements in γ are equal to one. In this sense, estimates of the inclusion probabilities reaching at least 50% are necessary to consider a variable as a part of the true underlying data generating process, see also Bottolo and Richardson (2010), Hans et al. (2007), and Russu et al. (2012). Finally, note that within the simulation experiments as well as within the empirical illustration, we set M to equal $20,000$ after a burn-in phase of $5,000$ was found sufficient to discard the effects of initialization both within the simulations study and the empirical illustration. Next, we will discuss variable selection and handling of missing values in the context of shrinkage estimators the relations to Bayesian estimation.

2.3 Shrinkage estimation for binary regression models with missing covariate data

Variable selection as a special case of model selection can be performed in terms of shrinkage estimators. Thereby, the task is to identify a subset of variables that are potential predictors for the dependent variable and is best with respect to predefined optimality criterion. Resulting reduced models give us a higher chance to interpret, visualize and handle the results suitably.

In general, the shrinkage or penalized estimation approach for variable selection is provided in terms of a loss function. Hence, with $\theta = (\alpha, \beta)$, the resulting can be defined as

$$\hat{\theta}_{\text{shrink}} = \arg\min_{\theta}\{\mathcal{LF}(D, \theta) + p_{\text{shrink}}(\beta, \lambda)\}. \tag{2.11}$$

Thereby, $\mathcal{LF}(D, \theta)$ measures the ability of the model to fit the data usually taking a form close to a likelihood function, whereas $p(\beta, \lambda)$ penalizes model complexity, i.e., the dimensionality of the parameter vector β and λ steers the magnitude of penalization. The different shrinkage estimators discussed in the literature differ with respect to alternative specifications of $\mathcal{LF}(D, \theta)$ and $p_{\text{shrink}}(\beta, \lambda)$. In general, the loss function resembles in structure the Mallows criterion, see Mallows (1973).

Prominent choices for measuring model fitness is the residual sum of squares for continuous dependent variables, where the sum of squared residuals is also a constituent part of the log likelihood function given the assumption of multivariate normality or more generally assuming an ellipsoid distribution. The penalization function should be monotonically increasing for larger dimension of β, where typical functions fulfilling this requirement are quadratic or absolute distance functions. Note that these also play a prominent role in logarithms of densities, for example the normal or Laplace density, respectively. Given this, log likelihood functions and log prior distributions operate in the same way as loss and penalty functions respectively, which in turn makes log likelihood functions and priors natural candidates for defining shrinkage estimators. These relationships will be highlighted in more detail, when discussing specific shrinkage estimators in the following. A final remark relates to the mechanism how the consideration of penalty functions causes the selection of a subset of variables. For illustration consider the case, where the function assessing model fitness takes the form of an ellipsoid with fitness decreasing with larger distance to the center of the ellipsoid. The penalty function contributes minimum loss when parameters take the value zero. The overall loss function is thereby minimized via balancing marginal increase in fitness with marginal loss arising from the penalization in a Lagrangean manner. The point where the marginal fitness and penalization contributions can be balanced depends on the chosen functional forms, where opting for absolute loss may provide shrinkage of single parameters exactly towards zero.

Different specifications of $p(\beta, \lambda)$ imply different shrinkage estimators. A general specification for the penalization function can be stated in form of a linear combination of different distance functions or norms, i.e.,

$$p(\beta, \lambda) = \sum_{j=1}^{J} \lambda_j |\beta'\beta|^{\frac{\kappa_j}{2}},$$

where for κ_j taking values $1, 2, \ldots$, i.e., the corresponding L_κ-norms are involved. The following special cases are prominent in the literature. For $J = 1$ and $\kappa_1 = 2$, the penalization function involves quadratic norms L_2 what is referred to as Ridge regression, i.e., $p_{\text{Ridge}}(\beta, \lambda) = \lambda_1 \beta'\beta$, see Hoerl and Kennard (1970) and Friedman et al. (2010) for a discussion in the context of generalized linear models. λ_1 controls the impact of the penalization,

with higher values pushing more coefficients towards zero. If $\lambda_1 \to \infty$, then $\hat{\beta}_{\text{Ridge}} \to 0$, so that the model finally consists only of the intercept. With Ridge regression no genuine variable selection is possible because all coefficients stay in the model but are more or less shrunk towards zero. In the context of the binary probit regression model, the loss function is typically chosen as $\mathcal{LF}(D,\theta) = -\ln \mathcal{L}(D|\theta) = -\ln \mathcal{L}(y|X,\beta) = -\sum_{i=1}^{N} \ln \Phi((2y_i - 1)(\alpha + X_i\beta))$. Given this, the Ridge regression approach towards binary dependent data resembles a Bayesian estimation approach with multivariate normal prior in the sense that the overall optimality function of the Ridge regression has a functional form that is proportional to the logarithm of the implied unnormalized posterior distribution.

A similar situation arises when considering the case with $J = 1$ and $\kappa_1 = 1$, implying the use of absolute values instead of squared ones. Tibshirani (1996) introduced this least absolute shrinkage and selection operator (Lasso) to the linear regression problem, extended to the generalized linear model by Park and Hastie (2007). Given this, the penalization function $p(\beta,\lambda)$ becomes $p_{\text{Lasso}}(\beta,\lambda) = \lambda_1(\beta'\beta)^{\frac{1}{2}}$. Increasing λ_1 sets more coefficients to zero, causing the selection of fewer variables with the selected being shrunk, and finally the number of nonzero coefficients decreases. The analytical solution of the Lasso due to the L_1-norm penalty is more complicated than with L_2-norm.[14] Again, similarity to a Bayesian setup with Laplace prior distribution should be noted when considering the functional forms of the logarithm of the unnormalized posterior density and the overall loss function. Friedman et al. (2010) show that both Lasso and Ridge regression have their drawbacks and advantages in the context of correlated variables and over-fitting. Therefore, Zou and Hastie (2005) proposed the Elastic net approach incorporating to include the best of both. This method uses both L_1- and L_2-penalties and thus is a convex combination of the Ridge regression and Lasso approach towards shrinkage. Friedman et al. (2010) extend the Elastic net to generalized linear models where $J = 2$, $\lambda_1 = 1$, and $\lambda_2 = 2$. After reparametrization the Elastic net manifests as

$$\beta_{\text{ElasticNetMod}} = \arg\min_{\beta}\{\mathcal{LF}(D,\theta) + \lambda(\varphi|\beta'\beta| + (1-\varphi)|\beta'\beta|)^{\frac{1}{2}}\}, \tag{2.12}$$

thereby using φ as control parameter for the weight given to the L_1- and L_2-norm driven penalty with $0 < \varphi < 1$ and $\lambda > 0$ as weighting parameter given to the penalty. The combination causes the Lasso penalization term to select among the variables thereby putting all weight on the set of selected variables, whereas the Ridge term shrinks coefficients towards each other. Hence, Elastic net finds a sparse model with typically high prediction accuracy. Framing this approach from a Bayesian perspective implies that the penalization function is similar to the kernel of the logarithm of a mixture distribution, where the two mixture components follow normal and Laplace distributions, respectively.

As discussed, the Bayesian approach to handling missing values and variable selection, as well as established methods, can set regression coefficients to zero if appropriate. This means that variables are selected by the different approaches, but without ranking the

[14] Note that for handling the problem of (high) correlation among the variables, fused Lasso (Tibshirani et al., 2005) or adaptive Lasso (Zou, 2006) are discussed in the literature.

importance of these variables. However, if using standardized variable values, the size of the regression coefficients can be used as a comfortable and simpler way to rank the variables. This ensures that all variables are on the same scale, which facilitates easy comparison of each variable. This approach is recommended by Kyung et al. (2010) for handling different variables with different measurements. Additionally, as discussed by Friedman et al. (2010), standardizing the variables simplifies the analysis. When ranking variables based on the size of the regression coefficients, it is essential to recognize that this ranking reflects the effect sizes. However, it is also crucial to consider other concepts of variable importance, such as significance and permutation feature importance. These alternative concepts can provide additional insights into the relationships between the variables and the response variable and should be taken into account when interpreting the results (Strobl et al., 2007). By considering multiple perspectives on variable importance, it becomes possible to gain a more nuanced understanding of the relationships in the data.

After standardization we resort to various R packages. For Elastic net estimation the glmnet-package (Friedman et al., 2010) provides many setting options for the as mentioned above control and penalty parameters. As hyperparameter, we set both $\varphi = 0.0$ for Ridge regression penalized estimation and $\varphi = 1.0$ for the Lasso penalty and additionally $\varphi = 0.5$. The strength of penalty is controlled by the tuning parameter λ which can also be set before estimation or via cross-validation which is most widely used and implemented (Friedman et al., 2010). The k-fold cross-validation is used with $k = 10$ folds and a squared loss to use for cross-validation by mean squared error. The shrinkage parameter λ is picked up out of the sequence of possible parameters as the within one standard error of the minimum mean cross-validated error value, so that the most regularized model is given. After chosen the shrinkage parameter the Elastic net is finally estimated with the set control and the cross-validated shrinkage parameter for each m data set. Then, the presented results are averaged over the M datasets.

However, the methods are typically discussed and evaluated under the assumption of fully observed covariate data, so that missing values can be handled in a variety of ways beyond the automated Bayesian approach. First, the data augmentation method within Gibbs sampling is particularly well suited to dealing with missing data problems, since the inclusion of missing values in the parameter vector results in the treatment of all other model quantities as if the data were fully observed. In addition, adding the data augmentation step to the Bayesian estimation routine allows for the avoidance of combination rules (Tanner & Wong, 1987). Second, to cope with missing values in covariate data in the other above-mentioned approaches, we make use of multiple imputation, see Rubin (1976), where we use the multiple imputation via chained equations (MICE) approach, following van Buuren and Groothuis-Oudshoorn (2011). Within the MICE approach, for each variable showing missing values, a full conditional model is specified, where imputations are generated via sampling from these full conditional distributions. Given an appropriate implementation scheme, the MICE algorithm repeatedly iterates over the sequence of assumed full conditional distributions, generates imputations via sampling, and hence updates the data. After an appropriately chosen burn-in phase, obtained draws can be used to build an imputed dataset

that can be used in subsequent analyses. Repeating the MICE algorithm M times provides M imputed datasets given those the shrinkage estimation routines are performed for each of the imputed datasets.

Furthermore, the treatment of missing values leads to further considerations regarding the evaluation of the quality aspects of the different selection and estimation procedures. The issues and challenges associated with multiple imputation appear widely, in some cases, as convergence issues, imputations model mis-specification, imputation of rare events, large amounts of missing data, challenges while reporting and interpreting or issues while pooling the results. Following van Buuren (2018) the general routine associated with MICE of imputing data, analyzing results and pooling results across all M imputed datasets becomes difficult in variable selection because the set of selected variables will differ more or less. In the literature of statistical and machine learning context exists no standard method to pool the results and to combine the information provided by the M different model results. Even for complete datasets, the likelihood-based variable selection methods reach to several limitations (Miller, 2019). In the literature different methods are discussed and for a current overview see Du et al. (2022). Brand (1999) presents a two-step solution which pools the results based on the pooled likelihood ratio p-values selecting insignificant variables after applying stepwise regression to each M imputed dataset and exclude variables from a combined supermodel if they have been selected at least less than half of the runs. Otherwise, (Bayesian) model averaging can be resorted to in order to account for the variability of the selected variables across all imputed datasets (Yang et al., 2005). van Buuren (2018) distinguishes the literature into three general approaches, referring to Wood et al. (2008) and Vergouwe et al. (2010): the Majority approach counting how often a variable is selected (at least half of the models applied to the imputed datasets), and the Stack approach, so that all imputed datasets are stacked into a single dataset applying variable selection methods with weights, and the Wald approach, especially for stepwise selection, pooling based on the Wald statistics.

Considering the above-mentioned variable selection approaches while handling missing data, we present the following routines. First, we present an average-based approach (Average), where the final shrinkage estimator is obtained as the arithmetic mean of the M estimators obtained for each imputed dataset and the combined variance estimator is given as the sum of within and between variance of the M estimators obtained from the imputed datasets.[15] Hence, this procedure is a rough way of pooling estimates, but common in daily practice. As a second approach (Majority), we set up an imputation based on the above-mentioned MICE settings, where we perform variable selection techniques on each imputed dataset m resulting to M different selection models with partly different selected variables. For pooling, we extract the selected variables from each model and sum across the imputations to identify variables that were selected in at least half of the imputed datasets. Then, we estimate a probit model as supermodel with the mostly selected variables and pool results according to the Rubin's rules for generalized linear models. However, this procedure implicitly involves a loss of information. Finally, following van Buuren (2018) we implement a pooling approach based on the Wald test (Wald) and expand the Majority-approach by

[15] Typically rules for asymptotically normally distributed estimators are considered.

extracting the redundant variable by testing. After counting who often a variable is selected in the M imputed datasets, we compare the variable appearing in more than 50% of the M models and apply a Wald test to determine which variable of the sorted counts.

2.4 Quality assessments of variable selection while handling missing values

Value of data is in general linked to the ability to form informed decisions based on the available data information. This value depends on the quality of statistical or machine learning algorithms to use the available data information thoroughly. The proposed Bayesian approach illustrates in the context of a binary regression model the possibilities to integrate all available information into the analysis of factors influencing the binary dependent variable. The proposed method covers the data constellation with many although incompletely observed variables and relatively few observations. The proposed algorithm learns in a machine learning like manner the best combination of covariate variables explaining the dependent variable thereby addressing the entailed problem to decide for a subset of variables and their corresponding influence. To assess the quality in terms of the statistical efficiency of the proposed method, we present a couple of statistical measures typically used to compare different statistical approaches. In the context of variable selection and missing value imputation, the quality aspects refer to the indicators used to assess the performance of different approaches in selecting the correct variables and handling missing values. These quality aspects are typically related to the prediction performance and accuracy of the model. This means that the quality of the approaches is evaluated based on how well they can identify the correct variables while imputing missing values. First, a common quality aspect is to evaluate the performance of variable selection approaches including the accuracy of selection in both model and variable, i.e., the ability of the approach to select the correct variables and exclude irrelevant ones. The following criteria are used to assess the performance of the different strategies within the different scenarios. For evaluation diagnostics of the different approaches, we use the precision rate (PR), the recall rate (RR), and the F-measure assessing the number of covariate variables (in-)correctly identified as (false) true, i.e., whether the decision to incorporate them in the model is in line with the DGP or not. A true positive (TP) is a correctly selected positive, in our case correctly selecting a variable which is indeed part of the assumed DGP. A true negative (TN) is a correctly non-selected non-important one. A false positive (FP) is an incorrect selection that a variable is important, when in fact, was non-important. Finally, false negative (FN) is an incorrect selection that a variable is non-important, when in fact is important. Based on these definitions the above-mentioned performance measurements are defined as

$$PR = \frac{TP}{TP + FP}, \quad RR = \frac{TP}{TP + FN}, \quad \text{and} \quad F = \frac{2 \cdot PR \cdot RR}{PR + RR} = \frac{2TP}{2TP + FP + FN}. \tag{2.13}$$

Note that the F-measure is given as the weighted harmonic mean of the precision and sensitivity. The F-measure balances hence both and is useful if the recall RR has large values,

but the precision PR has small ones.

In addition to the aforementioned considerations, the consistency of variable selection across different iterations in the case of M imputed datasets represents a further quality aspect in shrinkage estimation approaches. But as above mentioned the calculus of counting the selected variables seems not to be a suitable quality aspect. Due to space limitations only the average estimates over the $M = 100$ datasets are reported, thn the biases are straightforward. Note that the number of estimators per shrinkage approaches varies. For instance, the Ridge regression provides M estimates even for redundant variables because estimates are shrunk to zero. In contrast, the Elastic net and the Lasso set the impact of variables to zero. The Bayesian spike-and-slab approach, on the other hand, provides M estimates for all variables.

Generally, the root mean square error (RMSE) over the results from the M datasets is an quality indicator for the robustness not only in the complete cases, but also in the handling of missing values. For a parameter $\hat{\theta}_p$ the RMSE is calculated

$$\text{RMSE}(\hat{\theta}_p) = \sqrt{\text{MSE}(\hat{\theta}_p)} = \sqrt{E[(\beta_p^{\text{true}} - \hat{\theta}_p)^2]} = \sqrt{\frac{1}{D}\sum_{d=1}^{D}(\beta_p^{\text{true}} - \hat{\theta}_p^{(d)})^2}, \tag{2.14}$$

where MSE denotes the mean square error. In the context of Ridge regression and the Bayesian approach, the value of D is equal to M. However, in the case of the Elastic net, Lasso, and stepwise regression, the value of D differs from M for each variable. In these approaches, redundant variables are typically excluded from the model, and in a few datasets, important variables are attenuated, thus D varies.

2.5 Experimental study

The following simulation experiments aim at a comparison of different strategies to achieve variable selection within a binary probit regression framework. The simulations were implemented in R (R Core Team, 2020) and Julia (Bezanson et al., 2017) and compare the performance of the Bayesian variable selection approach - hereafter referred to as Bayesian Spike-and-Slab (SnS) - with the stepwise regression (SR) strategy as implemented in the MASS R package (Venables & Ripley, 2002)) and different variations of Elastic net (EN) like Lasso and Ridge regression as available within the R package glmnet (Friedman et al., 2010). For the shrinkage estimators, we consider for the Elastic net setup different control parameter $\varphi = 0.0$ for Ridge regression (EN.0), $\varphi = 0.5$ for a variation of Elastic net (EN.5), and $\varphi = 1.0$ for Lasso (EN1.). Via cross-validation λ is chosen such that the error is within one standard error of the minimum shrinkage parameter (Friedman et al., 2010). Thereby, stepwise regression strategy depends on the choice of the selection criterion, where we opt for the Bayesian information criterion (BIC). For the Bayesian estimation approach, estimates are each based on MCMC chains of length $40,000$. After discarding the first $10,000$ iterations as burn-in, inference is based on the remaining $30,000$ simulated draws from the joint posterior distribution. Convergence is monitored via Geweke statistics, the Gelman-Rubin statistics,

and the effective sample size, see Gelman et al. (2023) and Geweke (1991). The convergence diagnostics indicate overall convergence.

The simulated data is generated to follow a probit regression model with $N = 1,000$ and $P = 10$. The considered data generating process satisfies the following conditions. Next to a constant, the covariate data $X_{(p)}$ with $1, \ldots, 9$ is generated from standard normal distributions in each variable. Only the first five out of the total ten parameters are set to a non-zero value by setting the indicator vector, accordingly, including the intercept. The signs of the 10 parameters are chosen to alternate. For the sake of variation, X_1 to X_4 are drawn from a multivariate normal distribution with expectation $\mu = (0, 0, 0, 0)'$ and covariance $\text{vech}(\Sigma) = (1, .85, .65, .45, 1, .45, .35, 1, .25, 1)'$ mimicking a situation with correlated covariate data.[16] Finally, a total of 100 simulated datasets are generated through replication.

Based on this data generating process, we consider two different variations for missingness in the covariate data: missing completely at random (MCAR) and missing at random (MAR). Simulating missing values as MCAR, we randomly set 20% of the values in the covariates X_2 and X_3 to missing later named as experimental study 1 (Ex 1) and we randomly set 20% in X_2, 30% in X_3 and 10% in X_3 to missing as experimental study 2 (Ex 2). For the MAR variation (Ex 3), we consider a missing generating mechanism for $X_{2,i}$ where $X_{2,i}$ is missing if $F_U(U_i) > 0.75$, where $F_U(U_i)$ denotes the empirical distribution function of the random variable U_i which is

$$U_i = \frac{1}{1 + \exp\{\omega_{2,i}\}} \quad i = 1, \ldots, N,$$

with $\omega_{2,i} = 0.2X_{2,i} + \rho_i$ and ρ_i being standard normally distributed. Thus, a missing rate of 25% for X_2 is generated. For further details on the described missing designs, see table 2.2.

To assess the different estimation strategies in case of missing values, we designed a comprehensive simulation study which is split in different scenarios. First, we consider a benchmark estimation without missing values labeled as before deletion (BD), followed by estimation of complete cases (CC) only, and finally a scenario with missing values (MIS), where missing values are handled either via multiple imputation before estimation as for the shrinkage estimators or embedded within the MCMC algorithm as for the Bayesian approach.

Regarding estimation results, we provide the true parameter values used in the DGP, mean posterior medians and averaged estimates, respectively over the 100 replications obtained for the BD, CC, MIS scenarios. We report on both the regression coefficients and conditional variance parameters. Beside the averaged estimates, simulation results are also evaluated in terms of the root mean square error (RMSE) and the inclusion probabilities where possible. Therefore, it is important to bear in mind that we assess the models' performance on two levels. Firstly, the model selection, that is, regardless of the classification rule, how well do the different routines predict the initial model parameter based on the DGP. And secondly, how exactly the parameters underlying the data generating process are estimated.

[16] Note that $\text{vech}(\cdot)$ denotes the half-vectorization operator as defined in Lütkepohl (1996).

The results of the different variations are presented in tables 2.3 for experimental study 1 (Ex 1), 2.4 for experimental study 2 (Ex 2), and 2.5 for experimental study 3 (Ex 3) summarizing the three scenarios BD, CC, and MIS with results as average estimates and root mean square errors (RMSEs). Looking on Ex 1, the tables differ only in the results of estimation and pooling of the treatment of missing values, which are examined under quality aspects. As benchmark we report results for a simple probit model covering all important variables $(\alpha, \beta_1, \ldots, \beta_4)$ for the three different experimental studies. In the BD scenario we find overall unbiased results for all parameters, therefore we can assume that the considered routines are implemented correctly. The different pooling approaches produce very different results in terms of accuracy and variation in the estimation results. It is surprising that the results of simple averaging come to significant improvements in the estimation results compared to the before deletion values, which is set as a benchmark for the three pooling strategies. It seems that the majority method tends to over-fit, as the estimation results are clearly too unbiased. This can be seen in all three experimental studies. Compared to our Bayesian spike-and-slab approach, which also treats missing values, there are sometimes large differences in the estimation results compared to the respective pooling approaches. However, the average and Wald methods show comparable patterns in that the RMSEs are better compared to the complete cases, whereas they are in part more distorted compared to the before deletion results. When datasets with many of missing values in several variables are available, the Bayesian spike-and-slab method shows its strengths. The RMSE increases slightly for all methods in the CC scenario, whereas the absolute estimation bias decrease for the three Elastic net methods. The bias of the estimators increases for the stepwise regression method and for the Spike-and-Slab approach. In contrast, imputation shows that performance decreases for all five approaches measured with the RSME. With the Bayesian Spike-and-Slab, the inclusion probability for X_2 increases again, so that the variables are selected with a very high probability, but the bias increases so that the RSME is high compared to the stepwise regression. Interestingly, all three Elastic net approaches do not show altogether large deviations in the imputation scenario due to the combining rules. Thus, the results of stepwise regression and Bayesian Spike-and-Slab are quite similar. The experimental study 2 shows similar results, but due to the higher missing dropout, more biased results are to be expected. This shows the strength of our Bayesian approach, which does not only treat the missing values in the run-up to selection and estimation, making the estimation and selection results more precise and the selection of variables more correct in terms of precision, recall and F-measure. Experimental Study 3 shows less precise results in estimation and selection due to the MAR failure of data for all presented selection strategies, however, the Bayesian spike-and-slab approach can score here by having precision ahead of the other methods.

However, it must be noted that the estimation accuracy is more accurate with stepwise regression (SR) and Bayesian Spike-and-Slab (SnS), and thus the bias is small and the RMSE is half towards the Elastic net (EN) approaches. Table 2.6 shows the results of the calculations of Accuracy: precision, recall, and F-measure for before deletion, complete case and imputation obtained with Elastic net, stepwise regression, and Bayesian Spike-and-Slab. Looking on the important measurement, e.g., F-measure, shows in principle the same result. Accuracy, as

measured for Ridge regression (EN.0) , always yields the same results because all variables are always included in the model and no selection is made, only shrinkage. The spike-and-slab is almost among the top performers in terms of F-measure. This shows that the Ridge approach does not perform selection in the strict sense, but merely shrinks the values close to zero. Thus, the values for the Ridge results never reach the accuracy of the other estimation methods. On the other hand, it is striking that the accuracy values of the Lasso estimates do not approach those of stepwise regression or our Bayesian approach. However, stepwise regression also shows its strengths in correctly selecting appropriate variables that were involved in the data generating process. In general, the accuracy decreases when estimating the complete cases compared to the before deletion values and then increases when dealing with missing values, as intended. Here, it is also shown that the average and the Wald method of pooling yields comparable results to the Bayesian approach and provides the intended accuracy gain of handling missing values compared to the complete cases.

2.6 Empirical illustration

In order to illustrate the usefulness of the suggested Bayesian spike and slab approach in empirical analysis, we provide exemplary applications using the scientific data use file of the German National Educational Panel Study (NEPS), Starting Cohort Grade 9, see NEPS, National Educational Panel Study (2021) and Blossfeld and Roßbach (2019). For this purpose, a random sample of schools, stratified by school type, was drawn throughout Germany. Within the schools, all students of two randomly selected ninth grades were invited to participate in the survey. The technical details on weighting are reported for each wave, see among others Bergrab (2020). Here, variable selection finds its application in selecting the appropriate variables that describe a student's probability of participation in a specific wave, where we analyze wave twelve here. From the set of available covariate variables ($P = 16$), only a few have to be selected in order to determine the participation probabilities and subsequently prepare suitable weights for further analyses. In the starting cohort considered here all students, regardless of whether in vocational education or academic track, willing to participate in the NEPS are followed up over time. The students entering the academic track usually remain within their school context. In contrast, students entering the vocational education leave school for a vocational training. In wave twelve all students left their school context and are surveyed individually. To account for the wave-specific participation decision of students' response propensity re-weighting is used to provide corresponding weights. To model binary participation decisions a model with probit link function is used for all three variable selection methods: backward selection, Elastic net, Bayesian spike-and-slab. By wave twelve, the panel cohort has reduced to $n = 7,911$ students in the age of mean 26.76 (standard deviation 0.73). For our analysis, we included all students and $p = 16$ variables. In contrast to the weighting report, all variables were standardized before selection to show comparability with the above-mentioned experimental studies.

Table 2.7 summarizes the results for stepwise regression, Elastic net, and Bayesian spike-and-slab (Bayesian SnS) approach. To model individual participation, for the stepwise

regression the `glm`-function with a probit link provided in R (R Core Team, 2020) was used. BIC based backward selection was used and only significant coefficients are reported. For Elastic net only non-negligible variables are reported. The dash indicates that this variable was not included in the estimation model. The shrinkage parameter of Elastic net λ is set to the largest value such that the error is 1 standard error of the minimum: $\lambda_{\text{min}} = 0.015$ obtained with the `cv.glmnet`-function in R. The Elastic net represents a Lasso selection with a control parameter of 1.0. For the results of our Bayesian approach both, the estimates $\hat{\beta}$ as median of posterior and the corresponding marginal posterior inclusion probabilities $\hat{\gamma}$, are presented. The bold results in the last two columns show variables with an inclusion probability higher than 50%, where the other results are listed for the sake of completeness. The prior setting for the spike-and-slab were set to $\tau_2 = 1 >> \tau_1 = 0.015$ and the shrinkage $w = 0.015$. The posterior estimates are based on MCMC chains of length $20,000$. After discarding the first $5,000$ iterations as burn-in, inference is based on the remaining $15,000$ simulated draws from the joint posterior distribution. The convergence diagnostics indicate overall convergence.

Table 2.7 show less differences among the three approaches. Participation in the last waves is selected according to all three approaches, except for participation in wave 6, which is not selected by Elastic net. The total estimate and the selection of migration background vary across the three models. The most parsimonious model is determined by Elastic net and our spike-and-slab approach, which includes a total of six variables. In contrast, stepwise regression only includes one additional variable in the model. The selection of stepwise regression shows overall significant results, whereas the Bayesian spike-and-slab approach extracts migration background as a redundant variable with an inclusion probability of $\hat{\gamma} = .227$. The remaining inclusion probabilities indicate a markedly distinct decision to include. The model based on elastic net does not consider participation in Wave 6. However, it includes all other variables that are also selected by stepwise regression and the Bayesian approach. An analysis of the inclusion probabilities reveals that, in addition to the intercept, positive participation in the surveys in waves 9, 10, and 11 also strongly influences the probability of participating in wave 12. The inclusion probabilities are 1 or nearly 1 for all four variables. This demonstrates that stepwise regression and the Bayesian approach yield nearly identical outcomes, as the most crucial variables are selected in a comparable sequence. The advantage of the Bayesian approach is that, in addition to the assessment based on the significance level in stepwise regression, the inclusion probability offers a more direct approach to the evaluation.

It is important to recognize that the Bayesian approach is susceptible to the influence of prior beliefs, which requires further investigation. The impact of the variance priors on the values of β and γ is negligible, as well as the starting values. However, when holding $\tau_2 = 1$, the exploration of the gradual adjustment of the hyperparameter τ_1 from 0.050 to 0.200 in steps of 0.0.025 reveals sensitive changes. The results of this can be found in table 2.8. The Bayesian spike-and-slab algorithm is sensitive to the specific choice of spike, i.e., τ_2; in detail, climbing up the increments on τ_2 keeps the variables participation in the last wave and migration background in the model with constant high inclusion probabilities. Furthermore,

the estimates of the latter two variables alternate, but the estimate for participation in the previous wave is extremely constant. The other selected variables vanish out gradually. It is noteworthy that a similar outcome is achieved when the shrinkage parameter of Elastic net, defined as $\lambda_{\text{min}} = 0.015$, is employed in conjunction with the Bayesian spike-and-slab approach.

As previously discussed, these methods will set regression coefficients to zero, if necessary. The three approaches compared provide results that are essentially similar, although differences in depth are apparent. The results offer an intriguing perspective on the Bayesian approach to handling missing values, as well as variable selection and its performance relative to established methods for variable selection. Further investigations regarding the weights calculated with the models were not carried out.

2.7 Conclusion

This paper provides assessment of different strategies, i.e., applying statistical as well as machine learning algorithms, for variable selection in the context of binary regression models with missing values in the covariate variables. We show how our algorithmic strategies can be combined and how they can accommodate inference over the prior inclusion probability and which prior settings affect the posterior estimates. To handle missing values in the Bayesian estimation paradigm, the device of data augmentation can be used. The discussion of the various strategies highlights similarities and differences between shrinkage estimators and Bayesian estimation approaches. The choice of hyperparameters is in all methods a sensitive issue. The tuning parameter for using shrinkage estimators is typically estimated by cross-validation, whereas the hyperparameters in Bayesian estimation are fixed a priori.

From a methodological view, the conceptual strengths of the Bayesian spike-and-slab model and the Ridge regression are revealed in the matter that they do not exhaust predictive information in trying to determine which variables have exactly zero effect. Any attempt to select variables after the fact, as is done in Lasso or Elastic net, does not lead to the loss of information to worse predictions. Therefore, in the Bayesian setting it is important to understand that the set of selected variables that have no predictive effect has a probability of (approximately) zero. Whereas all variables that have an inclusion probability that is above a certain threshold can be considered truly predictive. The evidence provided in the empirical illustration suggests that for the purposes of weighting in the $N > P$ setting, all strategies work well, but with small advantages of the Bayesian approach, especially when missing values occur in the covariate data. As discussed in Du et al. (2022) our Bayesian approach, simultaneously imputing the missing values, selecting and estimating jointly the model parameters, is time-consuming and computationally intensive not only for a large amount of missingness. Therefore, we show that dealing with missing values in the context of statistics and machine learning requires a convincing strategy for imputing the missing values. Not considering the missing data pattern as well as the averaging of imputed estimation results without taking pooling rules into account leads to a loss of quality, which can be seen e.g., in biased estimation results or lower variances of the estimators. Hence, case-deletion

or complete-case strategies are frequently used where individuals are excluded from the analysis, if they are missing any of the variables or items. whereas the quality of the data analysis suffers as a result. As shown, non-Bayesian variable selection approaches cannot be easily applied within the framework of multiple imputation. The difficulty lies in how to combine the selection results across the multiply imputed data sets, because non-Bayesian variable selection approaches would commonly identify different redundant coefficients for the various imputed datasets and thus will lead to different numbers of coefficients to compare.

In conclusion, we present a strategy clearly combining estimation, shrinkage and handling of missing values. The Bayesian holistic method for combining variable selection and imputation of missing values in covariates offers several advantages over traditional methods. By treating missing values as parameters and assigning priors to them, this approach provides a more accurate and reliable estimation of regression coefficients. While it may appear to users that the Elastic net and stepwise regression approaches are assumption-free, there are nevertheless assumptions involved regarding the setting of the control and shrinkage parameters, which are set via the priors in the Bayesian approach. Likewise, the pooling step based on chained equation is non-trivial, while in the Bayesian approach this can be implemented as an additional step in the iterative Gibbs sampler.

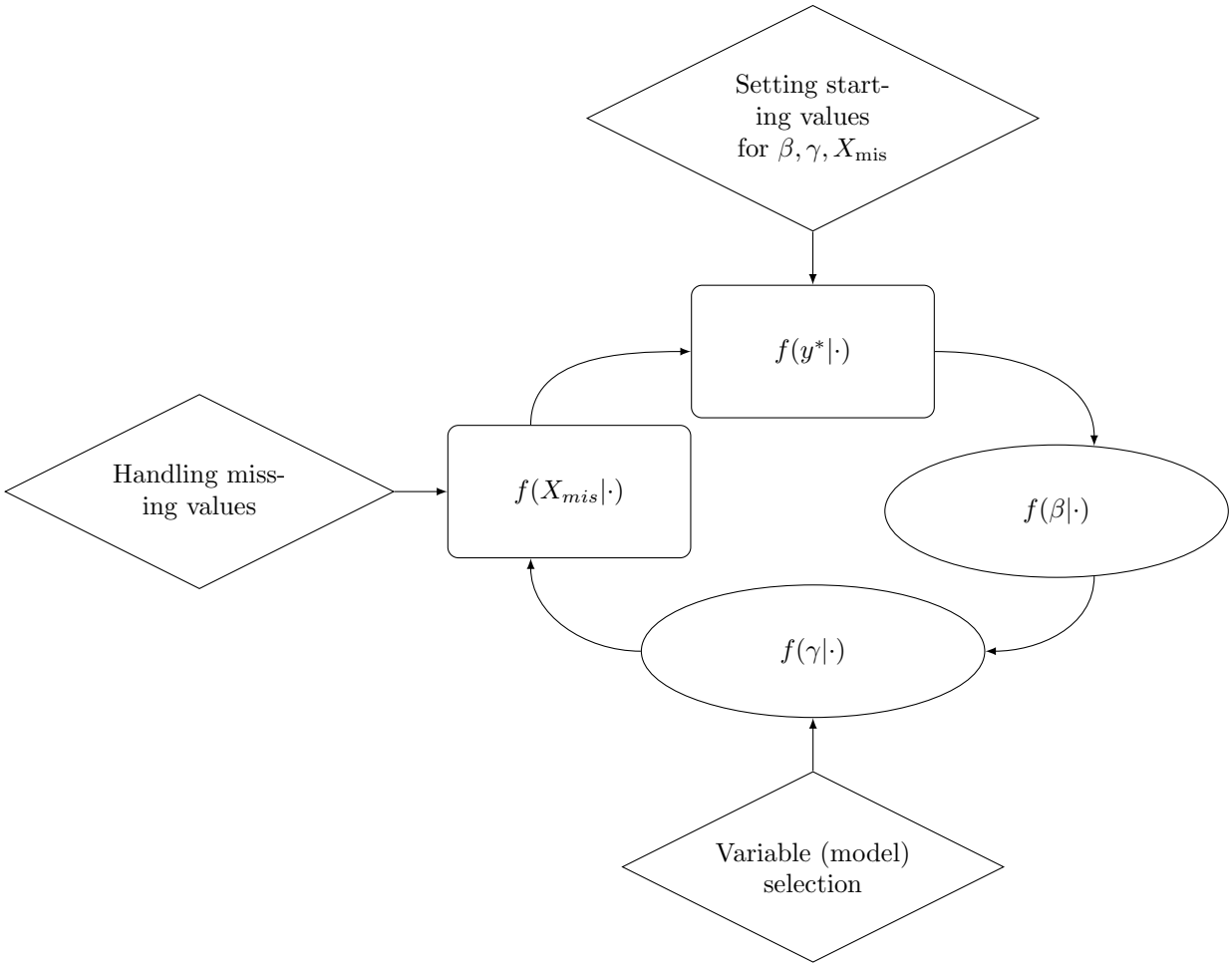
There are several approaches in the literature for imputing missing values and selecting variables. Note that Heymans et al. (2007) suggest a boot-strapped variable selection under multiple imputation to overcome the pitfalls of applying the combining rules to stepwise regression and Panken and Heymans (2022) provide similar frameworks for the logistic setup based on the majority based approach dividing the imputed datasets in test and train data. Chen and Wang (2013) extend the Lasso to multiple imputation with grouping the imputed data, combining multiple imputation and random Lasso see Liu et al. (2016), and combining Lasso with the Expectation-Maximization algorithm (Sabbe et al., 2013). The presented Bayesian holistic method highlights the potential of combining imputation with advanced techniques with accuracy and reliability in statistical results and offers promising avenues for future research.

2.8 Appendix section 2

2.8.1 Figures

FIGURE 2.1: Schematic progress of Gibbs sampler

Schematic progress of the sequential structure of the full conditional distributions within the Gibbs sampler. Note that the starting values are set once before starting the $m = 1, \ldots, M$ iterations. The full conditional distributions in the ellipses express extended data, whereas the rectangular blocks represent the full conditional distributions, which provide the output of interest. After subtracting an appropriate burn-in phase, both provide the corresponding estimators based on the median or mean.



2.8.2 Tables

TABLE 2.1: Prior specification and MCMC starting values

Parameter	Functional form	Probability distribution	Starting values
Intercept			
α	$\propto \phi(\alpha_0 = 0, \sigma_\alpha^2 h)$ depending on scaling parameter h wet set to $h = 1$, thus	Normal	-
σ_α	$\propto IG(c_1, c_2)$ with c_1 and c_2 const.	Inverse gamma	-
Regression vector			
β_p	$\propto (1-\gamma_j)\mathcal{N}(\boldsymbol{\beta}_0, \tau_1^2\sigma_\beta^2) + \gamma_j\mathcal{N}(\boldsymbol{\beta}_0, \tau_2^2\sigma_\beta^2)$ depending on $\boldsymbol{\beta}_0 = 0$	Mixing normal	$\{1\}_{p=1}^P$
	$\tau_2 >> \tau_1 > 0 \quad$ with $\quad \tau_2 = 1$ and τ_2	Constant	Set individually
σ_β	$\propto IG(d_1, d_2)$ with d_1 and d_2 const. mostly $d_1 = 100$, and $d_2 = 100$	Inverse gamma	-
Indicator vector, i.e., spike-and-slab			
γ	$\propto$ Bernoulli(w) depending on	Bernoulli	$\{1\}_{p=1}^P$
	$w \in (0, 1)$	Constant	Set individually
Missing values			
X_{mis}	$\propto$ observed sample distribution	Nonparametric	Random draws

The hyperparameters for the inverse gamma distribution are chosen to provide finite variance and smallest possible prior sample size.

TABLE 2.2: Overview of the missing design of the experimental studies

Design	Missing mechanism	Total missing rate (%)	Results
Ex 1	$Pr(X_{2,i} = missing) = 0.2$	35.99[1]	Table 2.3
MCAR	$Pr(X_{3,i} = missing) = 0.2$		
Ex 2	$Pr(X_{2,i} = missing) = 0.2$	49.51[1]	Table 2.4
MCAR	$Pr(X_{3,i} = missing) = 0.3$		
	$Pr(X_{4,i} = missing) = 0.1$		
Ex 3	$X_{2,i} = missing$ if $1/(1 + \exp(-\omega_{2,i}))$	25.00	Table 2.5
MAR	$w_{2,i} = 0.2X_{2,i} + \rho_i$ and $\rho_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$		

The experimental studies Ex 1 and Ex 2 can be characterized as missing completely at random (MCAR) and experimental study Ex 3, where the missing probability depends on the variable itself as missing at random (MAR). All simulation runs have been performed with $40,000$ Gibbs iterations with the first $10,000$ iterations as burn-in.

[1] Average over $M = 100$ datasets.

TABLE 2.3: Experimental study 1: Results

Results of MCAR with 20% missing rate in X_1 and X_2, and with correlation between $X_1, \ldots, X_4$, and with true parameter values, mean posterior medians and root mean squared errors (RMSEs) of important $(\alpha, \beta_1, \ldots, \beta_4)$ and non-important $(\beta_5, \ldots, \beta_9)$ regression coefficients over 100 replications obtained by Elastic net with control parameter $\varphi = 0.0$ for Ridge regression (EN.0), $\varphi = 0.5$ for a variation of Elastic net (EN.5), and $\varphi = 1.0$ for Lasso (EN1.) and stepwise regression (SR) and Bayesian spike-and-slab (SnS), where the first column presents the estimates $(\hat{\beta})$ and the second one the inclusion probabilities $(\hat{\gamma})$. The prior setting for the spike-and-slab were set to $\tau_2 = 2 >> \tau_1 = 0.2$ and controlling the number of selecting variables with $w = 0.5$.

	Average estimates							RMSEs					
true values	probit	EN.0	EN.5	EN1.	SR	SnS		probit	EN.0	EN.5	EN1.	SR	SnS
Before Deletion													
$\alpha = 0.5$	.513	.501	.509	.508	.515	.519	1.000	.046	.043	.045	.045	.047	.048
$\beta_1 = -0.5$	-.514	-.362	-.434	-.430	-.516	-.511	.993	.108	.156	.130	.129	.108	.106
$\beta_2 = 0.5$	.518	.384	.449	.447	.520	.518	.999	.096	.135	.109	.106	.096	.096
$\beta_3 = -0.5$	-.505	-.496	-.507	-.508	-.507	-.511	1.000	.058	.047	.056	.057	.058	.059
$\beta_4 = 0.5$	.513	.459	.489	.488	.514	.517	1.000	.055	.062	.056	.057	.056	.058
$\beta_5 = 0.0$	-	-.001	.000	.000	.000	.009	.060	-	.041	.040	.038	.088	.045
$\beta_6 = 0.0$	-	-.003	-.003	-.003	-.012	.007	.058	-	.041	.040	.038	.085	.045
$\beta_7 = 0.0$	-	.000	-.001	.000	.015	-.010	.070	-	.045	.044	.042	.091	.049
$\beta_8 = 0.0$	-	.008	.007	.007	.041	.019	.066	-	.043	.041	.039	.089	.048
$\beta_9 = 0.0$	-	-.002	-.001	-.001	.002	.008	.070	-	.046	.044	.043	.083	.050
Complete case													
$\alpha = 0.5$	.513	.499	.505	.505	.515	.522	1.000	.063	.060	.062	.062	.047	.064
$\beta_1 = -0.5$	-.514	-.359	-.399	-.399	-.516	-.514	.975	.137	.169	.173	.179	.108	.142
$\beta_2 = 0.5$	.523	.387	.425	.425	.520	.526	.992	.128	.145	.148	.153	.096	.131
$\beta_3 = -0.5$	-.512	-.503	-.514	-.515	-.507	-.521	1.000	.079	.064	.075	.077	.058	.082
$\beta_4 = 0.5$	.512	.458	.477	.479	.514	.519	1.000	.066	.070	.051	.066	.056	.069
$\beta_5 = 0.0$	-	-.002	-.002	-.001	.000	.012	.162	-	.054	.051	.049	.088	.059
$\beta_6 = 0.0$	-	.002	.002	.003	-.012	.015	.168	-	.053	.050	.048	.085	.059
$\beta_7 = 0.0$	-	-.003	-.001	-.001	.015	.011	.164	-	.053	.050	.049	.091	.058
$\beta_8 = 0.0$	-	.006	.005	.006	.041	.021	.160	-	.050	.047	.045	.089	.056
$\beta_9 = 0.0$	-	-.004	-.004	-.004	.002	-.008	.175	-	.056	.052	.050	.083	.060
Imputation - average method[a]													
$\alpha = 0.5$	.514	.501	.509	.509	.515	.511	1.000	.048	.044	.046	.046	.048	.045
$\beta_1 = -0.5$	-.512	-.361	-.432	-.431	-.516	-.344	.999	.135	.166	.155	.154	.135	.187
$\beta_2 = 0.5$	.515	.384	.447	.447	.518	.378	1.000	.124	.146	.133	.132	.124	.153
$\beta_3 = -0.5$	-.505	-.496	-.506	-.508	-.507	-.546	1.000	.061	.049	.058	.060	.061	.078
$\beta_4 = 0.5$	.512	.459	.488	.488	.513	.491	1.000	.058	.064	.059	.058	.059	.053
$\beta_5 = 0.0$	-	-.001	-.002	-.002	.004	.002	.175	-	.042	.040	.039	.076	.044
$\beta_6 = 0.0$	-	-.003	-.004	-.004	-.027	.000	.174	-	.041	.040	.039	.085	.044
$\beta_7 = 0.0$	-	.000	-.001	.000	.032	.003	.170	-	.045	.043	.042	.085	.048
$\beta_8 = 0.0$	-	.008	.007	.007	.014	.011	.170	-	.043	.042	.040	.076	.047
$\beta_9 = 0.0$	-	-.002	-.001	-.001	-.009	.001	.184	-	.046	.044	.042	.067	.050
Imputation - majority method													
$\alpha = 0.5$	-	.517	.517	.517	.515	-	-	-	.049	.049	.049	.048	-
$\beta_1 = -0.5$	-	-.516	-.522	-.526	-.519	-	-	-	.138	.134	.133	.132	-
$\beta_2 = 0.5$	-	.520	.517	.513	.514	-	-	-	.125	.130	.139	.128	-
$\beta_3 = -0.5$	-	-.508	-.509	-.511	-.508	-	-	-	.061	.065	.068	.064	-
$\beta_4 = 0.5$	-	.515	.514	.513	.512	-	-	-	.059	.060	.060	.059	-
$\beta_5 = 0.0$	-	-.001	-.001	-.002	-.023	-	-	-	.045	.051	.055	.091	-
$\beta_6 = 0.0$	-	-.004	-.006	-.008	-.023	-	-	-	.045	.052	.055	.089	-
$\beta_7 = 0.0$	-	.000	-.001	-.001	.021	-	-	-	.048	.054	.057	.096	-
$\beta_8 = 0.0$	-	.009	.011	.011	.050	-	-	-	.047	.054	.055	.092	-
$\beta_9 = 0.0$	-	-.002	-.002	-.002	.003	-	-	-	.050	.054	.056	.092	-
Imputation - Wald method													
$\alpha = 0.5$	-	.504	.505	.505	.515	-	-	-	.050	.051	.051	.048	-
$\beta_1 = -0.5$	-	-.359	-.412	-.412	-.516	-	-	-	.168	.168	.168	.130	-
$\beta_2 = 0.5$	-	.388	.480	.480	.515	-	-	-	.145	.143	.142	.128	-
$\beta_3 = -0.5$	-	-.503	-.538	-.543	-.509	-	-	-	.060	.055	.055	.065	-
$\beta_4 = 0.5$	-	.458	.458	.460	.513	-	-	-	.069	.065	.065	.059	-
$\beta_5 = 0.0$	-	-.002	-.001	-.001	-.015	-	-	-	.043	.043	.043	.091	-
$\beta_6 = 0.0$	-	-.002	-.004	-.004	-.027	-	-	-	.042	.042	.042	.089	-
$\beta_7 = 0.0$	-	-.003	-.001	-.001	.016	-	-	-	.045	.044	.043	.090	-
$\beta_8 = 0.0$	-	.006	.007	.007	.043	-	-	-	.043	.044	.044	.090	-
$\beta_9 = 0.0$	-	-.004	.000	.000	-.007	-	-	-	.045	.042	.038	.080	-

[a] The imputation results are presented for the spike-and-slab approach from the Gibbs sampler and for the probit from multiple imputation only once to avoid redundancy.

TABLE 2.4: Experimental study 2: Results

Results of MCAR with missings in X_1 (20%), X_2 (30%), and X_3 (10%), and with correlation between $X_1, \ldots, X_4$, and with true parameter values, mean posterior medians and root mean squared errors (RMSEs) of important $(\alpha, \beta_1, \ldots, \beta_4)$ and non-important $(\beta_5, \ldots, \beta_9)$ regression coefficients over 100 replications obtained by Elastic net with control parameter $\varphi = 0.0$ for Ridge regression (EN.0), $\varphi = 0.5$ for a variation of Elastic net (EN.5), and $\varphi = 1.0$ for Lasso (EN1.) and stepwise regression (SR) and Bayesian spike-and-slab (SnS), where the first column presents the estimates $(\hat{\beta})$ and the second one the inclusion probabilities $(\hat{\gamma})$. The prior setting for the spike-and-slab were set to $\tau_2 = 2 >> \tau_1 = 0.2$ and controlling the number of selecting variables with $w = 0.5$.

	Average estimates							RMSEs					
true values	probit	EN.0	EN.5	EN1.	SR	SnS		probit	EN.0	EN.5	EN1.	SR	SnS
Before Deletion													
$\alpha = 0.5$	.507	.495	.503	.503	.509	.513	1.000	.046	.043	.044	.044	.046	.047
$\beta_1 = -0.5$	-.484	-.365	-.440	-.436	-.520	-.516	.992	.119	.152	.130	.131	.107	.106
$\beta_2 = 0.5$	.494	.378	.446	.518	.513	.511	.997	.095	.141	.119	.120	.100	.100
$\beta_3 = -0.5$	-.511	-.493	-.504	-.470	-.504	-.507	1.000	.063	.056	.066	.067	.069	.069
$\beta_4 = 0.5$	.499	.450	.480	.437	.505	.508	1.000	.051	.069	.058	.058	.055	.056
$\beta_5 = 0.0$	-	.002	.002	.002	.012	.012	.069	-	.044	.043	.041	.096	.049
$\beta_6 = 0.0$	-	-.001	-.002	-.002	-.015	.008	.065	-	.044	.042	.040	.088	.047
$\beta_7 = 0.0$	-	.006	.006	.005	.036	-.016	.059	-	.040	.038	.037	.081	.045
$\beta_8 = 0.0$	-	.000	.000	.000	.016	-.010	.062	-	.042	.040	.039	.087	.046
$\beta_9 = 0.0$	-	.003	.003	.004	.003	.013	.084	-	.050	.049	.047	.091	.055
Complete case													
$\alpha = 0.5$	.501	.490	.493	.492	.503	.511	1.000	.062	.062	.063	.064	.064	.066
$\beta_1 = -0.5$	-.472	-.329	-.338	-.329	-.495	-.470	.929	.172	.208	.238	.260	.148	.179
$\beta_2 = 0.5$	.477	.350	.361	.352	.477	.479	.971	.140	.183	.202	.226	.143	.141
$\beta_3 = -0.5$	-.503	-.491	-.502	-.503	-.509	-.517	.998	.086	.071	.080	.084	.094	.093
$\beta_4 = 0.5$	.493	.440	.449	.446	.493	.500	1.000	.069	.086	.087	.093	.071	.073
$\beta_5 = 0.0$	-	-.004	-.004	-.003	.011	.012	.204	-	.057	.052	.050	.121	.063
$\beta_6 = 0.0$	-	.001	.003	.003	.015	.017	.242	-	.067	.061	.059	.123	.073
$\beta_7 = 0.0$	-	.009	.009	.008	.056	-.027	.215	-	.057	.051	.049	.111	.065
$\beta_8 = 0.0$	-	.004	.004	.004	.032	-.021	.226	-	.063	.058	.057	.126	.070
$\beta_9 = 0.0$	-	-.010	-.010	-.010	-.049	.006	.211	-	.059	.053	.050	.111	.064
Imputation - average method[a]													
$\alpha = 0.5$	.500	.488	.495	.495	.502	.492	1.000	.047	.047	.047	.047	.048	.055
$\beta_1 = -0.5$	-.468	-.333	-.389	-.388	-.485	-.391	.964	.156	.198	.194	.196	.136	.134
$\beta_2 = 0.5$	.480	.360	.412	.411	.483	.431	.995	.128	.167	.160	.160	.129	.089
$\beta_3 = -0.5$	-.514	-.502	-.514	-.515	-.516	-.522	1.000	.077	.063	.074	.076	.080	.064
$\beta_4 = 0.5$	.497	.447	.471	.472	.498	.471	1.000	.054	.072	.063	.062	.055	.050
$\beta_5 = 0.0$	-	-.005	-.005	-.004	-.002	.011	.089	-	.044	.041	.040	.081	.060
$\beta_6 = 0.0$	-	-.001	.000	.000	.040	.013	.083	-	.041	.038	.037	.074	.046
$\beta_7 = 0.0$	-	.007	.006	.005	.046	.029	.080	-	.039	.036	.034	.073	.072
$\beta_8 = 0.0$	-	.007	.007	.007	.020	.024	.101	-	.045	.044	.042	.083	.060
$\beta_9 = 0.0$	-	-.005	-.005	-.004	.003	.001	.054	-	.039	.037	.035	.064	.028
Imputation - majority method													
$\alpha = 0.5$	-	.503	.502	.502	.501	-	-	-	.048	.048	.048	.048	-
$\beta_1 = -0.5$	-	-.474	-.505	-.513	-.496	-	-	-	.157	.127	.123	.130	-
$\beta_2 = 0.5$	-	.485	.473	.467	.476	-	-	-	.128	.151	.160	.227	-
$\beta_3 = -0.5$	-	-.517	-.522	-.525	-.519	-	-	-	.080	.089	.092	.060	-
$\beta_4 = 0.5$	-	.500	.497	.496	.496	-	-	-	.055	.058	.059	.065	-
$\beta_5 = 0.0$	-	-.005	-.008	-.009	-.017	-	-	-	.048	.055	.056	.082	-
$\beta_6 = 0.0$	-	.000	.000	.001	.004	-	-	-	.044	.051	.054	.078	-
$\beta_7 = 0.0$	-	.008	.009	.010	.062	-	-	-	.042	.048	.051	.080	-
$\beta_8 = 0.0$	-	.008	.010	.011	.035	-	-	-	.050	.057	.060	.086	-
$\beta_9 = 0.0$	-	-.006	-.007	-.008	-.004	-	-	-	.043	.051	.053	.074	-
Imputation - Wald method													
$\alpha = 0.5$	-	.504	.505	.505	.515	-	-	-	.049	.051	.052	.050	-
$\beta_1 = -0.5$	-	-.340	-.385	-.386	-.496	-	-	-	.195	.193	.193	.130	-
$\beta_2 = 0.5$	-	.370	.420	.415	.481	-	-	-	.168	.165	.165	.128	-
$\beta_3 = -0.5$	-	-.503	-.512	-.512	-.510	-	-	-	.060	.060	.059	.065	-
$\beta_4 = 0.5$	-	.458	.466	.466	.499	-	-	-	.071	.068	.068	.059	-
$\beta_5 = 0.0$	-	-.003	-.002	-.002	-.015	-	-	-	.043	.043	.043	.091	-
$\beta_6 = 0.0$	-	-.001	-.002	-.002	-.007	-	-	-	.040	.039	.038	.089	-
$\beta_7 = 0.0$	-	-.003	-.001	-.001	.006	-	-	-	.044	.045	.043	.090	-
$\beta_8 = 0.0$	-	.006	.007	.007	.023	-	-	-	.042	.042	.041	.090	-
$\beta_9 = 0.0$	-	-.004	-.003	-.003	.009	-	-	-	.045	.042	.038	.080	-

[a] The imputation results are presented for the spike-and-slab approach from the Gibbs sampler and for the probit from multiple imputation only once to avoid redundancy.

TABLE 2.5: Experimental study 3: Results

Results of MAR with missings in X_1 (20%), and with correlation between $X_1, \ldots, X_4$, and with true parameter values, mean posterior medians and root mean squared errors (RMSEs) of important $(\alpha, \beta_1, \ldots, \beta_4)$ and non-important $(\beta_5, \ldots, \beta_9)$ regression coefficients over 100 replications obtained by Elastic net with control parameter $\varphi = 0.0$ for Ridge regression (EN.0), $\varphi = 0.5$ for a variation of Elastic net (EN.5), and $\varphi = 1.0$ for Lasso (EN1.) and stepwise regression (SR) and Bayesian spike-and-slab (SnS), where the first column presents the estimates $(\hat{\beta})$ and the second one the inclusion probabilities $(\hat{\gamma})$. The prior setting for the spike-and-slab were set to $\tau_2 = 2 >> \tau_1 = 0.2$ and controlling the number of selecting variables with $w = 0.5$.

	Average estimates							RMSEs					
true values	probit	EN.0	EN.5	EN1.	SR	SnS		probit	EN.0	EN.5	EN1.	SR	SnS
Before Deletion													
$\alpha = 0.5$	.507	.495	.502	.503	.509	.506	1.000	.048	.047	.047	.048	.048	.046
$\beta_1 = -0.5$	-.493	-.349	-.412	-.412	-.495	-.481	.981	.110	.168	.139	.140	.110	.121
$\beta_2 = 0.5$	.488	.363	.419	.420	.490	.494	.997	.086	.150	.116	.116	.086	.096
$\beta_3 = -0.5$	-.504	-.495	-.506	-.508	-.506	-.516	1.000	.058	.047	.056	.057	.057	.065
$\beta_4 = 0.5$	.501	.450	.477	.478	.503	.504	1.000	.052	.068	.057	.058	.053	.053
$\beta_5 = 0.0$	-	-.001	.003	-.001	-.032	.003	.062	-	.040	.038	.037	.082	.046
$\beta_6 = 0.0$	-	.000	-.003	.000	.004	.009	.057	-	.045	.043	.042	.089	.045
$\beta_7 = 0.0$	-	.001	-.009	.002	-.001	.016	.055	-	.041	.039	.038	.080	.044
$\beta_8 = 0.0$	-	-.002	-.002	-.002	-.003	.017	.074	-	.043	.041	.040	.081	.051
$\beta_9 = 0.0$	-	.002	-.005	.001	.010	.003	.050	-	.041	.039	.038	.080	.041
Complete case													
$\alpha = 0.5$	.512	.499	.506	.506	.514	. 520	1.000	.055	.053	.055	.055	.056	.058
$\beta_1 = -0.5$	-.501	-.354	-.406	-.407	-.504	-.503	.984	.134	.171	.164	.161	.133	.134
$\beta_2 = 0.5$	.495	.367	.415	.415	.498	.499	.995	.112	.154	.145	.139	.112	.111
$\beta_3 = -0.5$	-.509	-.499	-.512	-.513	-.512	-.517	1.000	.075	.061	.072	.074	.076	.077
$\beta_4 = 0.5$	.505	.453	.476	.478	.507	.513	1.000	.059	.070	.064	.063	.060	.061
$\beta_5 = -0.5$	-	.003	.003	.002	.004	.014	.126	-	.045	.043	.041	.094	.050
$\beta_6 = 0.5$	-	.003	.002	.002	.014	.014	.145	-	.051	.049	.049	.101	.057
$\beta_7 = 0.5$	-	-.001	-.001	-.000	.014	.010	.127	-	.047	.044	.043	.089	.051
$\beta_8 = 0.5$	-	-.006	-.007	-.006	-.017	.006	.137	-	.050	.048	.047	.094	.054
$\beta_9 = 0.5$	-	.002	.003	.002	.013	.014	.148	-	.051	.049	.047	.099	.057
Imputation - average method[a]													
$\alpha = 0.5$	.515	.501	.510	.510	.517	.520	1.000	.050	.046	.048	.048	.050	.052
$\beta_1 = -0.5$	-.498	-.354	-.419	-.417	-.501	-.408	.958	.137	.173	.156	.156	.138	.153
$\beta_2 = 0.5$	.490	.364	.422	.421	.492	.435	.991	.103	.153	.127	.127	.103	.116
$\beta_3 = -0.5$	-.504	-.494	-.505	-.507	-.505	-.545	1.000	.062	.049	.059	.060	.061	.073
$\beta_4 = 0.5$	.502	.450	.478	.478	.503	.477	1.000	.054	.069	.058	.058	.055	.057
$\beta_5 = -0.5$	-	-.001	-.001	-.001	-.039	.008	.085	-	.040	.038	.037	.082	.043
$\beta_6 = 0.5$	-	.000	.000	.001	-.017	.009	.101	-	.045	.044	.042	.084	.049
$\beta_7 = 0.5$	-	.001	.001	.001	.010	.010	.089	-	.042	.040	.038	.069	.046
$\beta_8 = 0.5$	-	-.002	-.003	-.002	-.030	-.006	.092	-	.043	.041	.040	.084	.047
$\beta_9 = 0.5$	-	.002	.001	.001	.011	.011	.088	-	.042	.039	.038	.069	.046
Imputation - majority method													
$\alpha = 0.5$	-	.518	.518	.518	.517	-	-	-	.051	.051	.051	.050	-
$\beta_1 = -0.5$	-	-.503	-.503	-.505	-.501	-	-	-	.137	.137	.137	.137	-
$\beta_2 = 0.5$	-	.493	.493	.488	.492	-	-	-	.103	.103	.116	.103	-
$\beta_3 = -0.5$	-	-.507	-.507	-.509	-.505	-	-	-	.061	.061	.063	.061	-
$\beta_4 = 0.5$	-	.505	.505	.503	.503	-	-	-	.055	.055	.056	.055	-
$\beta_5 = -0.5$	-	-.001	.000	.000	-.043	-	-	-	.044	.049	.050	.085	-
$\beta_6 = 0.5$	-	.000	.000	.000	.002	-	-	-	.049	.054	.056	.092	-
$\beta_7 = 0.5$	-	.001	-.002	.003	.004	-	-	-	.044	.050	.053	.089	-
$\beta_8 = 0.5$	-	-.003	-.003	-.004	-.020	-	-	-	.047	.055	.057	.087	-
$\beta_9 = 0.5$	-	.002	.001	.001	.024	-	-	-	.045	.050	.051	.083	-
Imputation - Wald method													
$\alpha = 0.5$	-	.505	.510	.510	.584	-	-	-	.046	.048	.048	.070	-
$\beta_1 = -0.5$	-	-.237	-.406	-.405	-.490	-	-	-	.274	.167	.167	.125	-
$\beta_2 = 0.5$	-	.361	.355	.358	.349	-	-	-	.149	.158	.159	.164	-
$\beta_3 = -0.5$	-	-.421	-.472	-.472	-.494	-	-	-	.090	.061	.063	.053	-
$\beta_4 = 0.5$	-	.352	.439	.438	.478	-	-	-	.158	.099	.099	.077	-
$\beta_5 = 0.5$	-	.002	.002	.001	.028	-	-	-	.039	.039	.039	.079	-
$\beta_6 = 0.5$	-	-.002	-.002	-.002	-.031	-	-	-	.038	.037	.037	.083	-
$\beta_7 = 0.5$	-	.009	.010	-.009	-.031	-	-	-	.040	.040	.039	.076	-
$\beta_8 = 0.5$	-	.002	.002	-.002	-.022	-	-	-	.037	.037	.036	.083	-
$\beta_9 = 0.5$	-	-.005	-.005	-.005	-.018	-	-	-	.037	.037	.036	.083	-

[a] The imputation results are presented for the spike-and-slab approach from the Gibbs sampler and for the probit from multiple imputation only once to avoid redundancy.

TABLE 2.6: Overview F-measurements for experimental studies 1, 2, and 3

For experimental study 1 (Ex1), experimental study 2 (Ex2), and experimental study 3 (Ex3) comparison of precision, recall, and F-measure obtained by Elastic net with control parameter $\varphi = 0.0$ for Ridge regression (EN.0), $\varphi = 0.5$ for a variation of Elastic net (EN.5), and $\varphi = 1.0$ for Lasso (EN1.) and stepwise regression (SR) and Bayesian spike-and-slab (SnS) over 100 replications.

	Average precision					Average recall					Average F-measure				
	EN0.0	EN0.5	EN1.0	SR	SnS	EN0.0	EN0.5	EN1.0	SR	SnS	EN0.0	EN0.5	EN1.0	SR	SnS
Experimental study 1															
BD	1.00	.99	.98	1.00	1.00	.50	.53	.54	.89	.98	.67	.68	.69	.93	.99
CC	1.00	.95	.93	.98	1.00	.50	.54	.57	.87	.95	.67	.68	.69	.91	.97
IMP-Average	1.00	.99	.99	.95	1.00	.50	.54	.55	.88	.99	.67	.70	.70	.93	1.00
IMP-Majority	1.00	.98	.99	1.00	1.00	.50	.52	.52	.89	.99	.67	.68	.69	.94	1.00
IMP-Wald	1.00	.97	.98	.98	1.00	.50	.55	.54	.89	.99	.67	.70	.70	.93	1.00
Experimental study 2															
BD	1.00	1.00	.99	1.00	1.00	.50	.52	.54	.88	.99	.67	.68	.69	.93	1.00
CC	1.00	.93	.91	.95	.99	.50	.53	.57	.85	.95	.67	.67	.69	.89	.97
IMP-Average	1.00	.99	.99	1.00	1.00	.50	.52	.53	.88	.99	.67	.68	.69	.93	1.00
IMP-Majority	1.00	.98	.99	1.00	1.00	.50	.52	.52	.89	.99	.67	.68	.69	.94	1.00
IMP-Wald	1.00	.97	.98	.98	1.00	.50	.55	.54	.90	.99	.67	.70	.70	.95	1.00
Experimental study 3															
BD	1.00	.99	.99	1.00	1.00	.50	.52	.53	.90	.99	.67	.69	.70	.94	1.00
CC	1.00	.95	.92	.98	.99	.50	.55	.57	.89	.96	.67	.68	.69	.93	.98
IMP-Average	1.00	.99	.99	1.00	.99	.50	.52	.53	.88	.98	.67	.68	.69	.93	.98
IMP-Majority	1.00	.98	.99	.99	.99	.50	.52	.52	.89	.98	.67	.68	.69	.94	.98
IMP-Wald	1.00	.97	.97	.98	.99	.50	.55	.54	.88	.98	.67	.70	.70	.94	.98

TABLE 2.7: Weighting model for NEPS-SC 4: Results

Model estimating the individual participation propensity for students in Wave 12 of SC 4 used to derive adjustment factors for adjusted wave-specific cross-sectional and longitudinal weights. From left-to-right the estimates for stepwise regression backwards, Elastic net with control mixing parameter $\alpha = 1.0$, i.e., Lasso penalty, and Bayesian Variable selection (BVS) with spike-and-slab prior. Additionally, the BVS estimates $\hat{\beta}$ are completed by the corresponding inclusion probabilities $\hat{\gamma}$.

Variables	Stepwise regression	Elastic net	Bayesian SnS	
			$\hat{\beta}$	$\hat{\gamma}$
Intercept	-1.452***	-.638	**-1.571**	**1.000**
	(.096)			
Migration Background	-.013***	-.017	-.037	.227
Yes	(.035)			
Student participated in	-.121***	-.004	**-.129**	**.693**
wave 6	(.031)			
Student participated in	.218***	-	**.129**	**.560**
wave 8	(.056)			
Student participated in	.371***	.069	**.343**	**1.000**
wave 9	(.053)			
Student participated in	.356***	.001	**.251**	**.973**
wave 10	(.056)			
Student participated in	1.278***	1.159	**1.214**	**1.000**
wave 11	(.036)			

Reference categories are: Migration background *no*.

To model individual participation, for the stepwise regression the `glm`-function with a probit link provided in R (R Core Team, 2020) was used. ***, **, and * denote significance at the 0.1%, 1%, and 5% level, respectively. Standard errors are given in parentheses. BIC based backward selection was used, and only significant coefficients are reported. BIC for the final model with selected variables: $BIC = 9,454.741$.

For Elastic net only non-negligible variables are reported. The shrinkage parameter λ is set to the largest value such that the error is 1 standard error of the minimum: $\lambda_{\min} = 0.015$ obtained with the `cv.glmnet`-function in R.

The bold results in the last two columns show variables with an inclusion probability higher than 50%. The other two variables are not selected by the Bayesian approach, but are reported for the sake of completeness. The prior setting for the spike-and-slab were set to $\tau_2 = 1 >> \tau_1 = 0.015$ and controlling the number of selecting variables with $w = 0.1$.

TABLE 2.8: Weighting model for NEPS-SC 4: Sensitivity analysis

Model estimating the individual participation propensity for students in Wave 12 of SC 4 used to derive adjustment factors for adjusted wave-specific cross-sectional and longitudinal weights. Sensitivity analysis - Bayesian Variable selection (BVS) with spike-and-slab prior: results of the estimates of regression coefficients $\hat{\beta}$ for different τ_1. Bold printed values have an inclusion probability $\hat{\gamma} > 0.50$. The prior setting for the spike-and-slab were set to $\tau_2 = 1$ and the spike prior tau_1 is set gradually in 0.025 steps from 0.0050 to 0.0375 controlling the number of selecting variables with $w = 0.1$.

	$\tau_0 = .050$		$\tau_0 = .075$		$\tau_0 = .100$		$\tau_0 = .125$		$\tau_0 = .150$		$\tau_0 = .175$		$\tau_0 = .200$	
	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\gamma}$
Intercept	-1.920	1.000	-1.970	1.000	-1.916	1.000	-1.917	1.000	-1.919	1.000	-1.921	1.000	-1.927	1.000
Age: younger half	.051	.014	.051	.014	.049	.015	.049	.014	.050	.017	.050	.019	.050	.022
Gender: female	.023	.009	.023	.009	.023	.011	.023	.014	.023	.017	.023	.020	.023	.023
Nationality: German	.064	.083	.064	.083	.063	.019	.064	.018	.063	.021	.063	.024	.063	.024
Primary language: German	.078	.148	.079	.045	.079	.024	.078	.022	.081	.021	.078	.023	.080	.024
Migration Background: yes	-.095	.106	-.095	.029	-.095	.019	-.095	.020	-.096	.018	-.094	.022	-.094	.026
Participation in wave 1	-.061	.548	-.057	.398	-.064	.283	-.065	.197	-.068	.142	-.063	.105	-.060	.087
Participation in wave 2	.200	.680	.196	.353	.196	.151	.198	.081	.197	.049	.196	.045	.198	.035
Participation in wave 3	.057	.056	.057	.021	.056	.014	.058	.018	.057	.021	.056	.022	.058	.020
Participation in wave 4	-.001	.028	-.003	.013	-.001	.012	-.001	.015	-.003	.016	-.000	.019	-.002	.020
Participation in wave 5	.155	.481	.156	.182	.157	.072	.157	.040	.155	.030	.157	.032	.157	.035
Participation in wave 6	-.130	.241	-.130	.052	-.131	.027	-.132	.025	-.130	.025	-.131	.026	-.130	.023
Participation in wave 7	.030	.030	.030	.014	.031	.014	.028	.019	.032	.017	.031	.020	.030	.024
Participation in wave 8	.196	.719	.197	.324	.197	.123	.197	.059	.197	.044	.195	.038	.197	.037
Participation in wave 9	.356	.999	.358	.971	.356	.771	.358	.464	.357	.239	.359	.154	.357	.103
Participation in wave 10	.337	.999	.337	.940	.337	.684	.337	.374	.339	.206	.339	.126	.339	.098
Participation in wave 11	1.269	1.000	1.270	1.000	1.269	1.000	1.267	1.000	1.268	1.000	1.269	1.000	1.269	1.000

Reference categories are: Age *older half*, Gender *male* Nationality *other than German*, primary language *other than German*, migration background *no*.

Chapter 3

Variable selection in statistical inference and machine learning

Variable selection is a critical aspect of both statistical inference and machine learning, albeit with differing emphases and methodologies. In addition to the considerations in chapter 2, in statistical inference, variable selection aims to identify a subset of predictors that are most relevant for explaining the response variable, while maintaining interpretability and minimizing bias. Traditional statistical methods, such as stepwise regression or information criteria like AIC and BIC, are commonly employed in inference to select variables based on their statistical significance and contribution to model fit.

Following generally James et al. (2013), in machine learning, variable selection is often integrated within the broader context of feature selection, where the focus is on optimizing predictive performance rather than inference. Machine learning algorithms incorporate mechanisms for automatic feature selection by assessing variable importance or applying regularization techniques to shrink less relevant features towards zero.[1] While both fields share many methodologies, they are distinguished by their primary objectives (Bzdok et al., 2018). Following Jordan and Mitchell (2015) in general, statistics traditionally focuses on making population inferences from sampled data. The used methods are designed to estimate parameters, test hypotheses, and quantify uncertainty in order to draw conclusions about the underlying population from which the sample was drawn. Statistical techniques often prioritize interpretability, robustness to assumptions, and inference over prediction accuracy alone. Examples include linear regression, hypothesis testing, and analysis of variance. In statistical inference, the goal of variable selection is often to enhance the interpretability of the model by isolating the most significant predictors, allowing for more reliable estimation of model parameters. On the other hand, machine learning is primarily concerned with building predictive models that can generalize well to unseen data. While machine learning methods can also perform inference, their main goal is to uncover patterns and relationships within data that enable accurate predictions or decisions. Machine learning algorithms prioritize predictive performance, scalability, and flexibility in handling diverse data types

[1] The terminology of the different applications (e.g., Lasso) is used in both statistical inference and machine learning. Machine learning draws on statistical considerations in many applications, although the exact description of the black box, i.e., the mechanisms behind it, is not usually described in any depth to ensure clarity. A clear semantic distinction between the terms of the different applications cannot be found in current scientific discourse, but there is a mixture of terms and approaches and frames.

and structures. Examples include decision trees, support vector machines, and neural networks.

Despite these distinctions, there is significant overlap between machine learning and statistics. Many statistical techniques, such as regression analysis and Bayesian inference, are foundational to machine learning algorithms. Likewise, machine learning methods, such as ensemble methods and regularization techniques, are increasingly being adopted in statistical modeling. In essence, while statistics and machine learning diverge in their primary goals—population inference versus predictive modeling—they share a common foundation of mathematical and computational techniques. This interplay between the two fields continues to drive innovation and advancement in both theory and practice, enriching our understanding of data and enhancing our ability to extract meaningful insights and make informed decisions (Murphy, 2012; Valiant, 1984).

Hence, variable selection in both domains shares common objectives, including improving model interpretability, reducing model complexity, and enhancing predictive accuracy. Moreover, recent advancements have seen a convergence of techniques between statistical inference and machine learning, with methods like regularized regression and Bayesian variable selection gaining popularity for their ability to balance predictive performance with interpretability and uncertainty quantification. As discussed in chapter 2, various methodologies have been developed, such as the Lasso (Least Absolute Shrinkage and Selection Operator) proposed by Tibshirani (1996), which simultaneously performs variable selection and regularization to improve the prediction accuracy and interpretability of the model. In machine learning, feature selection plays a crucial role in improving the performance of predictive models by reducing the dimensionality of the input space, which helps prevent overfitting and reduces computational cost. Regularization techniques, such as Ridge regression (Hoerl & Kennard, 1970) and Elastic net (Zou & Hastie, 2005), combine penalties to manage multicollinearity and encourage sparsity, which are important in both inference and machine learning contexts. Feature importance metrics, such as those derived from tree-based methods like Random forests (Breiman et al., 1984), have also become widely used in machine learning for their ability to rank the relevance of features without requiring strong model assumptions.[2] One of the key distinctions between statistical inference and machine learning is how they handle uncertainty and model assumptions. While distinguishing between stochastic data generated models and unknown data mechanisms, Breiman (2001) advocates for shifting from the statistical community's reliance on traditional data models to incorporating more flexible algorithmic models, which are better suited for solving complex problems across various data sets. Both modeling approaches, the generative approach favored in statistics versus the predictive focus of machine learning, emphasize different aspects of model building. In inference, parametric assumptions about the underlying data-generating process are often necessary to derive confidence intervals and p-values, which are used to assess the significance of selected variables. Following Hastie et al. (2009) machine learning methods are not only more agnostic about the data distribution but also prioritize

[2]Most techniques and methods can be found in literature in both cases, in the statistical framework and in the machine learning environment.

predictive performance over interpretability, emphasize algorithmic complexity over parametric simplicity, and adapt dynamically to patterns in data without relying on predefined functional forms. In contrast to statistical inference, which seeks to explain relationships and derive uncertainty measures under specific assumptions, machine learning focuses on optimizing predictions, often leveraging high-dimensional representations, regularization, and ensemble techniques to improve generalization. See e.g., Breiman (2001) who contrasts the generative (statistical inference) and algorithmic (machine learning) approaches, arguing for a shift toward flexible models that prioritize predictive accuracy or the differentiation of Shmueli (2010) between explanatory modeling (inference) and predictive modeling (machine learning), emphasizing their different goals and methodological considerations.

Recent research has focused on hybrid methods that bridge the gap between these two approaches. For instance, Hastie et al. (2009) discuss how statistical learning techniques like penalized regression (e.g., Lasso, Ridge regression) can be adapted for both inferential and predictive purposes. Bayesian methods, too, offer a natural framework for integrating prior knowledge with data-driven learning, as seen in the development of Bayesian variable selection approaches that account for uncertainty in both model parameters and the model selection process itself (George & McCulloch, 1993). These approaches have gained traction in both statistical and machine learning communities due to their flexibility and capacity to quantify uncertainty in a coherent manner. Recent research continues to explore hybrid methods that blend statistical inference and machine learning techniques for variable selection. Penalized regression methods remain central to this discourse. For example, newer adaptations of the Bayesian Lasso have emerged, where penalized Bayesian algorithms allow for more flexibility in handling uncertainty and model complexity in high-dimensional settings. This approach shows promise in fields like economic forecasting and disease prediction, where complex, correlated datasets need to be navigated effectively while maintaining interpretability (Pacifico & Pilone, 2024; Scheipl et al., 2013). In addition, automated and scalable feature selection methods, such as those embedded in algorithms like gradient boosting, are increasingly used in machine learning for their efficiency with large-scale data and their ability to manage complex feature interactions. For instance, recent advances in multiplicative regression models and adaptive Lasso approaches offer improved performance in high-dimensional and noisy environments (Chen, Ming, et al., 2024). These developments not only optimize predictive accuracy but also help strike a balance between interpretability and computational feasibility, addressing limitations seen in traditional statistical approaches.

Building on these advances, Bayesian Additive Regression Trees (BART) represents a cutting-edge method that combines the flexibility of machine learning with the interpretability sought in traditional statistical approaches. It is a Bayesian ensemble method that models the response as a sum of many shallow regression trees, with regularization priors to control complexity, enabling robust performance in high-dimensional setting and prevent overfitting. Unlike single-tree methods like Classification and Regression Trees (CART), BART provides posterior distributions for predictions, allowing for uncertainty quantification and improved predictive performance (Chipman et al., 2010; Linero, 2018). Therefore BART is a more sophisticated and powerful extension of CART by using a Bayesian ensemble approach to

improve stability, uncertainty quantification, and predictive accuracy. CART is simpler and more interpretable but can suffer from high variance unless used within an ensemble method like random forests or boosting. BARTs intrinsic ability to select relevant variables, manage complex feature interactions, and adapt to non-linear relationships makes it particularly valuable in applications such as economic forecasting and disease prediction (Bleich et al., 2014; Kapelner & Bleich, 2016). Furthermore, extensions to BART enhance its scalability and enable it to handle missing data effectively, offering a compelling balance between predictive accuracy and interpretability (Carnegie & Wu, 2019; Xu et al., 2016).

In summary, while the goals and methodologies of variable selection may vary between inference and machine learning, both domains recognize its importance in building parsimonious and effective models for understanding data and making predictions. The performance and implementation of the various methods is discussed.

3.1 Variable selection in machine learning

Variable selection is a fundamental aspect of machine learning methodologies, underpinning the construction of predictive models that effectively leverage the most relevant features within a dataset (Guyon & Elisseeff, 2003; Hastie et al., 2009). This process is crucial for optimizing model performance, enhancing interpretability, and addressing the challenges posed by high-dimensional data. In the realm of machine learning, variable selection encompasses techniques aimed at identifying subsets of features that significantly contribute to predictive accuracy while minimizing model complexity. These techniques serve to mitigate the curse of dimensionality, improve model efficiency, and facilitate the extraction of actionable insights from complex datasets.

Within the context of machine learning, variable selection methodologies often involve assessing the importance of individual features through measures such as feature importance scores or regularization techniques. These methods prioritize features based on their predictive power, allowing for the construction of parsimonious models that balance interpretability with predictive performance, see Guyon and Elisseeff (2003) and Kotsiantis (2011). Additionally, dimensionality reduction techniques, such as principal component analysis (PCA) and feature extraction, are employed to reduce the number of variables while retaining the most salient information, thereby streamlining the modeling process and improving generalization capabilities. Dimensionality reduction methods, such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA), are commonly employed to address the challenges posed by high-dimensional datasets. These techniques effectively reduce the number of variables under consideration to a set of principal components or discriminant functions, which capture the most important information in the data while minimizing redundancy. However, challenges arise in high-dimensional datasets, where issues like linear combinations, correlations, and singularity of the covariance matrix can complicate analysis. To mitigate these challenges, researchers have proposed various techniques. For instance, Witten and Tibshirani (2011) introduced a penalized LDA approach, which penalizes the coefficients of the discriminant functions to address collinearity concerns and improve model

stability. Additionally, alternative techniques such as independent components analysis, canonical correlation analysis, or partial least squares discriminant analysis have been utilized to address similar issues (Hastie et al., 2009). Moreover, variable selection in machine learning necessitates careful consideration of the trade-off between model interpretability and predictive accuracy. While simpler models with fewer variables may offer greater interpretability, they risk sacrificing predictive performance. Conversely, more complex models may achieve superior predictive accuracy but at the expense of interpretability. Balancing these competing objectives requires a nuanced understanding of the underlying data and domain-specific knowledge to guide the variable selection process effectively (Bishop, 2009; Cranmer & Desmarais, 2017).

In summary, variable selection in machine learning represents a multifaceted endeavor that encompasses both statistical techniques and domain expertise. By judiciously selecting informative features and balancing model complexity with interpretability, researchers can construct predictive models that not only capture the underlying patterns in the data but also yield actionable insights for decision-making and problem-solving in various domains: among others see in healthcare Austin and Tu (2004), in finance Crook et al. (2007) or in genomics Fan and Lv (2010).[3]

Following Dhal and Azad (2021) the dimensionality reduction can be split in feature selection and feature extraction, albeit the focus is on feature selection with the split in filter models, wrapper models and embedded models.

Machine learning algorithms, such as decision trees, random forests, and gradient boosting machines, can automatically handle feature selection to some extent by prioritizing informative features during model training. However, in high-dimensional datasets or when dealing with collinear predictors, explicit feature selection or dimensionality reduction techniques may be necessary to improve model performance and interpretability. Furthermore, machine learning techniques like neural networks and deep learning architectures can automatically learn hierarchical representations of the data, effectively performing feature extraction and dimensionality reduction as part of the model training process. These models can capture complex patterns and relationships in the data, making them powerful tools for predictive modeling in diverse domains.

Despite the benefits of variable extraction and dimensionality reduction in both statistical modeling and machine learning, it es essential to acknowledge that the newly generated variables or learned representations may lack the straightforward interpretability of the original variables. This introduces challenges in understanding and interpreting the results of the analysis, particularly in complex machine learning models. Nonetheless, variable extraction and dimensionality reduction methods remain valuable techniques for improving model performance, reducing computational complexity, and extracting meaningful insights from high-dimensional datasets. The next steps will focus on variations of Elastic net as well as boosting methods like XGBoost, which will be compared with Bayesian approaches.

[3]A review of the literature reveals that the techniques in question can be found in both machine learning and statistical modeling.

3.2 Details on variable selection in statistical modeling

The selection of variables is a pivotal stage in the process of statistical modeling, whereby the most pertinent variables from a larger set are identified and incorporated into the model. The objective is to enhance the performance of the model, facilitate interpretation, and circumvent overfitting. As previously stated in chapter 2, shrinkage procedures refer to a class of regularization methods that entail fitting a regression model using all P predictors, subject to a constraint on the magnitude of their estimated coefficients. The removal of predictors from the model can be conceptualized as the setting of their coefficients to zero. Rather than imposing an exact value of zero, these coefficients are subjected to a penalty for being too distant from zero, thereby ensuring their continuous reduction to a minimal value. The following section will present several optimization methods utilized in variable selection procedures. Key methods for variable selection include forward selection, backward elimination, combined in stepwise selection, and regularization techniques such as Lasso or Ridge regression.

3.2.1 Ridge regression

In Ridge regression, some variables are shrunk towards zero, thus model complexity can be decreased while keeping all variables in the model. The regression coefficients β shrink close to zero, so that variables, with minor contribution to the outcome, are slightly redundant and are shrunk closer to zero. Hence, the objective function minimizes both the likelihood and the penalty term:

$$\hat{\beta}_{\text{Ridge}} = \arg\min_{\beta} \left\{ L(D, \beta) + p_{\text{Ridge}}(\beta, \lambda) \right\} \tag{3.1}$$

with λ as the regularization parameter.

In the case of OLS, the loss function is augmented by minimizing the sum of squared residuals but also penalizing the size of parameter estimates, in order to shrink them towards zero:

$$\hat{\beta}_{\text{Ridge, OLS}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j=1}^{M} \hat{\beta}_j{}^2 \right\} = ||Y - X\hat{\beta}||^2 + \lambda ||\hat{\beta}||^2. \tag{3.2}$$

Solving this for $\hat{\beta}$ gives the Ridge regression estimates $\hat{\beta}_{Ridge,OLS} = (X'X + \lambda I)^{-1}(X'Y)$, where I denotes the identity matrix. Remark that as $\lambda \to 0$, $\hat{\beta}_{Ridge} \to \hat{\beta}_{OLS}$ and as $\lambda \to \infty$, $\hat{\beta}_{Ridge} \to 0$. Setting $\lambda = 0$ recovers the OLS solution, while increasing λ strengthens regularization, shrinking all coefficients toward zero except the intercept (if not regularized).

The bias and variance are given by:

$$\text{Bias}(\hat{\beta}_{\text{Ridge, OLS}}) = -\lambda(X'X + \lambda I)^{-1}\beta, \tag{3.3}$$

$$\text{Variance}(\hat{\beta}_{\text{Ridge, OLS}}) = \sigma^2(X'X + \lambda I)^{-1}X'X(X'X + \lambda I)^{-1}. \tag{3.4}$$

Minimizing the negative log-likelihood of the Probit model with added Ridge penalty yields:

$$\hat{\beta}_{\text{Ridge, Probit}} = \arg\min_{\beta} \left\{ -\sum_{i=1}^{N} y_i \log(\Phi(X_i\beta)) - (1-y_i)\log(1-\Phi(X_i\beta)) + \lambda \sum_{j=1}^{M} \beta_j^2 \right\} \quad (3.5)$$

Obtaining a precise closed-form expression for the bias in a Probit model involves complex mathematical analysis due to the nonlinearity introduced by the Probit link function. The Probit link function is defined as the derivative of the CDF of the standard normal distribution.

Thus, as λ increases, variance decreases while bias increases, illustrating the bias-variance trade-off. Therefore, an optimization routine for the shrinkage parameter λ is needed. There are two ways to optimize: after stepwise regression, optimization can be done using information criteria, such as AIC or BIC, or by cross-validation. The first approach emphasizes the model's fit to the data by choosing λ so that the information criterion is the smallest, while the second approach focuses more on its predictive performance. By performing cross-validation, which is often found in machine learning approaches, the selection process is forced to the shrinkage parameter λ by minimizing the cross-validated sum of squared residuals. In addition, Ridge regression assumes that the predictors are standardized and the response is centered.

Minimizing by information criteria

This approach requires estimating the model with many different values for λ and choosing that value minimizing the AIC and BIC, respectively:

$$\text{AIC}_{Ridge} = nlog(e'e) + 2df_{Ridge}, \quad (3.6)$$

$$\text{BIC}_{Ridge} = nlog(e'e) + df_{Ridge}log(n). \quad (3.7)$$

where df_{Ridge} denotes the effective of degrees of freedom, which is different in Ridge regression than in the regular OLS. For both methods, the degrees of freedom are given by the trace of the so-called Hat matrix, which maps the vector of observed response values y to the vector of fitted values $\hat{y}$:

$$\hat{y} = Hy. \quad (3.8)$$

In OLS, the Hat matrix is given by:

$$H_{OLS} = X(X'X)^{-1}X, \quad (3.9)$$

which results in degrees of freedom:

$$df_{OLS} = tr(H_{OLS}) = m, \quad (3.10)$$

where m is the number of predictor variables. In contrast, Ridge regression modifies the Hat matrix by incorporating the regularization penalty:

$$H_{Ridge} = X(X'X + \lambda I)^{-1}X, \tag{3.11}$$

which leads to:

$$df_{Ridge} = tr(H_{Ridge}), \tag{3.12}$$

which is generally less than m and decreases as λ increases. This means that Ridge regression typically yields lower AIC and BIC values compared to OLS due to its reduced complexity, though the exact comparison can depend on the data and the choice of λ for Ridge regularization. For Ridge regression, df_{Ridge} is less than mm because the regularization shrinks the coefficients, effectively reducing the model's complexity. The AIC penalizes model complexity by adding $2df_{Ridge}$. Since $df_{Ridge} < m$, the AIC for Ridge regression will generally be smaller than for OLS because Ridge has fewer effective degrees of freedom. The BIC also penalizes complexity, but more heavily than AIC due to the $log(n)$ factor. Since $df_{Ridge} < df_{OLS} = m$, the BIC for Ridge will be smaller than the BIC for OLS, as it penalizes complexity more strongly and Ridge has fewer degrees of freedom. Thus, AIC for Ridge will generally be smaller than AIC for OLS and BIC for Ridge will generally be smaller than BIC for OLS .

Minimizing by cross-validation

The shrinkage parameter λ can be chosen via cross-validation (CV), which is a resampling approach used to estimate the Mean Squared Error (MSE) of a model by repeatedly withholding a subset of observations and applying the chosen method to predict (Arlot & Celisse, 2010). The goal is to obtain a more robust estimate of the model's performance and to ensure that it generalizes well to unseen data. Both leave-one-out cross-validation (LOOCV) and k-fold cross-validation are techniques used to evaluate variable selection, but they differ in how they partition the dataset for training and validation (Efron & Tibshirani, 1997).

- **Leave-One-Out Cross-Validation (LOOCV)** For each data point in the dataset, a model is trained or fitted on all other, $n - 1$, data points, and its performance is evaluated on the single left-out data point. This step is repeated n times with the result of low bias and high variance.
- **K-Fold Cross-Validation** The dataset is divided into K equally sized folds or subsets. The model is trained on $K - 1$ folds and tested on the remaining fold. This process is repeated K times, each time using a different fold as the test set. The number of folds is pre-specified by the user (common choices include 5 or 10).

In the context of shrinking k-fold Cross-Validation is used to choose the value of λ that minimizes the estimated MSE. Values of k between 5 and 10 are typically indicated as good for computational burden and the bias trade-off. The key to the improvement of Ridge

regression over OLS is in the bias-variance trade-off: as λ increases, so does the bias, but the variance decreases due to the reduced flexibility. The dataset is split into K folds to test P different values for the shrinkage parameter. See 1 for details.

3.2.2 Elastic net

The shrinkage penalty term can be generalized to $\lambda \sum_{j=1}^{p} |\beta_j|^q$ for $q > 0$, where for $q = 2$, the Ridge regression (L2-norm penalty), for $q = 1$, the Lasso is specified (L1-norm penalty). An extension of these methods is the Elastic net, a regression model that combines both the L1-norm (Lasso) and the L2-norm (Ridge) penalties (Zou & Hastie, 2005). This dual penalty allows for the advantages of both methods: it shrinks coefficients (like Ridge) and can set some coefficients exactly to zero (like LASSO). In the Elastic net, the λ_2 is the regularization parameter for the Ridge (L2) part, and λ_1 is for the Lasso (L1) part, leading to the following formulation:

$$\hat{\beta}_{\text{Elastic net}} = \arg\min_{\beta} \left\{ L(D, \beta) + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right\} \tag{3.13}$$

The Elastic Net combines both L1 and L2 penalties to leverage the strengths of both Ridge and Lasso. The L1-norm encourages sparsity (setting some coefficients to zero), while the L2-norm shrinks coefficients toward zero, preventing overfitting when predictors are correlated. The objective function contains both regularization terms, controlled by λ_1 (L1) and λ_2 (L2), and a loss function $L(D, \beta)$ that could represent, for example, the residual sum of squares for linear regression or log-likelihood for logistic regression. Larger values of λ_1 lead to more coefficients being shrunk to zero (like Lasso). Larger values of λ_2 lead to more shrinking of the coefficients toward zero without setting them exactly to zero (like Ridge).

An elastic net is a regularization and variable selection procedure that makes use of the penalty

$$\lambda \left[\frac{1}{2}(1 - \alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right], \tag{3.14}$$

where $\alpha \in [0, 1]$ is called the mixing parameter and λ has the usual interpretation. Lasso and Ridge regression are special cases, respectively for $\alpha = 1$ and $\alpha = 0$.

Hence, Elastic Net is a regularization procedure that overcomes some of the limitations of the Lasso by borrowing strength from the Ridge regression. In detail, Elastic Net allows selecting more than n variables, can handle (highly) correlated variables by jointly selecting or leaving out groups, is readily extendable to use with more general methods (generalized linear models). Elastic nets are especially useful when a sparse solution is either necessary or desirable (such as in $p >> N$ problems) and small groups of highly correlated predictors are present.

Choice of mixing and shrinkage parameter

The mixing parameter α adjusts the extent to which the Elastic net behaves as a Ridge regression or Lasso. As $\alpha \to 0$, the Ridge penalty gains more weight than the lasso; the

opposite happens when $\alpha \to 1$. Hence, following (Zou & Hastie, 2005) a naive version finding the estimator is a two-stage procedure:

1. For each fixed λ_2 find the Ridge regression coefficients, and then
2. do a Lasso type shrinkage.

Applying this shrinkage version leads to an increase in bias and poor prediction. In practice, one usually constructs a grid of A α values, say $\alpha_1, \ldots, \alpha_A$, chooses a fold configuration, e.g., $k = 5$ or $k = 10$, and for each $a = 1, \ldots, A$:

1. k-fold cross-validates λ for a given $\alpha = \alpha_a$, and
2. stores the test MSE.

Each fold will be used as a validation set while the model is trained on the remaining data. Finally, the A MSEs are compared, and the α associated with the preferred one is chosen. The goal is to identify the α associated with the lowest test MSE so that this α is considered the preferred choice in terms of prediction performance. With the preferred α, the corresponding optimal λ is chosen whose value is determined based on the results of the k-fold cross-validation performed at that specific α value. The best λ within the selected profile is then used for inference and prediction, respectively (Zou & Hastie, 2005).

3.2.3 Bias-Variance trade-off and variable selection

Shrinkage methods play a critical role in addressing the bias-variance trade-off in statistical modeling. As described in chapter 1, the bias-variance trade-off is a fundamental concept in statistical modeling and machine learning: Bias refers to the error introduced by approximating a real-world problem, which can be quite complex, with a simplified model. High bias can lead to underfitting, where the model is too simple to capture the underlying patterns in the data. Variance measures the sensitivity of the model to fluctuations in the training data. High variance can lead to overfitting, where the model captures noise in the training data and performs poorly on new, unseen data.

The regularization term penalizes large coefficients, which affects the bias and the variance (Bühlmann & van de Geer, 2011; Friedman et al., 2010; Hastie et al., 2009): By high regularization (e.g., Ridge regression), results tend to have high bias, which means that the model simplifies the underlying relationships in the data too much, possibly leading to underfitting (Hoerl & Kennard, 1970). Ridge regression, with its penalty term, can prevent overfitting but may bias the estimates of the coefficients towards zero. Regularized models, such as Ridge regression, tend to have lower variance. The penalty term in Ridge regression helps stabilize coefficient estimates, reducing the sensitivity of the model to noise in the data. Elastic net combines both L1 and L2 regularization and is there more moderate in regularization because it strikes a balance between Ridge (L2) and Lasso (L1) regularization (Tibshirani, 2011). Depending on the mixing parameter α, Elastic net can have a moderate level of bias compared to Ridge or Lasso alone. Elastic net aims to combine the benefits of Ridge and Lasso regularization. It can lead to a reduction in variance compared to models

with no regularization (like OLS), especially when there is multicollinearity among predictors. When α is close to 0, elastic net behaves more like Ridge regression. It tends to have lower bias but may still have some bias compared to models without regularization. Instead when α is close to 1, elastic net behaves more like Lasso regression. It can lead to sparsity in the model but may introduce more bias. Intermediate values of α strike a balance, providing a compromise between the benefits of Ridge and Lasso regularization. This intermediate level of regularization can help manage the bias-variance trade-off effectively.

3.2.4 Implementation in standard software

In R (R Core Team, 2020), *glmnet* provides necessary functions to estimate via penalized ML for generalized linear models, which also includes, as shown above, Ridge regression and Lasso (Friedman et al., 2010), while in Julia (Bezanson et al., 2017) the *glmnet* and the *CrossValidation* packages are necessary.

The mixing parameter, with $0 \leq \alpha \leq 1$ corresponds to the Lasso ($\alpha = 1$) and to the Ridge regression penalty ($\alpha = 0$). Values of α between 0 and 1 are related to the Elastic Net regularization. Realizing that the degree of regularization depends on the regularization parameter λ results in evaluating the regression function for a sequence of λ. By default, the implemented function evaluates a sequence of 100 λ values[4]. But specification can be set up by the user. Cross-validation is used to find the value of λ, which minimizes the residuals. By default, 10-fold cross-validation is implemented and the evaluation of the model performance is calculated by mean squared error (MSE).

3.3 Handling missing values in variable selection algorithms

Handling missing values is an important consideration in any statistical or machine learning analysis, including shrinkage estimation methods. Handling missing values is a critical aspect of variable selection in machine learning, as missing data can impact the quality and performance of models. Missing values can introduce bias if the missingness is related to the target variable or other features. This can skew the results and affect the model's ability to learn accurate patterns. Missing data can also increase variance by introducing noise or uncertainty into the dataset, especially if not handled properly. There are also losses in the performance of the models and instability in the models. This again leads to incorrect or inaccurate conclusions as well as incorrect representations and visualizations. Following Joel et al. (2022)

Based on Friedman's test, Jerez et al. (2010) show that in prediction machine learning algorithms such as multi-layer perceptron, self-organization maps, or k-nearest neighbor outperform statistical methods, especially when handling missing or noisy data. These algorithms are more flexible and robust, able to adapt to complex, non-linear relationships, and they can handle missing data more effectively. In contrast, traditional statistical methods tend to rely on rigid assumptions about the data and may perform poorly when data is

[4]Note: in cases where no change occurs the function stops earlier.

incomplete or noisy. An overview for classification with missing values and machine learning methods is given by García-Laencina et al. (2009). See also Joel et al. (2022) for a general look at the state-of-art methods. Hasan et al. (2021) provide a distinction between statistical techniques and machine learning techniques for handling missing values. A review of the literature shows that the boundaries and classifications of machine learning and statistical methods are not always clear. Some results are also contradictory, as Elastic net methods are usually classified as machine learning methods.

3.3.1 Bayesian adaptive regression trees

Combining the strengths of tree-based models with the robustness of Bayesian inference, Bayesian Additive Regression Trees (BART) are flexible and powerful non-parametric[5] Bayesian modeling approaches. It was developed by Chipman et al. (2010) and has since gained popularity due to its ability to model complex, non-linear relationships in data. BART is particularly well-suited for situations where uncertainty quantification and prediction accuracy are crucial, see for more insights Linero (2018) and Ročková and van der Pas (2020), and for variable selection Bleich et al. (2014). In literature BART is among others discussed for causal inference (Hahn et al., 2020; Hill, 2011) and modeling longitudinal survey data (Zinn & Gnambs, 2020). The trees in BART therefore differ not only in their structure, but also in the decisions that are made randomly and Bayesian during the training process. The non-parametric nature of BART refers to the flexibility of the trees and their ability to model complex relationships without fixed functional assumptions.

BART models the conditional mean function $f(\mathbf{X})$ as the sum of a set of regression trees

$$f(\mathbf{X}) = \sum_{j=1}^{m} T_j(\mathbf{X}; \mathcal{M}_j, \mathcal{G}_j) \tag{3.15}$$

with $\mathbf{X}$ representing the predictor variables, $T_j(\mathbf{X}; \mathcal{M}_j, \mathcal{G}_j)$ being the j-th regression tree in the ensemble, with $\mathcal{M}_j$ and $\mathcal{G}_j$ denoting the tree structure (splits and depths), and the terminal node values (leaf values), respectively, and m being the total number of trees in the model.[6] The sum of these trees forms a flexible and adaptive model that can capture complex interactions and non-linearities. Each tree T_j contributes a small component to the overall prediction, ensuring that the influence of any single tree is limited, thereby reducing the risk of overfitting.

BART uses a probabilistic approach to build and combine trees, and this involves specifying both a likelihood function and a set of priors. Given a response variable Y and predictors $\mathbf{X}$, the likelihood function for BART is defined as

$$Y_i = f(\mathbf{X}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{3.16}$$

[5] BART do not assume a fixed functional form between predictors and the outcome. Instead, BART flexibly models complex, non-linear relationships by combining tree-based structures within a Bayesian framework, allowing it to adapt to the data without predefined parameter constraints.

[6] Typically set to a large number (e.g., $m = 200$).

where Y_i is the observed outcome for observation i, $\mathbf{X}_i$ is the vector of predictor values for observation i, and σ^2 is the error variance, assumed to be constant across observations.

To regularize the model and control complexity, BART imposes priors on the structure and terminal node values of the trees:

- **Tree depth prior**: A prior on the probability of splitting at a node is given by

$$\Pr(\text{split at node}) = \alpha(1+d)^{-\beta}, \tag{3.17}$$

  where α and β are hyperparameters that control the tendency of the trees to grow, and d is the depth of the node. This prior ensures that trees remain shallow, reducing the risk of overfitting.

- **Priors on terminal node values** For the terminal node values μ, BART places a normal prior

$$\mu_{j,k} \sim \mathcal{N}(0, \sigma_\mu^2), \tag{3.18}$$

  where $\mu_{j,k}$ represents the mean value at the k-th terminal node of tree T_j, and σ_μ^2 is a variance parameter that scales according to the total number of trees, ensuring that each tree's contribution to the model is small. Following Chipman et al. (2010) for binary outcomes $\sigma_\mu = \dfrac{e}{k\sqrt{m}}$ is recommended as $m = 200$, $k = 2$ and $e = 3$ to suggest a 95% prior probability that $E(Y_i|X)$ falls between the observed minimum and maximum values of an appropriately (rescaled) Y.

The goal of BART is to estimate the posterior distribution of the sum-of-trees model

$$p(T_1, \ldots, T_m, \sigma^2 | Y, \mathbf{X}). \tag{3.19}$$

This is achieved using a Bayesian backfitting Markov Chain Monte Carlo (MCMC) algorithm[7] designed to estimate the posterior distribution of the sum-of-trees model efficiently. The MCMC algorithm iteratively updates each tree T_j while keeping the others fixed, simulating from their full conditional distributions. This approach allows the model to refine each tree individually based on the residuals left after accounting for the contributions of the other trees. Specifically, each tree is updated using a Metropolis-Hastings step (Hastings, 1970), proposing new splits and terminal values based on the data, and the variance parameter σ^2 is updated using a Gibbs sampling step based on the residual sum of squares (Geman & Geman, 1984).

The backfitting MCMC algorithm for BART operates by iteratively updating each tree in the ensemble of m trees, $T_1, T_2, \ldots, T_m$, while holding the others fixed.

Let $\mathbf{T} = \{T_1, T_2, \ldots, T_m\}$ denote the entire ensemble of trees, and let $\mathbf{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_m\}$ represent their terminal node values. The posterior distribution of interest is

$$p(\mathbf{T}, \mathbf{G}, \sigma^2 | \mathbf{Y}, \mathbf{X}) \propto p(\mathbf{Y}|\mathbf{X}, \mathbf{T}, \mathbf{G}, \sigma^2) p(\mathbf{T}) p(\mathbf{G}) p(\sigma^2) \tag{3.20}$$

[7] In summary, the term *backfitting* MCMC highlights the combination of the iterative, component-wise fitting strategy (backfitting) and the probabilistic sampling approach (MCMC) that ensures convergence to the posterior distribution of the sum-of-trees model.

where $p(\mathbf{Y}|\mathbf{X}, \mathbf{T}, \mathbf{G}, \sigma^2)$ is the likelihood function based on the Gaussian noise assumption, and $p(\mathbf{T})$, $p(\mathbf{G})$ and $p(\sigma^2)$ are the priors on the trees, terminal values, and variance, respectively. The Metropolis-Hastings (MH) algorithm is used to update each tree T_j individually. Given the residuals R_j (the difference between the observed responses and the sum of the contributions from the other trees), the conditional posterior distribution of tree T_j is updated. The MH algorithm proceeds as a proposal step. A new tree structure T_j' is proposed based on the current tree T_j. This can involve growing a new branch, pruning a branch, changing a split point, or swapping split rules. The proposal distribution $q(T_j'|T_j)$ specifies the probability of proposing the new tree structure from the current one. Setting $p(T_j'|R_j, \sigma^2)$ is the posterior probability of the proposed tree given the residuals and variance, the new tree structure T_j' is accepted with probability

$$\alpha = \min\left(1, \frac{p(T_j'|R_j, \sigma^2)\, q(T_j|T_j')}{p(T_j|R_j, \sigma^2)\, q(T_j'|T_j)}\right). \tag{3.21}$$

The acceptance probability balances the likelihood of the data under the new and old trees with the probabilities of proposing each configuration, ensuring that the Markov chain converges to the posterior distribution. If the proposal is accepted, T_j is updated to T_j'. Otherwise, it remains unchanged.

Once the structure of T_j is updated, the values at its terminal nodes (leaf nodes) are updated. Given the structure of the tree, the terminal node values are sampled from their full conditional posterior distribution, which is normally distributed by

$$\mu_{j,k}|\mathbf{R}_j, \sigma^2 \sim \mathcal{N}\left(\frac{\sum_{i\in \text{Node}_k} R_{j,i}}{n_k + \frac{\sigma^2}{\sigma_\mu^2}}, \frac{\sigma^2}{n_k + \frac{\sigma^2}{\sigma_\mu^2}}\right) \tag{3.22}$$

with $\mathbf{R}_j$ being the vector of residuals associated with the treeT_j, n_k being the number of observations in the k-th terminal node, and σ_μ^2 being the prior variance for the terminal node values. This step ensures that the terminal node values are updated in a way that accounts for both the data and the prior distribution.

The variance parameter σ^2 is updated using a Gibbs sampling step, as its full conditional posterior distribution is an Inverse-Gamma distribution. Given the residuals $\mathbf{R} = \{R_1, R_2, \ldots, R_n\}$ after all trees have been updated and with n being the number of observations, a and b being hyperparameters for the Inverse-Gamma prior distribution on σ^2, and $R_i = (Y_i - f(\mathbf{X}_i))^2$ being the residual for observation i, the conditional posterior of σ^2 is

$$\sigma^2|\mathbf{R} \sim \text{Inverse-Gamma}\left(\frac{n}{2} + a, \frac{\sum_{i=1}^n R_i^2}{2} + b\right). \tag{3.23}$$

The Gibbs sampling step efficiently draws samples for σ^2 by directly using its conjugate prior relationship with the residual sum of squares. For a binary outcome, e.g., in case of a probit model, a unit variance with $\sigma^2 = 1$ is required to specify the prior settings.

BART is inherently capable of handling missing values due to its probabilistic nature and the way it fits trees (Carnegie & Wu, 2019; Xu et al., 2016). When missing values occur

in predictor variables, BART can address them through surrogate splits or data imputation methods. During tree construction, in surrogate splitting BART can utilize surrogate splits to handle missing data. If a primary split variable contains missing values, a secondary (surrogate) variable that best mimics the behavior of the primary split is used. This allows BART to continue splitting and building trees even when data is incomplete. BART selects surrogate splits based on similarity in data partitioning, the ability to reduce variability in the target variable, and the availability of non-missing data for the candidate surrogate variables. These surrogates ensure that the tree-building process remains robust and that missing values do not significantly hinder the algorithm's performance. Another common method for handling missing values with BART is to integrate it within a multiple imputation framework, as proposed by Kapelner and Bleich (2016). In this approach, BART is used to impute missing values and fit the model simultaneously. The steps are as follows:

1. Initialize missing values with plausible guesses, such as mean or median imputation.
2. Fit the BART model to the partially observed data.
3. Impute missing values based on posterior draws from the BART model.
4. Repeat the process, iteratively updating the imputations and model parameters until convergence.

BART's implementation typically involves setting a few key hyperparameters for the number of trees m, for the tree depth control parameters α and β, and for the prior variance σ_μ^2. Once the model is set up, the MCMC algorithm iterates through N of samples to approximate the posterior distribution, generating predictions for the response variable Y. The model's predictive uncertainty is quantified using the posterior distribution, allowing for Bayesian credible intervals to be formed around predictions.

BART's flexibility, its ability to model complex relationships, and its probabilistic treatment of missing values make it a powerful tool for modern statistical modeling. Its use of a Bayesian framework provides interpretable uncertainty quantification, and the tree-based structure offers a natural and efficient approach to dealing with missing predictor values through surrogate splits or imputation. Algorithm 2 includes steps for initializing the parameters, iteratively updating each tree using the Metropolis-Hastings step, updating terminal node values, and updating the variance parameter using Gibbs sampling. Note that the algorithm initializes the parameters, including the ensemble of trees and the variance parameter. The outer loop runs for N iterations, simulating the posterior distribution. The inner loop iterates through each tree in the ensemble, updating it using the Metropolis-Hastings step. Terminal node values are updated based on their conditional posterior distribution. The variance parameter σ^2 is updated using a Gibbs sampling step.

Algorithm 3 represents an expanded version of the BART algorithm that incorporates a robust procedure for handling missing values in the predictor matrix X. This modified algorithm integrates an imputation process into the Markov Chain Monte Carlo (MCMC) routine, allowing for simultaneous modeling and missing data imputation. The procedure begins by initializing the missing values in X using an initial imputation technique, such as

mean imputation, k-nearest neighbors imputation, or any other suitable imputation method that provides reasonable estimates for the missing entries. Once the data is initialized, the standard BART updates proceed: each tree is updated through Metropolis-Hastings sampling, where the tree structure is proposed and potentially accepted based on the residuals from the current predictions. The terminal node values G_j are adjusted to best fit the residuals, and the variance parameter σ^2 is updated using Gibbs sampling, as done in the standard BART routine. These steps ensure that the model adapts to the observed data iteratively. After completing the updates for the trees and variance, the algorithm then revisits the missing values in X by imputing them based on the current state of the model. This imputation is done using the predictions from the aggregated trees, which are summed to form an estimate for the missing entries. Importantly, these imputed values are updated iteratively across the MCMC iterations, allowing for refinement of the imputations as the model progresses. This iterative imputation process ensures that missing values are continuously updated in a manner consistent with the underlying patterns in the data, based on both the observed and predicted values. When dealing with different types of data, such as binary, count or categorical variables, several adjustments are required in the BART algorithm to appropriately model these outcomes. For classification tasks with binary outcomes, the residuals are calculated for example in a logistic regression framework, using the logit-transformed values. The residuals in this case are based on the difference between the observed binary outcome and the predicted probability, ensuring that the model outputs probabilities between 0 and 1. The loss function for binary outcomes is typically binary cross-entropy, which is used to guide model updates during the MCMC procedure. For count data (e.g., Poisson regression), the residuals are based on the predicted counts, calculated using the suitable distribution (e.g., Poisson distribution). The residuals represent the difference between the observed counts and the predicted rate of occurrence, adjusted for the nature of the data (e.g., using a log link function for Poisson regression). For nominal categorical variables, BART can be extended by using one-hot encoding or dummy variables to represent categories as binary variables. These can then be treated as separate predictors within the tree structure. If the categorical variable is ordinal, it can be treated as continuous by using its rank, or alternatively, modeled as a multiclass classification problem where the tree splits correspond to different categories. When imputing missing values for binary, count or categorical data, the imputation process must ensure that the imputed values are consistent with the data type: for binary data, the imputation could be based on the predicted class probabilities from the model's posterior distribution, for count data, imputation would sample from the posterior distribution of the counts, ensuring consistency with the data's distribution (e.g., using a Poisson distribution), and for categorical data, imputation would involve sampling from the predicted categorical distribution.

The BART algorithm can be adapted to handle these non-continuous data types by adjusting the form of the tree functions $f(T_k, \mathcal{G}_k, \mathbf{X})$, using appropriate generalized link functions for binary and count outcomes (such as logit for binary data or log for count data). The outputs of the trees would represent the log-odds or probabilities for classification tasks, and the model would aggregate these outputs to predict class membership or counts. In

conclusion, the key challenge when handling different data types (binary, count, categorical) in BART lies in modifying the residuals and prediction procedures to align with the nature of these outcomes. Specialized likelihood functions, such as logistic regression for binary outcomes or Poisson regression for count outcomes, are needed to compute residuals and fit the model properly. Additionally, appropriate imputation strategies for binary, count and categorical data ensure that missing values are filled in correctly and consistently with the underlying data distributions. By integrating these modifications, the BART framework can be effectively applied to non-continuous data types, offering a powerful tool for both predictive modeling and missing data imputation.

BART is implemented in standard statistical software, e.g., BART by Sparapani et al. (2021), bayesTree by Chipman and McCulloch (2024) and bartMachine by Kapelner and Bleich (2016), which is used in the following.

3.3.2 Extreme Gradient Boosting

XGBoost (Extreme Gradient Boosting), introduced by Chen and Guestrin (2016), is a highly efficient and scalable implementation of gradient boosting for machine learning. It is designed to efficiently handle large-scale data and complex, non-linear relationships in predictive modeling tasks. XGBoost is widely regarded for its flexibility and exceptional performance across diverse domains, including regression, classification, and ranking problems. For details see (Natekin & Knoll, 2013). Unlike standard gradient boosting, XGBoost integrates several innovations, such as regularization, parallelization, and sparsity-aware algorithms, making it a preferred choice in machine learning competitions and applications requiring accurate predictions with interpretability. XGBoost builds on the concept of gradient boosting proposed by Friedman (2001), where an ensemble of weak learners (usually decision trees) is used to iteratively minimize a differentiable loss function by learning residual errors from previous models.[8] The created ensemble of decision trees, where each subsequent tree corrects the residuals of the prior trees, are optimized by XGBoost expanding the training speed and accuracy, while also incorporating mechanisms to handle missing data and overfitting. Let the training dataset consist of n observations with p features, denoted as $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^p$ represents the feature vector for the i-th observation and $y_i \in \mathbb{R}$ is the corresponding target value. In gradient boosting, the model $\hat{y}_i^{(t)}$ at iteration t is given by:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F}, \tag{3.24}$$

where f_k represents the k-th decision tree in the ensemble, and $\mathcal{F}$ denotes the space of regression trees. Each f_k maps an input $\mathbf{x}_i$ to a predicted output through a sequence of splits.

XGBoost optimizes the following objective function at each iteration t:

$$\text{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k). \tag{3.25}$$

[8] For more details see 3.7.4.

where $L(y_i, \hat{y}_i^{(t)})$ is the loss function, typically the squared error for regression or logistic loss for classification:

$$L(y_i, \hat{y}_i^{(t)}) = (y_i - \hat{y}_i^{(t)})^2 \quad \text{(for regression with squared error)}, \tag{3.26}$$

$$L(y_i, \hat{y}_i^{(t)}) = -\left[y_i \log(\hat{y}_i^{(t)}) + (1 - y_i) \log(1 - \hat{y}_i^{(t)})\right] \quad \text{(for logistic loss)}, \tag{3.27}$$

with $y_i \in \{0, 1\}$ being the true label, and $\hat{y}_i^{(t)}$ being the predicted probability, computed from the log-odds (logit):

$$\hat{y}_i^{(t)} = \frac{1}{1 + \exp(-\hat{z}_i^{(t)})}, \tag{3.28}$$

and where $\Omega(f_k)$ is a regularization term to control model complexity:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2, \tag{3.29}$$

with T being the number of leaves in the tree, w_j being the weight of leaf j, γ penalizing the number of leaves (encouraging simpler trees), and λ regularizing the leaf weights to prevent overfitting. The loss function $L(y_i, \hat{y}_i^{(t)})$ measures the difference between the true target y_i and the model's prediction $\hat{y}_i^{(t)}$ at iteration t for regression with a squared error loss function. For binary classification, the logistic loss (or binary cross-entropy) is used to evaluate how close the predicted probability $\hat{y}_i^{(t)}$ is to the true label y_i. The regularization term $\Omega(f_k)$ controls the complexity of each tree f_k in the model by penalizing large weights and complex structures. This helps prevent overfitting and improves generalization.

The algorithm uses second-order Taylor expansion to approximate the objective function. This allows for a more efficient optimization process, where the first- and second-order gradients (denoted as g_i and h_i, respectively) with respect to the prediction $\hat{y}_i^{(t)}$ are computed:

$$g_i = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}, \quad h_i = \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)^2}}. \tag{3.30}$$

The model then minimizes the approximate objective:

$$\tilde{\text{Obj}} \approx \sum_{i=1}^{n} \left(g_i f_k(\mathbf{x}_i) + \frac{1}{2} h_i f_k(\mathbf{x}_i)^2 \right) + \Omega(f_k). \tag{3.31}$$

This quadratic approximation enables fast and accurate optimization during tree construction.

XGBoost grows trees using a greedy algorithm that selects splits based on maximizing the reduction in the loss function. For each potential split, XGBoost computes the gain, defined as the difference between the loss before and after the split. If a split results in a higher reduction in the loss, it is selected. With G_L and G_R being the sums of the gradients for the left and right nodes, and H_L and H_R being the sums of the second-order gradients for the left

and right nodes, respectively, the gain for a split at node j is calculated as:

$$\text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) - \gamma, \tag{3.32}$$

One of the critical innovations of XGBoost is its ability to handle missing values natively, without the need for imputation, see among other Zhang et al. (2020) for applying in regression settings and Deng and Lumley (2023) for combining mean matching and XGBoost. Missing values are treated as a separate case during tree splitting. When a feature contains missing values, XGBoost evaluates both possible split directions (left or right) for the missing data and assigns it to the direction that minimizes the loss.

For any split based on feature x_j, if x_j is missing for a particular observation, the model assigns the observation to the default branch - either left or right, depending on which option results in the lowest loss. This adaptive treatment of missing values prevents data loss and ensures robustness in the presence of incomplete datasets. Moreover, during the tree construction phase, XGBoost uses a sparsity-aware split finding algorithm that handles both missing values and sparsity in the feature matrix. This process dynamically adjusts the tree-building process to account for missing data, ensuring optimal splits are chosen without requiring explicit imputation.

Based on Rusdah and Murfi (2020), algorithm 4 incorporates the objective function and all the key steps for both regression and classification tasks. The algorithm outlines the iterative process of boosting and includes updates for handling the residuals and the regularization. To add missing data handling in XGBoost, the algorithm uses a strategy where the decision tree can create alternative splits (or paths) for missing values (Chen & Guestrin, 2016). When encountering missing values in the feature data, XGBoost chooses a default direction (left or right in a decision tree) based on maximizing the gain. For features with missing data, XGBoost chooses the optimal direction (either left or right) during the construction of the decision tree based on which direction provides the highest gain. This is determined by evaluating the potential impact on the objective function (usually based on improvement in squared error or logistic loss). Missing values are then automatically routed down the tree in the direction that maximizes the gain. This is handled during tree construction for each split. The rest of the algorithm proceeds as usual, with trees fit to residuals or pseudo-residuals, and regularization applied to control tree complexity. The predicted values are updated iteratively, as in standard boosting.

3.4 Experimental study

The following experimental study aims at a comparison of different strategies to achieve variable selection for a classification problem based on a binary probit framework with missing values. The simulated data is generated to follow a probit regression model with $N = 1,000$ and $P = 10$ with $y = (y_1, \ldots, y_N)'$ being a $N \times 1$ vector of binary dependent variables, a $N \times P$ matrix of covariate data $X = (X_1', \ldots, X_N')'$ not including a constant, α as corresponding constant, $e = (e_1, \ldots, e_N)'$ is a $N \times 1$ vector of independent standard normally

distributed error terms and $y^* = (y_1^*, \ldots, y_N^*)'$ a vector of latent variables. Thus, the binary regression model with probit link is given as $y_i = \mathbb{1}\{y_i^* = \alpha + X_i\beta + e_i > 0\}$.

The considered data generating process satisfies the following conditions. Next to a constant, the covariate data $X_{(p)}$ with $6, \ldots, 10$ is generated from standard normal distributions in each variable. Only the first five out of the total ten parameters are set to a non-zero value by setting the indicator vector, accordingly, including the intercept. The signs of the 10 parameters are chosen to alternate. For the sake of variation, X_1 to X_5 are drawn from a multivariate normal distribution with expectation $\mu = (0, 0, 0, 0, 0)'$ and covariance $\text{vech}(\Sigma) = (1, .85, .65, .45, .35, 1, .45, .35, .65, 1, .25, .20, 1, .30, 1)'$ mimicking a situation with correlated covariate data.[9] Finally, a total of $M = 100$ simulated datasets are generated through replication.

Building on the data-generating process described above, two distinct missing data mechanisms are introduced to examine the sensitivity of model performance under realistic conditions: missing completely at random (MCAR) and missing at random (MAR). These mechanisms reflect different assumptions about the dependence of missingness on observed and unobserved information, and they guide the subsequent design of the simulation scenarios. Simulating missing values as MCAR, the missing rate is set randomly to .35 and .20 for X_2 and X_3, respectively. For the MAR variation, a missing generating mechanism for $X_{2,i}$ and $X_{3,i}$ is considered where $X_{2,i}$ and $X_{3,i}$, respectively, is missing if $F_U(U_i) > .65$ and $F_U(U_i) > .80$, respectively. Whereby $F_U(U_i)$ denotes the empirical distribution function of the random variable U_i which is

$$U_i = \frac{1}{1 + \exp\{\omega_i\}} \quad i = 1, \ldots, N,$$

with $\omega_i = .2X_{u,i} + \rho_{u,i}$ with $u \in \{2, 3\}$ and $\rho_{u,i}$ being standard normally distributed. This results in a missing rate of 35% for X_2 and additional a missing rate of 20% for X_3. For further details on the described missing designs, see table 3.1.

Given the structure of the data and the defined missing data patterns, several machine learning and statistical models are employed to predict the binary outcome variable. These models differ in their underlying assumptions, flexibility and ability to handle incomplete data, making them particularly suitable for a comparative evaluation under varying missingness conditions. This simulation study evaluates the predictive performance of various machine learning and statistical models – XGBoost, Bayesian Additive Regression Trees (BART), Elastic net, i.e., Lasso and Ridge regression – across datasets with different treatments for missing values. Specifically, the model performance is performed across three data conditions: before deletion (BD) including the dataset before creating missing values, complete cases (CC) only including observations without missing values, and finally imputed data (IMP) the complete dataset with handling the missing values as well. The handling of missing values is described above for the XGBoost and BART, for the Elastic net the average approach described in chapter 2.

[9] Note that $\text{vech}(\cdot)$ denotes the half-vectorization operator as defined in Lütkepohl (1996).

To ensure methodological consistency and comparability across models, all algorithms are implemented in R with consistent data preparation and evaluation procedures as well as using standardized settings, with hyperparameters either fixed based on common practice or selected via internal cross-validation. Below, the key model-specific settings and tuning strategies used in this study are summarized. The XGBoost model was trained using the *XGBoost* package (Chen & Guestrin, 2016) with the *binary:logistic* objective for classification. Missing values were explicitly handled by setting *missing = NA*, allowing the algorithm to internally assign optimal default directions for missing entries during tree construction. The model was trained with the following fixed hyperparameters: maximum tree depth (*max.depth = 4*), learning rate (*eta = 1*), and number of boosting rounds (*nrounds = 2*). Although no extensive hyperparameter tuning was performed in this case, the chosen settings reflect a trade-off between computational efficiency and model expressiveness. BART models were estimated using the *bartMachine* package (Kapelner & Bleich, 2016), which supports missing data natively through a Bayesian splitting mechanism. A 30 regression trees (*num_tree = 30*) is used, with 1,000 burn-in iterations and 4,000 posterior samples retained after burn-in. The *use_missing_data = TRUE* option was enabled, allowing the model to learn from the structure of missingness directly. Elastic net models were fitted using the *glmnet* package (Friedman et al., 2010) with a probit link (*family = binomial(link = probit)*). Two variants of the elastic net were considered to assess the influence of different penalization strategies:

- Ridge regression ($\alpha = 0$): A purely L2-penalized model, favoring small but nonzero coefficients.
- Lasso ($\alpha = 1$): A purely L1-penalized model, promoting sparsity and automatic variable selection.

For both models, 100 values of the regularization parameter λ were evaluated using internal cross-validation via *cv.glmnet*. Final models were refitted using the optimal penalty parameter λ_{min} identified in this process to ensure stable coefficient estimation and predictive accuracy. Model evaluation was conducted using k-fold cross-validation, where each model was trained and validated on stratified subsets of the data.

For the imputed datasets, cross-validation was repeated independently for each of the $D = 10$ multiply imputed versions, and the results were subsequently averaged to obtain a single performance estimate. To facilitate a fair comparison across models, extensive hyperparameter tuning was deliberately restricted and was relied on standardized or commonly used settings where appropriate. While this approach enhances comparability, it may not reflect the absolute best performance of each individual method, highlighting the inherent trade-off between model-specific optimization and systematic evaluation. Consequently, the results should be interpreted as comparative rather than definitive, emphasizing relative differences in predictive performance rather than the superiority of any single model.

For M independent replications, the predictive performance of each model is assessed based on its ability to correctly classify the binary outcome variable. The primary evaluation criterion is classification accuracy, defined as the proportion of correctly predicted outcomes. To provide a more nuanced assessment of model performance, additional metrics are reported,

including precision, recall, and the F1-score, as discussed in Section 2.4 and in line with the evaluation framework proposed by Rainio et al. (2024). Furthermore, the root mean square error (RMSE) is included to quantify the average deviation between predicted probabilities and true outcomes. In addition, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is reported to account for the models' ability to discriminate between classes across different thresholds; see Section 3.7.5 for details. Together, these metrics provide a comprehensive view of predictive quality across models and data scenarios.

The evaluation results for all models across the different data scenarios–before deletion (BD), complete cases (CC), and imputed datasets (IMP) under both MCAR and MAR mechanisms–are summarized in Table 3.2. Overall, the performance differences between the four predictive approaches are relatively small in the before deletion scenario. Lasso and Ridge regression achieve the highest F-measure values (.8771 and .8776, respectively), while BART shows the highest recall (.9684) but slightly lower precision (.7972) and AUC (.7416). XGBoost exhibits slightly lower AUC (.6695) and F-measure (.8433) compared to the other methods. In the complete-case scenarios under both MAR and MCAR, Lasso and Ridge regression continue to provide the highest F-measure values, with BART maintaining the highest recall but lower precision. Notably, imputation generally increases recall for Lasso and Ridge to 1.0, while BART shows consistently high recall with slightly lower AUC. XGBoost remains the most conservative method, with lower AUC and F-measure values across all imputation and missing-data conditions. Regarding RMSE, BART consistently achieves the lowest values across almost all scenarios, indicating strong predictive accuracy in terms of continuous outcome reconstruction, whereas Elastic Net methods show slightly higher RMSE, particularly after imputation. In summary, all models display similar behavior across missing-data treatments, with BART favoring high recall and low RMSE, Elastic Net methods maximizing F-measure, and XGBoost exhibiting moderate performance across all metrics. These results highlight that the choice of predictive model and missing-data handling procedure can subtly affect performance, but overall differences remain modest.

To evaluate the robustness of the results with respect to key modeling decisions, several sensitivity analyses were conducted. These analyses examined how changes in core hyperparameters influenced predictive performance and whether the relative model rankings remained stable under alternative specifications. For XGBoost, the number of boosting iterations (*nrounds*) was varied across a typical range (e.g., 50, 100, 200), with other hyperparameters held constant. Performance improved moderately with increasing *nrounds*, as expected, but the overall ranking of models remained largely unaffected. This indicates that the conclusions regarding comparative performance are not highly sensitive to this setting. In the case of Bayesian Additive Regression Trees (BART), increasing the number of regression trees from 30 to 100 led to slight improvements in predictive accuracy, reflecting the increased capacity of the ensemble. However, the relative performance of BART compared to other models remained stable, suggesting that the method's effectiveness does not critically depend on this hyperparameter within the evaluated range. For the Elastic Net models, two regularization extremes were compared: Ridge regression ($\alpha = 0$) and Lasso ($\alpha = 1$). While both variants yielded similar overall accuracy, their behavior in terms of variable

selection differed substantially. This highlights the role of regularization choice depending on whether prediction accuracy or model interpretability is prioritized. Sensitivity analyses were conducted primarily under the MAR-imputation scenario, as it represents the most realistic and methodologically relevant condition in applied research. Additional results under the idealized before deletion setting are reported to offer a comparative reference. Complete case scenarios were not included in the sensitivity analysis, since their limited data structure and potential for bias under MAR render hyperparameter tuning less informative. The full set of results are presented in table 3.3 for BART and in tables 3.4 and 3.5 for XGBoost. Although further gains in predictive accuracy might be achievable through more extensive tuning, the chosen configurations represent a balanced compromise between model complexity, computational feasibility, and methodological comparability. Across all sensitivity checks, the relative differences in model performance remained consistent, supporting the robustness of the main findings.

Overall, the simulation study establishes a controlled and transparent framework to evaluate predictive performance under varying missing data conditions. In the next step, the outlined modeling procedures are applied to empirical data from the German National Educational Panel Study (NEPS) to assess their practical relevance and generalizability in a real-world setting.

3.5 Participation in NEPS starting cohort 4

A common issue encountered in panel studies, especially in the social science, is that of unit non-response resulting from a refusal to participate in a study (Lupu & Michelitch, 2018; Peytchev, 2009). Panel studies necessitate the continued involvement of sample units over extended periods. A significant challenge arises when a substantial number of respondents are unwilling to commit to the requisite sustained effort and decline to participate in subsequent surveys (Kleinert et al., 2019; Yan & Williams, 2022; Zinn & Gnambs, 2020). The starting cohort 4 (SC 4) is a panel study published by the National Educational Panel Study (NEPS) which commenced in 2010/11 to follow a representative sample of German students over their trajectories; for details see Blossfeld and Roßbach (2019). It is a longitudinal study of students starting in grade 9 in regular and special schools in the state. Zinn et al. (2020) discuss the attrition in all starting cohorts of the NEPS and emphasize that up to wave 9 the propensity to drop-out depends on the test person being a male student, or showing lower achievement in the mathematical competence tests. Furthermore, the level of parental education exerts a considerable influence on attrition. Students whose parents have a higher casmin score tend to demonstrate a greater willingness to participate in the panel. From wave 10 onwards, the school context is no longer a factor; instead, the individual field is the sole consideration. Consequently, the question of attrition can be re-examined in the context of this new behavior.

Based on the considerations in Zinn and Gnambs (2020) the participation in SC 4, here scientific use field (SUF) 14, is analyzed with the above mentioned methods (NEPS Network, 2024a). Tree-based methods, encompassing both machine learning and statistical learning,

can be employed to forecast future values based on a range of covariates, which can be selected from a vast array of variables (Kern et al., 2019). In the context of surveys, the anticipated response rate can be regarded as a key example of such a prediction. The participation in wave 14 is the variable of interest which is predicted by XGBoost, BART and Elastic net. The missings occurring in the SUF 14 are assumed as missing at random (Rubin, 1976). Table 3.6 provides an overview of the utilized variables and their respective missing rates. In wave 14, 3,891 of the 16,425 original participants complete the survey. The proportion of female respondents is 50.4% and the average age 28.513 years. The participation status is recorded at each wave as either participation, temporary drop-out or permanent drop-out. In the presented analysis, the participation in wave 14 is predicted, and a total of 52 variables are used based on the considerations in previous non-responses analyses or in the weighting reports, see among others Bergrab and Aßmann (2024), Zinn and Gnambs (2020), and Zinn et al. (2018). The covariates can be split in three groups. The first one contains all participation $(0 = \text{not participate in wave } t, 1 = \text{ participate in wave } t)$ up to the current wave. In the second group the individual characteristics such as age (in years), sex $(0 = \text{male}, 1 = \text{female})$, mother tongue $(0 = \text{German}, 1 = \text{non German})$ and migration background $(0 = \text{yes}, 1 = \text{no})$[10] are subsumed. Another group contains various self-reported psychological characteristic, such as satisfaction with life, current living standards, health, family life, acquaintances and friends, and school on a eleven-point response scale form 0 (completely dissatisfied) to 10 (completely satisfied). The self-rated health is measured on a five-point scale from 1 (very good) to 5 (very bad), and global self-esteem on a scale based on Rosenberg (von Collani & Herzberg, 2003). Every item in this subgroup is picked up with latest response in the survey. A bunch of competence measurements compose the last group, which is surveyed in waves 9, 10, 11, 12, and 14.

Table 3.7 presents a summary of the results obtained from the various models. The participation in wave 14 is analyzed across 13 previous waves by the above mentioned methods. The XGBoost is set up with the above-mentioned pre-defines, while for the BART $m = 300$ trees for $N = 1,000$ iterations are pre-defined as well as with prior setting $\alpha = .95$ and $\beta = 2.0$. The BART uses a MCMC algorithm with 100 burn-in and 400 posterior draws. Following Geweke (1991) the convergence is evaluated by means of visual inspections of auto-correlation and trace plots, see for details figure 3.2. The corresponding tests yielded unremarkable results and indicate stable chains, thus robust resulting posterior estimates. The Elastic net is set up with different control parameters $\varphi = 1.0$ for Lasso and $\varphi = 0.0$ for Ridge regression. To validate the results a 10-fold cross-validation is implemented.

In wave 14 the participation rate is 54.2%, whereby only permanent drop-outs are excluded from the analysis. Wave 4 and wave 6 was a path-specific survey questionnaire for special-need school, thus the missing rate is even higher than in the other waves. The results of XGBoost and BART are comparable, exhibiting a similar level of accuracy in prediction. In contrast, the two elastic net methods, Lasso and Ridge regression, demonstrate a considerably higher degree of inaccuracy, as evidenced by their lower AUC values, F-measures, and RMSE.

[10]Migration background means the person has at least between a 1.0 and 2.5 generation background, i.e., the target person is born in Germany, one parent born abroad; other parent born in Germany and one of that parent's parents born abroad.

This can be attributed to the insufficient treatment of missing values. Figure 3.1 shows the variable inclusion proportion for the BART model, i.e., the relative influence of the different covariates ranked from their importance. The red bars correspond to the standard error of the variable inclusion proportion estimates, i.e., they represent 95% credible intervals. Hence, only the participation in wave 13 has an inclusion proportion significantly higher than .10, where the corrected WLE-score for the scientific literacy and the participation in wave 6 are over the .10 bounder, but show no significant result. The weighting report by Bergrab and Aßmann (2024) lists the participation in wave 13 and in wave 6 as important and significant, too. Instead, participating in other previous waves such as waves 12, 11, and 9 has a significant influence on the weights, but don't show up in a measurable manner or are selected out by BART. Figure 3.3 shows the variable importance for XGBoost results where also participation in wave 13 has the greatest influence on the participation status in the next wave. Furthermore, the the corrected WLE-score for ICT literacy and participation in wave 12 have an influence in a measurable level.

3.6 Conclusion

This chapter has focused on the performance and handling of missing values across machine learning and statistical models. The comparison of the different methods highlights stable results for the boosting algorithm and the Bayesian regression tree, while exposing certain weaknesses in the imputation settings used by the elastic net approaches. These findings are illustrated by the AUC and MSE figures presented in tables 3.2 and 3.7. The results underline the need for further research aimed at fine-tuning analysis and prediction methodologies and identifying additional weaknesses in the methods studied. Specifically, the challenges associated with handling missing values in elastic net models warrant more detailed investigation. Future work should explore alternative imputation strategies, such as mixgb – a multiple imputation implementation leveraging XGBoost, subsampling, and predictive mean matching – as demonstrated by Trinkaus and Kauermann (2023) and Deng and Lumley (2023). Their findings indicate that mixgb achieves less biased estimates compared to other multiple imputation implementations.

In addition, model averaging across the evaluated methods offers a promising avenue for enhancing prediction accuracy and robustness. The idea that the truth does not lie in a single model suggests the potential value of combining predictions from multiple approaches to better capture the complexities of real-world phenomena. From a critical perspective, the algorithms used in this study should be further scrutinized to ensure that their assumptions and limitations are well-understood. This includes considering the specific conditions under which each method performs optimally and addressing any biases introduced by their design.

Overall, this chapter contributes to a growing body of literature that aims to improve the application of machine learning and statistical methods in the presence of missing data. However, significant opportunities remain for advancing the field, particularly in refining methods like elastic net and exploring the synergies between different predictive frameworks.

3.7 Appendix section 3

3.7.1 Algorithms

Algorithm 1 Cross-validation for shrinkage parameter

for p in 1:P **do**
  for k in 1:K **do**
    Keep fold k as hold-out data
    Use the remaining folds and $\lambda = \lambda_p$ to estimate $\hat{\beta}_{Ridge}$
    Predict hold-out data: $y_{test,k} = X_{test,k}\hat{\beta}_{Ridge}$
    Compute sum of squared residuals: $\text{SSR}_k = ||y - y_{test,k}||2$
  end for
  Average SSR over the folds: $SSR_p = \frac{1}{K}\sum_{k=1}^{K} SSR_k$
end for
Choose optimal value $\lambda_{opt} = argmin_p SSR_p$

Algorithm 2 Bayesian Additive Regression Trees (BART) routine

Input: Data $\mathbf{X}, \mathbf{Y}$, number of trees m, number of iterations N
Initialize: Set initial trees $\{T_1, T_2, \ldots, T_m\}$, terminal node values $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_m\}$, and variance parameter σ^2
for iteration $= 1$ to N **do**
    for each tree $j = 1$ to m **do**
        Compute residuals $R_j = \mathbf{Y} - \sum_{k \neq j} f(T_k, \mathcal{G}_k, \mathbf{X})$
        Propose a new tree structure T_j' using one of the following moves: grow, prune, change split, or swap
        Compute the acceptance probability:

$$\alpha = \min\left(1, \frac{p(T_j'|R_j, \sigma^2)\, q(T_j|T_j')}{p(T_j|R_j, \sigma^2)\, q(T_j'|T_j)}\right)$$

        Accept T_j' with probability α. If accepted, set $T_j \leftarrow T_j'$
        Update terminal node values $\mathcal{G}_j$ based on the new tree structure using:

$$\mu_{j,k}|\mathbf{R}_j, \sigma^2 \sim \mathcal{N}\left(\frac{\sum_{i \in \text{Node}_k} R_{j,i}}{n_k + \frac{\sigma^2}{\sigma_\mu^2}}, \frac{\sigma^2}{n_k + \frac{\sigma^2}{\sigma_\mu^2}}\right)$$

    end for
    Update the variance parameter σ^2 using the Gibbs sampling step:

$$\sigma^2 \sim \text{Inverse-Gamma}\left(\frac{n}{2} + a, \frac{\sum_{i=1}^{n} R_i^2}{2} + b\right)$$

end for
Output: Posterior samples of $\{T_j\}, \{\mathcal{G}_j\}, \sigma^2$

Algorithm 3 Bayesian Additive Regression Trees (BART) Routine with Missing Values

Input: Data $\mathbf{X}, \mathbf{Y}$ with missing values, number of trees m, number of iterations N
Initialize:
    Impute missing values in $\mathbf{X}$ using an initial method (e.g., mean imputation or k-nearest neighbors)
    Set initial trees $\{T_1, T_2, \ldots, T_m\}$, terminal node values $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_m\}$, and variance parameter σ^2
for iteration $= 1$ to N **do**
    for each tree $j = 1$ to m **do**
        Compute residuals $R_j = \mathbf{Y} - \sum_{k \neq j} f(T_k, \mathcal{G}_k, \mathbf{X})$
        Propose a new tree structure T_j' using one of the following moves: grow, prune, change split, or swap
        Compute the acceptance probability:

$$\alpha = \min\left(1, \frac{p(T_j'|R_j, \sigma^2)\, q(T_j|T_j')}{p(T_j|R_j, \sigma^2)\, q(T_j'|T_j)}\right)$$

        Accept T_j' with probability α. If accepted, set $T_j \leftarrow T_j'$
        Update terminal node values $\mathcal{G}_j$ based on the new tree structure using:

$$\mu_{j,k}|\mathbf{R}_j, \sigma^2 \sim \mathcal{N}\left(\frac{\sum_{i \in \text{Node}_k} R_{j,i}}{n_k + \frac{\sigma^2}{\sigma_\mu^2}}, \frac{\sigma^2}{n_k + \frac{\sigma^2}{\sigma_\mu^2}}\right)$$

    end for
    Update the variance parameter σ^2 using the Gibbs sampling step:

$$\sigma^2 \sim \text{Inverse-Gamma}\left(\frac{n}{2} + a, \frac{\sum_{i=1}^{n} R_i^2}{2} + b\right)$$

    Impute Missing Values:
        For each observation i with missing values, impute based on the current state of the model:
            Update $\mathbf{X}_{\text{miss},i}$ using predictions from the sum of trees:

$$\mathbf{X}_{\text{miss},i} \sim p(\mathbf{X}_{\text{miss},i}|\mathbf{X}_{\text{obs},i}, \mathbf{Y}, \mathbf{T}, \mathbf{G}, \sigma^2)$$

        Update $\mathbf{X}$ with the newly imputed values for the next iteration
end for
Output: Posterior samples of $\{T_j\}, \{\mathcal{G}_j\}, \sigma^2$, and imputed values in $\mathbf{X}$

Algorithm 4 XGBoost algorithm with handling missing values

Input: Training data $(\mathbf{x}_i, y_i)$ for $i = 1, \ldots, n$, number of iterations T, learning rate η, regularization parameters λ, γ

Initialize: Set initial prediction $\hat{y}_i^{(0)} = \bar{y}$ (mean of the target) for regression, or initial log-odds for classification.

for t = 1 to T **do**

Compute residuals (for regression):

$$r_i^{(t)} = -\frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} = y_i - \hat{y}_i^{(t-1)}$$

or pseudo-residuals (for classification):

$$r_i^{(t)} = -\frac{\partial L(y_i, \hat{z}_i^{(t-1)})}{\partial \hat{z}_i^{(t-1)}} = y_i - \frac{1}{1 + \exp(-\hat{z}_i^{(t-1)})}$$

where L is the squared error for regression or logistic loss for classification.

Handle missing values:

When constructing a decision tree $f_t(\mathbf{x})$, for any feature x_{ij} with missing values, identify the optimal direction (left or right) for missing data splits based on maximizing gain.

If x_{ij} is missing for a data point, send it in the direction that yields the highest gain during tree construction.

Fit the decision tree $f_t(\mathbf{x})$ to the residuals $r_i^{(t)}$, incorporating the missing value split strategy.

Compute the regularized objective function:

$$\text{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k),$$

where $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_j w_j^2$ penalizes tree complexity.

Update the model's prediction:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(\mathbf{x}_i)$$

for regression.

For classification, update log-odds:

$$\hat{z}_i^{(t)} = \hat{z}_i^{(t-1)} + \eta f_t(\mathbf{x}_i)$$

and compute the predicted probability:

$$\hat{y}_i^{(t)} = \frac{1}{1 + \exp(-\hat{z}_i^{(t)})}.$$

end for

Output: Final model with prediction $\hat{y}_i^{(T)}$.

3.7.2 Figures

FIGURE 3.1: Variable inclusion proportions of BART

Variable inclusion proportion for the BART model shows the relative influence of the different covariates ranked from the important ones on the left based on participation in wave 14 in starting cohort 4. The bars correspond to the standard error of the variable inclusion proportion estimates, i.e., they represent 95% credible intervals.

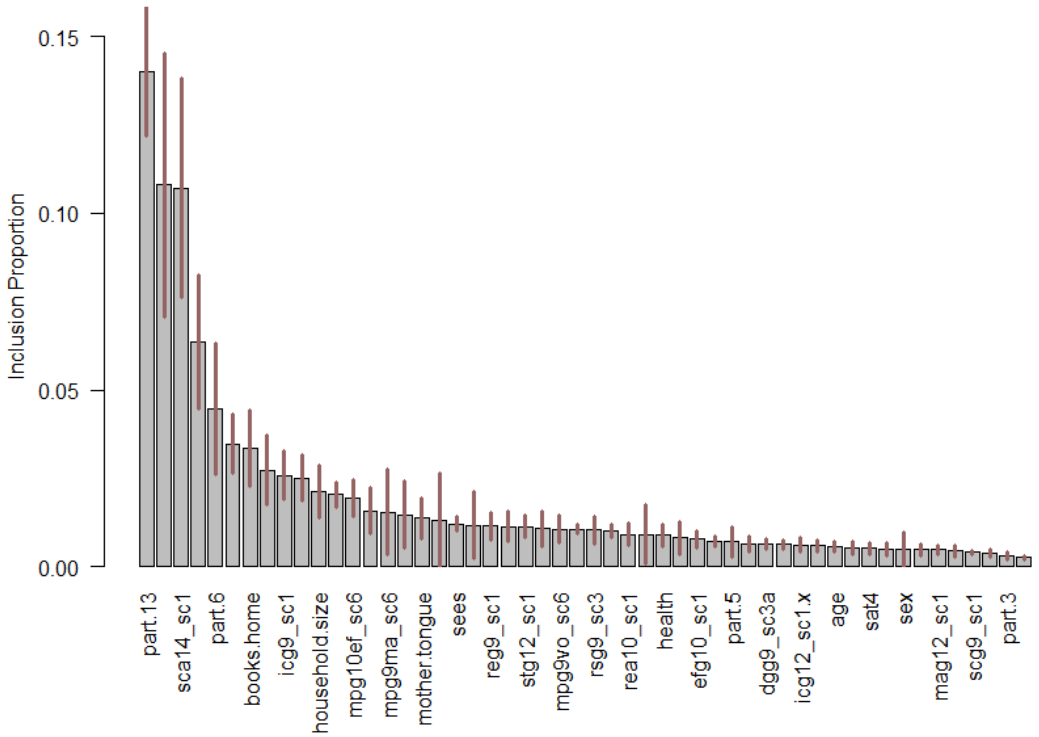


FIGURE 3.2: Convergence plots of BART

Convergence plot of BART model for participation in wave 14 in starting cohort 4.

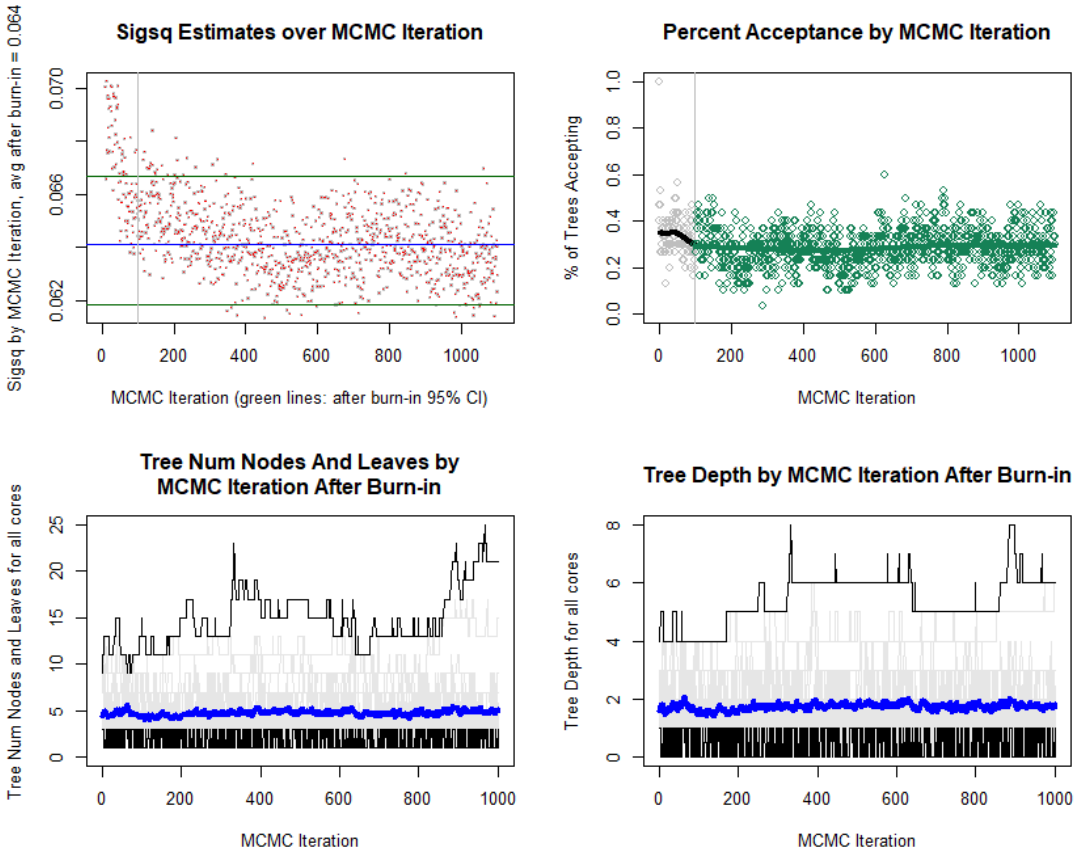
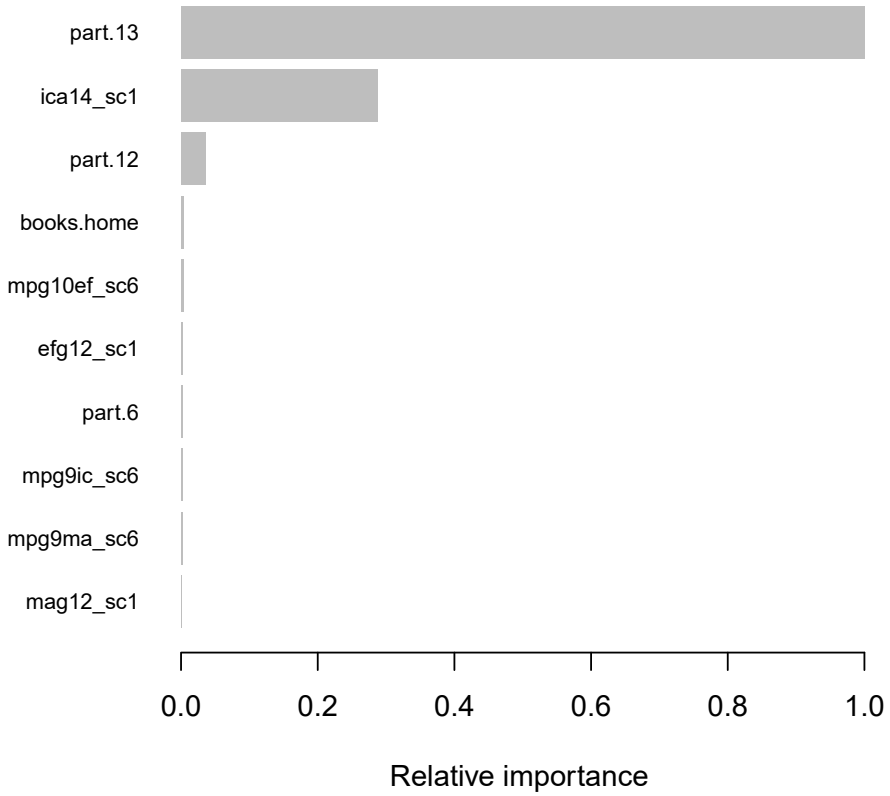


FIGURE 3.3: Variable inclusion proportions of XGBoost

Variable importance for XGBoost model shows the relative influence of the different covariates ranked from the important ones atop based on participation in wave 14 in starting cohort 4.



3.7.3 Tables

TABLE 3.1: Experimental study: Overview of the missing design

The experimental study (Ex) is split in missing completely at random (MCAR) and missing at random (MAR). The average missing rate is calculated over the $M = 100$ datasets. For MAR the missing rate is set to .35 for X_2 and .20 for X_3 as the same in the MCAR case. The complete case rate is presented as an average over the $M = 100$ datasets.

Design	Missing mechanism	Complete case rate
MCAR	$Pr(X_{2,i} = missing) = .35$ $Pr(X_{3,i} = missing) = .2$	.52
MAR	$X_{2,i} = missing$ if $1/(1 + \exp(-\omega_{2,i}))$ $w_{2,i} = .2X_{2,i} + \rho_{1,i}$ and $\rho_{1,i} \overset{\text{i.i.d.}}{\sim} N(0,1)$ $X_{3,i} = missing$ if $1/(1 + \exp(-\omega_{3,i}))$ $w_{3,i} = .2X_{3,i} + \rho_{2,i}$ and $\rho_{2,i} \overset{\text{i.i.d.}}{\sim} N(0,1)$	.52

TABLE 3.2: Experimental study: Results

Experimental study: Results of the experimental studies with missing completely at random (MCAR) and missing at random (MAR). The data generation has a correlation between $X_1, \ldots, X_5$, and additionally five redundant variables. For MCAR and MAR missings are generated in X_2 with .35 and in X_3 with .20, for details in the missing process see table 3.1. The AUC, RMSE, precision, recall and F-measure are presented after averaging over 100 replications for validating the XGBoost, BART, Lasso and Ridge regression.

	AUC	RMSE	precision	recall	F-measure
Before deletion					
XGBoost	.6695	.5020	.8120	.8789	.8433
BART	.7416	.4618	.7972	.9684	.8735
Lasso	.7711	.3796	.8138	.9530	.8771
Ridge regression	.7691	.3802	.8072	.9633	.8776
MAR: complete case					
XGBoost	.6676	.4926	.8184	.8889	.8506
BART	.7233	.4451	.8043	.9861	.8846
Lasso	.7626	.3745	.8178	.9607	.8821
Ridge regression	.7623	.3743	.8157	.9655	.8831
MAR: imputation					
XGBoost	.6412	.5126	.8042	.8741	.8368
BART	.7145	.4669	.7864	.9864	.8751
Lasso	.7779	.4758	.7730	1.0000	.8718
Ridge regression	.7816	.4592	.7867	1.0000	.8800
MCAR: complete case					
XGBoost	.6626	.5014	.8115	.8806	.8429
BART	.7306	.4696	.7860	.9762	.8695
Lasso	.7633	.3825	.8088	.9544	.8740
Ridge regression	.7620	.3826	.8056	.9590	.8743
MCAR: imputation					
XGBoost	.6399	.5100	.8054	.8767	.8386
BART	.7157	.4640	.7909	.9828	.8765
Lasso	.7710	.4697	.7778	1.000	.8746
Ridge regression	.7797	.4653	.7820	1.000	.8772

TABLE 3.3: Experimental study: Sensitivity analysis for BART

Sensitivity analysis for BART: variation of `num_tree` under MAR (imputation) and before deletion scenarios.

`number of tree`	MAR: imputation			before deletion		
	AUC	RMSE	F-measure	AUC	RMSE	F-measure
30	.7145	.4669	.8751	.7416	.4618	.8735
50	.7134	.4680	.8748	.7481	.4603	.8768
100	.7140	.4702	.8737	.7491	.4606	.8771
200	.7135	.4700	.8739	.7473	.4619	.8766

Note. BART estimated using `bartMachine` with missing data support and fixed burn-in. Other parameters unchanged.

TABLE 3.4: Experimental study: Sensitivity analysis for XGBoost

Sensitivity analysis for XGBoost: variation of `nrounds` under MAR (imputation) and before deletion scenarios.

`nrounds`	MAR: imputation			before deletion		
	AUC	RMSE	F-measure	AUC	RMSE	F-measure
22	.6676	.4926	.8506	.6695	.5020	.8433
50	.6271	.4480	.8742	.6532	.4415	.8710
100	.6362	.4372	.8838	.6642	.4375	.8807
200	.6609	.5071	.8424	.6897	.4933	.8505

Note. Results for XGBoost with fixed depth and learning rate. Other parameters held constant.

TABLE 3.5: Sensitivity analysis for XGBoost over multiple datasets

This table reports the results of a comprehensive sensitivity analysis of key parameters in the XGBoost algorithm, evaluated across multiple datasets under the MAR (Missing At Random) scenarios. The parameters examined include: maximum depth, which controls the maximum depth of each individual tree and thereby the model complexity; number of boosting rounds, indicating the total number of boosting iterations (i.e., the number of trees); learning rate η, the learning rate that determines the contribution of each tree to the final model; subsample ratio, the proportion of the training data randomly sampled for growing each tree; and column subsample by tree, the fraction of features randomly sampled for each tree. Performance is assessed using three metrics: the area under the ROC curve (AUC), the root mean squared error (RMSE), and the F1 score (F-measure), which balances precision and recall. These results allow insight into the stability and predictive quality of XGBoost under various regularization and stochasticity configurations.

Max. depth	Number of boosting rounds	Learning rate (η)	Subsample ratio	Column subsample by tree	AUC	RMSE	F-measure
5	150	0.10	1.0	0.8	0.6412	0.4827	0.8607
3	150	0.10	0.8	1.0	0.6582	0.4990	0.8500
5	100	0.01	1.0	0.8	0.6539	0.4701	0.8720
7	50	0.10	0.6	1.0	0.6512	0.4754	0.8655
3	50	0.01	1.0	0.8	0.6780	0.4658	0.8745
3	100	0.01	0.6	1.0	0.6858	0.4648	0.8744
5	100	0.10	1.0	1.0	0.6528	0.4837	0.8605
5	50	0.30	0.8	0.8	0.6378	0.4950	0.8509
7	150	0.01	0.8	0.6	0.6581	0.4637	0.8767
7	150	0.10	0.8	0.8	0.6346	0.4930	0.8539
3	100	0.01	0.8	0.8	0.6878	0.4680	0.8733
3	150	0.30	0.8	0.6	0.6346	0.5089	0.8407
5	150	0.30	0.6	1.0	0.6168	0.5050	0.8442
3	150	0.30	0.6	1.0	0.5995	0.5158	0.8348
7	150	0.30	0.6	0.8	0.6179	0.5196	0.8339
5	100	0.30	0.6	0.8	0.6218	0.4940	0.8509
5	50	0.10	0.6	1.0	0.6601	0.4899	0.8575
5	50	0.01	0.8	0.8	0.6693	0.4680	0.8734
7	50	0.30	0.6	1.0	0.6112	0.5040	0.8446
7	50	0.10	1.0	0.6	0.6507	0.4754	0.8673
3	100	0.01	1.0	0.6	0.6736	0.4669	0.8747
5	50	0.30	1.0	1.0	0.6399	0.5000	0.8499
3	50	0.10	1.0	0.8	0.6728	0.4680	0.8729
5	100	0.30	0.6	0.6	0.5945	0.5089	0.8424
7	100	0.01	1.0	0.8	0.6519	0.4743	0.8688
5	150	0.10	1.0	0.6	0.6280	0.4909	0.8562
3	150	0.10	0.6	1.0	0.6435	0.4950	0.8521
7	50	0.01	0.6	1.0	0.6613	0.4754	0.8676

TABLE 3.5: Sensitivity analysis for XGBoost parameters over multiple datasets *(continued)*

Max. depth	Number of boosting rounds	Learning rate (η)	Subsample ratio	Column subsample by tree	AUC	RMSE	F-measure
5	150	0.30	0.8	1.0	0.6159	0.5167	0.8360
7	150	0.30	0.8	0.6	0.6153	0.5070	0.8437
5	150	0.10	1.0	0.8	0.6702	0.4722	0.8681
3	150	0.10	0.8	1.0	0.6898	0.4806	0.8631
5	100	0.01	1.0	0.8	0.7234	0.4583	0.8795
7	50	0.10	0.6	1.0	0.6959	0.4593	0.8765
3	50	0.01	1.0	0.8	0.7235	0.4483	0.8857
3	100	0.01	0.6	1.0	0.7273	0.4583	0.8791
5	100	0.10	1.0	1.0	0.6723	0.4837	0.8616
5	50	0.30	0.8	0.8	0.6625	0.4858	0.8599
7	150	0.01	0.8	0.6	0.6949	0.4539	0.8833
7	150	0.10	0.8	0.8	0.6738	0.4764	0.8659
3	100	0.01	0.8	0.8	0.7255	0.4494	0.8852
3	150	0.30	0.8	0.6	0.6596	0.4743	0.8650
5	150	0.30	0.6	1.0	0.6656	0.4868	0.8568
3	150	0.30	0.6	1.0	0.6591	0.4827	0.8606
7	150	0.30	0.6	0.8	0.6559	0.4775	0.8630
5	100	0.30	0.6	0.8	0.6534	0.4899	0.8569
5	50	0.10	0.6	1.0	0.6975	0.4712	0.8702
5	50	0.01	0.8	0.8	0.7183	0.4517	0.8835
7	50	0.30	0.6	1.0	0.6806	0.4909	0.8547
7	50	0.10	1.0	0.6	0.6641	0.4669	0.8737
3	100	0.01	1.0	0.6	0.7270	0.4494	0.8860
5	50	0.30	1.0	1.0	0.6746	0.4796	0.8630
3	50	0.10	1.0	0.8	0.7130	0.4583	0.8794
5	100	0.30	0.6	0.6	0.6711	0.4743	0.8647
7	100	0.01	1.0	0.8	0.7114	0.4572	0.8792
5	150	0.10	1.0	0.6	0.6581	0.4754	0.8672
3	150	0.10	0.6	1.0	0.6941	0.4817	0.8628
7	50	0.01	0.6	1.0	0.7151	0.4637	0.8749
5	150	0.30	0.8	1.0	0.6718	0.4775	0.8639
7	150	0.30	0.8	0.6	0.6542	0.4733	0.8672
5	150	0.10	1.0	0.8	0.6258	0.5000	0.8513
3	150	0.10	0.8	1.0	0.6314	0.5030	0.8480
5	100	0.01	1.0	0.8	0.6805	0.4919	0.8583
7	50	0.10	0.6	1.0	0.6498	0.4889	0.8569
3	50	0.01	1.0	0.8	0.6955	0.4909	0.8614
3	100	0.01	0.6	1.0	0.7091	0.4796	0.8667
5	100	0.10	1.0	1.0	0.6234	0.4940	0.8554

TABLE 3.5: Sensitivity analysis for XGBoost parameters over multiple datasets *(continued)*

Max. depth	Number of boosting rounds	Learning rate (η)	Subsample ratio	Column subsample by tree	AUC	RMSE	F-measure
5	50	0.30	0.8	0.8	0.6258	0.5050	0.8448
7	150	0.01	0.8	0.6	0.6650	0.4848	0.8659
7	150	0.10	0.8	0.8	0.6326	0.5020	0.8475
3	100	0.01	0.8	0.8	0.7062	0.4817	0.8674
3	150	0.30	0.8	0.6	0.6116	0.5215	0.8327
5	150	0.30	0.6	1.0	0.6139	0.5187	0.8349
3	150	0.30	0.6	1.0	0.6031	0.5119	0.8396
7	150	0.30	0.6	0.8	0.5977	0.5357	0.8243
5	100	0.30	0.6	0.8	0.6095	0.5187	0.8362
5	50	0.10	0.6	1.0	0.6707	0.4970	0.8529
5	50	0.01	0.8	0.8	0.6936	0.4796	0.8674
7	50	0.30	0.6	1.0	0.6134	0.5215	0.8348
7	50	0.10	1.0	0.6	0.6457	0.4899	0.8588
3	100	0.01	1.0	0.6	0.6930	0.4837	0.8673
5	50	0.30	1.0	1.0	0.6209	0.5079	0.8432
3	50	0.10	1.0	0.8	0.6898	0.4785	0.8655
5	100	0.30	0.6	0.6	0.5713	0.5225	0.8328
7	100	0.01	1.0	0.8	0.6558	0.4909	0.8580
5	150	0.10	1.0	0.6	0.6284	0.5000	0.8516
3	150	0.10	0.6	1.0	0.6571	0.5050	0.8458
7	50	0.01	0.6	1.0	0.6923	0.4827	0.8627
5	150	0.30	0.8	1.0	0.6015	0.5109	0.8411
7	150	0.30	0.8	0.6	0.6040	0.5089	0.8422
5	150	0.10	1.0	0.8	0.6802	0.5000	0.8454
3	150	0.10	0.8	1.0	0.6955	0.4980	0.8473
5	100	0.01	1.0	0.8	0.7111	0.4722	0.8675
7	50	0.10	0.6	1.0	0.6949	0.4980	0.8484
3	50	0.01	1.0	0.8	0.7067	0.4754	0.8658
3	100	0.01	0.6	1.0	0.7147	0.4817	0.8616
5	100	0.10	1.0	1.0	0.6934	0.4960	0.8487
5	50	0.30	0.8	0.8	0.6673	0.5099	0.8390
7	150	0.01	0.8	0.6	0.6909	0.4837	0.8637
7	150	0.10	0.8	0.8	0.6813	0.5040	0.8433
3	100	0.01	0.8	0.8	0.7094	0.4743	0.8678
3	150	0.30	0.8	0.6	0.6678	0.5225	0.8305
5	150	0.30	0.6	1.0	0.6682	0.5215	0.8302
3	150	0.30	0.6	1.0	0.6681	0.5167	0.8327
7	150	0.30	0.6	0.8	0.6702	0.5099	0.8390
5	100	0.30	0.6	0.8	0.6545	0.5187	0.8313

TABLE 3.5: Sensitivity analysis for XGBoost parameters over multiple datasets *(continued)*

Max. depth	Number of boosting rounds	Learning rate (η)	Subsample ratio	Column subsample by tree	AUC	RMSE	F-measure
5	50	0.10	0.6	1.0	0.7025	0.4990	0.8475
5	50	0.01	0.8	0.8	0.7023	0.4701	0.8698
7	50	0.30	0.6	1.0	0.6544	0.5089	0.8388
7	50	0.10	1.0	0.6	0.6772	0.4960	0.8529
3	100	0.01	1.0	0.6	0.7146	0.4858	0.8644
5	50	0.30	1.0	1.0	0.6834	0.5128	0.8359
3	50	0.10	1.0	0.8	0.7177	0.4827	0.8607
5	100	0.30	0.6	0.6	0.6625	0.5196	0.8309
7	100	0.01	1.0	0.8	0.6988	0.4940	0.8526
5	150	0.10	1.0	0.6	0.6689	0.5109	0.8398
3	150	0.10	0.6	1.0	0.6900	0.4970	0.8473
7	50	0.01	0.6	1.0	0.6980	0.4806	0.8617
5	150	0.30	0.8	1.0	0.6835	0.5020	0.8447
7	150	0.30	0.8	0.6	0.6591	0.5235	0.8301
5	150	0.10	1.0	0.8	0.6853	0.4980	0.8523
3	150	0.10	0.8	1.0	0.7007	0.4817	0.8609
5	100	0.01	1.0	0.8	0.7007	0.4722	0.8720
7	50	0.10	0.6	1.0	0.7064	0.4919	0.8564
3	50	0.01	1.0	0.8	0.6994	0.4712	0.8736
3	100	0.01	0.6	1.0	0.7240	0.4690	0.8741
5	100	0.10	1.0	1.0	0.6986	0.4879	0.8591
5	50	0.30	0.8	0.8	0.6737	0.5010	0.8481
7	150	0.01	0.8	0.6	0.7008	0.4722	0.8729
7	150	0.10	0.8	0.8	0.6822	0.4980	0.8510
3	100	0.01	0.8	0.8	0.7203	0.4669	0.8760
3	150	0.30	0.8	0.6	0.6569	0.5138	0.8370
5	150	0.30	0.6	1.0	0.6650	0.5030	0.8446
3	150	0.30	0.6	1.0	0.6680	0.5000	0.8457
7	150	0.30	0.6	0.8	0.6629	0.4990	0.8491
5	100	0.30	0.6	0.8	0.6554	0.5050	0.8446
5	50	0.10	0.6	1.0	0.7053	0.4930	0.8566
5	50	0.01	0.8	0.8	0.7132	0.4743	0.8712
7	50	0.30	0.6	1.0	0.6817	0.4990	0.8488
7	50	0.10	1.0	0.6	0.6766	0.4868	0.8623
3	100	0.01	1.0	0.6	0.7011	0.4754	0.8719
5	50	0.30	1.0	1.0	0.6749	0.4950	0.8522
3	50	0.10	1.0	0.8	0.7212	0.4785	0.8670
5	100	0.30	0.6	0.6	0.6726	0.5010	0.8490
7	100	0.01	1.0	0.8	0.6965	0.4775	0.8674

TABLE 3.5: Sensitivity analysis for XGBoost parameters over multiple datasets *(continued)*

Max. depth	Number of boosting rounds	Learning rate (η)	Subsample ratio	Column subsample by tree	AUC	RMSE	F-measure
5	150	0.10	1.0	0.6	0.6827	0.4980	0.8529
3	150	0.10	0.6	1.0	0.7139	0.4848	0.8587
7	50	0.01	0.6	1.0	0.7102	0.4775	0.8666
5	150	0.30	0.8	1.0	0.6584	0.5040	0.8456
7	150	0.30	0.8	0.6	0.6393	0.5079	0.8432
5	150	0.10	1.0	0.8	0.6541	0.4722	0.8698
3	150	0.10	0.8	1.0	0.6597	0.4680	0.8732
5	100	0.01	1.0	0.8	0.6952	0.4528	0.8842
7	50	0.10	0.6	1.0	0.6665	0.4637	0.8762
3	50	0.01	1.0	0.8	0.7089	0.4405	0.8914
3	100	0.01	0.6	1.0	0.7096	0.4450	0.8884
5	100	0.10	1.0	1.0	0.6569	0.4680	0.8724
5	50	0.30	0.8	0.8	0.6449	0.4712	0.8701
7	150	0.01	0.8	0.6	0.6629	0.4483	0.8873
7	150	0.10	0.8	0.8	0.6421	0.4722	0.8703
3	100	0.01	0.8	0.8	0.7044	0.4427	0.8905
3	150	0.30	0.8	0.6	0.6360	0.4879	0.8591
5	150	0.30	0.6	1.0	0.6288	0.4899	0.8585
3	150	0.30	0.6	1.0	0.6064	0.4879	0.8586
7	150	0.30	0.6	0.8	0.6137	0.4775	0.8654
5	100	0.30	0.6	0.8	0.6171	0.4990	0.8522
5	50	0.10	0.6	1.0	0.6784	0.4604	0.8772
5	50	0.01	0.8	0.8	0.6866	0.4472	0.8877
7	50	0.30	0.6	1.0	0.6442	0.4879	0.8599
7	50	0.10	1.0	0.6	0.6412	0.4572	0.8810
3	100	0.01	1.0	0.6	0.6958	0.4427	0.8905
5	50	0.30	1.0	1.0	0.6631	0.4879	0.8605
3	50	0.10	1.0	0.8	0.6997	0.4506	0.8851
5	100	0.30	0.6	0.6	0.6262	0.4930	0.8576
7	100	0.01	1.0	0.8	0.6841	0.4561	0.8814
5	150	0.10	1.0	0.6	0.6432	0.4785	0.8672
3	150	0.10	0.6	1.0	0.6554	0.4680	0.8726
7	50	0.01	0.6	1.0	0.6789	0.4517	0.8838
5	150	0.30	0.8	1.0	0.6331	0.4785	0.8656
7	150	0.30	0.8	0.6	0.6310	0.4669	0.8730
5	150	0.10	1.0	0.8	0.7104	0.4764	0.8649
3	150	0.10	0.8	1.0	0.7143	0.4733	0.8661
5	100	0.01	1.0	0.8	0.7265	0.4561	0.8797
7	50	0.10	0.6	1.0	0.7243	0.4438	0.8831

TABLE 3.5: Sensitivity analysis for XGBoost parameters over multiple datasets *(continued)*

Max. depth	Number of boosting rounds	Learning rate (η)	Subsample ratio	Column subsample by tree	AUC	RMSE	F-measure
3	50	0.01	1.0	0.8	0.7136	0.4669	0.8760
3	100	0.01	0.6	1.0	0.7276	0.4637	0.8759
5	100	0.10	1.0	1.0	0.7162	0.4604	0.8747
5	50	0.30	0.8	0.8	0.7173	0.4764	0.8628
7	150	0.01	0.8	0.6	0.7237	0.4615	0.8790
7	150	0.10	0.8	0.8	0.7163	0.4669	0.8702
3	100	0.01	0.8	0.8	0.7214	0.4680	0.8751
3	150	0.30	0.8	0.6	0.7074	0.4785	0.8605
5	150	0.30	0.6	1.0	0.6952	0.4743	0.8637
3	150	0.30	0.6	1.0	0.6880	0.4858	0.8568
7	150	0.30	0.6	0.8	0.7046	0.4680	0.8679
5	100	0.30	0.6	0.8	0.6920	0.4827	0.8591
5	50	0.10	0.6	1.0	0.7142	0.4583	0.8760
5	50	0.01	0.8	0.8	0.7220	0.4648	0.8765
7	50	0.30	0.6	1.0	0.7057	0.4743	0.8628
7	50	0.10	1.0	0.6	0.7160	0.4648	0.8730
3	100	0.01	1.0	0.6	0.7084	0.4680	0.8765
5	50	0.30	1.0	1.0	0.7219	0.4680	0.8684
3	50	0.10	1.0	0.8	0.7217	0.4593	0.8776
5	100	0.30	0.6	0.6	0.6904	0.4733	0.8649
7	100	0.01	1.0	0.8	0.7220	0.4626	0.8756
5	150	0.10	1.0	0.6	0.7124	0.4690	0.8699
3	150	0.10	0.6	1.0	0.7227	0.4583	0.8747
7	50	0.01	0.6	1.0	0.6957	0.4604	0.8756
5	150	0.30	0.8	1.0	0.6981	0.4733	0.8650
7	150	0.30	0.8	0.6	0.7082	0.4817	0.8602
5	150	0.10	1.0	0.8	0.6843	0.4848	0.8621
3	150	0.10	0.8	1.0	0.6958	0.4785	0.8650
5	100	0.01	1.0	0.8	0.6934	0.4669	0.8747
7	50	0.10	0.6	1.0	0.6947	0.4690	0.8704
3	50	0.01	1.0	0.8	0.6960	0.4669	0.8756
3	100	0.01	0.6	1.0	0.7134	0.4583	0.8796
5	100	0.10	1.0	1.0	0.6868	0.4785	0.8655
5	50	0.30	0.8	0.8	0.6767	0.4848	0.8601
7	150	0.01	0.8	0.6	0.7024	0.4648	0.8775
7	150	0.10	0.8	0.8	0.6871	0.4806	0.8636
3	100	0.01	0.8	0.8	0.7063	0.4583	0.8801
3	150	0.30	0.8	0.6	0.6753	0.4950	0.8531
5	150	0.30	0.6	1.0	0.6824	0.4909	0.8558

TABLE 3.5: Sensitivity analysis for XGBoost parameters over multiple datasets *(continued)*

Max. depth	Number of boosting rounds	Learning rate (η)	Subsample ratio	Column subsample by tree	AUC	RMSE	F-measure
3	150	0.30	0.6	1.0	0.6623	0.5070	0.8431
7	150	0.30	0.6	0.8	0.6683	0.4970	0.8518
5	100	0.30	0.6	0.8	0.6863	0.4970	0.8526
5	50	0.10	0.6	1.0	0.7020	0.4701	0.8702
5	50	0.01	0.8	0.8	0.7006	0.4658	0.8756
7	50	0.30	0.6	1.0	0.6885	0.4879	0.8578
7	50	0.10	1.0	0.6	0.6743	0.4712	0.8720
3	100	0.01	1.0	0.6	0.7091	0.4615	0.8798
5	50	0.30	1.0	1.0	0.6773	0.4879	0.8583
3	50	0.10	1.0	0.8	0.6952	0.4593	0.8779
5	100	0.30	0.6	0.6	0.6656	0.4930	0.8538
7	100	0.01	1.0	0.8	0.6929	0.4733	0.8705
5	150	0.10	1.0	0.6	0.6786	0.4858	0.8610
3	150	0.10	0.6	1.0	0.6926	0.4817	0.8631
7	50	0.01	0.6	1.0	0.7099	0.4572	0.8795
5	150	0.30	0.8	1.0	0.6769	0.4868	0.8581
7	150	0.30	0.8	0.6	0.6339	0.4858	0.8592
5	150	0.10	1.0	0.8	0.6542	0.5167	0.8362
3	150	0.10	0.8	1.0	0.6771	0.5099	0.8409
5	100	0.01	1.0	0.8	0.6885	0.5050	0.8490
7	50	0.10	0.6	1.0	0.6735	0.5187	0.8354
3	50	0.01	1.0	0.8	0.7035	0.4950	0.8571
3	100	0.01	0.6	1.0	0.7065	0.4899	0.8586
5	100	0.10	1.0	1.0	0.6743	0.5128	0.8384
5	50	0.30	0.8	0.8	0.6557	0.5148	0.8348
7	150	0.01	0.8	0.6	0.6771	0.4990	0.8564
7	150	0.10	0.8	0.8	0.6561	0.5206	0.8338
3	100	0.01	0.8	0.8	0.7074	0.4990	0.8556
3	150	0.30	0.8	0.6	0.6582	0.5273	0.8257
5	150	0.30	0.6	1.0	0.6589	0.5329	0.8217
3	150	0.30	0.6	1.0	0.6591	0.5263	0.8276
7	150	0.30	0.6	0.8	0.6513	0.5148	0.8347
5	100	0.30	0.6	0.8	0.6439	0.5301	0.8259
5	50	0.10	0.6	1.0	0.6920	0.5050	0.8433
5	50	0.01	0.8	0.8	0.6940	0.5079	0.8491
7	50	0.30	0.6	1.0	0.6482	0.5282	0.8264
7	50	0.10	1.0	0.6	0.6656	0.5040	0.8490
3	100	0.01	1.0	0.6	0.6966	0.4980	0.8573
5	50	0.30	1.0	1.0	0.6493	0.5320	0.8245

TABLE 3.5: Sensitivity analysis for XGBoost parameters over multiple datasets *(continued)*

Max. depth	Number of boosting rounds	Learning rate (η)	Subsample ratio	Column subsample by tree	AUC	RMSE	F-measure
3	50	0.10	1.0	0.8	0.7067	0.5010	0.8503
5	100	0.30	0.6	0.6	0.6427	0.5225	0.8300
7	100	0.01	1.0	0.8	0.6823	0.5050	0.8488
5	150	0.10	1.0	0.6	0.6685	0.5158	0.8376
3	150	0.10	0.6	1.0	0.6807	0.5138	0.8382
7	50	0.01	0.6	1.0	0.6910	0.5070	0.8464
5	150	0.30	0.8	1.0	0.6411	0.5263	0.8278
7	150	0.30	0.8	0.6	0.6438	0.5394	0.8190
5	150	0.10	1.0	0.8	0.7082	0.4806	0.8598
3	150	0.10	0.8	1.0	0.7172	0.4806	0.8602
5	100	0.01	1.0	0.8	0.7052	0.4858	0.8606
7	50	0.10	0.6	1.0	0.7213	0.4680	0.8676
3	50	0.01	1.0	0.8	0.7110	0.4754	0.8696
3	100	0.01	0.6	1.0	0.7220	0.4796	0.8645
5	100	0.10	1.0	1.0	0.7061	0.4909	0.8538
5	50	0.30	0.8	0.8	0.6785	0.4950	0.8502
7	150	0.01	0.8	0.6	0.7142	0.4743	0.8706
7	150	0.10	0.8	0.8	0.7074	0.4785	0.8615
3	100	0.01	0.8	0.8	0.7222	0.4754	0.8695
3	150	0.30	0.8	0.6	0.6991	0.4940	0.8503
5	150	0.30	0.6	1.0	0.6751	0.4960	0.8489
3	150	0.30	0.6	1.0	0.6890	0.4919	0.8517
7	150	0.30	0.6	0.8	0.6787	0.4817	0.8590
5	100	0.30	0.6	0.8	0.6831	0.4970	0.8478
5	50	0.10	0.6	1.0	0.7069	0.4837	0.8584
5	50	0.01	0.8	0.8	0.7151	0.4775	0.8669
7	50	0.30	0.6	1.0	0.7076	0.4785	0.8593
7	50	0.10	1.0	0.6	0.7039	0.4658	0.8712
3	100	0.01	1.0	0.6	0.7040	0.4733	0.8731
5	50	0.30	1.0	1.0	0.7018	0.4858	0.8558
3	50	0.10	1.0	0.8	0.7262	0.4775	0.8654
5	100	0.30	0.6	0.6	0.6920	0.4970	0.8468
7	100	0.01	1.0	0.8	0.6963	0.4858	0.8593
5	150	0.10	1.0	0.6	0.7060	0.4701	0.8663
3	150	0.10	0.6	1.0	0.7106	0.4889	0.8552
7	50	0.01	0.6	1.0	0.7110	0.4690	0.8682
5	150	0.30	0.8	1.0	0.6858	0.4827	0.8571
7	150	0.30	0.8	0.6	0.7058	0.4837	0.8575

TABLE 3.6: NEPS-SC4: Overview of variables

Overview of the variables from the NEPS-SC4 (wave 14) summarized by mean, standard error, range and missing rate of the various variables.

Variable name	Mean	Standard error	Range from	Range to	Missing rate
Participation (1 = participate in)					
Wave 1	.987	.113	.000	1.000	.000
Wave 2	.945	.229	.000	1.000	.000
Wave 3	.901	.299	.000	1.000	.000
Wave 4	.084	.277	.000	1.000	.000
Wave 5	.932	.252	.000	1.000	.000
Wave 6	.405	.491	.000	1.000	.000
Wave 7	.911	.284	.000	1.000	.000
Wave 8	.862	.345	.000	1.000	.000
Wave 9	.918	.275	.000	1.000	.000
Wave 10	.845	.362	.000	1.000	.000
Wave 11	.692	.462	.000	1.000	.000
Wave 12	.546	.498	.000	1.000	.000
Wave 13	.519	.500	.000	1.000	.000
Wave 14	.542	.498	.000	1.000	.000
Individual characteristics					
Sex (0 = male, 1 = female)	.504	.500	.000	1.000	.004
Age in years	28.513	.676	25.000	34.000	.042
Mother's tongue (0 = German, 1 = Non-German)	.159	.366	.000	1.000	.026
Migration background (0 = no, 1 = yes)	.162	.369	.000	1.000	.011
Household size (absolute)	2.624	1.902	1.000	99.000	.003
Books at home (absolute)	4.002	1.471	1.000	6.000	.056
Psychological characteristics					
Satisfaction with life (0 = low, 10 = high)	7.276	2.055	.000	10.00	.032
Satisfaction with standard living (0 = low, 10 = high)	7.948	1.990	.000	10.00	.032
Satisfaction with health (0 = low, 10 = high)	7.856	2.247	.000	10.00	.032
Satisfaction with family life (0 = low, 10 = high)	7.715	2.396	.000	10.00	.032
Satisfaction acquaintances and friends (0 = low, 10 = high)	8.241	1.964	.000	10.00	.032
Satisfaction with school (0 = low, 10 = high)	6.553	2.351	.000	10.00	.032
Self-rated health (1 = very good, 5 = very bad)	1.958	.829	1.000	5.000	.054
Global self-esteem (sum score; 10–50, higher=more)	38.978	6.683	10.000	50.000	.102
competence measurements wave 1					
Scientific literacy: WLE (corrected, higher=more)	.209	1.006	-2.712	5.287	.093
Procedural metacognition (vocabulary)	.660	.206	.000	1.000	.111
Procedural metacognition (mathematics)	.665	.222	.000	1.000	.111
Procedural metacognition (ICT)	.645	.176	.000	1.000	.117
Mathematical competence: WLE (corrected, higher=more)	.256	1.246	-4.066	4.619	.091
ICT literacy: WLE (corrected, higher=more)	.178	.921	-3.288	4.128	.092
Procedural metacognition (science)	.072	.206	-.857	.857	.127
Reading speed (sum score; 0–51)	34.791	8.581	.000	51.000	.091
DGCF (perceptual speed; score; 3–93)	59.374	13.397	3.000	93.000	.117
Declarative metacognition	.821	.122	.000	1.000	.117
Procedural metacognition (reading total)	.735	.168	.000	1.000	.135

TABLE 3.6: NEPS-SC4: Overview of variables *(continued)*

Variable name	Mean	Standard error	Range from	Range to	Missing rate
Reading competence: WLE (corrected, higher=more)	.223	1.235	-4.746	3.300	.119
competence measurements wave 3					
Procedural metacognition (English)	.628	.194	.000	1.000	.294
English reading competence: WLE (corrected)	.268	1.607	-6.993	6.344	.277
competence measurements wave 5					
Scientific literacy: WLE (corrected, higher=more)	.088	.850	-2.916	5.290	.650
competence measurements wave 7					
ICT literacy: WLE (corrected, higher=more)	.757	.795	-3.741	4.111	.540
Mathematical competence: WLE (corrected, higher=more)	.113	1.123	-3.524	4.125	.545
Reading competence: WLE (corrected, higher=more)	.117	.984	-3.546	4.314	.540
Scientific thinking: WLE (corrected, higher=more)	.070	.647	-2.246	3.093	.682
English reading competence: WLE (corrected, higher=more)	.184	1.595	-5.038	5.383	.684
competence measurements wave 10					
Mathematical competence: WLE (corrected, higher=more)	-.128	.870	-6.608	3.139	.267
DGCF (perceptual speed; score; 0–58)	38.365	7.656	.000	58.000	.965
Reading competence: WLE (corrected, higher=more)	-.052	.770	-2.418	3.551	.269
competence measurements wave 14					
ICT literacy: WLE (corrected, higher=more)	.441	.790	-4.700	4.606	.707
Scientific literacy: WLE (corrected, higher=more)	.148	.713	-2.148	3.099	.709

TABLE 3.7: NEPS-SC4: Participation results

Results of the analysis of participation in starting cohort (SC) 4 of the NEPS data in SUF 14. The AUC, RMSE, precision, recall and F-measure is presented for the XGBoost, BART, Lasso and Ridge regression. For each method a 10-fold cross-validation is used.

	AUC	RMSE	precision	recall	F-measure
XGBoost	.9159	.2970	.8627	.9488	.9037
BART	.9214	.2871	.8681	.9665	.9146
Lasso	.6733	.6174	.6601	.4123	.5076
Ridge regression	.6749	.6206	.5954	.4701	.5254

3.7.4 Standard gradient boosting

Gradient boosting is a powerful ensemble learning technique introduced by Friedman (2001). It builds a predictive model by sequentially combining multiple weak learners—typically decision trees—into a single strong model. Each new tree is added to correct the errors of the previous ensemble of trees, improving the overall performance. In gradient boosting, the goal is to minimize a loss function, $L(y, \hat{y})$, which measures the difference between the true target values y and the model predictions $\hat{y}$. The key idea is to fit a new model to the residuals (the prediction errors) of the previous model. The process is iterative, with each step involving the following:

1. **Initialize** the model with a constant prediction, often the mean of the target values $\hat{y}_i^{(0)} = \bar{y}$.
2. **Iterative updates**: For each iteration t:
   - Compute the residuals or pseudo-residuals, which are the negative gradients of the loss function with respect to the current predictions:
   $$r_i^{(t)} = -\left[\frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}\right], \quad i = 1, \ldots, n.$$
   - Fit a new weak learner (e.g., a decision tree) $f_t(\mathbf{x})$ to the residuals.
   - Update the model's predictions:
   $$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(\mathbf{x}_i),$$
   where $\eta \in (0, 1]$ is the learning rate that controls the contribution of each tree to the final model.
3. **Repeat** this process for T iterations, gradually improving the model by reducing the prediction errors.

The loss function $L(y, \hat{y})$ can be tailored to different problems. For regression, the squared error loss is often used:

$$L(y, \hat{y}) = (y - \hat{y})^2.$$

For binary classification, the logistic loss is common:

$$L(y, \hat{y}) = \log(1 + \exp(-y\hat{y})).$$

At each iteration, the new tree is fitted to the gradients (residuals), which represent the direction in which the model's predictions should be adjusted to minimize the loss.

To prevent overfitting, gradient boosting incorporates several regularization techniques. Firstly, the parameter η reduces the impact of each tree, requiring more trees to fully correct errors, which can improve generalization (shrinkage learning rate). Secondly, limiting the

depth of trees reduces their complexity, controlling overfitting. Thirdly, randomly sampling a subset of data at each iteration can introduce randomness, reducing variance.

Algorithm 5 presents the gradient boosting that includes updates at each iteration t, including both regression and classification with logistic loss. The algorithm is structured to handle the iterative nature of gradient boosting and includes the specific updates for both cases.

Algorithm 5 Gradient Boosting Algorithm

1: **Input:** Training data $(\mathbf{x}_i, y_i)$ for $i = 1, \ldots, n$, number of iterations T, learning rate η

2: **Initialize:** Set initial prediction $\hat{y}_i^{(0)} = \bar{y}$ (mean of the target) for regression, or log-odds for classification.

3: **for** t = 1 to T **do**

4: Compute residuals (for regression):

$$r_i^{(t)} = -\frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} = y_i - \hat{y}_i^{(t-1)}$$

or pseudo-residuals (for classification):

$$r_i^{(t)} = -\frac{\partial L(y_i, \hat{z}_i^{(t-1)})}{\partial \hat{z}_i^{(t-1)}} = y_i - \frac{1}{1 + \exp(-\hat{z}_i^{(t-1)})}$$

where L is the loss function (squared error for regression, logistic loss for classification).

5: Fit a decision tree $f_t(\mathbf{x})$ to the residuals $r_i^{(t)}$.

6: Update the model's prediction:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(\mathbf{x}_i)$$

where η is the learning rate.

7: For classification, update log-odds:

$$\hat{z}_i^{(t)} = \hat{z}_i^{(t-1)} + \eta f_t(\mathbf{x}_i)$$

and compute the predicted probability:

$$\hat{y}_i^{(t)} = \frac{1}{1 + \exp(-\hat{z}_i^{(t)})}$$

8: **end for**

9: **Output:** Final model with prediction $\hat{y}_i^{(T)}$.

3.7.5 Area Under the ROC Curve (AUC-ROC)

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a performance measure commonly used to evaluate binary classification models, seeBowers and Zhou

(2019) and Junge and Dettori (2018). The ROC curve itself is a plot of the true positive rate (TPR) against the false positive rate (FPR) across various classification thresholds. The AUC quantifies the overall ability of the model to distinguish between the positive and negative classes by summarizing the ROC curve's performance.

For a binary classifier, TPR and FPR are defined as:

$$\text{True Positive Rate (TPR)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

and

$$\text{False Positive Rate (FPR)} = \frac{\text{False Positives (FP)}}{\text{False Positives (FP)} + \text{True Negatives (TN)}}.$$

The AUC represents the probability that a randomly selected positive example ranks higher than a randomly selected negative example according to the model's output probability scores. The AUC value ranges from 0 to 1 with an AUC of 1 indicating a perfect classifier, and an AUC of .5 suggesting no discrimination capability, equivalent to random guessing.

Mathematically, AUC is the integral of the ROC curve:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(x))\, dx$$

In practice, AUC can be computed by numerically approximating this integral, often using the trapezoidal rule based on the TPR and FPR values at various thresholds.

Chapter 4

Model comparison and selection

4.1 Overview and research context

The challenge of selecting one model among different models appears in inference statistic as well as in machine learning (Hastie et al., 2009), and in each case model comparison is an essential step to evaluate the performance of different models and choose the one that best fits your problem. The selection of a model from a range of candidates, based on the analysis of given data, is known as model selection. It is important to note that one theoretical phenomenon or problem may be described by different models. As discussed in chapter 2, it can be shown that variable selection is a special case of model selection and looking at different models, not at variable selection leads to different considerations. Foremost, a clear understanding of the problem or phenomenon is necessary to formulate the goals and the solution method. Different models might be better suited for different types of problems so that the models trying to describe the real-world situations can result in the choice of one type of model with different assessments or of different model types. Again, since the seminal paper of Schwarz (1978) providing a benchmark criterion for model complexity obeying Occam's razor in both statistical and machine learning model selection, criteria are necessary for selecting given candidate models of similar explanatory or predictive power. Both frequentist and Bayesian approaches can be applied to model comparison, but they differ in their methodologies and in their benchmark criteria.

Using p-values[1] is a common practice to perform carefully variable selection and model evaluation (Lehmann & Lösler, 2016) or likelihood ratio test (LRT) (Lewis et al., 2010; Sixt & Aßmann, 2020) or cross-validation (Arlot & Celisse, 2010; Zhang & Yang, 2015). The LRT is a powerful tool in frequentist model comparison. It involves comparing the likelihoods of two nested models: a smaller (null) model that includes a subset of variables, and a larger (alternative) model that additionally includes further variables. The test evaluates whether the improvement in model fit justifies the increase in model complexity. Under the null hypothesis that the additional parameters in the larger model are equal to zero (i.e., that the smaller model is sufficient), the test statistic follows a chi-square distribution (Casella & Berger, 2002; Cox & Hinkley, 1979). The hypothesis testing only holds for these nested models, for non-nested models see Vuong (1989), who recommends a framework for

[1]The drawbacks of using p-values are widely discussed, e.g., by Goodman (2008), for a detailed overview see Held and Ott (2018) and Wagenmakers (2007).

solving this problem and as for the discussion of the use of Vuong's test on special cases like zero-inflation (non-nested) models, see among others Clarke (2001) and Wilson (2015). As an alternative, information criteria, among others Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), are commonly used for model comparison based on the Likelihood. The BIC was originally derived by Schwarz (1978) and is a method for model selection that is derived from an asymptotic approximation of the marginal likelihood from a Bayesian perspective. An essential property of the BIC is its asymptotic consistency: under regularity conditions and assuming that the true model is included in the set of candidate models, the probability that BIC selects the true model approaches one as the sample size grows (Kass & Raftery, 1995; Schwarz, 1978). A model selection criterion is said to be consistent if it selects the true model with a high probability as the sample size increases, provided that this is contained in the set of models under consideration. This consistency can be attributed to the enhanced penalization for supplementary parameters by the penalty term $p\ln(n)$, where p denotes the number of parameters and n signifies the sample size. Systematically excluding over-parameterized models as the sample size increases, the penalty term becomes larger, meaning that additional parameters are increasingly penalized. Consequently, over-parameterized models, which possess an enhanced likelihood but do not deliver a substantial enhancement, are eliminated from the model selection process. This property ensures that BIC selects the model with the minimum number of parameters required. Schwarz (1978) demonstrated that the probability that BIC selects the true model converges to 1 if the number of observations, n, tends to infinity and the true model is contained within the model class. This indicates that BIC is asymptotically guaranteed to select the correct model, provided that the model assumptions are correct. In comparison, the AIC is based on a different theoretical foundation, i.e., minimizing the expected Kullback-Leibler divergence between the true data generation process and the selected model (Akaike, 1974). The AIC is therefore asymptotically efficient, but not consistent, as it tends to favor more complex models and is not guaranteed to select the true model even for large n (Burnham & Anderson, 2004). Following Burnham and Anderson (2002), both criteria provide a quantitative measure to compare models, with per definition lower values indicating better-fitting models. The consistency of BIC only applies under the assumption that the true model is contained in the model class (Yang, 2005). If the true model is not in the considered set of models, BIC can select a sub-optimal model. While BIC aims to identify the true model, AIC aims to select the model with the best predictive quality. In situations where the goal is not to identify the true model but to minimize the prediction error, AIC may be more advantageous. Cross-validation, especially techniques like k-fold cross-validation, is another frequentist approach to model comparison. By partitioning the data into multiple subsets and iteratively training and testing the model, cross-validation provides insights into how well each model generalizes unseen data (Arlot & Celisse, 2010; Zhang & Yang, 2015).

Following Gelman et al. (2023), Jeffreys (1961), and Kass and Raftery (1995), consideration of model comparison in a Bayesian perspective leads at first to the Bayes factor, which is a key tool in Bayesian model comparison. It quantifies the evidence in favor of one model over another by comparing the marginal likelihoods of the models. It encapsulates the trade-off

between model fit and complexity, providing a more nuanced measure than p-values, see more details in subsection 4.3.

The Deviance Information Criterion (DIC) is a Bayesian analogue to AIC. It evaluates model fit while penalizing for model complexity. DIC considers both likelihood and a measure of effective parameters, providing a balance that helps prevent overfitting. Following Spiegelhalter et al. (2002) the connection between DIC and the marginal likelihood lies in the components used to compute DIC. DIC incorporates both the likelihood of the data given the model and a measure of effective parameters. The effective parameters essentially capture the complexity of the model, including the number of parameters and their effective degrees of freedom. The marginal likelihood, also known as the model evidence, is a fundamental quantity in Bayesian inference. It represents the likelihood of the observed data averaged over all possible parameter values in the model, weighted by the prior distribution. In essence, it quantifies the fit of the model to the data while penalizing for model complexity through the prior distribution. The connection between DIC and the marginal likelihood arises from the fact that DIC can be expressed as the difference between the posterior mean of the deviance and the deviance at the posterior mean of the parameters. The deviance is essentially a measure of model fit, similar to the negative log-likelihood, but it also takes into account the complexity of the model. Therefore, DIC can be seen as a way to approximate the marginal likelihood by balancing model fit (deviance) with model complexity (effective parameters). While DIC is not exactly equal to the marginal likelihood, they are related through their shared consideration of both fit and complexity in evaluating Bayesian models. DIC aims to strike a balance between goodness of fit and model complexity, similar to AIC.

Bayesian model comparison often involves Posterior Predictive Checks (PPC), where simulated datasets are generated from the posterior predictive distribution of each model. By comparing these simulated datasets to the observed data, researchers can assess the ability of each model to reproduce key features of the real-world data. PPC entails the generation of simulated datasets under each model's assumptions, utilizing parameter values sampled from the posterior distribution. These simulated datasets serve as proxies for the observed data, enabling a direct comparison between model predictions and empirical observations. Key to the PPC process is the computation of summary statistics, capturing essential features of the data, such as means, variances, and quantiles. To implement PPC, researchers first derive the posterior distribution of model parameters using Bayesian inference techniques such as Markov chain Monte Carlo (MCMC) sampling. Subsequently, simulated datasets are generated by drawing parameter values from the posterior distribution and simulating data according to the model's specifications. Summary statistics are then computed for both simulated and observed data, facilitating a quantitative comparison (Berkhof et al., 2000; Gelman & Shalizi, 2012; Rubin, 1984). The crux of PPC lies in assessing the congruence between simulated and observed data through statistical measures. Discrepancies between summary statistics of simulated and observed data may indicate inadequacies or misspecifications in the model. Conversely, a close match between simulated and observed data across relevant summary statistics suggests a good fit of the model to the data. PPC serves as a diagnostic tool for iterative model refinement, guiding researchers in identifying areas for

model enhancement or modification. By iteratively adjusting model parameters, incorporating additional complexity, or exploring alternative model structures, researchers can strive towards a more accurate representation of the underlying phenomena.

While BIC and AIC provide a useful frequentist approximation for model selection, the Bayesian framework naturally accommodates the full posterior uncertainty via the Deviance Information Criterion (DIC) (Kaplan & Chen, 2014). A Bayesian approach allows for model averaging, where multiple models are considered, and their models are weighted based on their posterior probabilities. Following Burnham and Anderson (2002), Hoeting et al. (1999), Kaplan (2021), and Raftery and Zheng (2003) Bayesian model averaging (BMA) serves as a sophisticated tool for addressing uncertainty not only in parameter estimation but also in the selection of the appropriate statistical model. Moreover, Bayesian model averaging allows inference to incorporate both parameter and model uncertainty, avoiding overconfidence in a single selected model. This accounts for uncertainty not only in parameter values but also in the choice of the model itself. A single model chosen as a preferable one is typically chosen based on some criterion such as Maximum Likelihood Estimation or goodness-of-fit statistics. However, this approach may disregard the uncertainty inherent in selecting the true model, particularly when there are multiple plausible models that could explain the data. BMA adopts a more nuanced perspective by considering a collection of candidate models, each representing a distinct hypothesis regarding the underlying data-generating process. Instead of committing to a single model, BMA assigns posterior probabilities to each model based on their compatibility with the observed data and prior beliefs about their plausibility. These posterior probabilities capture the updated uncertainty about the models after observing the data. Once the posterior probabilities of the candidate models are established, BMA combines their predictions by weighing them according to these probabilities. Consequently, models with higher posterior probabilities exert more influence on the overall inference, while those with lower probabilities contribute less. This approach offers a comprehensive representation of uncertainty by addressing both parameter uncertainty within each model and model uncertainty across the candidate models. One of the key advantages of BMA in inference is its ability to incorporate prior knowledge or beliefs about the models and parameters. This feature can enhance the stability of inference, particularly in experimental studies with limited data. Moreover, BMA facilitates model comparison by providing a principled framework for assessing the relative plausibility of different models based on the observed data. Bayesian model averaging is a flexible and robust approach that integrates uncertainty at both the parameter and model levels, providing a more comprehensive and reliable inference in complex statistical modeling problems. By acknowledging and quantifying uncertainty in model selection, BMA enables more informed decision-making and enhances the credibility and generalizability of statistical analyses.

In both Bayesian and frequentist paradigms, model comparison presents distinct challenges and strengths. Bayesian approaches naturally account for parameter uncertainty through posterior distributions (Gelman et al., 2023), but their dependence on prior beliefs requires careful selection and justification. Following Berger (2006a) sensitivity analysis

helps mitigate subjectivity and assess robustness, yet the computational demands of techniques such as Markov Chain Monte Carlo (Hoffman et al., 2013) can pose practical barriers, particularly in high-dimensional settings. Advanced techniques like Markov Chain Monte Carlo (MCMC) or variational inference are widely used to approximate posteriors for complex models. Despite their effectiveness, these methods require significant computational resources and expertise. In contrast, frequentist methods provide computationally efficient tools like the Akaike Information Criterion (Hastie et al., 2009) and cross-validation but may underestimate true uncertainty by relying on point estimates. The philosophical distinction between the paradigms – Bayesianism's subjective interpretation of probability versus frequentism's long-run frequency perspective – further influences model comparison choices (Wasserman, 2008). Ultimately, researchers benefit from synthesizing the strengths of both frameworks, tailoring their approach to the data's nature, the complexity of the models, and the research objectives. Leveraging robust sensitivity analyses and hybrid methodologies offers a pathway to reconciling the epistemological divide, enriching the practice of model comparison.

In the following, model comparison is carried out in order to be able to select different models of a model type, each with a different number and composition of covariates, and not to be able to perform model averaging or prediction. For this purpose, the Bayesian approach is chosen in order to be able to carry out model selection using marginal likelihood. This Bayesian approach also includes the treatment of missing values in the covariates, so that no additional data preparation steps are necessary before estimating and evaluating the different models, see also Aßmann et al. (2023) and Aßmann and Preising (2020). The missing data problem is thus integrated into the respective model estimation. Therefore MCMC techniques offers the appropriate options for this approach, so that this additional step of handling missing values can be easily incorporated. In the presence of missing data, Bayesian inference treats mainly unobserved values as latent variables and naturally propagates the associated uncertainty throughout the posterior distribution. This coherent treatment contrasts with ad hoc methods such as (single) imputation or complete-case analysis (Kaplan & Chen, 2014), and allows bridging the gap between falsifiability and empirical reality.

Both the Bayes factor and the marginal likelihood are powerful tools in Bayesian model comparison and selection. The marginal likelihood, i.e., the normalizing constant of the posterior density, can be extended to the Bayes factor for directly comparing two or more models. The calculation of the marginal likelihood can be challenging and various methods are used as an alternative, see Pajor (2017), and the model evidence is determined by integrating the likelihood function (Chib, 1995). Model comparison using p-values or information criteria and using marginal likelihood are quite distinct methodologies on the same aim: compare models with different complexities and different numbers of parameters. In the Bayesian estimation approach conjugate prior distributions are chosen for the model parameters and MCMC methods (see e.g., Geweke (1989)), namely Gibbs Sampling (e.g., Gelfand and Smith (1990) and Geman and Geman (1984)), are used to provide the approximations of the corresponding posterior densities. The treatment of missing values as additional step is also derived as full conditional distribution implied by the model setup as

compared with the regression estimates and augments the data accordingly compared to the latent augmentation step caused by the binary dependent variable (Li, 1988; Tanner & Wong, 1987). This additional augmentation step considers non-parametric approximations of the full conditional distributions of the missing values provided by classification and regression trees (CART) (Burgette & Reiter, 2010). This is in line with non-parametric prior distributions for the missing values in the covariates.

After clearly defining competing models which are to compare, each model represents a distinct hypothesis or set of assumptions. Then, the prior distributions for the parameters of each model are specified so that the choice of priors reflects the prior beliefs or knowledge about the parameters.[2] In a Bayesian view, the likelihood function describes how the observed data is generated given the parameters of the model and is specified for the estimation routine, but can be the crucial component of Bayesian inference. The marginal likelihood, also known as the evidence, is calculated for each model and is the probability of observing the data given the model, marginalized over all possible parameter values. Out of this, the Bayes factor is computed by taking the ratio of the marginal likelihoods of the two models. The Bayes factor quantifies the relative evidence in favor of one model over the other and provides an interpretable translation of the marginal likelihoods (Jeffreys, 1961).

In this chapter, a binary dependent variable is considered, which may represent the decision of individuals in a given group, e.g., the decision of attending a distinct school type in a given region, the covariates represent individual-specific and group-specific differences, which is additionally specified as individual heterogeneity captured by random effects (Gu et al., 2009; Heckman & Willis, 1976). For details of the Bayesian model formulation see section 4.2, for details of the corresponding model comparison and evaluation see section 4.3, and for the details of the implementation and quality aspects see section 4.4. Within the Bayesian literature estimation approaches for probit models extended to heterogeneity are typically discussed for completely observed, or in the case of panel data unbalanced, datasets. Aßmann and Preising (2020) provide a description of dynamic linear models that are able to accommodate missing values. The accuracy of the presented approach in terms of parameter estimation and model comparison is evaluated by a simulation study and the relevance is presented with an empirical application where the results are presented in sections 4.5.1 as well as 4.5.4. The results are compared with the p-value tests based on Maximum Likelihood Estimation and handling missing values before estimation via multiple imputation (MICE). Comparison of different models that are describing one (theoretical) problem or a phenomenon in different nuances means also comparing non-nested models, which is straight forward in terms of Bayes factors (Jeffreys, 1961; Kass & Raftery, 1995). In general, the explanatory power of two or more models is compared by identifying the model that best explains the variance in Y and should satisfy the criteria of accuracy and parsimony, i.e. the best model should describe and predict accurately, and it should be the one with fewer assumptions, e.g., parameters.[3] Otherwise, a statistical model can include terms describing some theoretical constructs and thus can influence the variance of explanation. A set of

[2]Sensitivity analyses may be conducted to assess the impact of different priors.

[3]The parsimonious argument of Occam's Razor is described above and in literature aesthetic up to pragmatic arguments can be found which should underline the parsimonious argument. Including too many parameters

parameters $\boldsymbol{\theta}$ will be found that minimizes the explanation or prediction error that describes the theoretical problem or phenomenon.

4.2 Model formulation and estimation

4.2.1 Model setup

Let y_{ij} denote the observed dichotomous outcome with $i = 1, \ldots, N_j$ in group $j = 1, \ldots, J$, where N_j denotes the number of individuals in group j and $N = \sum_{j=1}^{J} N_j$ is the total sample size. The structure between the observed binary variables y and the observed explaining factors X is given via the latent variable y_{ij}^* and is described in a probit link function:

$$y_{ij} = \begin{cases} 1 & \text{if} \quad y_{ij}^* = X_{ij}\beta + a_j + e_{ij} > 0, \\ 0 & \text{if} \quad y_{ij}^* = X_{ij}\beta + a_j + e_{ij} \leq 0, \end{cases} \tag{4.1}$$

where X_{ij} is a $a \times K$ row vector of covariates, β is the $K \times 1$ regression coefficient vector, and a_j denotes a group-specific random intercept. For identification of the probit scale, it is set: $e_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ independent of $a_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_a^2)$.

Conditional on the random effects $a = (a_1, \ldots, a_J)'$, the success probability is

$$\Pr(y_{ij} = 1 \mid X_{ij}, a_j, \beta) = \Phi\big(X_{ij}\beta + a_j\big)\,, \tag{4.2}$$

and the corresponding conditional likelihood and log-likelihood are

$$\mathcal{L}(y \mid X, \beta, a) = \prod_{j=1}^{J} \prod_{i=1}^{N_j} \Phi\big((2y_{ij} - 1)\,(X_{ij}\beta + a_j)\big)\,, \tag{4.3}$$

$$\ln \mathcal{L}(y \mid X, \beta, a) = \sum_{j=1}^{J} \sum_{i=1}^{N_j} \ln \Phi\big((2y_{ij} - 1)\,(X_{ij}\beta + a_j)\big)\,. \tag{4.4}$$

In the context of Bayesian inference, a Gibbs sampler routine is utilized to draw random samples from the joint posterior distribution. Involving $r = 1, \ldots, R$ iterations, it is employed to generate random samples from the joint full conditional distributions of the considered parameter blocks. This approach is outlined in the works of Casella and George (1992), Gelfand and Smith (1990), and Geman and Geman (1984).

4.2.2 Data augmentation and missing data

Following Aßmann and Preising (2020), a missing pattern in the covariates is suggested assuming that the missingness mechanism has negligible effect on the estimation (Little & Rubin, 2002). Throughout this chapter the response, y, is assumed fully observed. The data expansion tool is particularly well suited to dealing with missing data, since adding the

can lead to overfitting the real-world data which also can be (very) complex and choosing a more complex model can be worth considering.

missing values to the parameter vector allows all other model variables to be treated as if they were fully observed. Moreover, data expansion fits well within the Bayesian framework of an estimation routine, where no combining rules are required to obtain valid estimators and corresponding variances. To deal with missing values, the Gibbs sampler is augmented to include draws from the corresponding full conditional distributions, whose functional forms are in principle the same for the random effects or only fixed effects specification.

Tanner and Wong (1987) present the idea of data augmentation, i.e., the introduction of a latent quantity L^* without substantial interest, which simplifies handling the missing data problem immediately. Generally the data augmentation is composed by,

$$\begin{aligned} f(\theta|y,X) &= \int (\theta, L^*|y,X)\mathrm{d}L^* \\ &= \int f(\theta|y,X,L^*)f(L^*|y,X)\mathrm{d}L^* \end{aligned} \tag{4.5}$$

where $f(\theta|y,X)$ denotes the posterior density of θ given the data, $f(L^*|y,X)$ the predictive density of the latent data and $f(\theta|y,X,L^*)$ the conditional density of θ given the augmented data (cf. Tanner and Wong (1987)). Recall that marginal posterior distribution of latent L^*, $f(L^*|y,X)$, can be augmented with the parameter vector θ, so that is valid to:

$$\begin{aligned} f(L^*|y,X) &= \int f(L^*,\theta|y,X)\mathrm{d}\theta \\ &= \int f(L^*|y,X,\theta)f(\theta|y,X)\mathrm{d}\theta. \end{aligned} \tag{4.6}$$

The joint posterior distribution decomposes into $f(L^*|y,X,\theta)$, i.e., the conditional distribution of the latent data L^* given the parameters θ and in the case of data augmentation for handling missing values in X this corresponds to an imputation of L^* under the current parameter values, as well as into $f(\theta|y,X)$, i.e., the posterior distribution of the parameters, marginalized over L^*. This decomposition leads to the fundamental part of data augmentation and Gibbs sampling: by drawing L^* from $f(L^*|y,X,\theta)$ and then θ from $f(\theta|y,X)$, the correct joint posterior distribution of (θ, L^*) can be iteratively sampled. Marginalization ensures that all uncertainties are properly accounted for, even though only one variable is simulated at a time. Drawing from L^* (or θ) (E-Step) yields a draw of θ (or L^*) from $f(\theta|y,X,L^*)$ (or $f(L^*|y,X,\theta)$) (M-Step) depending on the E-Step.[4]

In the context of missing values the above discussed iterative process appears in drawing missing values conditioned on current parameters θ (E-Step or step I of a Gibbs sampler) and drawing model parameter θ conditioned on the current draw of missing values (M-Step or step II of a Gibbs sampler). Let $X = [X_{obs}, X_{mis}]$ represent all covariates which occur as observed (obs) and missing (mis) data. Let assume the augmentation steps as follows:

$$f(\theta|X_{obs}) = \int f(\theta|X_{obs},X_{mis})f(X_{mis}|X_{obs})\mathrm{d}X_{mis} \tag{4.7}$$

[4]Interestingly, this iterative procedure for handling the missing data problem can also be interpreted as a Gibbs sampler or, equivalently, as an Expectation-Maximization (EM) algorithm (Dempster et al., 1977; Tanner & Wong, 2010).

and

$$f(X_{mis}|X_{obs}) = \int f(X_{mis}|X_{obs}, \theta) f(\theta|X_{obs}) \mathrm{d}\theta \tag{4.8}$$

Correspondingly, the posterior distribution $f(\theta|X_{obs})$ can be drawn with the data augmentation step: firstly, draw missing values from their full conditional posterior distribution conditioned on observed data and model parameters $f(X_{mis}|X_{obs}, \theta)$ and secondly, draw the model parameters from the augmented dataset $f(\theta|X_{obs}, X_{mis})$. Thus, the posterior distribution can easily be marginalized via Gibbs Sampling (Boone et al., 2007; Gelman et al., 2023). By contrast, Little and Rubin (2002) present multiple imputation as an approach to impute the missing values and the parameter estimates multiple times and calculate the average using necessary formulas and combing rules.

4.2.3 Bayesian estimation routine with missingness

In case of Bayesian estimation, the prior specification is essential for the full conditional distributions which are used in the Gibbs sampler while identifying the joint posterior distributions.

Prior settings

Instead of sampling directly from this joint posterior distribution an adequate Gibbs sampling scheme is implemented to estimate the marginal posterior distributions via drawing from the full conditional distributions (s). According to Tanner and Wong (1987) and Nasrollahzadeh (2007) the full conditional distributions are as follows. Assuming the above mentioned probit link function 4.3 with the parameters $\theta = [A', \beta']$ the prior is set by conjugate normal priors for β and a_j:

$$\beta \sim \mathcal{N}(\mu_\beta, \Sigma_\beta) \tag{4.9}$$

$$a_j \sim \mathcal{N}(\mu_a = 0, \Sigma_a) \tag{4.10}$$

with μ_β denoting the mean and Σ_β the covariance-matrix of the normal prior for the regression coefficients and μ_a denoting the mean and Σ_a the variance of the normal prior for the random coefficients, respectively. The conjugate prior for the random effects variance is set to an inverse-gamma distribution with c_0 being the shape parameter and d_0 being the scale parameter:

$$\Sigma_a = IG(c_0, d_0). \tag{4.11}$$

Following Aßmann and Preising (2020), it is suggested to implement non-parametric approximations for the full conditional distributions of missing values.[5] These can be provided in the form of classification and sequential regression trees (CART), as discussed in Burgette and Reiter (2010), while essentially using a tree-based method to impute missing values. CART models can be useful for capturing complex relationships in the data and are

[5]The prior for the missing variables X_{mis} is discussed in detail in the posterior section above.

commonly used in the imputation process. CART divides the data into disjoint subsets at each partitioning step. Maximizing the reduction of heterogeneity is the partitioning criterion for these partitioning steps. Thus, the observations are partitioned into highly homogeneous groups. Therefore, the variance (continuous variables) or the entropy (discrete variables) is used. In each Gibbs iteration m, the draws from these distributions are obtained by a Bayesian bootstrap. This corresponds to an uninformative prior setting, where all observed values of a covariate ($X_{obs,k}$) are considered as potential donors for missing values in this covariate ($X_{mis,k}$). The prior distributions are based on this donor sets, which are obtained by the computation of a Gaussian kernel density for continuous variables or an empirical frequency distribution for categorical variables.

These expressions involve updating the priors with the observed data and adjusting for the likelihood term. The specific form of the posterior distribution depends on the choice of priors, likelihood, and the assumptions made in the model.

The joint posterior distribution represents the complete uncertainty about the parameters given the data and the prior information. By sampling from this distribution using MCMC methods, estimates of the parameters can be obtained and the uncertainty in these estimates is quantified.

Posterior computation and inference

A Gibbs sampler (e.g., Aßmann and Preising (2020), Gelfand and Smith (1990), and Geman and Geman (1984)) is implemented to generate the posterior distributions of interest including the structural model parameters and the missings in the covariates. In the context of these Bayesian setups that are mutually independent, the aforementioned conjugate prior distributions are assumed. According to Bayes' Theorem the joint posterior distribution is proportional to:

$$\pi(\beta, a_j, \sigma_a^2, X_{\text{mis}}|X, y) \propto \mathcal{L}(\beta, a_j, \sigma_a^2, X_{\text{mis}}|X, y)\pi(\beta)\pi(a_j|\sigma_a^2)\pi(\sigma_a^2)\pi(X_{\text{mis}}) \tag{4.12}$$

where β is the vector of regression coefficients, a_j is the vector of random effects for group j, σ_a^2 is the variance of the random effects, X is the observed data, y is the observed binary variable, X_{mis} is the missing data, $L(\beta, a_j, \sigma_a^2, X_{\text{mis}}|X, y)$ is the likelihood function, $\pi(\beta)$ is the prior distribution of the regression coefficients, $\pi(a_j|\sigma_a^2)$ is the prior distribution of the random effects given the variance, $\pi(\sigma_a^2)$ is the prior distribution of the variance, and $\pi(X_{\text{mis}})$ is the prior distribution of the missing data. The full conditional distributions for each parameter are derived from the joint posterior distribution and the prior distributions, and are used in the Gibbs sampler to generate samples from the joint posterior distribution. The Gibbs sampler iteratively samples from the full conditional distributions of each parameter, given the current values of all other parameters and the observed data. The resulting samples are then used to estimate the posterior distributions of interest. The conjugate prior distributions are assumed for the Bayesian setup, which means that the prior distributions are chosen so that the full conditional distributions are known standard distributions, and sampling can be

performed using standard methods. This simplifies the implementation of the Gibbs sampler and improves the efficiency of the algorithm.

The method of data augmentation is especially suited to deal with missing data problems because the inclusion of the missing value in the parameter space results in handling all other model quantities as in completely observed data. Adding the data augmentation step into the Bayesian estimation routine allows for avoiding combining rules (Tanner & Wong, 1987). Hence, the quantities of interest are y^*, θ, σ_a^2, and X_{mis}, where X_{mis} denotes the missing values of the covariates with $X = [X_{obs}, X_{mis}]$. The core function of Bayesian inference is the augmented posterior distribution (Aßmann et al., 2023) given as

$$\begin{aligned} p(y^*, \beta, a_j, \sigma_a^2, X_{mis}|y, X_{obs}) \propto & f(y^*|y, X, \beta, a_j) \times f(\beta|y, X, a, \sigma_a^2) \times \\ & f(a_j|y, X, \beta, \sigma_a^2) \times f(\sigma_a^2|y, X, \beta, a) \times \\ & f(X_{\text{mis}}|X_{obs}, \beta, a_j, \sigma_a^2). \end{aligned}$$

The Gibbs sampler uses the following steps and the posterior values are drawn iteratively $r = 1, \ldots, R$ from the respective full conditional distributions. For compact notation, stack the latent variables into $\mathbf{y}^* \in \mathbb{R}^N$ and let $X \in \mathbb{R}^{N \times K}$ denote the (currently completed) design matrix. Define the group-incidence matrix $Z \in \{0,1\}^{N \times J}$ such that each row contains a single one indicating the group membership of the corresponding observation. Hence,$(Za)_n = a_{j(n)}$ for observation n belonging to group $j(n)$, and the latent model can be written as

$$\mathbf{y}^* = X\beta + Za + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_N).$$

After setting appropriate starting values for $\beta, a_j, \sigma_a^2, X_{\text{mis}}$, the set of full conditional distributions in the Gibbs sampler is set up as follows.

$f(y^*|\cdot)$ The full conditional distribution of the latent variables y^* corresponds to a product of truncated normal distributions $y^* \mid \beta, a, X \sim \mathcal{TN}(m_{y_{ij}^*}, V_{y_{ij}^*})$, since the single elements $y_{ij}^*, i = 1, \ldots, N_j$ and $j = 1, \ldots, J$, are mutually independent. Sampling for each element is hence performed from a truncated normal distribution with moments $m_{y_{ij}^*} = X_{ij}\beta + a_j$ and variance equal to one ($V_{y_{ij}^*} = 1$) with the truncated support ranging from $-\infty$ to 0 in case $y_{ij} = 0$ and in case $y_{ij} = 1$ ranging from 0 to ∞.

$f(\beta|\cdot)$ In accordance with the findings of Albert and Chib (1993) and Gilks et al. (1993), the full conditional distribution is, in principle, said to follow the Bayesian linear regression given the latent continuous variable y^*. The estimator of the covariates are assumed to be drawn from a multivariate normal distribution $\beta \mid \mathbf{y}^*, a, X \sim \mathcal{N}(m_\beta, V_\beta)$ with mean m_β and variance V_β:

$$V_\beta = (\Sigma_\beta^{-1} + X^\top X)^{-1} \quad \text{and} \quad m_\beta = V_\beta\big(\Sigma_\beta^{-1}\mu_\beta + X^\top(\mathbf{y}^* - Za)\big).$$

$f(a_j|\cdot)$ The full conditional distribution for a_j is drawn analogously to beta, namely from a multivariate normal distribution $a \mid \mathbf{y}^*, \beta, \sigma_a^2, X \sim \mathcal{N}(m_a, V_a)$ with mean m_{a_j} and

variance V_{a_j}:

$$V_a = (Z^\top Z + \sigma_a^{-2} I_J)^{-1}, \qquad m_a = V_a Z^\top (\mathbf{y}^* - X\beta),$$

equivalently, for each group $j = 1, \ldots, J$,

$$V_{a_j} = \left(N_j + \frac{1}{\sigma_a^2} \right)^{-1}, \qquad m_{a_j} = V_{a_j} \sum_{i=1}^{N_j} \left(y_{ij}^* - X_{ij}\beta \right).$$

$f(\sigma_a^2|\cdot)$ The random effects variance σ_a^2 is drawn from a Inverse-gamma distribution $a \mid \sigma_a^2 \sim \mathcal{N}(0, \sigma_a^2 I_J)$ with shape c and rate d

$$c = c_0 + \frac{J}{2} \quad \text{and} \quad d = d_0 + \frac{1}{2} a'a$$

$f(X_{\mathbf{mis}}|\cdot)$ Values of X_{mis} are sampled sequentially for each column vector of X, i.e., $X = (X^{(1)}, \ldots, X^{(P)})$, based on the non-parametric approximation suggested in the form of classification and sequential regression trees (CART), see Burgette and Reiter (2010). Let $X_{\text{com}}^{(k)} = (X_{\text{obs}}^{(k)}, X_{\text{mis}}^{(k)})$, $k = 1, \ldots, P$, denotes the completed variables, and $X_{\text{com}}^{(\backslash k)}$ denotes the completed matrix of variables except column p. The major advantage of this approach is that the latent variable, which may serve as a kind of sufficient statistic, can be used to approximate the full conditional distribution of missing values. In a first step, a decision tree is built for $X_{\text{com}}^{(k)}$ conditional on the corresponding values of all remaining variables $X_{\text{com}}^{(\backslash k)}$ as well as conditional on (θ, y^*).To incorporate prior uncertainty in the hyperparameters of the sequential partitioning regression trees, trees are constructed with a randomly varying minimum number of elements within nodes. For continuous outcomes,a larger minimum leaf size to ensure stable estimates is used and the complexity parameter is set to allow sufficient tree growth. For categorical outcomes, standard control parameters were applied to balance bias and variance in the imputed values. Each missing observation can then be assigned to a node, and thus a grouping of observations is implied by the binary partition in terms of the conditioning variables. The values within each node provide access to an empirical distribution function that serves as an approximation to the full conditional distribution of missing values, and thus as the key element for running the missing value data generation mechanism. This modeling approach is consistent with prior distributions of missing values that are proportional to observed data densities. Drawings from the empirical distribution function within a node correspond to draws from the full conditional distributions of missing values, with sampling performed via the Bayesian bootstrap to account for the estimation uncertainty of the full conditional distribution, see Rubin (1981). The considered approach further offers the flexibility to consider any function of observed or augmented data within the set of conditioning variables as well. The sampled X_{mis} values allow referring to an updated completed matrix of covariates in all other steps of the MCMC algorithm.

Given missing data in the covariates, the Gibbs sampler is extended to include draws from the full conditional distribution of missing values. In each iteration $r = 1, \ldots, R$, the missing values in X, i.e., X_{mis}, are drawn from their corresponding full conditional distributions. First, the parameter vector $\theta^{(r)} = (\beta^{(r)}, a_j^{(r)}, \sigma_a^{2(r)})$ is sampled from its corresponding full conditional distribution as mentioned above, then the missing values $X_{mis}^{(r)}$ are sampled from its full conditional posterior distribution $f(X_{mis}^{(r)}|y, z^{(r)}, \theta^{(r)}, X_{obs})$. These methods do not directly incorporate the uncertainty associated with the missing value into the analysis. In contrast, multiple imputation approaches attempt to incorporate such uncertainty by imputing missing values via some distribution. A Bayesian approach to missing covariates is considered and the intrinsic properties of the Markov Chain Monte Carlo (MCMC) technique, i.e., Gibbs sampling, are used to impute values for the missing covariates. Using the sampling properties of the Gibbs sampler naturally incorporates covariate structure and missing data mechanisms into the imputed values.

4.3 Model evaluation, comparison and selection

4.3.1 Marginal likelihood

In Bayesian model comparison, a central quantity is the marginal likelihood (Gelman et al., 2023). Simple approaches compare marginal likelihoods of different models directly, more complex ones are based on the fraction of two marginal likelihoods (Bayes factor) or on calculus on it (Bayesian Information Criterion (BIC)). The marginal likelihood, also known as the evidence or model likelihood, is a key quantity in Bayesian statistics, formally also known as the normalizing constant that ensures that the posterior distribution integrates to 1. It represents the probability of the observed data under a given model, averaged over all possible parameter values in that model. Mathematically, for a model M with parameters θ and observed data D, the marginal likelihood $P(D|M)$ is

$$P(D|M) = \int P(D|\theta, M)P(\theta|M)d\theta \tag{4.13}$$

where $P(D|\theta, M)$ is the sampling density (proportional to the likelihood)[6], representing how probable the observed data is given a specific set of parameter values, $P(\theta|M)$ is the prior distribution, representing the beliefs about the parameter values before observing any data, and θ is the vector of model parameters (Newton & Raftery, 1994). Marginal likelihoods form the basis of Bayes factors and related criteria; for instance, BIC can be interpreted as an asymptotic approximation to $-2 \log m(D \mid M)$ under regularity conditions.

The estimation of marginal likelihood is a crucial aspect of Bayesian model comparison, and various methods have been proposed to tackle this challenging problem. Llorente et al. (2020) provide a comprehensive overview of different approaches for computing marginal likelihoods, e.g., Perrakis et al. (2014) discuss an importance sampling, Pajor (2017) an arithmetic mean, and Reichl (2020) a geometric identity approach. It is important to note that

[6]To be precise, $P(D|\theta, M)$ is the sampling density, which is proportional to the model likelihood.

each method has its strengths and limitations, and the choice of the method may depend on the specific characteristics of the model and data. Some methods may perform well for certain types of models but not for others. Additionally, computational efficiency is often a consideration. In the context of Gibbs sampling the marginal likelihood can be quantified as a by-product of the above-shown MCMC estimation procedure. Chib (1995) shows that the marginal likelihood can be estimated with the results from the Gibbs sampler by so-called additional reduced Gibbs samplings, which refer to an extra step in the Gibbs sampling process used to sample from the conditional distribution of some parameters. Gibbs sampling is a Markov Chain Monte Carlo (MCMC) technique used to draw samples from the joint posterior distribution of all the parameters in a Bayesian model. In a standard Gibbs sampler, you iteratively sample from the conditional distributions of each parameter given the current values of the other parameters. The reduced part implies that in this estimation process, Gibbs sampling is performed on a subset of the parameters, typically conditioning on fixed values for other parameters. Therefore, after obtaining Gibbs samples from the joint posterior, an additional Gibbs sampling step is introduced. This step focuses on sampling from the conditional distribution of some specific parameters (e.g., precision parameters, variances, or latent variables) while keeping other parameters fixed at their current values.

The normalizing constant of the posterior density, necessary for the marginal likelihood and the Bayes factor, is given as

$$f(y|X) = \frac{f(y|X,\theta)f(\theta)}{f(\theta|y,X)}. \tag{4.14}$$

According to Chib (1995) and Chib and Jeliazkov (2001) the marginal likelihood $\ln\hat{m}(y|X_{obs})$ can be generally approximated by the natural logarithm so that

$$\ln\hat{m}(y|X) = \ln f(y|\hat{\theta},X) + \ln\pi(\hat{\theta}) - \ln\hat{\pi}(\hat{\theta}|y,X), \tag{4.15}$$

where $\ln f(y|\hat{\theta},X)$ being the approximated log-likelihood function evaluated at the posterior mode $\hat{\theta}$, $\ln\pi(\hat{\theta})$ being the log-prior density evaluated at the posterior mode $\hat{\theta}$, and $\ln\hat{\pi}(\hat{\theta}|y,X)$ being the log-posterior density evaluated at the posterior mode $\hat{\theta}$ with $\hat{\theta}$ is chosen as a point within the highest density region, i.e., the median or mean of all Gibbs sampler iterations less a previously defined burn-in.

In the above-mentioned setting, the covariate matrix contains missing entries. Therefore $X = (X_{\text{obs}}, X_{\text{mis}})$ is written and the observed data as $D_{\text{obs}} = (y, X_{\text{obs}})$ is defined. The relevant quantity for model comparison is the observed-data marginal likelihood

$$m(y \mid X_{\text{obs}}, M) = \iint p(y \mid X_{\text{obs}}, X_{\text{mis}}, \theta, M)\; p(X_{\text{mis}} \mid X_{\text{obs}}, \theta, M)\; p(\theta \mid M)\, dX_{\text{mis}}\, d\theta. \tag{4.16}$$

In the Gibbs sampler, X_{mis} is treated as an additional unknown quantity and is updated jointly with the model parameters, so that uncertainty due to missing covariates ispropagated through the posterior draws.

Chib's identity (Chib, 1995; Chib & Jeliazkov, 2001) provides

$$m(y \mid X_{\text{obs}}, M) = \frac{p(y \mid X_{\text{obs}}, \theta^\star, M)\, p(\theta^\star \mid M)}{p(\theta^\star \mid y, X_{\text{obs}}, M)}, \tag{4.17}$$

which yields the estimator

$$\log \widehat{m}(y \mid X_{\text{obs}}, M) = \log p(y \mid X_{\text{obs}}, \theta^\star, M) + \log p(\theta^\star \mid M) - \log \widehat{p}(\theta^\star \mid y, X_{\text{obs}}, M). \tag{4.18}$$

The point $\theta^\star$ is chosen in a high posterior density region (e.g., posterior mean or median from the Gibbs output after burn-in).

To evaluate (4.18) in latent-variable and missing-data models, Chib (1995) proposes estimating the posterior ordinate $p(\theta^\star \mid y, X_{\text{obs}}, M)$ by factorizing it into products of full conditional ordinates and estimating these ordinates using additional reduced Gibbs runs in which blocks of parameters are held fixed at $\theta^\star$ while sampling the remaining blocks. In our probit model with data augmentation, the reduced Gibbs samplers naturally condition on $\theta^\star$ and repeatedly update the augmented quantities (e.g., y^*, a, and X_{mis}) to obtain $\widehat{p}(\theta^\star \mid y, X_{\text{obs}}, M)$.

Posterior component of the marginal likelihood

Posterior inference is conducted via a Gibbs sampler, iteratively sampling from the conditional distributions of each parameter given the others parameters and the (observed) data. To estimate the posterior component of the marginal likelihood following Chib (1995), a reduced Gibbs sampling approach is employed, which involves evaluating the posterior ordinate at a fixed point $\hat{\theta}$:

$$\ln p(y|\mathcal{M}) = \ln L(y|\hat{\theta}) + \ln \pi(\hat{\theta}) - \ln \hat{\pi}(\hat{\theta}|y), \tag{4.19}$$

where $\ln L(y|\hat{\theta})$ denotes the log-likelihood evaluated at $\hat{\theta}$, $\ln \pi(\hat{\theta})$ the log-prior, and $\hat{\pi}(\hat{\theta}|y)$ the estimated posterior ordinate. The latter is computed by splitting the Gibbs sampler into reduced conditional distributions, fixing certain parameters sequentially and averaging over draws of the remaining parameters.

For example, the posterior ordinate for the fixed effects $\hat{\beta}$ is obtained as:

$$\hat{\pi}(\hat{\beta}^*|y) = \frac{1}{R} \sum_{r=1}^{R} p(\hat{\beta}^*|\hat{a}^{(r)}, \hat{\sigma}_a^{2(r)}, \hat{X}_{\text{mis}}^{(r)}, y), \tag{4.20}$$

where $\{\hat{a}^{(r)}, \hat{\sigma}_a^{2(r)}, \hat{X}_{\text{mis}}^{(r)}\}_{r=1}^{R}$ are draws from the Gibbs sampler conditional on $\hat{\beta}$. Analogous reduced conditionals are computed for $\hat{a}$, $\hat{\sigma}_a^2$, and $\hat{X}_{\text{mis}}$.

This reduced (or shrinked) Gibbs sampling approach ensures a numerically stable and consistent estimate of the posterior ordinate, allowing the marginal likelihood to incorporate uncertainty from both missing covariates and random effects in a coherent Bayesian manner.

Generally, the posterior distribution $\ln\hat{\pi}(\hat{\theta}|y,X)$ including the normalizing constant can be approximated by recursively shortened Gibbs sampler decomposed as follow:

$$\ln\hat{\pi}(\hat{\theta}|y,X) = \ln\hat{\pi}(\hat{\theta}_1|y,X) + \ln\hat{\pi}(\hat{\theta}_2|y,X,\hat{\theta}_1) + \cdots + \hat{\pi}(\hat{\theta}_P|y,X,\hat{\theta}_1,\ldots,\hat{\theta}_{P-1}) \tag{4.21}$$

with $p = 1,\ldots,P$ covariates. Each component of the posterior distribution derivation can be obtained by

$$\hat{\pi}(\hat{\theta}_p|y,X,\hat{\theta}_1,\ldots,\hat{\theta}_{p-1}) = \frac{1}{R}\sum_{r=1}^{R}\pi(\hat{\theta}_p|y,X,\hat{\theta}_1,\ldots,\hat{\theta}_{p-1},\theta_{p+1}^{(r)},\ldots,\theta_P^{(r)}) \tag{4.22}$$

with $r = 1,\ldots,R$ Gibbs sampling iterations as shortened Gibbs samplers for each parameter, which corresponds to evaluating the posterior ordinate at the fixed point $\hat{\theta}$ by averaging over the conditional distributions of the remaining parameters, thereby ensuring a consistent and computationally feasible estimate of the marginal posterior density component.

Considering the probit model with random effects and missing covariate values X_{mis}, shown in 4.3, so that the full parameter vector is given by: $\theta = (\beta, a, \sigma_a^2, X_{\text{mis}})$, then let $\hat{\theta} = (\hat{\beta}, \hat{a}, \hat{\sigma}_a^2, \hat{X}_{\text{mis}})$ denote a fixed point of interest (e.g., the posterior mode). The marginal likelihood can be expressed as:

$$\ln\hat{m}(y|X_{\text{obs}}) = \ln f(y|\hat{\theta}, X_{\text{obs}}) + \ln\pi(\hat{\theta}) - \ln\hat{\pi}(\hat{\theta}|y, X_{\text{obs}}), \tag{4.23}$$

where $\ln f(y|\hat{\theta}, X_{\text{obs}})$ is the approximated log-likelihood function evaluated at the fixed point (e.g., posterior mode) $\hat{\theta}$, $\ln\pi(\hat{\theta})$ is the log-prior density evaluated at $\hat{\theta}$, and $\ln\hat{\pi}(\hat{\theta}|$ $y, X_{\text{obs}})$ is the log-posterior ordinate at $\hat{\theta}$, which includes the normalizing constant. The posterior density at the fixed point $\hat{\theta} = (\hat{\beta}, \hat{a}, \hat{\sigma}_a^2, \hat{X}_{\text{mis}})$ can be decomposed using a recursive application of conditional densities:

$$\begin{aligned}\ln\hat{\pi}(\hat{\theta}|y, X_{\text{obs}}) = {} & \ln\hat{\pi}(\hat{X}_{\text{mis}}|\hat{\beta}, \hat{a}, \hat{\sigma}_a^2, X_{\text{obs}}, y) \\ & + \ln\hat{\pi}(\hat{\beta}|\hat{a}, \hat{\sigma}_a^2, \hat{X}_{\text{mis}}, X_{\text{obs}}, y) \\ & + \ln\hat{\pi}(\hat{a}|\hat{\beta}, \hat{X}_{\text{mis}}, X_{\text{obs}}, y) \\ & + \ln\hat{\pi}(\hat{\sigma}_a^2|\hat{\beta}, \hat{a}, \hat{X}_{\text{mis}}, X_{\text{obs}}, y).\end{aligned} \tag{4.24}$$

This decomposition ensures that missing covariate values are handled first in the Gibbs sampling sequence, following advice that prioritizes removing uncertainty in the design matrix before estimating structural parameters. To approximate the density of X_{mis}, for continuous components of X_{mis}, kernel density estimation (KDE) is applied to approximate $\hat{\pi}(\hat{X}_{\text{mis}}|\cdot)$, while empirical frequency tables are used for discrete components. These estimates are obtained from draws produced by a CART-based sequential regression imputation within the Gibbs sampler. The estimated density is computed by averaging the pointwise posterior kernel or frequency densities over R iterations. In detail, the average of the posterior density ordinates, which are calculated by the ordinates of the Gaussian kernel densities in case of continuous variables and by the empirical frequency distributions in case of discrete variables

within each iteration of the CART imputation step via Gibbs sampling, approximates the estimated missing values posterior density ordinates.[7] Overall, this approach provides a flexible and efficient way to estimate the posterior density of missing covariates and evaluate the marginal likelihood of the observed data given the imputed values. Each component can be approximated by:

$$\hat{\pi}(\hat{\theta}_p|\cdot) \approx \frac{1}{R}\sum_{r=1}^{R} \pi(\hat{\theta}_p|\cdot, \theta_{-p}^{(r)}), \tag{4.25}$$

where $\theta_{-p}^{(r)}$ denotes sampled values for all parameters other than θ_p, and R is the number of Gibbs iterations. This corresponds to evaluating each full conditional density at the fixed value $\hat{\theta}_p$, while integrating (via averaging) over the posterior samples of the remaining parameters. Specifically, the posterior ordinate of $\hat{a}$ is approximated by:

$$\ln\hat{\pi}(\hat{a}_i|y, X_{\text{obs}}) = \sum_{i=1}^{N} \ln\left(\frac{1}{R}\sum_{r=1}^{R} N(\hat{a}_i|\mu_{a_i}^{(m)}, \Sigma_{a_i}^{(r)})\right), \tag{4.26}$$

where $\mu_{a_i}^{(r)}$ and $\Sigma_{a_i}^{(r)}$ are posterior moments from the m-th iteration of the Gibbs sampler conditional on all other parameters fixed at $\hat{\theta}_{-a}$.

This strategy avoids the need for full MCMC over X_{mis} and allows efficient comparison of imputation strategies by directly integrating the imputed values into the marginal likelihood approximation via Chib's method (see Aßmann and Preising (2020), Chib (1995), and West (1993)).

Prior component of the marginal likelihood

In the random-effects probit model, the joint prior distribution over the parameters $\hat{\theta} = (\hat{\beta}, \hat{a}, \hat{\sigma}_a^2, \hat{X}_{\text{mis}})$ is factorized as follows:

$$\ln \pi(\hat{\theta}) = \ln \pi(\hat{\beta}) + \ln \pi(\hat{a}|\hat{\sigma}_a^2) + \ln \pi(\hat{\sigma}_a^2) + \ln \pi(\hat{X}_{\text{mis}}|X_{\text{obs}}) \tag{4.27}$$

where $\ln \pi(\hat{\beta})$ is the log-prior for the vector of fixed effects. A common choice is a weakly informative Gaussian prior, e.g., $\beta \sim \mathcal{N}(0, \tau^2 I)$ with large τ^2. $\ln \pi(\hat{a}|\hat{\sigma}_a^2)$ denotes the log-prior for the individual-specific random effects. These are typically assumed to follow a normal distribution: $a_i \sim \mathcal{N}(0, \hat{\sigma}_a^2)$, independently across individuals. $\ln \pi(\hat{\sigma}_a^2)$ is the log-prior for the random effect variance component. A standard choice is an inverse-gamma or a half-Cauchy prior to allow for uncertainty over the scale of heterogeneity. $\ln \pi(\hat{X}_{\text{mis}}|X_{\text{obs}})$ represents the conditional prior for the missing covariate values. In the above-mentioned model, this distribution is defined implicitly via a data augmentation step using regression trees (CART), integrated into the Gibbs sampling procedure. It reflects a

[7] This approach allows to estimate the posterior density of the missing values without having to run a full MCMC simulation, which can be computationally expensive. Instead, the imputed values obtained from CART as a starting point are used and shortened MCMC sequences are performed to estimate the posterior density. The estimated posterior density ordinates can then be used to evaluate the marginal likelihood of the observed data given the imputed values of the missing covariates. Hence, this allows to compare different imputation methods and select the one that provides the best fit to the observed data.

flexible, model-driven imputation approach based on the observed data. The inclusion of $\pi(\hat{X}_{\text{mis}}|X_{\text{obs}})$ in the joint prior accounts for uncertainty due to missing covariates and can be interpreted in light of the assumed missing data mechanism (MCAR, MAR, or MNAR). While this distribution is not specified analytically, its structure is learned empirically within the MCMC procedure. The approach differentiates between factor and continuous covariates: for categorical variables the empirical frequencies of each category in the observed data are used to define the prior probabilities. For a missing entry assigned to category c, the log-prior contribution is $\log(p_c)$, where p_c is the empirical probability of category c. And for continuous variables a nonparametric kernel density estimate (KDE) of the observed values is employed. The log-density of the imputed value under this KDE is taken as the log-prior contribution. If a value lies outside the observed range, a very small log-probability is assigned to prevent numerical issues.

Recent research highlights the robustness of tree-based imputation methods under a Bayesian framework. For instance, Doove et al. (2014) show that using nonparametric models such as CART or random forests within a multiple imputation framework can effectively capture complex dependencies and interactions, especially under missing at random (MAR) conditions. Moreover, such models can be embedded into fully Bayesian Gibbs samplers to propagate imputation uncertainty in posterior inference (see also Hahn et al. (2020)).

Careful specification of the prior distributions is especially important in settings with small sample sizes or large proportions of missing data. In such cases, the employment of weakly informative or informative priors informed by subject-matter knowledge has been demonstrated to assist in stabilizing inference and avoiding overfitting (Gelman et al., 2023).

(Log)-Likelihood component of the marginal likelihood

For the final component of the marginal likelihood computation the log-likelihood component, $\ln f(y|\hat{\theta}, X)$, is approximated. In the case of complete covariate information, the log-likelihood of the observed data under a random-intercept probit model can be approximated using Gauss–Hermite quadrature (GHQ). Assuming one normally distributed random effect $a_j \sim \mathcal{N}(0, \sigma_a^2)$ per cluster or region j, the marginal likelihood for cluster j is

$$p(y_j|X_j, \hat{\beta}, \hat{\sigma}_a^2) = \int \prod_{j=1}^{J} \Phi\left(x_{ij}^\top \hat{\beta} + a_j\right)^{y_{ij}} \left[1 - \Phi\left(x_{ij}^\top \hat{\beta} + a_j\right)\right]^{1-y_{ij}} \cdot \phi(a_j; 0, \hat{\sigma}_a^2)\, da_j. \tag{4.28}$$

Here, $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) and $\phi(\cdot)$ the probability density function (PDF) of the standard normal distribution. Using the substitution $a_j = \sqrt{2\hat{\sigma}_a^2} \cdot z_q$, the integral is approximated by

$$p(y_i|X_i, \hat{\beta}, \hat{\sigma}_a^2) \approx \frac{1}{\sqrt{\pi}} \sum_{q=1}^{Q} w_q \prod_{j=1}^{J} \Phi\left(x_{ij}^\top \hat{\beta} + \sqrt{2\hat{\sigma}_a^2} z_q\right)^{y_{ij}} \left[1 - \Phi\left(x_{ij}^\top \hat{\beta} + \sqrt{2\hat{\sigma}_a^2} z_q\right)\right]^{1-y_{ij}}, \tag{4.29}$$

where $\{z_q\}$ and $\{w_q\}$ denote the Gauss–Hermite nodes and weights. This deterministic integration method is computationally efficient and well-suited for low-dimensional integrals arising from models with a single random effect per unit.

In models with normally distributed random effects and binary outcomes, the evaluation of the marginal likelihood becomes particularly challenging when covariates are partially missing. If these missing values are imputed within the Gibbs sampler – e.g., via Classification and Regression Trees (CART) combined with Bayesian bootstrap – the complete data likelihood is no longer directly available, and its approximation must respect the underlying latent structure of the model. A feasible solution in this context is to condition the likelihood on the imputed covariates $\hat{\mathbf{X}}_{\text{mis}}$, treating them as fixed when computing the log-likelihood:

$$\log p(\mathbf{y}|\hat{\mathbf{X}}, \hat{\beta}, \hat{\mathbf{a}}) = \sum_{j=1}^{J} \log \left\{ \Phi\left(\hat{\mathbf{x}}_i^\top \hat{\beta} + \hat{a}_{j[i]}\right)^{y_i} \cdot \left[1 - \Phi\left(\hat{\mathbf{x}}_i^\top \hat{\beta} + \hat{a}_{j[i]}\right)\right]^{1-y_i} \right\}, \tag{4.30}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function and $j[i]$ indicates the cluster assignment. However, since the random effects $\hat{a}_j$ follow a latent distribution, the likelihood must integrate over their posterior uncertainty. When Gauss–Hermite quadrature is no longer practical[8] the Geweke–Hajivassiliou–Keane (GHK) simulator provides an efficient method to approximate the required high-dimensional integrals (Geweke & Keane, 2001; Hajivassiliou & McFadden, 1998; Hajivassiliou & Ruud, 1994; Keane, 1994).[9] The GHK simulator approach recursively simulates from a sequence of truncated univariate normal distributions, based on a Cholesky decomposition of the covariance matrix. The GHK simulator is an alternative approach often used for computing the likelihood in multivariate probit models or models with multiple correlated latent variables per observation. It simulates draws from truncated multivariate normal distributions and relies on recursive conditioning (Keane, 1994). In contrast, GHQ is deterministic and suitable when the integral dimension is low — as in our case with one random effect per group. The Gauss–Hermite Quadrature instead is opted for GHK simulator because complete case and before deletion contains only one normally distributed random intercept per group, and no time-varying or multidimensional latent structures. This makes GHQ both computationally efficient and numerically stable, avoiding the Monte Carlo error inherent in GHK simulations. For each cluster i, the contribution to the marginal likelihood is approximated as

$$\hat{p}(y_j|\hat{X}_j, \hat{\beta}, \hat{\sigma}_a^2) \approx \frac{1}{R} \sum_{r=1}^{R} \prod_{j=1}^{J} \left[\mathbb{I}(y_{ij} = 1)\Phi(x_{ij}^\top \hat{\beta} + a_j^{(r)}) + \mathbb{I}(y_{ij} = 0) \left(1 - \Phi(x_{ij}^\top \hat{\beta} + a_j^{(r)})\right) \right], \tag{4.31}$$

where $a_j^{(r)} \sim \mathcal{N}(0, \hat{\sigma}_a^2)$ are sampled random intercepts.

In this framework, the GHK simulator is applied post-estimation using posterior point estimates $\hat{\beta}$ and $\hat{\sigma}_a^2$. It enables the construction of a consistent log-likelihood estimate even when

[8] E.g., in the presence of multiple random effects, interactions, or complex data structures.

[9] The GHK simulator is particularly advantageous in settings where the latent structure induces multivariate normal truncation, such as in multivariate probit or panel models with individual and time random effects

covariates were imputed non-parametrically. When embedded into a marginal likelihood estimator this approximation contributes the term $\log p(y|\hat{\theta})$, where $\hat{\theta} = (\hat{\beta}, \hat{a}, \hat{\sigma}_a^2)$. Averaging over multiple imputations $\hat{X}^{(r)}$ further improves robustness and reflects uncertainty due to missingness. This approach to approximating the log-likelihood via the GHK simulator is consistent with the methodology proposed by Aßmann and colleagues, who applied GHK-based simulation techniques for marginal likelihood estimation in multivariate and hierarchical probit models with latent structures and clustered random effects (see Aßmann (2007) and Steinhauer and Aßmann (2018), while in the absence of missing values the computationally more efficient and numeric stable version, the GHQ, is used. Similar strategies have since been adopted in more recent work on simulated Bayesian inference in non-linear panel models and discrete choice frameworks, see Daziano and Achtnicht (2014) for Bayesian estimates of a multinomial probit model, Lucchetti and Pedini (2023) for estimation in the ML framework or Liesenfeld and Richard (2010) for probit models with correlated errors.

This recursive formulation allows us to isolate each component of the posterior and evaluate the density at the fixed point $\hat{\theta}$ using Gibbs output. It thus yields an accurate estimate of the marginal likelihood, even in the presence of missing data and latent structures such as random effects.

4.3.2 Bayes factor

In direct conjunction with the above mentioned, for some data X and a parameter vector θ the Bayes factor (BF_{12}) of comparing model $\mathcal{M}_1$ with model $\mathcal{M}_2$ is the fraction of both marginal likelihoods:

$$BF_{12} = \frac{p(X|\mathcal{M}_1)}{p(X|\mathcal{M}_2)} \tag{4.32}$$

and in the case of using the log of the marginal likelihoods:

$$\ln BF_{12} = \ln p(X|\mathcal{M}_1) - \ln p(X|\mathcal{M}_2). \tag{4.33}$$

Comparing two models yields to these odds:

$$\underbrace{\frac{p(\mathcal{M}_1|X)}{p(\mathcal{M}_2|X)}}_{\text{Posterior odds}} = \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}}_{\text{Prior odds}} \times \underbrace{\frac{p(X|\mathcal{M}_1)}{p(X|\mathcal{M}_2)}}_{\text{Bayes factor}} \tag{4.34}$$

The objective of model comparison is to ascertain which model is more likely to have produced the observed data. In other words, it is a question of which model is more likely to have produced the observed data. When marginal probabilities are compared, the Bayes factor is a measure of the relative strength of one of the compared models over the other(s). It is evident that the Bayesian approach, which incorporates uncertainty in the model-building process, bears a close relationship to the methodology of hypothesis testing (Kass & Raftery, 1995).

The Bayes factor provides a measure of the evidence in favor of one hypothesis over the other. Typically, a Bayes factor greater than 1 favors model 1, while a Bayes factor less than

1 favors model 2. By using the thumb of rule: if $\ln BF_{12} > 1$, model M_1 is preferred, and if $\ln BF_{12} < 1$, model M_2 is preferred. Categorizations of the Bayes factor are found by Jeffreys (1961), Goldman and Whelan (2000) and Held and Ott (2018). The BF can be used to compare nested and unnested models. See table 4.1 for an overview of the scales provided by Jeffreys (1961) and Kass and Raftery (1995).

4.3.3 Bayesian Information Criterion

The Bayesian Information Criterion (BIC)[10] is a special case of the Information Criterion by Schwarz (1978).

In general, with n being the number of observations and k being the number of model parameters the log-likelihood function $lnf(y|X,\theta)$ is defined as follows:

$$BIC = -2\ln f(y|X,\theta) + k \cdot \ln n. \tag{4.35}$$

Thus, in model comparison "the model with the highest posterior probability is the one that minimizes" (Kass & Raftery, 1995) the BIC. If the data is completely observed, then the log-likelihood function is evaluated at $\hat{\theta}$, i.e., the posterior mean or median derived from the Gibbs sampler. If missing values appear in the covariates, the BIC in equation 4.35 can calculated by

$$BIC = -2\ln \hat{f}(y_{\text{obs}}|X_{\text{obs}},\hat{\theta}) + k \cdot \ln n. \tag{4.36}$$

with the approximated observed data log-likelihood function

$$\ln \hat{f}(y_{\text{obs}}|X_{\text{obs}},\hat{\theta}) \approx \ln\left[\frac{1}{R}\sum_{r=1}^{R}\hat{f}(y_{\text{obs}}|X_{\text{obs}},\hat{\theta},X_{\text{mis}}^{(r)})\right] \tag{4.37}$$

where M is the number of Gibbs sampling iterations.

In addition to the above mentioned Bayes factor criterion, the strength of the information criterion is that it can be used for non-nested models and that two or more models can be clearly defined and thus distinguished: always choose the one with the lower BIC (Kass & Raftery, 1995). However, the focus of the BIC is on the balance between the fit of the model and the complexity and can be used to compare several non-nested models. BICs can be seen as a substitute for Bayes factors only when no reliable prior information is available and when the sample size is quite large.

4.3.4 Median p-value approach for Maximum Likelihood Estimation

In contrast to the Bayesian augmented data approach mentioned above, a Maximum Likelihood approach can be used, e.g., for based on quadrature integration see among others Butler and Moffitt (1982) and Sixt and Aßmann (2020). Missing values are treated before estimation and not while estimating. According to Rubin (1981), the missing values are often imputed, so-called multiple imputation. Multiple imputation via chained equations (MICE,

[10]The BIC is also called Schwarz Information Criterion (SIC), cf. Frühwirth-Schnatter (2010).

see Rubin (1976) and van Buuren and Groothuis-Oudshoorn (2011)) handles the uncertainty in parameter estimation associated with missing covariates. Instead of imputing a single value, multiple datasets, R, are created, each with different imputed values. This captures the uncertainty associated with missing data. The analysis is then performed separately on each imputed dataset, and the results are combined to produce more accurate and reliable estimates, accounting for the variability introduced by the imputation process. Multiple imputation is particularly useful in preserving the uncertainty inherent in missing data experimental studies and is widely employed in various fields such as epidemiology, social sciences, and clinical research.

Likelihood ratio tests (LRT) are statistical tests used to compare the fit of two nested models, typically a null model and an alternative model (Greene, 2003). The test statistic is based on the ratio of the likelihoods of the two models. Consider two nested models, denoted as M_0 (null model) and M_1 (alternative model), where M_1 is a special case of M_0 and thus nested within it. The likelihood ratio test statistic (LR) is calculated as the ratio of the likelihoods under M_1 and M_0:

$$LR = -2 \times \log\left(\frac{\mathcal{L}(M_0)}{\mathcal{L}(M_1)}\right). \tag{4.38}$$

Under the null hypothesis (H_0), the test statistic LR follows a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between M_0 and M_1. Therefore, the observed LR is compared with the critical value from the chi-squared distribution, if LR is greater than the critical value, the H_0 is rejected otherwise the decision is not for model M_1. Likelihood ratio tests are only valid for nested models in which the smaller model can be obtained from the larger one by constraining parameters. For models with identical random-effects structure and differing only in fixed effects, the test statistic follows a chi-squared distribution with degrees of freedom equal to the difference in the number of fixed-effect parameters (Greene, 2003). However, when testing whether variance components are zero, the null hypothesis lies on the boundary of the parameter space, leading to a non-standard distribution (typically a mixture of chi-squared distributions), and the naive chi-squared approximation can be anticonservative (Goldman & Whelan, 2000; Pinheiro & Bates, 2000; Self & Liang, 1987; Stram & Lee, 1994). In such cases, parametric bootstrap versions of the LRT are recommended (Crainiceanu & Ruppert, 2003). For non-nested models, likelihood-ratio-based inference is not appropriate. Instead, information criteria (AIC, BIC) and their transformations into Akaike weights (Burnham & Anderson, 2002) or BIC-based log Bayes factors (as described above) can be employed.

To obtain an overall summary measure, for example after multiple imputation, the median pooling rule refers to a method of combining p-values from multiple tests (Eekhout et al., 2017; Marshall et al., 2009). This approach is often used when conducting multiple hypothesis tests and involves taking the median of the individual p-values. First, LRT are performed for each hypothesis of interest, obtaining p-values $p_1, p_2, \ldots, p_R$. The individual p-values are

then compared using the median pooling rule:

$$\text{total p-value} = \text{median}(p_1, p_2, \ldots, p_R).$$

If the overall p-value is below a specified significance level (e.g., 0.05), the null hypothesis is rejected. Compared to the marginal likelihood approach above, the median p-value (Eekhout et al., 2017) and the LR test (Chan & Meng, 2022; Meng & Rubin, 1992) can be used to compare nested models. Both multiple imputation and the median p-rule address uncertainty whereby multiple imputation acknowledges uncertainty about missing data, and the median p-rule accounts for variability in the individual p-values. By combining results from multiple imputed datasets, the validity of hypothesis tests can be improved, especially when missing data are a concern.

Additionally, it should be mentioned that it is not clear which degrees of freedom are taken into account for terms with random effects. The LRT produces biased estimates because the derivation of the LRT depends on a Taylor expansion of the log-likelihood around the null parameters (Goldman & Whelan, 2000; Pinheiro & Bates, 2000; Self & Liang, 1987; Stram & Lee, 1994). Random effects models do not have the clarity of fixed effects models, where a solid geometric interpretation leads to an explicit definition of degrees of freedom. If the parameters to be estimated are counted, then random effects for the levels of a factor only cost one degree of freedom, regardless of the number of levels, here J additional levels that can cause underestimation, if degrees of freedom are used as a measurement of explanatory power of models. Hence, in the presented model comparisons study only models with random effects are compared so that the use of random effects in each model causes a cost of one degree of freedom overall.[11] This contrasts with the consideration that following the geometric argument of the fixed effects model, the degrees of freedom would be the number of levels or that number minus 1 (Li & Redden, 2015).[12]

As a side note, the so-called median pooling rule (Eekhout et al., 2017; Marshall et al., 2009) can be formulated in exact finite-sample terms. Consider R p-values obtained from repeated analyses (e.g., across R imputations) when comparing a model of interest against a reference model (in the experimental study later on model 2) via likelihood-ratio tests. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(R)}$ denote the order statistics of these p-values, and let $p_{(r)}$ denote the r-th smallest value. Under the null hypothesis H_0, and assuming the p-values are independent and identically distributed as $U(0,1)$ variables, the event $\{p_{(r)} \leq x\}$ is equivalent to having at least r of the R values less than or equal to x, which follows a Binomial(R, x) distribution (cf. Mittelhammer, 2013). Its cumulative probability is therefore

$$P\big(p_{(r)} \leq x\big) = \sum_{j=r}^{R} \binom{R}{j} x^j (1-x)^{R-j}. \tag{4.39}$$

[11] The truth probably lies between these two extremes, so that many researchers take the path of simple solutions and apply the naive LRT.

[12] Imaging the OLS regression, counting the number of estimates results in the degree of freedom. But the estimates do not have all the degrees of freedom associated with the geometric subspace. The estimators for random effects refer to a penalized regression problem, which means that they have a moderating effect on the estimators for the coefficients.

It is well known that this sum equals the cumulative distribution function (CDF) of a Beta$(r, R-r+1)$ distribution (David & Nagaraja, 2003), whose density is

$$f(x) = \frac{R!}{(r-1)!(R-r)!}\, x^{r-1}(1-x)^{R-r}, \quad 0 < x < 1. \tag{4.40}$$

Consequently, the exact pooled p-value is given by

$$p_{\text{pool}} = F_{\text{Beta}}\big(p_{(r)};\ r,\ R-r+1\big), \tag{4.41}$$

where $r = \left\lceil \frac{R}{2} \right\rceil$[13] for the median pooling rule. This exact finite-sample formulation replaces the common practice of simply taking the arithmetic median of the R p-values and treating it as uniformly distributed. The Beta-based adjustment accounts for the non-uniform sampling distribution of order statistics, thereby yielding exact type-I error control for given R and r. Figure 4.1 illustrates the median pooling for $R = 20$ and $r = 10$.

4.3.5 BIC-based log Bayes factors for Maximum Likelihood Estimation

As outlined in Section 4.3.3, the BIC can be interpreted as a large-sample approximation to twice the negative log of the marginal likelihood (Kass & Raftery, 1995; Schwarz, 1978). Hence, the BIC difference between two models, say M_r (reference model) and M_j (comparison model), provides an estimate of the logarithm of the Bayes factor $\ln BF$ without the need to specify prior distributions explicitly:

$$\ln BF_{j,r} \ \approx\ -\frac{1}{2}\left[BIC_j - BIC_r\right]. \tag{4.42}$$

Positive values of $\ln BF_{j,r}$ indicate evidence in favor of model M_j, while negative values support model M_r. Following Kass and Raftery (1995), the strength of evidence can be classified as: $0 < |\ln BF| < 0.5$ (barely worth mentioning), $0.5 < |\ln BF| < 1.0$ (substantial), $1.0 < |\ln BF| < 2.0$ (strong), and $|\ln BF| > 2.0$ (decisive).

In the case of multiple imputation, equation (4.42) can be applied to each imputed dataset separately, yielding $\ln BF_{j,r}^{(m)}$, $r = 1, \ldots, R$. To incorporate imputation uncertainty, the mean and standard deviation of these $\ln BF_{j,r}^{(m)}$ is reported across imputations. This approach retains the interpretability of the Bayes factor scale while being computationally efficient and applicable in the ML framework. An additional advantage of the $\ln BF$ based on BIC in the ML framework is its applicability to both nested and non-nested model comparisons. Unlike the LRT, which is restricted to nested models and can produce ambiguous results in the presence of boundary issues or small samples, the $\ln BF$ delivers a single interpretable measure of relative model support that is consistent across different model structures.

[13] $\lceil \cdot \rceil$ denotes the ceiling function, i.e., rounding up to the nearest integer. It specifies here the index of the median p-value in the ordered list $p_{(1)}, p_{(2)}, \ldots, p_{(R)}$.

4.4 Implementation and quality aspects

To evaluate the proposed procedures and ensure the reliability of the resulting estimates and imputations, three aspects are addressed: (i) implementation details in R and Julia, and (ii) model comparison strategies, including hypothesis testing after multiple imputation.

4.4.1 Implementation in R and Julia

The Bayesian estimation approach, based on a Gibbs sampler, was implemented both in R and Julia. While both versions yielded virtually identical parameter estimates, the Julia implementation was substantially faster (Table 4.2). The R code relied on `rpart` (Therneau & Atkinson, 2018) for CART-based imputation, `truncnorm` (Mersmann et al., 2023) for truncated normals, `MASS` (Venables & Ripley, 2002) for multivariate normals, `invgamma` (Martin et al., 2011) for inverse gamma draws, and `mvtnorm` (Genz et al., 2021) for multivariate normal densities. In each Gibbs iteration $m = 1, \ldots, M$, missing covariate values were imputed using a Bayesian bootstrap with the `mice` package (van Buuren & Groothuis-Oudshoorn, 2011). For continuous variables, Gaussian kernel densities (bandwidth `bw.nrd0`) were fitted to the donor values $X_{\text{obs},k}$, while categorical variables were imputed using empirical frequency distributions. The ML estimation was implemented fully in R, with random-effects models estimated via `lme4` (Bates et al., 2015) and missing values handled through `mice`.

The Julia Gibbs sampler (Bezanson et al., 2017) made use of specialized packages for numerical integration, automatic differentiation, and probabilistic programming, ensuring efficient matrix operations and sampling.

4.4.2 Model comparison and pooling after multiple imputation

For the ML analyses, probit generalized linear mixed models (GLMMs) with random intercepts $(1|\text{id})$ are estimated. The models differed only in their fixed-effects specification; the random-effects structure was held constant across models to ensure comparability in likelihood-ratio tests (LRTs). Estimation was performed using `lme4::glmer` with the probit link, Laplace approximation (default, `nAGQ` = 1), and the `bobyqa` optimizer with `maxfun` = 10^5, which is standard for stable GLMM fitting (Bates et al., 2015). Within each dataset $d = 1, \ldots, M$ and for each model, results across the R imputations were combined using Rubin's rules (Rubin, 1987). For GLMMs, pooling was implemented using the per-imputation estimates $\hat{\beta}^{(r)}$ and their variances $\widehat{\text{Var}}\left(\hat{\beta}^{(r)}\right)$, applying the small-sample correction of Barnard and Rubin (1999) where appropriate. Random-effects variance components $\hat{\sigma}^2_{u,(r)}$ were summarized by their mean across imputations (no Rubin pooling was applied to variance components). Model fit statistics (log-likelihood, AIC, and BIC) were averaged across imputations, with the standard deviation of the log-likelihood also reported. The AIC and BIC were computed from the marginal log-likelihood as provided by `glmer` (Laplace).

Across the M simulated datasets, the mean estimate, empirical standard deviation, bias,

$$\text{Bias} = \frac{1}{M} \sum_{m=1}^{M} \left(\hat{\beta}^{(m)} - \beta_{\text{true}} \right), \tag{4.43}$$

and root mean squared error (RMSE),

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\beta}^{(m)} - \beta_{\text{true}}\right)^2}, \tag{4.44}$$

as well as the empirical coverage rate of Wald-type 95% confidence intervals are summarized for each parameter. At the model level, the mean random-effects variance, mean and standard deviation of the log-likelihood, and the mean AIC and BIC across datasets are reported.

For nested model comparisons, per imputation the likelihood-ratio statistic is computed by

$$\text{LRT}^{(r)} = -2\left(\ell_0^{(r)} - \ell_1^{(r)}\right), \quad \text{df} = k_1 - k_0, \tag{4.45}$$

with $\ell_0^{(r)}$ and $\ell_1^{(r)}$ denoting the maximized log-likelihoods of the restricted and full model, respectively. From this, the χ^2_{df} p-value $p^{(r)}$ was obtained. Following Marshall et al. (2009) and Eekhout et al. (2017), the R per-imputation p-values were then combined using the *median p-rule*: Let $p_{(1)} \leq \cdots \leq p_{(R)}$ denote the ordered p-values and $r = \lceil R/2 \rceil$ the index of the median. Specifically, the r-th order statistic $p_{(r)}$ with $r = \lceil R/2 \rceil$ was identified and transformed via equation 4.41. This adjustment accounts for the non-uniform sampling distribution of order statistics under H_0, providing exact Type I error control for independent Uniform(0,1) p-values and remaining a good approximation when imputations are not strictly independent. Over the M datasets, p_{pool} by its median, mean, and the proportion of values below $\alpha = 0.05$ is summarized. In the experimental study, model 2 always served as the reference specification for LRT-based comparisons. For non-nested model comparisons, it is referred to information criteria and Bayes factors. For nested as well as non-nested model comparisons, the Bayesian Information Criterion (BIC) was computed per imputation and used to approximate the log Bayes factor (lnBF) via equation 4.42 following Kass and Raftery (1995). The log Bayes factors were computed per imputation and summarized by the mean and a t-based 95% confidence interval. When multiple datasets ($M > 1$) were available, these summaries were averaged across datasets to yield an overall evidence measure.

This workflow was implemented in R using `lme4` (Bates et al., 2015) for GLMM estimation, `mice` (van Buuren & Groothuis-Oudshoorn, 2011) for multiple imputation, and custom functions for pooling and model comparison. All hypothesis tests after multiple imputation were conducted relative to the designated reference model.

4.5 Evaluation

4.5.1 Design of the experimental study

This section presents the simulation design used to evaluate the performance of the Gibbs sampler approach and the ML approach described above. The entire data generation process is implemented in R, while estimation is performed in both R and Julia (see section 4.4). One experimental study is conducted with setting missing values and correlation among the covariates where $M = 100$ independently generated datasets are analyzed.

Data generating process

The binary outcome Y_{ij} is generated according to the probit random-intercept model

$$Y_{ij}^* = X_{ij}^\top \beta + a_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0,1), \tag{4.46}$$

where $Y_{ij} = \mathbb{1}(Y_{ij}^* > 0)$, $i = 1, \ldots, N_j$ denotes individuals, and $j = 1, \ldots, J$ denotes groups (e.g., imaging a region). The number of groups is fixed at $J = 50$, with a total of $N = 2{,}000$ individuals, randomly assigned to regions using the `sample()` function in R (sampling with replacement). Group-specific random effects a_j are drawn from $\mathcal{N}(0, \sigma_a^2)$ with $\sigma_a^2 = 1.00$.

For each replication $m = 1, \ldots, M$, a design matrix X of covariates was generated as follows. First, a constant term $X_1 \equiv 1$ was included for all N individuals. The three primary continuous predictors X_2, X_3, X_4 were drawn jointly from a trivariate normal distribution with mean vector $\boldsymbol{\mu} = (0, 0, 0)^\top$ and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & .45 & -.15 \\ .45 & 1 & .25 \\ -.15 & .25 & 1 \end{pmatrix}, \tag{4.47}$$

thus inducing moderate positive and negative correlations between covariates. A categorical variable X_5 with three equally likely levels (A, B, C) was generated by independent sampling from the discrete uniform distribution. An additional continuous noise variable X_7 was generated from $\mathcal{N}(0,1)$, and a second categorical noise variable X_8 was sampled analogously to X_5. Finally, a further continuous noise covariate $X_{10} \sim \mathcal{N}(0,1)$ was generated. The categorical variable X_5 disintegrates into X_{5A}, X_{5B}, X_{5C} referred below to X_5, X_6, whereas the X_{5A} is skipped as a reference category. The same applies to X_7, which disintegrates into X_{8A}, X_{8B}, X_{8C} referred below to X_8, X_9, whereas the X_{8A} is skipped as a reference category.

Categorical covariates were transformed into dummy variables via one-hot encoding, producing a full design matrix including all continuous predictors and binary indicators for the factor levels. Group-specific random intercepts a_j were generated for $j = 1, \ldots, J$ and assigned to individuals according to their group membership. The linear predictor for individual i in group j was then

$$\mu_{ij} = X_{ij}^\top \beta_{\text{true}} + a_j, \tag{4.48}$$

where $\beta_{\text{true}} = (.20, -1.50, .80, -.30, -.40, -.30)^\top$ denotes the true regression coefficients. Thus, variables X_7, X_8, X_9, and X_{10} are redundant variables. A latent variable representation was employed, with

$$z_{ij} = \mu_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1), \tag{4.49}$$

and the binary outcome defined as $y_{ij} = \mathbb{1}(z_{ij} > 0)$, as mentioned in section 4.2.

Missing data design

The complete data matrix for each replication is stored for subsequent analyses. From each baseline dataset, two versions with missing values were created, see table 4.4:

1. **MCAR scenario:** Missing values were introduced by replacing a fixed fraction of entries in selected covariates with `NA` values, under the missing completely at random (MCAR) assumption. Specifically, in each simulated dataset, the following proportions of observations were set to missing via simple random sampling without replacement: 20% of X_2, 30% of X_3, 10% of the dummy variable X_5 as well as 15% of X_6, 20% of the variable X_7, and 30% of X_{10}. The selection of rows for deletion was independent of both the outcome variable and the remaining covariates, ensuring the MCAR property. This design generates varying amounts of missingness across variables, allowing the robustness of the estimation procedures to different missing rates to be assessed. Across the $M = 100$ simulated datasets, the mean proportion of complete cases across datasets was approximately 25%, reflecting the overlap of missingness across covariates.

2. **MAR scenario:** Missingness indicators were generated using a logistic selection model depending on observed covariates and independent noise. For example, for X_3, an auxiliary score
$$w_i = \frac{1}{1 + \exp\{0.2 \cdot X_{3,i} + \eta_i\}}, \quad \eta_i \sim \mathcal{N}(0, 1), \tag{4.50}$$
was computed for each individual i. Observations with w_i exceeding the empirical 90th percentile of the score distribution were set to missing. Analogous mechanisms were applied to the dummy variables (thresholds at the 85th and 75th percentiles for X_5 and X_6 , respectively) and to X_7 (80th percentile). This design induces missingness that is related to observed covariates but not to the unobserved values themselves, thereby satisfying the MAR assumption while producing heterogeneous missing rates across variables. Across the $M = 100$ simulated datasets, the mean proportion of complete cases across datasets was approximately 45%, reflecting the overlap of missingness across covariates.

Both MCAR and MAR datasets were saved separately for later analysis, maintaining alignment with the corresponding before deletion and complete cases dataset for each replication.

Model specifications

Six (non-) nested model specifications are considered (see Table 4.3). All models include a constant X_1 and a random intercept a_j for regions j:

- **Model I**: X_1 (intercept only),
- **Model II**: $X_1, X_2, X_3, X_4, X_5, X_6$ (reference model for the data generating process),
- **Model III**: X_1, X_2, X_5, X_6,

- **Model IV**: $X_1, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$,
- **Model V**: $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$,
- **Model VI**: $X_1, X_2, X_5, X_6, X_7, X_{10}$.

4.5.2 Sensitivity to prior specification

Model robustness was assessed by examining the sensitivity of results to alternative prior choices, as is common in Bayesian analysis. In the simulation studies, the priors for β (cf. equation 4.9) were specified with mean $\mu_\beta = \mathbf{0}$ and covariance matrices $\Sigma_\beta \in \{1 \times \mathbf{I}, 10 \times \mathbf{I}, 100 \times \mathbf{I}, 1000 \times \mathbf{I}\}$,[14] and the hyperparameters for σ_a were chosen to mimic an exponential distribution with scale 3 by setting $c_0 = 1$ and $d_0 = 3$ in the gamma prior. Additional scenarios varied $c_0 \in \{1\}$ and $d_0 \in \{1, 2\}$. For each setting, the analysis was repeated, and posterior summaries as well as convergence diagnostics were compared. Overall, the posterior estimates were only marginally affected by prior changes; differences were mainly observed in convergence speed. Table 4.17 reports the effect of different β prior specifications on the marginal likelihood. Therefore, to ensure stability, it is determined that $\Sigma_\beta = 100 \cdot I$ and $(c_0, d_0) = (1, 3)$.

4.5.3 Discussion of the evaluation of experimental study

The accuracy of parameter estimation was assessed separately for the ML and Bayesian approaches. For each estimation framework, results are reported for five data scenarios: (i) before deletion of any observations (BD), (ii) complete case analysis under missing completely at random (CC–MCAR), (iii) multiple imputation under MCAR (IMP–MCAR), (iv) complete case analysis under missing at random (CC–MAR), and (v) multiple imputation under MAR (IMP–MAR).

For each scenario, the results are presented in a table with model specifications model I to model VI in the columns and the following quantities in the rows: posterior mean (Bayes) or point estimate (ML), standard error (SE), (absolute) bias, root mean squared error (RMSE), and coverage of the nominal 95% credible or confidence intervals averaged over the $M = 100$ datasets.

Table 4.5 displays the results for Bayesian estimation for (i), table 4.6 for (ii), table 4.7 for (iii), table 4.12 for (iv), table 4.13 for (v) and table 4.8 displays the results for ML estimation for (i), table 4.9 for (ii), table 4.10 for (iii), table 4.14 for (iv), table 4.15 for (v), respectively. Across all scenarios, the ML estimates exhibited a clear degradation in performance under complete case analysis with MAR missingness: bias increased substantially, RMSE values were larger, and coverage often fell below the nominal 95% level. Comparing to complete case scenario multiple imputation mitigated these effects, reducing bias and restoring coverage close to the nominal level, particularly under MCAR. The Bayesian approach produced estimates that were broadly comparable to the ML results in terms of bias, albeit with a slightly greater shrinkage toward the prior mean, resulting in a small additional bias in some scenarios. This,

[14] $\mathbf{I}$ denotes the $k \times k$ identity matrix.

however, was compensated by consistently smaller posterior standard deviations, yielding more stable and precise estimates across all conditions. Even under MAR, the combination of data augmentation within the Gibbs sampler and multiple imputation effectively maintained coverage near nominal levels while reducing variance relative to the ML estimates.

The performance of the model selection process was evaluated by comparing each alternative specification model (I, II, III, IV, V, and VI) to the reference model (II). The same five data scenarios as above were considered. As illustrated in table 4.11 the results for the MCAR scenario and for each model are reported. Similarly, as demonstrated in table 4.16, the results for the MAR scenario and for each model are reported, with log Bayes factors (lnBF) for the Bayesian estimation results and median pooled p-values for each model pair (LR Test) and log Bayes factors (lnBF) for the ML estimation results. For the ML framework, model comparisons relied on likelihood ratio tests (LRTs) with the random-effects structure held constant across all models. Per imputation, the LRT statistic and its corresponding chi-squared p-value were computed; these p-values were then combined across imputations using the exact median-p pooling method, based on the Beta null distribution (Eekhout et al., 2017; Marshall et al., 2009) as mentioned above. The resulting pooled p-values were then summarized across the Monte Carlo replications. Note that likelihood-ratio tests are only valid for nested model comparisons. For non-nested pairs (here: models IV vs II and VI vs II), the LR statistic does not follow the usual asymptotic χ^2 reference distribution; consequently, the corresponding p-values are not inferentially meaningful and are reported, if at all, for descriptive completeness only. For the Bayesian framework, following equation 4.33 model comparisons were based on Bayes factors (BFs) computed from the BIC approximation of the marginal likelihood (Kass & Raftery, 1995; Schwarz, 1978). For each imputed dataset, the BIC for each model was obtained, and the corresponding lnBF relative to model II was computed. These were then summarized across replications. Interpretation follows the scale of Kass and Raftery (1995) with $lnBF$ values above 5 indicating strong evidence in favor of model II.

In the MCAR case, across all scenarios – before deletion, complete cases, and multiple imputation – the Bayes factors indicate decisive evidence in favor of model II over all other specifications. Negative lnBF values consistently favor model II, with values often exceeding 100 in magnitude, reflecting extremely strong support for this model. Notably, the evidence remains robust across the three data scenarios, demonstrating the stability of Bayesian model selection even when missing values are present. For ML results, the LRT similarly favor model II in most comparisons, particularly under the before deletion and imputation scenarios. The median p-values for nested models are extremely small (e.g., 1.148×10^{-186} for model I vs model II before deletion), corroborating the decisive evidence observed in the Bayesian framework. For non-nested comparisons, Bayes factors based on the ML log-likelihoods were computed, and these results align closely with the Bayesian analysis, again highlighting model II as strongly preferred. Accordingly, conclusions are not reported for these non-nested comparisons on LRT p-values. Only for model V under complete case and imputation analysis does the LR test yield a p-value closer to conventional significance thresholds ($p = 0.053$) and above (.147), reflecting the reduced power due to the deletion of incomplete cases and imputation. Comparing these two mostly similar models (II vs V)

shows a more interpretable evidence for the data generating model. For both frameworks the evidence diminishes in the complete cases scenario. In the MAR case, across all scenarios – before deletion, complete cases, and multiple imputation – the Bayes factors indicate decisive evidence in favor of model II over all other specifications. Conversely, the decision following imputation in the maximum likelihood estimation does not support model II. Despite the Bayes factors calculated for the ML results demonstrating clear evidence in favor of model II, they are not as significant in terms of magnitude as the Bayesian results, i.e., a median p-value if .3170 as result of the comparison of model V against model II and a logarithmic Bayes factor of -13 in comparison to -25 from the Bayesian results. The information criteria, especially the AIC, are also closer together in the ML results. It is noteworthy that the ML values subsequent to imputation indicate decisions for model II that are equally as clear as those observed in the scenario prior to deletion. However, the evidence for model in the Bayesian results following imputation and data augmentation is weaker.

A direct comparison between ML and Bayesian model comparison revealed several consistent patterns: (1) ML under complete case analysis with MAR missingness produced higher bias, lower coverage, and reduced power in LRTs; (2) the Bayesian framework, particularly when combined with imputation, delivered more stable Bayes factors and parameter estimates; and (3) the BD scenario provided optimistic performance estimates for both frameworks, which are not representative when missingness is present. These findings suggest that, for the settings considered here, Bayesian inference with multiple imputation or data augmentation is more robust to the adverse effects of missingness, particularly under MAR, both in terms of parameter recovery and in maintaining discriminatory power in model comparisons. Finally, Bayesian estimation combined with imputation yields more stable model selection results across missing data mechanisms, whereas ML performance under complete case analysis degrades notably for MAR settings.

4.5.4 Empirical illustration - NEPS starting cohort 6

The National Educational Panel Study (NEPS) is a large-scale, longitudinal study designed to investigate educational trajectories and their determinants in Germany. It is conducted by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide research network (Blossfeld & Roßbach, 2019). The NEPS collects data across multiple dimensions, including educational, social, and economic factors, providing a rich dataset for multidisciplinary research. The Starting Cohort 6 (SC 6), also referred to as the adult cohort, focuses on individuals born between 1944 and 1986. These participants represent a broad age range and diverse life stages, making SC 6 an invaluable resource for studying transitions into and within the labor market, educational attainment, and their interrelation with family dynamics and regional economic conditions. The SC 6 cohort employs a longitudinal design, with repeated measures collected in multiple survey waves. Respondents provide detailed information about their educational and occupational pathways, including transitions between employment, unemployment, and training. Additionally, the study captures demographic, socioeconomic, and regional factors, enabling comprehensive analysis of individual trajectories within their broader societal and economic

contexts. Thus, the data of SC 6 offers unique opportunities to examine employment dynamics in Germany, particularly: the impact of human capital (e.g., education, competencies, and training) on employment probabilities, as well as longitudinal data facilitates the study of employment trajectories over time, including transitions between employment and unemployment. The data used in this study are from SUF version 15.0.0 (NEPS Network, 2024b).

The analyses use data from the National Educational Panel Study (NEPS), starting cohort 6 (SC 6), wave 15, comprising $n = 3{,}623$ interviewed persons. Table 4.18 summarizes the categorical covariates. The education classification (CASMIN) is well spread across levels (obs. shares: 1c/2a ≈ 0.41, 2c ≈ 0.22, 3a/3b ≈ 0.35), the sample is gender-balanced (female 0.501), and about 14.6% report a migration background. Unemployment duration is available in ordered categories (reference: <1 year) and exhibits substantial missingness in the observed data. The imputed distributions (ML only) indicate that short spells (<1 year) are most common on average across imputations. Parental unemployment (mother/father) is rare in the observed data, with small between–imputation variability. Table 4.19 reports the metric covariates. Years of education centers around 15 years (median $= 15$, $M \approx 14.75$). The competence scores (WLE for ICT literacy, mathematics, and reading) are approximately standardized around zero with standard deviations close to unity. The procedural metacognition (ICT) measure is a proportion–correct index. For each metric variable we display the observed summary as well as the average across imputations (ML only) together with the between–imputation standard deviation, which is uniformly small—indicating stable imputations. Missingness is handled via multiple imputation by chained equations (MICE) with a CART imputation model in case of the ML or via data augmentation for the Bayesian approach. For the ML specifications, estimates are fit in each imputed dataset and pooled using Rubin's rules. The imputed sample columns in tables 4.18 and 4.19 therefore pertain to the ML analysis. The Bayesian specifications are reported for the observed data and do not rely on the imputed summaries. The tables include the imputed columns for comparability and transparency of the ML workflow.

To examine individual employment probabilities, different (static) probit models are estimated with the above mentioned Bayesian framework as wells as likelihood-based approaches. These approaches are particularly suited for binary outcome variables, such as employment status, and incorporates random effects to account for unobserved heterogeneity across regions. The models leverage rich individual-level data from the NEPS Starting Cohort 6 (SC 6), which captures education, skills, employment history, and socioeconomic factors, alongside contextual information on regional labor markets. Each model has unique theoretical underpinnings and is designed to emphasize different mechanisms driving employment outcomes that could be used to explain individual employment probabilities.

Model 1 As a baseline, the null model is estimated that includes only an intercept and regional random effects. The null model provides a reference point for assessing the explanatory power of subsequent models with additional predictors. Specifically, it allows

us to evaluate the extent to which regional labor market heterogeneity alone accounts for variations in individual employment probabilities. The null model is specified as:

$$P(Y_i = 1) = \Phi(\beta_0 + a_j), \tag{4.51}$$

where Φ is the cumulative distribution function of the standard normal distribution, β_0 represents the overall average employment probability across all individuals, and a_j denotes random effects for regions, capturing unobserved heterogeneity in labor market conditions. The null model acts as a baseline for comparison with more complex models, enabling the evaluation of the incremental explanatory power of additional predictors.

Model 2 This model describes the human capital theory that suggests that education and skills increase productivity, leading to higher employability. Becker (1970) argues that investments in human capital (e.g., education, training) improve workers' productivity and labor market outcomes. Hence, model 2 is specified by variables describing the education status, and competencies. The education status is measured as years of schooling and highest qualification achieved. the competencies are assessed through literacy and numeracy scores derived from NEPS skill assessments based on the last competence testing in mathematical and reading competence, computer literacy as well as procedural metacognition (ICT).[15]

Model 3 Based on search and matching theory (Mortensen & Pissarides, 1994), model 3 incorporates labor market dynamics, emphasizing the impact of unemployment duration and job search behavior on employment probabilities. Key predictors include the unemployment History as the duration of prior unemployment spells which are categorized lass than one year, between one and two years, up to five years, and more than five years.

Model 4 Grounded in theories of labor market discrimination (Arrow, 1973; Phelps, 1972), this model focuses on the specific challenges faced by migrants. Additional to the individual education stati and the competencies the migration background measured by the generation status is picked up in model 4. A person with migration status is classified if born abroad (1st generation), born abroad but immigrated as children, i.e., less then 18 years old (1.5 generation), born in Germany but with at least one parent born abroad (2nd generation), or born in Germany with one parent born abroad and the other born in Germany (2.5 generation).

Model 5 Model 5 as full model contains all above-mentioned variables.

The ML results are presented in table 4.20 for the coefficients and p-values and the 95% confidence intervals. The Bayesian estimates are presented in table 4.21 for the coefficients and in the 95% credible intervals. The results of the model comparison is presented in table 4.22.

[15] All competence variables are based on corrected WLE.

Across specifications, the signs and magnitudes of the main covariates are broadly consistent between maximum–likelihood (ML) and Bayesian estimation, with the clearest and most stable patterns for unemployment history and ICT skills. Years of education enter with a small positive coefficient in all enriched specifications, but the 95% intervals often touch zero (ML: Models 2, 4, 5; Bayesian: Models 2–5), pointing to at best a modest association once other covariates are controlled. The CASMIN categories relative to the reference group (1a/1b/2b) are predominantly negative but imprecisely estimated; the wide intervals that straddle zero indicate that compositional differences captured elsewhere (skills, unemployment history) likely account for most of the variation. ICT literacy (WLE) is positively associated with employment. In the Bayesian fits the 95% credible intervals exclude zero in all enriched models, suggesting a robust effect; in the ML fits the point estimates are similar but intervals are borderline in some specifications. The procedural metacognition (ICT) measure shows a robust positive association in both frameworks (ML Models 2–4; Bayesian Models 2–5), with intervals comfortably above zero, indicating that better task strategies around ICT are linked to higher employment chances. By contrast, the WLE scores for mathematics and reading do not show clear independent associations once ICT variables are included; intervals generally cover zero in both ML and Bayesian results. Relative to spells shorter than one year, longer unemployment durations are clearly and consistently associated with lower employment probabilities. In both ML (Model 3; replicated in Model 5) and Bayesian (Models 3 and 5) results, the coefficients for 1–2 years, 2–5 years, and > 5 years are negative, and the corresponding intervals are predominantly below zero, with the largest reductions for very long spells. This pattern is monotone and precisely estimated in the Bayesian models and largely mirrored in the ML models, underscoring the substantive importance of unemployment duration. Migration background (yes) shows no robust association: intervals include zero in all specifications under both estimators. Parental unemployment is mostly imprecise; there is some evidence in the ML results for a small negative association for maternal unemployment (Model 5), whereas the Bayesian intervals remain wide and include zero. The female indicator is negative in sign in both frameworks, but intervals overlap zero in several specifications; any gender gap appears small once covariates are controlled. Intercepts are positive on the link scale, as expected given the overall employment rate. The random–effects variance is small in the ML fits and larger in the Bayesian fits; because the estimators rely on different likelihood approximations and (for the Bayesian models) weakly informative priors, the raw magnitudes are not directly comparable. Substantively, both frameworks account for unobserved heterogeneity at the individual level. Adding unemployment duration (Model 3) yields the largest improvement in fit (substantial drops in AIC/BIC and higher marginal likelihood), with Model 5 performing similarly once further controls are included. Specifications that add skills without unemployment history (Models 2 and 4) improve fit much less. This ranking aligns with the Bayes–factor and LRT evidence reported in the model–comparison table.

Pairwise log Bayes factors (upper triangles; computed from BIC via the Laplace approximation) and likelihood–ratio tests with median p-values across imputations (lower triangles) consistently rank the five random–effects specifications. Relative to model 1, the log Bayes

factors are +61.127 (vs. M2), +13.449 (vs. M3), +59.485 (vs. M4), and +27.960 (vs. M5), which constitutes strong to decisive evidence in favor of model 1 over all alternatives on the Kass–Raftery scale (Kass & Raftery, 1995). Comparisons among the remaining models indicate that model 3 is preferred to model 2 and model 4 ($|\ln \mathrm{BF}| \approx 47.678$ and 46.036, respectively) and is also favored over model 5 ($\ln \mathrm{BF} \approx 14.511$), while Model 5 is preferred to Model 4 ($\ln \mathrm{BF} \approx 31.525$). The implied overall ordering by fit is

$$\text{Model 1} \succ \text{Model 3} \succ \text{Model 5} \succ \text{Model 4} \approx \text{Model 2}.$$

Where likelihood ratio tests are applicable, the median p-values corroborate the Bayes–factor ranking. Enlarging model 1 to models 2–5 yields highly significant improvements (M2 vs. M1: $p = 1.05 \times 10^{-17}$, $df = 8$; M3 vs. M1: $p = 9.89 \times 10^{-47}$, $df = 11$; M4 vs. M1: $p = 3.84 \times 10^{-17}$, $df = 9$; M5 vs. M1: $p = 3.01 \times 10^{-45}$, $df = 15$). By contrast, comparisons such as M4 vs. M2 show no improvement (median $p = 0.975$, $df = 1$), and M5 vs. M3 is not significant at conventional levels (median $p = 0.098$, $df = 4$). For pairs that are not nested, LRTs are not interpretable; in those cases the (ML based) Bayes–factor evidence should be used.

Finally, pairwise log Bayes factors (upper triangle) decisively favor model 1 over all alternatives and further support model 3 over models 2, 4, and (more modestly) 5. Model 5 is preferred to model 4. Median LRT p-values (lower triangle; nested pairs only) agree with this ranking. In sum, evidence from both Bayes factors and nested LRTs decisively favors Model 1, followed by model 3, with most gains attributable to the inclusion of unemployment-duration terms—an ordering that is consistent across ML and Bayesian estimation.

4.6 Conclusion and outlook

Model selection remains a central challenge in both statistical inference and machine learning, demanding the careful identification of the most suitable model among a set of competing candidates As discussed in chapter 2, this task transcends mere variable selection, requiring clear research objectives, a precise understanding of the data generating process, and careful consideration of appropriate performance metrics. The analyses underscore that no single model uniformly dominates across all scenarios; rather, distinct models can provide complementary insights, each with specific strengths and limitations. Bayesian approaches, in particular, offer a coherent and flexible framework for model evaluation. Through marginal likelihoods and Bayes factors, Bayesian model selection allows for comparison across both nested and non-nested models, quantifies the strength of evidence, and maintains robustness in the presence of missing or incomplete data. By contrast, maximum likelihood based techniques, such as likelihood ratio tests, are confined to nested models and can be sensitive to case deletion or data sparsity, especially under MAR conditions.

Empirically, in the experimental study both Bayesian and frequentist approaches consistently identify model II as the most plausible data-generating specification under MCAR. However, Bayesian inference provides additional practical advantages: it enables stable and interpretable posterior summaries, facilitates probabilistic statements about parameters and

model plausibility, and naturally incorporates prior knowledge to improve estimation, particularly in sparse or small-group settings. From a computational perspective, the scalability and efficiency of Bayesian model selection can be enhanced through modern approaches such as importance sampling, variational bayes, or hybrid algorithms, enabling application to high-dimensional and complex hierarchical settings.

The data augmentation (DA) approach employed in this study to impute missing values within the Gibbs sampler provides a flexible framework for handling latent variables in hierarchical models. Beyond classical Bayesian estimation, this strategy can be naturally extended to a variety of modern inference techniques. In variational Bayes approaches, missing values or other latent variables can be incorporated directly into the variational optimization, allowing simultaneous estimation of model parameters and latent quantities while preserving uncertainty. Here, DA draws provide a principled initialization and update mechanism for the latent variational distributions (Chen, Liu, et al., 2024; Kingma & Welling, 2013). Overall, the DA framework implemented here offers a versatile foundation that can be extended to contemporary Bayesian computation methods, enabling more efficient, stable, and interpretable inference in models with latent structures or incomplete data. DA can be leveraged in posterior predictive checks or predictive cross-validation, providing a robust mechanism for evaluating model adequacy under uncertainty and avoiding biases associated with listwise deletion or simplistic imputations (Piironen & Vehtari, 2016).

Finally, the results demonstrate that Bayesian methods offer a theoretically grounded, robust, and practically relevant framework for model selection in the presence of missing data, providing stable, interpretable, and decisive evidence. These findings have broad implications for applied research, highlighting the potential of Bayesian approaches to deliver rigorous inference while addressing the inherent uncertainties of real-world data.

4.7 Appendix chapter 4

4.7.1 Figures

FIGURE 4.1: Illustration of median pooling rule for combining R p-values

Illustration of the exact finite-sample median pooling rule for combining R p-values obtained from repeated analyses (e.g., multiple imputation). Panel 1 shows the empirical distribution of simulated p-values under H_0 (*Uniform*[0,1]); Panel 2 displays the ordered p-values $p_{(1)}, \ldots, p_{(R)}$ with the median $p_{(r)}$ highlighted; Panel 3 depicts the corresponding Beta$(r,\ R-r+1)$ density with the cumulative probability up to the observed median p-value, yielding the exact pooled p-value.

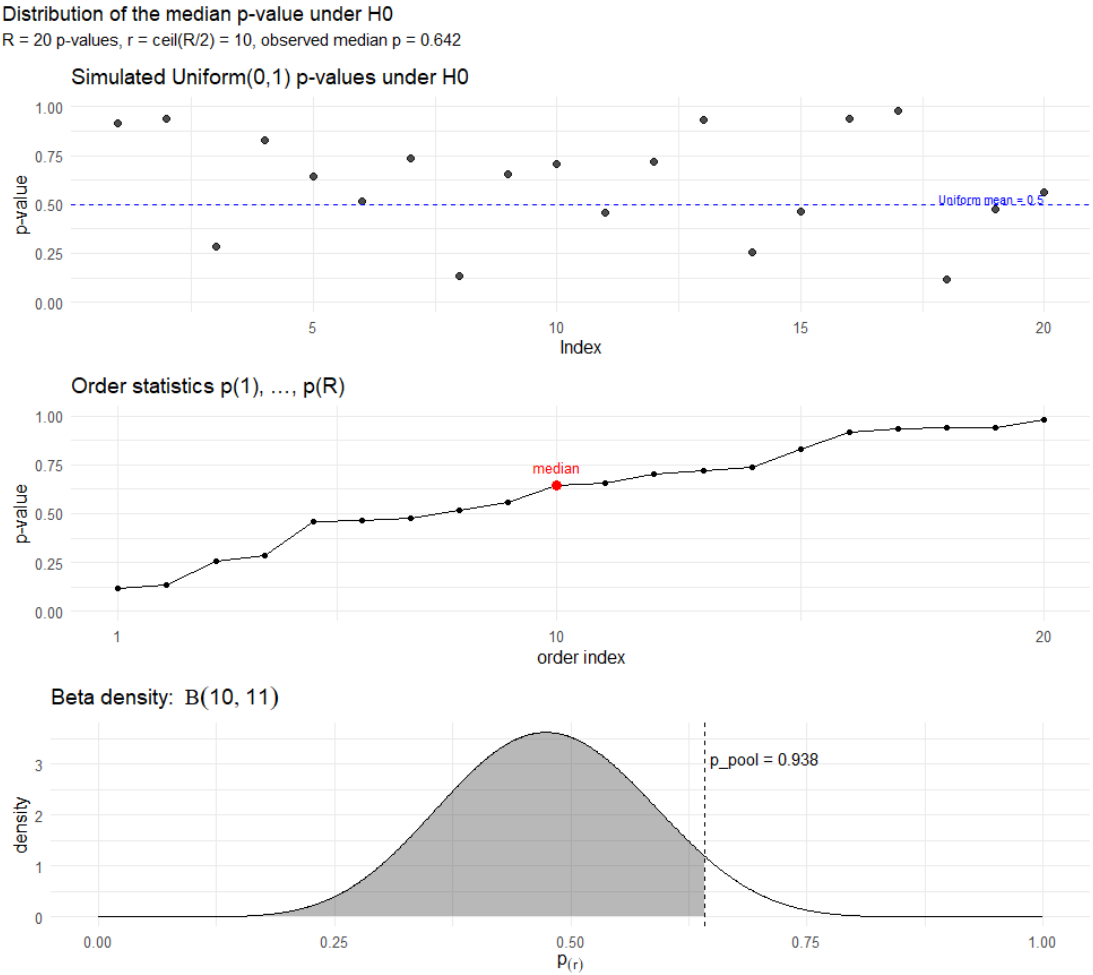


FIGURE 4.2: Autocorrelation functions of the estimated beta parameters

Autocorrelation functions (ACF) of the estimated beta parameters $\hat{\beta}_1, \ldots, \hat{\beta}_6$ from the Gibbs sampler: The plots display the autocorrelation at different lags for each parameter of the data-generating process. The initial burn-in iterations were discarded to stabilize the chains (first 10,000 of 40,000 iterations). High autocorrelation, particularly for the intercept β_1, indicates slower mixing of the chain.

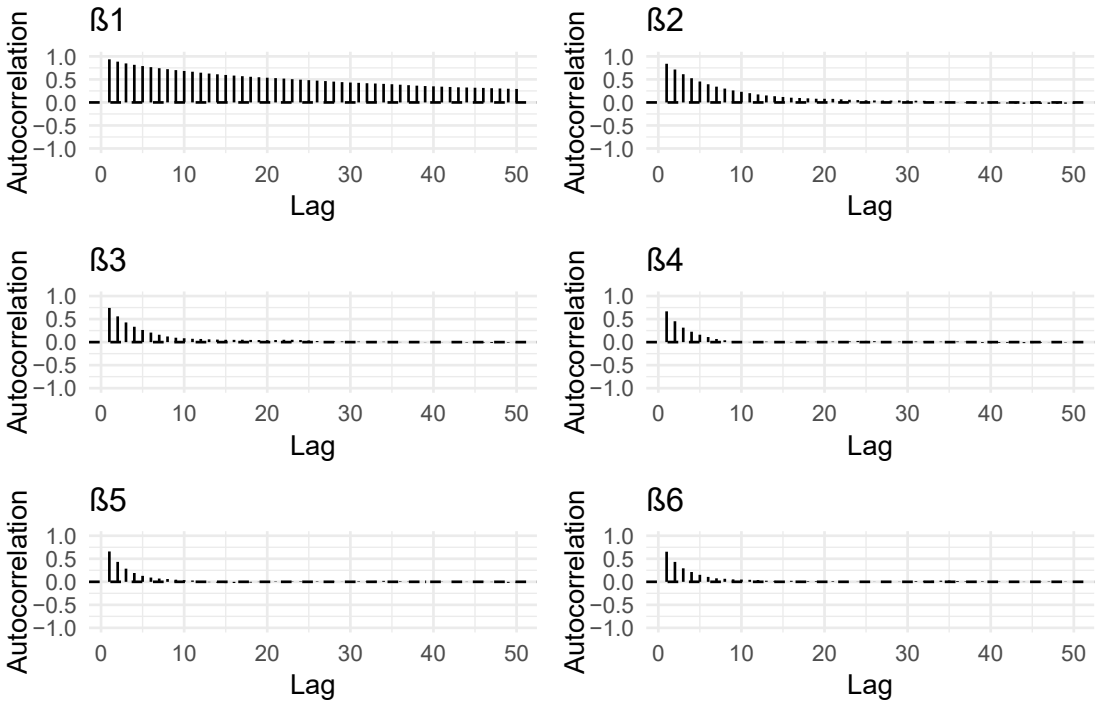
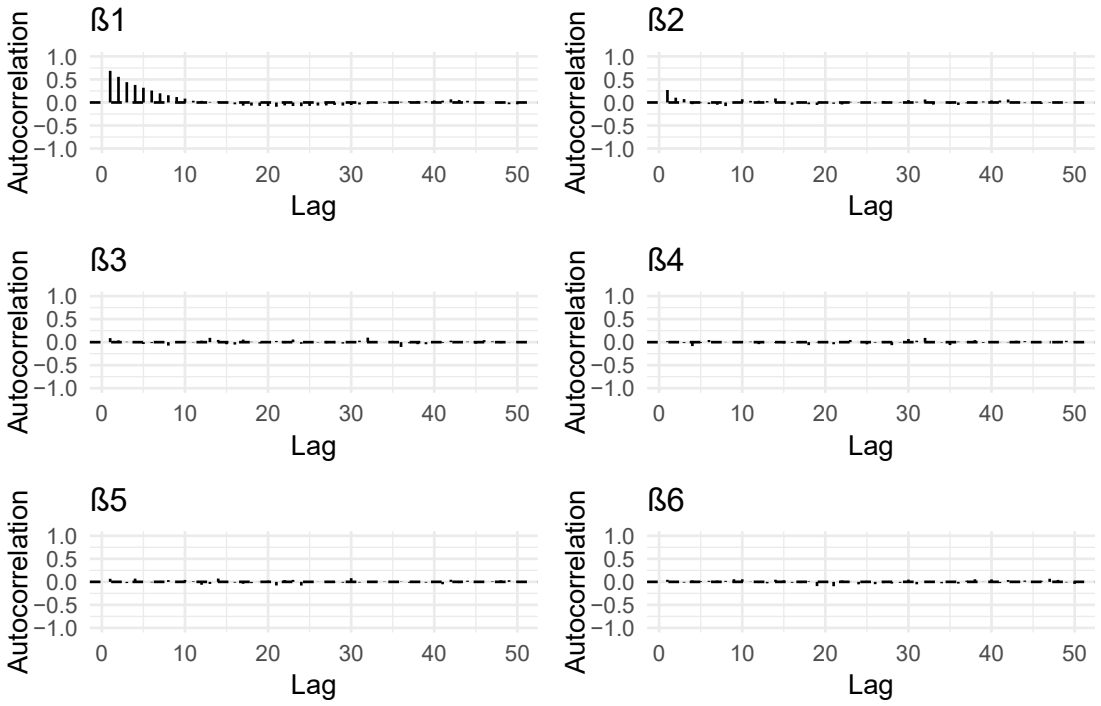


FIGURE 4.3: Autocorrelation functions of the estimated beta parameters after pruning

Autocorrelation functions (ACF) of the estimated beta parameters $\hat{\beta}_1, \ldots, \hat{\beta}_6$ from the Gibbs sampler: The plots display the autocorrelation at different lags for each parameter of the data-generating process. The initial burn-in iterations were discarded to stabilize the chains (first 10,000 of 40,000 iterations), and every 10th iteration was retained (thinning) to reduce autocorrelation. High autocorrelation, particularly for the intercept β_1, indicates slower mixing chains.



4.7.2 Tables

TABLE 4.1: Interpretation of Bayes factors

Interpretation of Bayes factors according to Jeffreys (1961) and Kass & Raftery (1995), shown in Bayes factor (BF) and natural logarithm (ln BF) scales. Negative values indicate evidence in favor of the competing model.

Jeffreys (1961)

BF range	ln BF range	Interpretation
< 1	< 0	Evidence for alternative model
1 – 3.2	0 – 1.15	Barely worth mentioning
3.2 – 10	1.15 – 2.30	Substantial evidence
10 – 32	2.30 – 3.47	Strong evidence
> 32	> 3.47	Decisive evidence

Kass & Raftery (1995)

BF range	ln BF range	Interpretation
< 1	< 0	Evidence for alternative model
1 – 2.7	0 – 1	Not worth more than a bare mention
2.7 – 20	1 – 3	Positive evidence
20 – 150	3 – 5	Strong evidence
> 150	> 5	Very strong evidence

TABLE 4.2: Julia and R: runtime comparison

The runtime results of implementing the Gibbs sampler are presented in Julia and R without and with missing value handling for the different models tested in the simulation study. All computations reported for $M = 100$ datasets were performed on the same notebook computer with 16GB of RAM and an i7 Intel processor, using the R programming language (R Core Team, 2020) and Julia (Bezanson et al., 2017).

	runtime Gibbs sampler			
	without missings		with handling missings	
Model	in Julia	in R	in Julia	in R
Model 1	2,484 sec	36,566 sec	4,776 sec	69,475 sec
Model 2	2,661 sec	32,431 sec	5,056 sec	61,618 sec
Model 3	2,835 sec	37,867 sec	5,389 sec	71,947 sec
Model 4	2,998 sec	38,423 sec	5,697 sec	73,003 sec
Model 5	3,222 sec	39,223 sec	6,270 sec	74,523 sec
Model 6	2,602 sec	33,041 sec	4,942 sec	62,777 sec

TABLE 4.3: Model specification for simulation study

Model specifications used for simulation-based model comparison. Each model includes an intercept and individual-specific random effects. Most models are nested within Model IV, except for Model VI, which deviates in structure. Covariates include both continuous (normally distributed) and categorical variables.

Variable	Model I	Model II	Model III	Model IV	Model V	Model VI	True values
X_1 (Intercept)	✓	✓	✓	✓	✓	✓	.20
X_2 (normal)		✓	✓		✓	✓	-1.50
X_3 (normal)		✓			✓		.80
X_4 (normal)		✓		✓	✓		-.30
X_5 (factor)		✓	✓	✓	✓	✓	-.40
X_6 (factor)		✓	✓	✓	✓	✓	-.30
X_7 (normal)				✓	✓	✓	not included
X_8 (factor)				✓	✓		not included
X_9 (factor)				✓	✓		not included
X_{10} (normal)					✓	✓	not included
Variance of a_i	✓	✓	✓	✓	✓	✓	1.00

Notes: X_2, X_3, X_4, X_7, and X_{10} are continuous covariates drawn from normal distributions with differing variances. X_5, and X_6 represent dummy-coded levels B and C of a categorical variable with reference level A. X_8, and X_9 are dummy-coded levels E and F from a second categorical variable with reference D.

TABLE 4.4: Overview of the missing designs of the experimental studies

The experimental study (Ex) is split in missing completely at random (MCAR) and missing at random (MAR). The average missing rate is calculate over the $M = 100$ datasets. In accordance with the MAR model, the missing rate is established at 0.35 for X_2 and 0.20 for X_3. The complete case rate is presented as an average over the $M = 100$ datasets.

Design	Missing description	Complete cases rate in (%)	Results presented
MCAR	$Pr(X_{2,i} = missing) = .20$ $Pr(X_{3,i} = missing) = .30$ $Pr(X_{5,i} = missing) = .10$ $Pr(X_{6,i} = missing) = .15$ $Pr(X_{7,i} = missing) = .20$ $Pr(X_{10,i} = missing) = 0.30$	24.04[1]	Tables 4.5, 4.6, 4.7 (Bayesian) 4.8, 4.9, 4.10 (ML)
MAR	$X_{3,i} = missing$ if $1/(1 + \exp(-\omega_{3,i}))$ $w_{3,i} = .1X_{3,i} + \rho_i$ and $\rho_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ $X_{5,i} = missing$ if $1/(1 + \exp(-\omega_{5,i}))$ $w_{5,i} = .15X_{5,i} + \rho_i$ and $\rho_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ $X_{6,i} = missing$ if $1/(1 + \exp(-\omega_{6,i}))$ $w_{6,i} = .25X_{6,i} + \rho_i$ and $\rho_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ $X_{7,i} = missing$ if $1/(1 + \exp(-\omega_{7,i}))$ $w_{7,i} = .20X_{7,i} + \rho_i$ and $\rho_i \overset{\text{i.i.d.}}{\sim} N(0,1)$	45.75[1]	Tables[2] 4.12, 4.13 (Bayesian) 4.14, 4.15 (ML)

Notes:
[1] Average over $M = 100$ datasets.
[2] The before deletion tables for MAR are not printed because of the same underlying data generating process, so for results see tables 4.5 and 4.8.

TABLE 4.5: Experimental study (MCAR) Bayesian results: Before deletion

Posterior summary statistics for Bayesian data augmentation estimation for before deletion. Across $D = 100$ simulated datasets, each with $N = 2,000$ individuals and $J = 50$ groups the reported values are posterior means, their standard deviations (SD), root mean squared errors, and empirical 95% credible interval coverage rates (averaged across replications). For each model information criteria are calculated and random effects are estimated. The data generating parameters are $\beta_1, \ldots, \beta_6$.

Variables		Model I	Model II	Model III	Model IV	Model V	Model VI
β_1 (Intercept)	Mean	-.029	.198	.161	.126	.194	.153
	SD	(.089)	(.158)	(.139)	(.113)	(.156)	(.137)
	Bias	.226	.002	.039	.074	.006	.047
	RMSE	.246	.158	.144	.135	.156	.145
	Coverage	.400	.920	.960	.990	.990	.920
β_2 (normal)	Mean	–	-1.510	-.906	–	-1.510	-.912
	SD	–	(.065)	(.045)	–	(.070)	(.044)
	Bias	–	.007	.594	–	.014	.588
	RMSE	–	.065	.595	–	.072	.590
	Coverage	–	.920	.000	–	.900	.000
β_3 (normal)	Mean	–	.814	–	–	.816	–
	SD	–	(.053)	–	–	(.053)	–
	Bias	–	.014	–	–	.016	–
	RMSE	–	.054	–	–	.056	–
	Coverage	–	.990	–	–	.990	–
β_4 (normal)	Mean	–	-.318	–	.067	-.318	–
	SD	–	(.042)	–	(.031)	(.041)	–
	Bias	–	.018	–	.367	.018	–
	RMSE	–	.046	–	.368	.045	–
	Coverage	–	.850	–	.000	.820	–
β_5 (factor)	Mean	–	-.350	-.296	-.234	-.357	-.299
	SD	–	(.093)	(.085)	(.076)	(.091)	(.085)
	Bias	–	.050	.104	.166	.043	.101
	RMSE	–	.106	.134	.183	.101	.132
	Coverage	–	.990	.820	.210	.980	.910
β_6 (factor)	Mean	–	-.323	-.280	-.220	-.324	-.281
	SD	–	(.094)	(.086)	(.075)	(.092)	(.085)
	Bias	–	.023	.002	.080	.024	.101
	RMSE	–	.096	.088	.110	.095	.132
	Coverage	–	.980	.950	.980	.990	.990
β_7 (normal)	Mean	–	–	–	.004	.012	.004
	SD	–	–	–	(.030)	(.038)	(.034)
	Bias	–	–	–	.004	.012	.004
	RMSE	–	–	–	.031	.040	.035
	Coverage	–	–	–	.990	.990	.990
β_8 (factor)	Mean	–	–	–	-.031	-.043	–
	SD	–	–	–	(.074)	(.093)	–
	Bias	–	–	–	.031	.043	–
	RMSE	–	–	–	.080	.103	–
	Coverage	–	–	–	.990	.980	–
β_9 (factor)	Mean	–	–	–	.007	.010	–
	SD	–	–	–	(.074)	(.092)	–
	Bias	–	–	–	.007	.010	–
	RMSE	–	–	–	.074	.093	–
	Coverage	–	–	–	.990	.990	–
β_{10} (normal)	Mean	–	–	–	–	.017	.014
	SD	–	–	–	–	(.037)	(.034)
	Bias	–	–	–	–	.017	.014
	RMSE	–	–	–	–	.041	.037
	Coverage	–	–	–	–	.990	.820
$Var(a_j)$	Mean	.665	1.060	.884	.673	1.066	.889
Log marg. likelihood	Mean	-1299.342	-925.453	-1025.893	-1331.072	-914.992	-1038.101
	SD	(27.755)	(23.862)	(23.605)	(28.188)	(23.897)	(22.757)
BIC	Mean	2,606.285	1,896.512	2,082.189	2,715.351	1,905.992	2,121.807

TABLE 4.6: Experimental study (MCAR) Bayesian results: Complete cases

Posterior summary statistics for Bayesian data augmentation estimation for complete cases. Across $D = 100$ simulated datasets, each with $N = 2,000$ individuals and $J = 50$ groups the reported values are posterior means, their standard deviations (SD), root mean squared errors, and empirical 95% credible interval coverage rates (averaged across replications). For each model information criteria are calculated and random effects are estimated. The data generating parameters are $\beta_1, \ldots, \beta_6$.

Variables		Model I	Model II	Model III	Model IV	Model V	Model VI
β_1 (Intercept)	Mean	.024	.199	.184	.178	.210	.178
	SD	(.112)	(.202)	(.169)	(.174)	(.240)	(.172)
	Bias	.176	.001	.016	.022	.010	.022
	RMSE	.208	.202	.170	.176	.240	.145
	Coverage	.610	.970	.910	.990	.990	.920
β_2 (normal)	Mean	–	-1.56	-.916	–	-1.600	-.925
	SD	–	(.154)	(.097)	–	(.157)	(.097)
	Bias	–	.061	.584	–	.098	.575
	RMSE	–	.166	.592	–	.185	.583
	Coverage	–	.930	.000	–	.950	.000
β_3 (normal)	Mean	–	.861	–	–	.889	–
	SD	–	(.119)	–	–	(.121)	–
	Bias	–	.061	–	–	.089	–
	RMSE	–	.133	–	–	.150	–
	Coverage	–	.990	–	–	.990	–
β_4 (normal)	Mean	–	-.310	–	.119	-.323	–
	SD	–	(.093)	–	(.065)	(.095)	–
	Bias	–	.010	–	.419	.023	–
	RMSE	–	.094	–	.424	.098	–
	Coverage	–	.910	–	.000	.950	–
β_5 (factor)	Mean	–	-.405	-.354	-.270	-.406	-.355
	SD	–	(.202)	(.180)	(.160)	(.207)	(.183)
	Bias	–	.005	.046	.130	.006	.045
	RMSE	–	.203	.186	.207	.207	.188
	Coverage	–	.990	.950	.910	.980	.930
β_6 (factor)	Mean	–	-.222	-.195	-.143	-.212	-.188
	SD	–	(.199)	(.176)	(.158)	(.200)	(.178)
	Bias	–	.078	.105	.157	.088	.112
	RMSE	–	.214	.205	.223	.219	.210
	Coverage	–	.980	.960	.980	.990	.990
β_7 (normal)	Mean	–	–	–	-.032	-.066	-.033
	SD	–	–	–	(.065)	(.084)	(.075)
	Bias	–	–	–	.032	.066	.033
	RMSE	–	–	–	.073	.107	.082
	Coverage	–	–	–	.980	.720	.920
β_8 (factor)	Mean	–	–	–	-.023	-.015	–
	SD	–	–	–	(.157)	(.203)	–
	Bias	–	–	–	.023	.015	–
	RMSE	–	–	–	.158	.204	–
	Coverage	–	–	–	.990	.980	–
β_9 (factor)	Mean	–	–	–	-.047	-.023	–
	SD	–	–	–	(.158)	(.205)	–
	Bias	–	–	–	.047	.023	–
	RMSE	–	–	–	.165	.206	–
	Coverage	–	–	–	.880	.990	–
β_{10} (normal)	Mean	–	–	–	–	-.009	-.008
	SD	–	–	–	–	(.084)	(.075)
	Bias	–	–	–	–	.009	.008
	RMSE	–	–	–	–	.084	.075
	Coverage	–	–	–	–	.990	.920
$Var(a_j)$	Mean	.647	1.008	.809	.676	1.040	.824
Log marg. likelihood	Mean	-363.563	-285.051	-313.032	-391.418	-302.563	-323.028
	SD	(6.999)	(14.288)	(10.794)	(6.075)	(15.684)	(11.207)
BIC	Mean	733.282	607.043	650.692	825.934	666.696	682.998

TABLE 4.7: Experimental study (MCAR) Bayesian results: Imputation

Posterior summary statistics for Bayesian data augmentation estimation for imputation. Across $D = 100$ simulated datasets, each with $N = 2,000$ individuals and $J = 50$ groups the reported values are posterior means, their standard deviations (SD), root mean squared errors, and empirical 95% credible interval coverage rates (averaged across replications). For each model information criteria are calculated and random effects are estimated. The data generating parameters are $\beta_1, \ldots, \beta_6$.

Variables		Model I	Model II	Model III	Model IV	Model V	Model VI
β_1 (Intercept)	Mean	.000	.246	.199	.135	.255	.005
	SD	(.097)	(.165)	(.143)	(.130)	(.182)	(.146)
	Bias	.200	.046	.001	.065	.055	.195
	RMSE	.223	.171	.143	.145	.190	.244
	Coverage	.490	.920	.960	.880	.940	.670
β_2 (normal)	Mean	–	-1.520	-.912	–	-1.530	-1.490
	SD	–	(.092)	(.061)	–	(.091)	(.090)
	Bias	–	.015	.588	–	.009	.009
	RMSE	–	.094	.591	–	.009	.009
	Coverage	–	.920	.000	–	.950	.930
β_3 (normal)	Mean	–	.814	–	–	.820	.800
	SD	–	(.072)	–	–	(.072)	(.070)
	Bias	–	.014	–	–	.020	.000
	RMSE	–	.074	–	–	.074	.070
	Coverage	–	.950	–	–	.960	.950
β_4 (normal)	Mean	–	-.303	–	-.250	-.304	-.298
	SD	–	(.057)	–	(.101)	(.057)	(.056)
	Bias	–	.003	–	.150	.004	.002
	RMSE	–	.058	–	.180	.057	.056
	Coverage	–	.930	–	.690	.920	.920
β_5 (factor)	Mean	–	-.411	-.335	-.250	-.304	–
	SD	–	(.126)	(.115)	(.101)	(.057)	–
	Bias	–	.011	.065	.150	.004	–
	RMSE	–	.127	.132	.180	.057	–
	Coverage	–	.890	.910	.690	.920	–
β_6 (factor)	Mean	–	-.305	-.254	-.181	-.308	–
	SD	–	(.126)	(.113)	(.100)	(.126)	–
	Bias	–	.005	.046	.119	.008	–
	RMSE	–	.126	.122	.156	.126	–
	Coverage	–	.920	.930	.770	.920	–
β_7 (normal)	Mean	–	–	–	-.002	.000	.000
	SD	–	–	–	(.041)	(.051)	(.050)
	Bias	–	–	–	.002	.000	.008
	RMSE	–	–	–	.041	.051	.051
	Coverage	–	–	–	.930	.960	.960
β_8 (factor)	Mean	–	–	–	.017	.005	–
	SD	–	–	–	(.101)	(.126)	–
	Bias	–	–	–	.017	.005	–
	RMSE	–	–	–	.102	.127	–
	Coverage	–	–	–	.970	.970	–
β_9 (factor)	Mean	–	–	–	.005	-.008	–
	SD	–	–	–	(.100)	(.126)	–
	Bias	–	–	–	.005	.008	–
	RMSE	–	–	–	.100	.126	–
	Coverage	–	–	–	.970	.970	–
β_{10} (normal)	Mean	–	–	–	–	.009	.008
	SD	–	–	–	–	(.051)	(.050)
	Bias	–	–	–	–	.009	.008
	RMSE	–	–	–	–	.052	.051
	Coverage	–	–	–	–	.950	.960
$Var(a_j)$	Mean	.647	1.020	.857	.659	1.029	1.006
Log marg. likelihood	Mean	-1299.342	-1128.766	-1197.652	-1443.069	-1170.888	-1220.467
	SD	(27.755)	(62.455)	(46.421)	(18.265)	(64.187)	(52.964)
BIC	Mean	2606.285	2303.138	2425.708	2939.344	2417.785	2486.540

TABLE 4.8: Experimental study (MCAR) ML results: Before deletion

Posterior summary statistics for Maximum Likelihood Estimation for before deletion. Across $D = 100$ simulated datasets, each with $N = 2,000$ individuals and $J = 50$ groups the reported values are posterior means, their standard deviations (SD), root mean squared errors, and empirical 95% credible interval coverage rates (averaged across replications). For each model information criteria are calculated and random effects are estimated. The data generating parameters are $\beta_1, \ldots, \beta_6$.

Variables		Model I	Model II	Model III	Model IV	Model V	Model VI
β_1 (Intercept)	Mean	-.034	.171	.140	.104	.177	.140
	SD	(.092)	(.152)	(.128)	(.109)	(.160)	(.128)
	Bias	.234	.029	.060	.096	.023	.060
	RMSE	.251	.154	.141	.145	.161	.141
	Coverage	.270	.960	.930	.860	.970	.940
β_2 (normal)	Mean	–	-1.503	-.915	–	-1.508	-.916
	SD	–	(.068)	(.047)	–	(.068)	(.047)
	Bias	–	.003	.585	–	.008	.584
	RMSE	–	.068	.587	–	.068	.586
	Coverage	–	.950	.000	–	.950	.000
β_3 (normal)	Mean	–	.798	–	–	.801	–
	SD	–	(.052)	–	–	(.053)	–
	Bias	–	.002	–	–	.001	–
	RMSE	–	.052	–	–	.052	–
	Coverage	–	.940	–	–	.940	–
β_4 (normal)	Mean	–	-.300	–	.076	-.301	–
	SD	–	(.044)	–	(.030)	(.044)	–
	Bias	–	.000	–	.376	.001	–
	RMSE	–	.043	–	.377	.044	–
	Coverage	–	.950	–	.000	.950	–
β_5 (factor)	Mean	–	-.397	-.333	-.238	-.397	-.332
	SD	–	(.083)	(.077)	(.070)	(.084)	(.077)
	Bias	–	.003	.067	.162	.003	.068
	RMSE	–	.083	.102	.177	.083	.102
	Coverage	–	.970	.870	.410	.970	.890
β_6 (factor)	Mean	–	-.278	-.225	-.164	-.279	-.225
	SD	–	(.099)	(.090)	(.082)	(.100)	(.090)
	Bias	–	.022	.075	.136	.021	.075
	RMSE	–	.101	.117	.158	.102	.117
	Coverage	–	.890	.830	.510	.900	.830
β_7 (normal)	Mean	–	–	–	.001	.004	.004
	SD	–	–	–	(.031)	(.041)	(.034)
	Bias	–	–	–	.001	.004	.004
	RMSE	–	–	–	.031	.041	.035
	Coverage	–	–	–	.950	.920	.950
β_8 (factor)	Mean	–	–	–	-.006	-.010	–
	SD	–	–	–	(.077)	(.098)	–
	Bias	–	–	–	.006	.010	–
	RMSE	–	–	–	.077	.098	–
	Coverage	–	–	–	.940	.950	–
β_9 (factor)	Mean	–	–	–	-.008	-.007	–
	SD	–	–	–	(.076)	(.093)	–
	Bias	–	–	–	.008	.007	–
	RMSE	–	–	–	.076	.093	–
	Coverage	–	–	–	.980	.940	–
β_{10} (normal)	Mean	–	–	–	–	-.002	-.002
	SD	–	–	–	–	(.036)	(.033)
	Bias	–	–	–	–	.002	.002
	RMSE	–	–	–	–	.036	.033
	Coverage	–	–	–	–	.940	.940
$Var(a_j)$	Mean	.368	.990	.681	.376	.996	.683
Log-Likelihood	Mean	-1250.533	-813.552	-954.348	-1239.182	-811.511	-953.398
	SD	(34.446)	(30.717)	(31.078)	(34.621)	(30.603)	(30.988)
AIC	Mean	2505.066	1641.103	1918.696	2494.363	1645.022	1920.795
BIC	Mean	2516.268	1680.309	1946.701	2539.171	1706.634	1960.001

TABLE 4.9: Experimental study (MCAR) ML results: Complete cases

Posterior summary statistics for Maximum Likelihood Estimation for complete cases. Across $D = 100$ simulated datasets, each with $N = 2,000$ individuals and $J = 50$ groups the reported values are posterior means, their standard deviations (SD), root mean squared errors, and empirical 95% credible interval coverage rates (averaged across replications). For each model information criteria are calculated and random effects are estimated. The data generating parameters are $\beta_1, \ldots, \beta_6$.

Variables		Model I	Model II	Model III	Model IV	Model V	Model VI
β_1 (Intercept)	Mean	-.039	.173	.137	.108	.182	.136
	SD	(.112)	(.204)	(.175)	(.168)	(.235)	(.176)
	Bias	.239	.027	.063	.092	.018	.064
	RMSE	.264	.205	.186	.191	.235	.186
	Coverage	.380	.960	.940	.880	.940	.950
β_2 (normal)	Mean	–	-1.541	-.936	–	-1.559	-.941
	SD	–	(.159)	(.094)	–	(.165)	(.095)
	Bias	–	.041	.564	–	.059	.559
	RMSE	–	.164	.572	–	.174	.567
	Coverage	–	.970	.000	–	.930	.950
β_3 (normal)	Mean	–	.810	–	–	.820	–
	SD	–	(.112)	–	–	(.115)	–
	Bias	–	.010	–	–	.020	–
	RMSE	–	.112	–	–	.116	–
	Coverage	–	.950	–	–	.910	–
β_4 (normal)	Mean	–	-.298	–	.087	-.302	–
	SD	–	(.105)	–	(.067)	(.109)	–
	Bias	–	.002	–	.387	.002	–
	RMSE	–	.105	–	.393	.108	–
	Coverage	–	.920	–	.000	.890	–
β_5 (factor)	Mean	–	-.422	-.351	-.350	-.422	-.350
	SD	–	(.197)	(.181)	(.153)	(.199)	(.182)
	Bias	–	.022	.049	.150	.022	.050
	RMSE	–	.197	.187	.213	.199	.188
	Coverage	–	.930	.950	.820	.900	.950
β_6 (factor)	Mean	–	-.282	-.221	-.165	-.281	-.219
	SD	–	(.201)	(.192)	(.169)	(.204)	(.192)
	Bias	–	.018	.079	.135	.019	.081
	RMSE	–	.201	.206	.216	.203	.208
	Coverage	–	.950	.920	.810	.920	.920
β_7 (normal)	Mean	–	–	–	.008	.013	-.017
	SD	–	–	–	(.062)	(.084)	(.074)
	Bias	–	–	–	.008	.013	.017
	RMSE	–	–	–	.062	.084	.075
	Coverage	–	–	–	.960	.900	.950
β_8 (factor)	Mean	–	–	–	-.015	-.025	–
	SD	–	–	–	(.169)	(.204)	–
	Bias	–	–	–	.015	.025	–
	RMSE	–	–	–	.169	.205	–
	Coverage	–	–	–	.930	.920	–
β_9 (factor)	Mean	–	–	–	-.013	-.005	–
	SD	–	–	–	(.168)	(.198)	–
	Bias	–	–	–	.013	.005	–
	RMSE	–	–	–	.167	.197	–
	Coverage	–	–	–	.910	.940	–
β_{10} (normal)	Mean	–	–	–	–	-.004	-.009
	SD	–	–	–	–	(.083)	(.073)
	Bias	–	–	–	–	.004	.009
	RMSE	–	–	–	–	.082	.073
	Coverage	–	–	–	–	.940	.980
$Var(a_j)$	Mean	.347	1.001	.677	.367	1.025	.683
Log-Likelihood	Mean	-314.180	-213.730	-245.410	-308.934	-211.858	-244.442
	SD	(12.409)	(13.061)	(12.942)	(12.636)	(13.184)	(12.999)
AIC	Mean	632.360	441.459	500.821	633.869	445.716	502.885
BIC	Mean	640.710	470.684	521.696	667.369	491.641	532.111

TABLE 4.10: Experimental study (MCAR) ML results: Imputation

Posterior summary statistics for Maximum Likelihood Estimation for imputation. Across $D = 100$ simulated datasets, each with $N = 2,000$ individuals and $J = 50$ groups the reported values are posterior means, their standard deviations (SD), root mean squared errors, and empirical 95% credible interval coverage rates (averaged across replications). For each model information criteria are calculated and random effects are estimated. The data generating parameters are $\beta_1, \ldots, \beta_6$.

Variables		Model I	Model II	Model III	Model IV	Model V	Model VI
β_1 (Intercept)	Mean	-.034	.166	.138	.106	.176	.139
	SD	(.092)	(.160)	(.132)	(.113)	(.167)	(.133)
	Bias	.234	.033	.061	.094	.026	.061
	RMSE	.251	.162	.145	.147	.169	.146
	Coverage	.270	.970	.940	.890	.960	.930
β_2 (normal)	Mean	–	-1.450	-.910	–	-1.459	-.912
	SD	–	(.085)	(.051)	–	(.086)	(.051)
	Bias	–	.049	.590	–	.041	.588
	RMSE	–	.098	.592	–	.100	.590
	Coverage	–	.890	.000	–	.910	.000
β_3 (normal)	Mean	–	.767	–	–	.772	–
	SD	–	(.068)	–	–	(.069)	–
	Bias	–	.033	–	–	.279	–
	RMSE	–	.075	–	–	.074	–
	Coverage	–	.900	–	–	.910	–
β_4 (normal)	Mean	–	-.287	–	.076	-.289	–
	SD	–	(.049)	–	(.030)	(.049)	–
	Bias	–	.013	–	.376	.011	–
	RMSE	–	.050	–	.377	.050	–
	Coverage	–	.910	–	.000	.920	–
β_5 (factor)	Mean	–	-.386	-.333	-.242	-.386	-.334
	SD	–	(.097)	(.090)	(.080)	(.098)	(.090)
	Bias	–	.014	.066	.158	.014	.066
	RMSE	–	.098	.111	.177	.098	.111
	Coverage	–	.990	.910	.500	.990	.910
β_6 (factor)	Mean	–	-.263	-.220	-.165	-.264	-.219
	SD	–	(.118)	(.102)	(.090)	(.120)	(.103)
	Bias	–	.037	.080	.134	.035	.081
	RMSE	–	.123	.129	.161	.124	.130
	Coverage	–	.910	.820	.610	.920	.830
β_7 (normal)	Mean	–	–	–	.001	.002	.000
	SD	–	–	–	(.032)	(.045)	(.050)
	Bias	–	–	–	.001	.002	.008
	RMSE	–	–	–	.032	.045	.051
	Coverage	–	–	–	.970	.950	.960
β_8 (factor)	Mean	–	–	–	-.006	-.010	–
	SD	–	–	–	(.077)	(.106)	–
	Bias	–	–	–	.007	.010	–
	RMSE	–	–	–	.077	.106	–
	Coverage	–	–	–	.930	.960	–
β_9 (factor)	Mean	–	–	–	-.007	-.009	–
	SD	–	–	–	(.076)	(.104)	–
	Bias	–	–	–	.007	.009	–
	RMSE	–	–	–	.076	.104	–
	Coverage	–	–	–	.950	.920	–
β_{10} (normal)	Mean	–	–	–	–	-.001	-.003
	SD	–	–	–	–	(.047)	(.041)
	Bias	–	–	–	–	.001	.003
	RMSE	–	–	–	–	.047	.041
	Coverage	–	–	–	–	.960	.970
$Var(a_j)$	Mean	.368	.945	.677	.377	.954	.680
Log-Likelihood	Mean	-1,250.533	-820.150	-954.380	-1238.557	-816.7484	-952.684
	SD	(34.446)	(34.649)	(31.552)	(34.846)	(34.820)	(31.549)
AIC	Mean	2505.066	1654.301	1918.761	2493.115	1655.497	1919.369
BIC	Mean	2516.268	1693.507	1946.765	2537.922	1717.107	1958.576

TABLE 4.11: Experimental study (MCAR): Model comparison

Bayesian results are reported as $\ln \mathrm{BF}$; ML results as LRT median p-values and $\Delta\mathrm{BIC}_{\mathrm{II}-k}$. Based on $D = 100$ simulated datasets.

Evidence for	Before deletion	Complete cases	Imputation
Bayesian results			
	lnBF	lnBF	lnBF
Model I vs II	Model II -409	Model II -214	Model II -170
Model III vs II	Model II -135	Model II -70	Model II -68
Model IV vs II	Model II -440	Model II -245	Model II -314
Model V vs II	Model II -25	Model II -22	Model II -42
Model VI vs II	Model II -148	Model II -1.35	Model II -92
Maximum Likelihood results			
	LRT p-value	LRT p-value	LRT p-value
Model I vs II	Model II $1.148e-186$	Model II $1.822e-41$	Model II $8.234e-184$
Model III vs II	Model II $7.123e-62$	Model II $1.742e-14$	Model II $5.067e-59$
Model IV vs II†	Model – $2.318e-18$	Model – $2.588e-43$	Model – $3.525e-90$
Model V vs II	Model II .043	not model II .053	not model II .147
Model VI vs II†	Model – 0	Model – 0	Model – 0
Maximum Likelihood results (information criterion)			
	ΔBIC (II$-k$)	ΔBIC (II$-k$)	ΔBIC (II$-k$)
Model I vs II	Model II -418	Model II -85	Model II -411
Model III vs II	Model II -133	Model II -26	Model II -126
Model IV vs II	Model II -429	Model II -98	Model II -422
Model V vs II	Model II -13	Model II -10	Model II -12
Model VI vs II	Model II -140	Model II -31	Model II -133

Note (Bayesian results): Bayes factor interpretation: following Kass and Raftery (1995) with $0 < \ln\mathrm{BF} < 0.5$ *Not worth more than a bare mention*, $0.5 < \ln\mathrm{BF} < 1.0$ *substantial evidence*, $1.0 < \ln\mathrm{BF} < 2.0$ *strong evidence*, $\ln\mathrm{BF} > 2.0$ *decisive*. A negative sign indicates evidence for the second model.

Remark to †: Models IV and VI (same parameter dimension) are not nested within model II. For completeness, the LRT p-values are still reported which are obtained by applying the usual χ^2 reference distribution (as if the models were nested). However, because the nesting condition is violated, the LR statistic does not follow the standard asymptotic χ^2 distribution; consequently, these p-values are not valid and should be interpreted descriptively only (i.e., not as evidence for or against a model in an inferential sense).

Note (ML, ΔBIC): We report $\Delta\mathrm{BIC}_{\mathrm{II}-k} = \mathrm{BIC}_{\mathrm{II}} - \mathrm{BIC}_k$ for the comparison "Model k vs II". Negative values indicate a lower BIC for Model II (preference for Model II). Under standard regularity conditions, this is approximately related to Bayes factors via $\ln \mathrm{BF}_{k,\mathrm{II}} \approx \frac{1}{2}\Delta\mathrm{BIC}_{\mathrm{II}-k}$, so that negative values correspond to evidence in favour of Model II, consistent with the sign convention used for $\ln \mathrm{BF}$.

TABLE 4.12: Experimental study (MAR) Bayesian results: Complete cases

Posterior summary statistics for Bayesian data augmentation estimation for complete cases. Across $D = 100$ simulated datasets, each with $N = 2,000$ individuals and $J = 50$ groups the reported values are posterior means, their standard deviations (SD), root mean squared errors, and empirical 95% credible interval coverage rates (averaged across replications). For each model information criteria are calculated and random effects are estimated. The data generating parameters are $\beta_1, \ldots, \beta_6$.

Variables		Model I	Model II	Model III	Model IV	Model V	Model VI
β_1 (Intercept)	Mean	.024	.199	.184	.178	.210	.178
	SD	(.112)	(.202)	(.169)	(.174)	(.240)	(.172)
	Bias	.176	.001	.016	.022	.010	.022
	RMSE	.208	.202	.170	.176	.240	.145
	Coverage	.610	.970	.910	.990	.990	.920
β_2 (normal)	Mean	–	-1.56	-.916	–	-1.600	-.925
	SD	–	(.154)	(.097)	–	(.157)	(.097)
	Bias	–	.061	.584	–	.098	.575
	RMSE	–	.166	.592	–	.185	.583
	Coverage	–	.930	.000	–	.950	.000
β_3 (normal)	Mean	–	.861	–	–	.889	–
	SD	–	(.119)	–	–	(.121)	–
	Bias	–	.061	–	–	.089	–
	RMSE	–	.133	–	–	.150	–
	Coverage	–	.990	–	–	.990	–
β_4 (normal)	Mean	–	-.310	–	.119	-.323	–
	SD	–	(.093)	–	(.065)	(.095)	–
	Bias	–	.010	–	.419	.023	–
	RMSE	–	.094	–	.424	.098	–
	Coverage	–	.910	–	.000	.950	–
β_5 (factor)	Mean	–	-.405	-.354	-.270	-.406	-.355
	SD	–	(.202)	(.180)	(.160)	(.207)	(.183)
	Bias	–	.005	.046	.130	.006	.045
	RMSE	–	.203	.186	.207	.207	.188
	Coverage	–	.990	.950	.910	.980	.930
β_6 (factor)	Mean	–	-.222	-.195	-.143	-.212	-.188
	SD	–	(.199)	(.176)	(.158)	(.200)	(.178)
	Bias	–	.078	.105	.157	.088	.112
	RMSE	–	.214	.205	.223	.219	.210
	Coverage	–	.980	.960	.980	.990	.990
β_7 (normal)	Mean	–	–	–	-.032	-.066	-.033
	SD	–	–	–	(.065)	(.084)	(.075)
	Bias	–	–	–	.032	.066	.033
	RMSE	–	–	–	.073	.107	.082
	Coverage	–	–	–	.980	.720	.920
β_8 (factor)	Mean	–	–	–	-.023	-.015	–
	SD	–	–	–	(.157)	(.203)	–
	Bias	–	–	–	.023	.015	–
	RMSE	–	–	–	.158	.204	–
	Coverage	–	–	–	.990	.980	–
β_9 (factor)	Mean	–	–	–	-.047	-.023	–
	SD	–	–	–	(.158)	(.205)	–
	Bias	–	–	–	.047	.023	–
	RMSE	–	–	–	.165	.206	–
	Coverage	–	–	–	.880	.990	–
β_{10} (normal)	Mean	–	–	–	–	-.009	-.008
	SD	–	–	–	–	(.084)	(.075)
	Bias	–	–	–	–	.009	.008
	RMSE	–	–	–	–	.084	.075
	Coverage	–	–	–	–	.990	.920
$Var(a_j)$	Mean	.647	1.008	.809	.676	1.040	.824
Log marg. likelihood	Mean	-363.563	-285.051	-313.032	-391.418	-302.563	-323.028
	SD	(6.999)	(14.288)	(10.794)	(6.075)	(15.684)	(11.207)
BIC	Mean	733.282	607.043	650.692	825.934	666.696	682.998

TABLE 4.13: Experimental study (MAR) Bayesian results: Imputation

Posterior summary statistics for Bayesian data augmentation estimation for imputation. Across $D = 100$ simulated datasets, each with $N = 2,000$ individuals and $J = 50$ groups the reported values are posterior means, their standard deviations (SD), root mean squared errors, and empirical 95% credible interval coverage rates (averaged across replications). For each model information criteria are calculated and random effects are estimated. The data generating parameters are $\beta_1, \ldots, \beta_6$.

Variables		Model I	Model II	Model III	Model IV	Model V	Model VI
β_1 (Intercept)	Mean	.000	.068	.080	.056	.064	.078
	SD	(.097)	(.142)	(.127)	(.117)	(.151)	(.082)
	Bias	.200	.132	.120	.144	.136	.122
	RMSE	.223	.194	.175	.186	.203	.174
	Coverage	.490	.930	.890	.730	.860	.730
β_2 (normal)	Mean	–	-1.440	-.896	–	-1.440	-.894
	SD	–	(.063)	(.044)	–	(.065)	(.045)
	Bias	–	.061	.604	–	.058	.606
	RMSE	–	.088	.606	–	.087	.607
	Coverage	–	.910	.000	–	.890	.000
β_3 (normal)	Mean	–	.764	–	–	.766	-.229
	SD	–	(.050)	–	–	(.052)	(.082)
	Bias	–	.036	–	–	.034	.171
	RMSE	–	.062	–	–	.062	.190
	Coverage	–	.940	–	–	.950	.510
β_4 (normal)	Mean	–	-.371	–	.093	-.272	-.171
	SD	–	(.042)	–	(.090)	(.042)	(.083)
	Bias	–	.029	–	.393	.028	.129
	RMSE	–	.051	–	.394	.050	.154
	Coverage	–	.890	–	.000	.920	.730
β_5 (factor)	Mean	–	-.266	-.229	-.152	-.261	–
	SD	–	(.090)	(.081)	(.072)	(.089)	–
	Bias	–	.134	.171	.248	.139	–
	RMSE	–	.162	.190	.259	.165	–
	Coverage	–	.870	.560	.210	.750	–
β_6 (factor)	Mean	–	-.196	-.177	-.105	-.189	–
	SD	–	(.091)	(.084)	(.073)	(.910)	–
	Bias	–	.104	.123	.195	.111	–
	RMSE	–	.138	.149	.208	.144	–
	Coverage	–	.910	.810	.230	.750	–
β_7 (normal)	Mean	–	–	–	-.013	-.020	-.006
	SD	–	–	–	(.032)	(.040)	(.036)
	Bias	–	–	–	.031	.020	.006
	RMSE	–	–	–	.034	.044	.037
	Coverage	–	–	–	.940	.860	.990
β_8 (factor)	Mean	–	–	–	-.009	-.028	–
	SD	–	–	–	(.074)	(.093)	–
	Bias	–	–	–	.009	.028	–
	RMSE	–	–	–	.075	.097	–
	Coverage	–	–	–	.890	.930	–
β_9 (factor)	Mean	–	–	–	-.011	-.016	–
	SD	–	–	–	(.074)	(.092)	–
	Bias	–	–	–	.011	.016	–
	RMSE	–	–	–	.075	.093	–
	Coverage	–	–	–	.990	.990	–
β_{10} (normal)	Mean	–	–	–	–	-.001	-.003
	SD	–	–	–	–	(.038)	(.035)
	Bias	–	–	–	–	.001	.003
	RMSE	–	–	–	–	.038	.035
	Coverage	–	–	–	–	.990	.990
$Var(a_j)$	Mean	.647	.980	.833	.656	.979	.834
Log marg. likelihood	Mean	-1,299.342	-1,159.482	-1,222.002	-1,447.780	-1,184.751	-1,223.991
	SD	(27.755)	(57.106)	(47.478)	(11.074)	(65.357)	(47.699)
BIC	Mean	2606.285	2364.569	2474.409	2948.765	2445.512	2493.587

TABLE 4.14: Experimental study (MAR) ML results: Complete cases

Posterior summary statistics for Maximum Likelihood Estimation for complete cases. Across $D = 100$ simulated datasets, each with $N = 2,000$ individuals and $J = 50$ groups the reported values are posterior means, their standard deviations (SD), root mean squared errors, and empirical 95% credible interval coverage rates (averaged across replications). For each model information criteria are calculated and random effects are estimated. The data generating parameters are $\beta_1, \ldots, \beta_6$.

Variables		Model I	Model II	Model III	Model IV	Model V	Model VI
β_1 (Intercept)	Mean	-.039	.159	.144	.091	.167	.145
	SD	(.096)	(.181)	(.150)	(.138)	(.208)	(.150)
	Bias	.239	.041	.056	.109	.033	.055
	RMSE	.258	.185	.159	.175	.210	.159
	Coverage	.380	.910	.930	.860	.910	.930
β_2 (normal)	Mean	–	-1.529	-.935	–	-1.541	-.938
	SD	–	(.109)	(.081)	–	(.111)	(.082)
	Bias	–	.029	.565	–	.041	.562
	RMSE	–	.113	.571	–	.118	.568
	Coverage	–	.930	.000	–	.910	.000
β_3 (normal)	Mean	–	.807	–	–	.813	–
	SD	–	(.083)	–	–	(.085)	–
	Bias	–	.007	–	–	.013	–
	RMSE	–	.082	–	–	.085	–
	Coverage	–	.960	–	–	.960	–
β_4 (normal)	Mean	–	-.300	–	.078	-.302	–
	SD	–	(.072)	–	(.051)	(.072)	–
	Bias	–	.000	–	.378	.002	–
	RMSE	–	.071	–	.382	.072	–
	Coverage	–	.920	–	.000	.910	–
β_5 (factor)	Mean	–	-.382	-.317	-.221	-.384	-.317
	SD	–	(.150)	(.134)	(.109)	(.151)	(.135)
	Bias	–	.018	.083	.179	.016	.083
	RMSE	–	.150	.157	.209	.151	.158
	Coverage	–	.940	.880	.600	.940	.880
β_6 (factor)	Mean	–	-.269	-.214	-.151	-.271	-.215
	SD	–	(.156)	(.136)	(.126)	(.159)	(.136)
	Bias	–	.031	.086	.149	.029	.085
	RMSE	–	.159	.160	.195	.161	.160
	Coverage	–	.890	.860	.750	.890	.860
β_7 (normal)	Mean	–	–	–	-.002	-.004	-.001
	SD	–	–	–	(.050)	(.068)	(.057)
	Bias	–	–	–	.002	.004	.001
	RMSE	–	–	–	.108	.068	.057
	Coverage	–	–	–	.960	.880	.930
β_8 (factor)	Mean	–	–	–	-.002	-.009	–
	SD	–	–	–	(.109)	(.153)	–
	Bias	–	–	–	.002	.009	–
	RMSE	–	–	–	.108	.153	–
	Coverage	–	–	–	.960	.940	–
β_9 (factor)	Mean	–	–	–	-.015	-.014	–
	SD	–	–	–	(.110)	(.144)	–
	Bias	–	–	–	.015	.014	–
	RMSE	–	–	–	.110	.144	–
	Coverage	–	–	–	.960	.950	–
β_{10} (normal)	Mean	–	–	–	–	-.000	-.002
	SD	–	–	–	–	(.055)	(.050)
	Bias	–	–	–	–	.000	.002
	RMSE	–	–	–	–	.055	.050
	Coverage	–	–	–	–	.970	.980
$Var(a_j)$	Mean	.364	1.012	.691	.375	1.029	.696
Log-Likelihood	Mean	-585.310	-388.840	-451.372	-578.814	-386.693	-450.374
	SD	(17.292)	(18.388)	(20.047)	(17.802)	(18.483)	(20.116)
AIC	Mean	1174.621	791.680	912.744	1173.628	795.386	914.748
BIC	Mean	1184.258	825.412	936.838	1212.179	848.392	948.480

TABLE 4.15: Experimental study (MAR) ML results: Imputation

Posterior summary statistics for Maximum Likelihood Estimation for imputation. Across $D = 100$ simulated datasets, each with $N = 2,000$ individuals and $J = 50$ groups the reported values are posterior means, their standard deviations (SD), root mean squared errors, and empirical 95% credible interval coverage rates (averaged across replications). For each model information criteria are calculated and random effects are estimated. The data generating parameters are $\beta_1, \ldots, \beta_6$.

Variables		Model I	Model II	Model III	Model IV	Model V	Model VI
β_1 (Intercept)	Mean	-.034	.151	.144	.105	.157	.143
	SD	(.092)	(.160)	(.133)	(.114)	(.169)	(.133)
	Bias	.234	.049	.056	.095	.043	.057
	RMSE	.251	.166	.144	.148	.174	.144
	Coverage	.270	.950	.940	.900	.940	.940
β_2 (normal)	Mean	–	-1.490	-.915	–	-1.490	-.916
	SD	–	(.072)	(.047)	–	(.071)	(.047)
	Bias	–	.015	.585	–	.010	.584
	RMSE	–	.073	.587	–	.072	.586
	Coverage	–	.920	.000	–	.940	.000
β_3 (normal)	Mean	–	.789	–	–	.792	–
	SD	–	(.056)	–	–	(.056)	–
	Bias	–	.011	–	–	.008	–
	RMSE	–	.057	–	–	.057	–
	Coverage	–	.900	–	–	.910	–
β_4 (normal)	Mean	–	-.293	–	.076	-.295	–
	SD	–	(.044)	–	(.030)	(.044)	–
	Bias	–	.007	–	.376	.005	–
	RMSE	–	.045	–	.378	.045	–
	Coverage	–	.960	–	.000	.950	–
β_5 (factor)	Mean	–	-.384	-.330	-.231	-.384	-.330
	SD	–	(.100)	(.093)	(.083)	(.101)	(.093)
	Bias	–	.016	.070	.168	.016	.070
	RMSE	–	.101	.116	.187	.102	.116
	Coverage	–	.950	.920	.510	.950	.930
β_6 (factor)	Mean	–	-.263	-.221	-.160	-.265	-.221
	SD	–	(.119)	(.105)	(.098)	(.121)	(.106)
	Bias	–	.036	.079	.140	.035	.079
	RMSE	–	.124	.131	.171	.125	.131
	Coverage	–	.920	.870	.650	.910	.870
β_7 (normal)	Mean	–	–	–	.002	.001	.003
	SD	–	–	–	(.038)	(.048)	(.042)
	Bias	–	–	–	.002	.001	.003
	RMSE	–	–	–	.038	.048	.042
	Coverage	–	–	–	.910	.930	.930
β_8 (factor)	Mean	–	–	–	-.006	-.008	–
	SD	–	–	–	(.077)	(.097)	–
	Bias	–	–	–	.006	.008	–
	RMSE	–	–	–	.077	.096	–
	Coverage	–	–	–	.940	.960	–
β_9 (factor)	Mean	–	–	–	-.008	-.007	–
	SD	–	–	–	(.076)	(.093)	–
	Bias	–	–	–	.008	.007	–
	RMSE	–	–	–	.076	.092	–
	Coverage	–	–	–	.960	.950	–
β_{10} (normal)	Mean	–	–	–	–	-.002	-.003
	SD	–	–	–	–	(.035)	(.033)
	Bias	–	–	–	–	.002	.003
	RMSE	–	–	–	–	.035	.033
	Coverage	–	–	–	–	.940	.940
$Var(a_j)$	Mean	.368	.973	.681	.377	.979	.683
Log-Likelihood	Mean	-1,250.533	-816.594	-954.316	-1,238.790	-814.233	-953.069
	SD	(34.446)	(31.553)	(31.552)	(34.691)	(31.579)	(31.144)
AIC	Mean	2505.066	1647.188	1918.633	2493.580	1650.468	1920.138
BIC	Mean	2516.268	1686.395	1946.638	2538.387	1712.078	1959.345

TABLE 4.16: Experimental study (MAR): Model comparison

Bayesian results are reported as $\ln \mathrm{BF}$; ML results as LRT median p-values and $\Delta\mathrm{BIC}_{\mathrm{II}-k}$. Based on $D = 100$ simulated datasets.

Evidence for	Before deletion	Complete cases	Imputation
Bayesian results			
	lnBF	lnBF	lnBF
Model I vs II	Model II -409	Model II -214	Model II -139
Model III vs II	Model II -135	Model II -70	Model II -63
Model IV vs II	Model II -440	Model II -245	Model II -288
Model V vs II	Model II -25	Model II -22	Model II -25
Model VI vs II	Model II -148	Model II -105	Model II -64
Maximum Likelihood results			
	LRT p-value	LRT p-value	LRT p-value
Model I vs II	Model II $1.148e-186$	Model II $9.854e-83$	Model II $2.380e-185$
Model III vs II	Model II $7.123e-62$	Model II $6.959e-28$	Model II $1.542e-60$
Model IV vs II†	Model – $2.318e-18$	Model – $1.277e-84$	Model – $7.813e-173$
Model V vs II	Model II .043	Model II .038	not model II .3170
Model VI vs II†	Model – 0	Model – 0	Model – 0
Maximum Likelihood results (information criterion)			
	ΔBIC (II$-k$)	ΔBIC (II$-k$)	ΔBIC (II$-k$)
Model I vs II	Model II -418	Model II -179	Model II -414
Model III vs II	Model II -133	Model II -55	Model II -130
Model IV vs II	Model II -429	Model II -193	Model II -425
Model V vs II	Model II -13	Model II -11	Model II -13
Model VI vs II	Model II -140	Model II -62	Model II -136

Note (Bayesian results): Bayes factor interpretation: following Kass and Raftery (1995) with $0 < \ln\mathrm{BF} < 0.5$ *Not worth more than a bare mention*, $0.5 < \ln\mathrm{BF} < 1.0$ *substantial evidence*, $1.0 < \ln\mathrm{BF} < 2.0$ *strong evidence*, $\ln\mathrm{BF} > 2.0$ *decisive*. A negative sign indicates evidence for the second model.

Remark to †: Models IV and VI (same parameter dimension) are not nested within model II. For completeness, the LRT p-values are still reported which are obtained by applying the usual χ^2 reference distribution (as if the models were nested). However, because the nesting condition is violated, the LR statistic does not follow the standard asymptotic χ^2 distribution; consequently, these p-values are not valid and should be interpreted descriptively only (i.e., not as evidence for or against a model in an inferential sense).

Note (ML, ΔBIC): We report $\Delta\mathrm{BIC}_{\mathrm{II}-k} = \mathrm{BIC}_{\mathrm{II}} - \mathrm{BIC}_k$ for the comparison "Model k vs II". Negative values indicate a lower BIC for Model II (preference for Model II). Under standard regularity conditions, this is approximately related to Bayes factors via $\ln \mathrm{BF}_{k,\mathrm{II}} \approx \frac{1}{2}\Delta\mathrm{BIC}_{\mathrm{II}-k}$, so that negative values correspond to evidence in favour of Model II, consistent with the sign convention used for $\ln \mathrm{BF}$.

TABLE 4.17: Experimental study: Sensitivity results

Sensitivity analysis - effect of prior uncertainty for results with model II (before deletion), i.e., variation in prior variance of β and variation in (c_0, d_0) for σ_a on log-marginal likelihood and BIC. Additionally, convergence diagnostics for the MCMC estimates are $\hat{R}_{\max}$ that is the maximum rank-normalized split-$\hat{R}$ across all monitored parameters and $\text{ESS}_{\min}$ that is the minimum (bulk) effective sample size across parameters. Smaller $\hat{R}$ and larger ESS indicate better convergence ($\hat{R}_{\max} \leq 1.01$, $\text{ESS}_{\min} \geq 400$). Results are averaged over $D = 100$ datasets.

Prior for β Σ_β	Prior for σ_a^2 (c_0, d_0)	log likelihood $\ln L(y\|\hat{\theta})$	log prior $\ln \pi(\hat{\theta})$	log posterior $\hat{\pi}(\hat{\theta}\|y)$	log marginal likelihood $\ln f(y\|\hat{\theta}, X)$	BIC	$\widehat{R}_{\max}$	$\text{ESS}_{\min}$
$1 \cdot I$	(1, 1)	-848.369	-80.764	-4.899	-924.420	1,894.445	1.000	723
$1 \cdot I$	(1, 3)	-848.369	-79.835	-4.766	-923.326	1,892.257	1.022	730
$10 \cdot I$	(1, 3)	-848.369	-82.069	-5.069	-925.453	1,896.512	1.001	714
$100 \cdot I$	(1, 1)	-848.369	-88.654	-4.945	-932.061	1,909.727	1.003	704
$100 \cdot I$	(1, 2)	-848.369	-88.695	-5.189	-932.301	1,910.208	1.019	748
$100 \cdot I$	(1, 3)	-848.369	-89.259	-5.041	-933.103	1,911.811	1.008	758
$1000 \cdot I$	(1, 3)	-848.369	-95.068	-5.021	-938.416	1,922.436	1.001	718

TABLE 4.18: NESP SC 6 - employment status: categorical variables

Overview of the categorical variables from the NEPS SC 6. Due to rounding differences, deviations in the sum of the individual values are possible. For the imputations required for the ML estimation, M denotes the arithmetic sample mean and SD the sample standard deviation.

	observed sample		*imputed sample*	
Variable	n	proportion	M of imputed samples	SD between imputations
CASMIN				
1a, 1b, 2b	72	.020	-	-
1c, 2a	1,482	.409	-	-
2c	792	.219	-	-
3a, 3b	1,277	.352	-	-
missing	-	-	-	-
Unemployment duration				
less then one year	1,232	.340	.547	.004
1 to 2 years	259	.071	.111	.002
2 to 5 years	309	.085	.132	$<$ .001
more then 5 years	502	.139	.210	.002
missing	1321	.365	-	-
Migration background				
mo	3,093	.854	-	-
yes	530	.146	-	-
missing	-	-	-	-
Parental unemployment (father)				
no	3,311	.914	.961	.001
yes	135	.037	.039	.001
missing	177	.049	-	-
Parental unemployment (mother)				
no	2,169	.599	.610	$<$ .001
yes	1,390	.384	.390	$<$ .001
missing	64	.018	-	-
Gender of interviewed person				
male	1,808	.499	-	-
female	1,815	.501	-	-
missing	-	-	-	-

TABLE 4.19: NESP SC 6 - employment status: metric variables

Overview of the metric variables from the NEPS SC 6. Due to rounding differences, deviations in the sum of the individual values are possible. For the imputations required for the ML estimation, M denotes the arithmetic sample mean and SD the sample standard deviation.

		observed sample	*imputed sample*	
Variable			M of imputed samples	SD between imputations
Years of education				
	median	15.000	15.000	$< .001$
	M	14.750	14.743	$< .001$
	SD	2.223	2.231	.002
	missing	7	-	-
ICT literacy: WLE (corrected)				
	median	.345	.326	.006
	M	.380	.345	.002
	SD	1.183	1.196	.004
	missing	1,280	-	-
Procedural metacognition (ICT): proportion correct				
	median	-.183	.655	$< .001$
	M	-.090	.615	.002
	SD	.911	.196	.006
	missing	1,299	-	-
Mathematical competence: WLE (corrected)				
	median	-.176	-.182	$< .001$
	M	-.088	-.090	.002
	SD	.899	.911	.006
	missing	1,576	-	-
Reading competence: WLE (corrected)				
	median	.115	.114	.008
	M	.148	.145	.004
	SD	.837	.836	.010
	missing	862	-	-

TABLE 4.20: NESP SC 6 - employment status: ML results

Random effects models for the probability of being employed in wave 15 in starting cohort 6 of the NEPS. Entries are posterior means with 95% confidence intervals in brackets. Fit statistics include averaged log likelihood and BIC. Source: NEPS starting cohort 6, wave 15; $n = 3{,}623$.

	Dependent variable: employed in wave 15				
	Model 1 estimate [95% CI]	Model 2 estimate [95% CI]	Model 3 estimate [95% CI]	Model 4 estimate [95% CI]	Model 5 estimate [95% CI]
Years of education		.073* [-.003; .149]	.069* [-.010; .148]	.073* [-.003; .149]	.075** [-.005; .156]
CASMIN (ref. 1a, 1b, 2b)					
1c, 2a		-.163 [-.569; .243]	-.226 [-.655; .204]	-.163 [-.569; .244]	-.237 [-.669; .195]
2c		-.073 [-.590; .445]	-.092 [-.638; .444]	-.072 [-.590; .445]	-.114 [-.659; .430]
3a, 3b		-.357 [-1.021; .307]	-.460 [-1.161; .240]	-.357 [-1.021; .307]	-.500 [-1.206; .207]
ICT literacy: WLE (corrected)		.072* [-.004; .147]	.056 [-.025; .138]	.072 [-.004; .148]	.055 [-.028; .138]
Procedural metacognition (ICT): proportion correct		.437** [.127; .748]	.395** [.068; .721]	.438*** [.127; .749]	.355 [.021; .689]
Mathematical competence: WLE (corrected)		.041 [-.051; .133]	.035 [-.051; .120]	.041 [-.051; .133]	.021 [-.070; .112]
Reading competence: WLE (corrected)		.039 [-.045; .123]	.023 [-.064; .110]	.039 [-.045; .123]	.038 [-.050; .126]
Unemployment duration (ref. less then one year)					
1 to 2 years			-.252*** [-.909; -.614]		-.247* [-.509; .015]
2 to 5 years			-.268* [-.511; -.006]		-.266** [-.470; -.061]
More then 5 years			-.761 [-.473; -.064]		-.754*** [-.901; -.607]
Migration background (ref. No)					
Yes				.004 [-.141; .149]	.008 [-.141; .157]
Parental unemployment (ref. No)					
Father: Yes					.000 [-.265; .264]
Parental unemployment (ref. No)					
Mother: Yes					-.120 [-.230; -.009]
Gender of interviewed person (ref. male)					
Female					-.094** [-.210; .022]
Intercept	1.035 [.963; 1.106]	-.096 [-.921; .730]	.340 [-.511; 1.192]	-.097 [-.923; .730]	.384 [-.473; 1.241]
Variance parameter for random effects	.007	.007	.007	.006	.004
Median log likelihood	-1,523.764	-1,474.725	-1,400.076	-1,474.725	-1,397.271
AIC	3,051.528	2,969.449	2,826.151	2,971.449	2,828.541
BIC	3,063.918	3,031.400	2,906.687	3,039.595	2,933.857

Note: *p<0.1; **p<0.05; ***p<0.01

TABLE 4.21: NESP SC 6 - employment status: Bayesian results

Random effects models for the probability of being employed in wave 15 in starting cohort 6 of the NEPS. Entries are posterior means with 95% credible intervals in brackets. Fit statistics include median log marginal likelihood and BIC. Source: NEPS starting cohort 6, wave 15; $n = 3{,}623$.

	Dependent variable: employed in wave 15				
	Model 1 estimate [95% CI]	Model 2 estimate [95% CI]	Model 3 estimate [95% CI]	Model 4 estimate [95% CI]	Model 5 estimate [95% CI]
Years of education		.061 [-.015; .138]	.056 [-.021; .133]	.061 [-.016; .138]	.070 [-.008; .148]
CASMIN (ref. 1a, 1b, 2b)					
1c, 2a		-.152 [-.558; .255]	-.218 [-.632; .196]	-.158 [-.569; .254]	-.254 [-.678; .170]
2c		-.069 [-.588; .450]	-.099 [-.625; .427]	-.071 [-.596; .454]	-.143 [-.683; .397]
3a, 3b		-.327 [-.993; .339]	-.426 [-1.095; .243]	-.328 [-1.001; .344]	-.509 [-1.192; .175]
ICT literacy: WLE (corrected)		.124 [.053; .194]	.112 [.040; .185]	.121 [.050; .192]	.086 [.013; .159]
Procedural metacognition (ICT): proportion correct		.275 [-.059; .609]	.225 [-.110; .561]	.272 [-.058; .602]	.262 [-.078; .603]
Mathematical competence: WLE (corrected)		.015 [-.062; .092]	.008 [-.071; .087]	.015 [-.060; .090]	.011 [-.068; .089]
Reading competence: WLE (corrected)		.029 [-.052; .110]	.014 [-.070; .099]	.029 [-.052; .110]	.023 [-.064; .109]
Unemployment duration (ref. less then one year)					
1 to 2 years			-.740 [-.871; -.610]		-.736 [-.868; -.603]
2 to 5 years			-.216 [-.400; -.033]		-.204 [-.386; -.022]
More then 5 years			-.263 [-.429; -.098]		-.258 [-.426; -.090]
Migration background (ref. No)					
Yes				-.012 [-.159; .135]	-.012 [-.160; .136]
Parental unemployment (ref. No)					
Father: Yes					-.109 [-.222; .003]
Parental unemployment (ref. No)					
Mother: Yes					-.019 [-.288; .249]
Gender of interviewed person (ref. male)					
Female					-.103 [-.217; .012]
Intercept	1.037 [.805; 1.269]	.144 [-.720; 1.008]	.614 [-.256; 1.484]	.148 [-.712; 1.007]	.517 [-.357; 1.390]
Variance parameter for random effects	.178	.410	.412	.412	.412
Median log marginal likelihood	-1,514.836	-1,575.963	-1,528.285	-1,574.321	-1,542.796
BIC	3,037.868	3,160.120	3,064.765	3,156.838	3,093.787

TABLE 4.22: NESP SC 6 - employment status: model comparison

Results of logarithm of Bayes factors (lnBF) and of Likelihood ratio Test (LRT) with median p-value for comparing the different models with dataset is the starting cohort 6. The upper triangle shows the lnBF for Bayesian and ML results as well as the lower triangle shows the median p values (LRT) for the ML results.

Bayesian results

versus	Model 1	Model 2	Model 3	Model 4	Model 5
Model 1	-	Model 1 61.127	Model 1 13.449	Model 1 59.485	Model 1 27.960
Model 2		-	Model 3 −47.678	Model 4 −1.642	Model 5 −33.167
Model 3			-	Model 3 46.036	Model 3 14.511
Model 4				-	Model 5 −31.525
Model 5					-

ML results

versus	Model 1	Model 2	Model 3	Model 4	Model 5
Model 1	-	Model 2 −16.259	Model 3 −78.616	Model 4 −12.162	Model 5 −65.031
Model 2	Model 2 $1.05e-17$ $(df=8)$	-	Model 3 −62.357	Model 2 4.097	Model 5 −48.772
Model 3	Model 3 $9.89e-47$ $(df=11)$	Model 3 $1.24e-30$ $(df=3)$	-	Model 3 66.454	Model 3 13.585
Model 4	Model 4 $3.84e-17$ $(df=9)$	Not model 4 .975 $(df=1)$	Model 4 $1.30e-31$ $(df=2)$	-	Model 5 −52.869
Model 5	Model 5 $3.01e-45$ $(df=15)$	Model 5 $1.14e-28$ $(df=7)$	Not model 5 .098 $(df=4)$	Model 5 $2.21e-29$ $(df=6)$	-

Remark: model IV and model VI are not nested with model II.
Bayes factor interpretation: following Kass and Raftery (1995) with $0 <$ lnBF < 0.5 *Not worth more than a bare mention*, $0.5 <$ lnBF < 1.0 *substantial evidence*, $1.0 <$ lnBF < 2.0 *strong evidence*, lnBF > 2.0 *decisive*. A negative sign indicates evidence for the second model.

Chapter 5

Conclusion

In this thesis, I conducted a comprehensive exploration of Bayesian variable and model selection techniques, focusing on the critical challenge of handling missing values within this framework. My primary objective was to achieve a delicate balance between bias and variance, acknowledging the indispensable nature of this trade-off robust and accurate inferences. The integration of a sophisticated review of the literature and the development of novel methodologies has resulted in significant advancements within the field. These advancements have elucidated the interplay between variable selection, model complexity, and the treatment of missing data. One of the central contributions of this work is the Bayesian approach developed in collaboration with my doctoral supervisor, which enables principled uncertainty quantification and facilitates the incorporation of prior knowledge. This dual capability enhances both the reliability and interpretability of inference. A particularly significant strength of the presented methodology lies in its effective handling of missing values, see Aßmann et al. (2023) and Aßmann and Preising (2020). Traditional methods often resort to imputation techniques that may introduce bias or ignore uncertainty, or they rely solely on complete-case analysis, which can reduce the dataset's representativeness. In contrast, the Bayesian framework integrates missing data mechanisms seamlessly, acknowledging the uncertainty inherent in the missing information and leveraging probabilistic modeling to make informed decisions.

Addressing missing data is a pervasive challenge in real-world datasets. Traditional imputation approaches, while widely used, frequently fail to capture the uncertainty associated with missingness. Bayesian methods, such as data augmentation (Tanner & Wong, 1987) and multiple imputation (Rubin, 1976), offer more principled solutions by incorporating the missing data mechanism into the model and enabling coherent uncertainty quantification. Recent advances – including joint modeling for longitudinal and survival data with missing covariates (Ibrahim et al., 2001) and the incorporation of auxiliary information via informative priors (Gelman et al., 2023) – underscore the importance of treating missing data as an integral aspect of the modeling process rather than a secondary concern.

The foundation of Bayesian variable selection lies in the principled incorporation of prior knowledge and uncertainty into the modeling framework and process. Early works on Bayesian model averaging, see among others George and McCulloch (1993) and Mitchell and Beauchamp (1988), provided a foundation for considering multiple models and variables,

each weighted by their posterior probabilities. Building on these concepts, this thesis employed advanced techniques such as stochastic search variable selection (SSVS) algorithms (George & McCulloch, 1997) and spike-and-slab priors (Ishwaran & Rao, 2005) to facilitate automatic variable selection while addressing multicollinearity and managing model complexity. These methods empower researchers to achieve a balanced compromise between bias and variance, safeguarding against overfitting while ensuring that essential features are not overlooked. The presentation and application of the Bayesian approach is supplemented by the implementation of various machine learning methods, which yield results that are analogous to those obtained through the use of the presented, extended Bayesian methods. However, there are notable deficiencies in the application of Elastic net methods and their imputation approaches. In accordance with the assertions put forth by Friedman et al. (2010), the estimates are found to be (significantly) biased, as evidenced by the preceding analysis. Moreover, the Elastic net method appears to be incapable of accounting for the inherent uncertainty associated with the missing data mechanism, as noted by Gelman et al. (2023). The utilization of these methodologies engenders a distorted depiction of reality, particularly in simulated datasets, and a substantial discrepancy is observed when they are employed on real datasets.

In a Bayesian view picking out one model among a bunch of different models may lead to averaging across all models (Koop et al., 2010). Model averaging, as a Bayesian technique, transcends the immediate purview of variable selection, offering a potent paradigm for fortifying predictive accuracy and robustness across a multitude of applications. The conceptual underpinning of model averaging posits that no individual model can fully encapsulate the intricacies of real-world data. Aggregating predictions across a repertoire of models, each endowed with distinct strengths and weaknesses, engenders a more stable and accurate estimation of the latent data-generating process. The seminal work by Hoeting et al. (1999) laid the foundation for Bayesian model averaging, accentuating its efficacy in augmenting predictive performance and mitigating the impact of model uncertainty. As Kaplan notes, Bayesian model averaging provides a coherent framework for incorporating model uncertainty into inference, rather than conditioning on a single selected model (Kaplan & Chen, 2014; Kaplan & Lee, 2018).

Empirical results from the analyses conducted as part of this thesis demonstrate the practical implications of these methodologies. By carefully selecting relevant variables and optimizing model complexity through Bayesian model selection, the results achieved a harmonious balance between bias and variance. This equilibrium mitigated the risks of overfitting while guarding against oversimplified models, thereby enhancing predictive accuracy and robustness across diverse datasets.

Despite the advancements made, several challenges remain. For instance, computational demands can increase significantly for Bayesian methods when applied to large datasets or highly complex models. Additionally, the effectiveness of the approach depends on the appropriateness of the prior specifications and assumptions regarding the missing data mechanism. These limitations provide fertile ground for future research, including the development of scalable algorithms, exploration of non-standard missing data mechanisms, and integration

with emerging high-dimensional data techniques. Looking forward, this thesis opens several avenues for further investigation. Future work could extend the methodology to accommodate dynamic models, incorporate auxiliary information from external data sources, or explore its applicability to specific domains such as economics, healthcare, and environmental science. The integration of Bayesian frameworks with machine learning techniques also holds significant promise for addressing contemporary challenges in data science. In future topics these two frameworks can be evaluated in more complex computational challenges for larger datasets or extreme cases of missingness or non-random missing mechanisms.

An essential philosophical and methodological insight underpinning this thesis is the role of distributional assumptions in the construction of statistical knowledge. In both Bayesian and classical inference, probability distributions serve as structured representations of uncertainty and as operational tools to bridge the gap between observed data and latent processes. These assumptions are not merely mathematical conveniences; they embody our scientific beliefs about how the world behaves and provide a falsifiable scaffold for empirical investigation. Following the spirit of Popperian philosophy, distributional assumptions render scientific models testable and revisable in the light of new evidence, while Bayesian reasoning allows their continuous refinement. By formally acknowledging and modeling uncertainty through probability distributions, we are better equipped to make sense of incomplete or noisy information – a challenge that is central to the treatment of missing data and to model-based inference at large.

At a foundational level, this thesis emphasizes that probability distributions are not merely tools for statistical inference, but fundamental conceptual instruments for describing and interrogating the empirical world. Particularly in the context of missing data and model uncertainty, it is through explicit distributional assumptions that ignorance can be coherently represented, uncertainty propagated, and empirically grounded conclusions drawn. Against this philosophical and methodological backdrop, the thesis offers a comprehensive Bayesian framework for variable and model selection, enriched by advanced strategies for handling missing data. By embracing uncertainty and navigating the bias–variance trade-off with care, the proposed methods provide robust and interpretable inference across a range of applications.

Empirical results from the analyses conducted as part of this thesis demonstrate the practical implications of different methodologies. By carefully selecting relevant variables and optimizing model complexity through Bayesian model selection, the results achieved a harmonious balance between bias and variance. This equilibrium mitigated the risks of overfitting while guarding against oversimplified models, thereby enhancing predictive accuracy and robustness across diverse datasets. Despite the advancements made, several challenges remain. Computational demands can increase substantially for Bayesian methods when applied to large datasets or highly complex models. Additionally, the effectiveness of the approach depends on the appropriateness of the prior specifications and assumptions regarding the missing data mechanism. These limitations provide fertile ground for future research, including the development of scalable algorithms, exploration of non-standard missing data mechanisms, and integration with emerging high-dimensional data techniques.

Future work could extend the methodology to accommodate dynamic models, incorporate auxiliary information from external data sources, or explore its applicability in domains such as economics, healthcare, and environmental science. Furthermore, the integration of Bayesian frameworks with modern machine learning techniques holds significant promise for addressing contemporary challenges in data science, particularly in settings with large-scale or highly incomplete datasets.

From a methodological standpoint, this thesis situates Bayesian model selection within a broader landscape of model evaluation criteria. Standard tools such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the Deviance Information Criterion (DIC), and the Widely Applicable Information Criterion (WAIC) provide practical approximations of predictive accuracy, often asymptotically equivalent to leave-one-out cross-validation. While AIC relies on the ML and penalizes model complexity by the number of parameters, DIC and WAIC extend this rationale to the Bayesian setting, incorporating posterior uncertainty in evaluating fit and model flexibility. In addition, fully Bayesian model comparison can be conducted via the marginal likelihood, which integrates the likelihood over the prior distribution of the parameters, naturally accounting for parameter uncertainty. Differences in marginal likelihoods between models can be expressed as Bayes factors, offering a probabilistic measure of evidence favoring one model over another (Kass & Raftery, 1995; Koop et al., 2010). Together, these tools provide a coherent framework for balancing fit, complexity, and uncertainty, complementing each other in both practical and theoretical terms.

At a philosophical and conceptual level, this thesis emphasizes that probability distributions and model choices are not mere technical conveniences but represent structured ways to encode and reason about uncertainty in the empirical world. It is worth noting, however, that all these models, criteria, and measures are ultimately human constructs designed to make sense of the world. As Nietzsche moodily observed, everything lies in a meadow "like lazy cows" — abundant, varied, and independent of our descriptions. Our statistical frameworks are simply ways to navigate this landscape, to choose lenses that allow us to approximate reality as faithfully as possible. By consciously reflecting on the assumptions and criteria we employ, we can better balance rigor, interpretability, and predictive accuracy, and engage with the rich complexity of empirical phenomena with both care and intellectual humility. In this sense, Bayesian model selection, together with principled information criteria and Bayes-factor-based comparisons, provides not only a robust methodological framework but also a reflective tool for understanding the scope, limitations, and epistemological significance of statistical modeling. By explicitly acknowledging and propagating uncertainty, this thesis bridges the gap between observed data and latent processes, offering a comprehensive approach for inference in the presence of missing data and model uncertainty.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/tac.1974.1100705

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using gibbs sampling. *Journal of Educational Statistics*, *17*(3), 251–269. https://doi.org/10.2307/1165149

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669–679. https://doi.org/10.1080/01621459.1993.10476321

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. https://doi.org/10.1214/09-ss054

Arrow, K. J. (1973). Higher education as a filter. *Journal of Public Economics*, *2*(3), 193–216. https://doi.org/10.1016/0047-2727(73)90013-3

Aßmann, C. (2007). *Determinants and costs of current account reversalsunder heterogeneity and serial correlation* (Economics Working Paper, No. 2007-17). Kiel University, Department of Economics. https://www.econstor.eu/bitstream/10419/22033/1/EWP-2007-17.pdf

Aßmann, C. (2012). Determinants and costs of current account reversals under heterogeneity and serial correlation. *Applied Economics*, *44*(13), 1685–1700. https://doi.org/10.1080/00036846.2011.554370

Aßmann, C., & Boysen-Hogrefe, J. (2011). A Bayesian approach to model-based clustering for binary panel probit models. *Computational Statistics & Data Analysis*, *55*(1), 261–279. https://doi.org/10.1016/j.csda.2010.04.016

Aßmann, C., Gaasch, J.-C., & Stingl, D. (2023). A Bayesian approach towards missing covariate data in multilevel latent regression models. *Psychometrika*, *88*, 1495–1528. https://doi.org/10.1007/s11336-022-09888-0

Aßmann, C., & Preising, M. (2020). Bayesian estimation and model comparison for linear dynamic panel models with missing values. *Australian and New Zealand Journal of Statistics*, *62*(4), 536–557. https://doi.org/10.1111/anzs.12316

Austin, P. C., & Tu, J. V. (2004). Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of Clinical Epidemiology*, *57*(11), 1138–1146. https://doi.org/10.1016/j.jclinepi.2004.04.003

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, *86*(4), 948–955.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Becker, G. S. (1970). *Human capital: A theoretical and empirical analysis, with special reference to education* (4th ed.). Columbia University Press.

Berger, J. O. (2006a). The case for objective bayesian analysis. *Bayesian Analysis*, *1*(3). https://doi.org/10.1214/06-ba115

Berger, J. O. (2006b). *Statistical decision theory and bayesian analysis* (2nd ed.). Springer.

Bergrab, M. (2020). *Samples, weights, and nonresponse: The sample of starting cohort 4 of the national educational panel study (wave 12)* (tech. rep.). Leibniz Institute for Eduational Trajectories, National Educational Panel Study. Bamberg. https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/12-0-0/SC4_12-0-0_W.pdf

Bergrab, M., & Aßmann, C. (2024). Automated bayesian variable selection methods for binary regression models with missing covariate data. *AStA Wirtschafts- und Sozialstatistisches Archiv*. https://doi.org/10.1007/s11943-024-00345-1

Berkhof, J., van Mechelen, I., & Hoijtink, H. (2000). Posterior predictive checks: Principles and discussion. *Computational Statistics*, *15*(3), 337–354. https://doi.org/10.1007/s001800000038

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10.1137/141000671

Bhattacharya, A., Chakraborty, A., & Mallick, B. K. (2016). Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, *103*(4), 985–991. https://doi.org/10.1093/biomet/asw042

Bishop, C. M. (2009). *Pattern recognition and machine learning* (8th ed.). Springer.

Bissiri, P. G., & Walker, S. G. (2019). On general bayesian inference using loss functions. *Statistics & Probability Letters*, *152*, 89–91. https://doi.org/10.1016/j.spl.2019.04.005

Biswas, N., Mackey, L., & Meng, X.-L. (2022). Scalable spike-and-slab. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *Proceedings of the 39th international conference on machine learning* (pp. 2021–2040). PMLR. https://proceedings.mlr.press/v162/biswas22a.html

Bleich, J., Kapelner, A., George, E. I., & Jensen, S. T. (2014). Variable selection for bart: An application to gene regulation. *The Annals of Applied Statistics*, *8*(3). https://doi.org/10.1214/14-aoas755

Blossfeld, H.-P., & Roßbach, H.-G. (2019). *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (2nd ed.). Springer VS. https://doi.org/10.1007/978-3-658-23162-0

Boone, E. L., Ye, K., & Smith, E. P. (2007). Using data augmentation via the gibbs sampler to incorporate missing covariate structure in linear models for ecological assessments. *Environmental and Ecological Statistics*, *16*(1), 75–87. https://doi.org/https://doi.org/10.1007/s10651-007-0050-z

Bottolo, L., & Richardson, S. (2010). Evolutionary stochastic search for bayesian model exploration. *Bayesian Analysis*, *5*(3), 583–618. https://doi.org/10.1214/10-BA523

Bowers, A. J., & Zhou, X. (2019). Receiver operating characteristic (roc) area under the curve (auc): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. *Journal of Education for Students Placed at Risk (JESPAR)*, *24*(1), 20–46. https://doi.org/10.1080/10824669.2018.1523734

Box, G. E., & Tiao, G. C. (1992). *Bayesian inference in statistical analysis* (1st ed.). Wiley. https://doi.org/10.1002/9781118033197

Brand, J. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets (phd. thesis)*. TNO Prevention and Health (Erasmus University Rotterdam).

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Chapman & Hall/CRC.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3). https://doi.org/10.1214/ss/1009213726

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer Berlin Heidelberg.

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, *172*(9), 1070–1076. https://doi.org/10.1093/aje/kwq260

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer New York. https://doi.org/10.1007/b97636

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding aic and bic in model selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Butler, J. S., & Moffitt, R. (1982). A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica*, *50*(3), 761–764. https://doi.org/10.2307/1912613

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233–234. https://doi.org/10.1038/nmeth.4642

Carnegie, N. B., & Wu, J. (2019). Variable selection and parameter tuning for bart modeling in the fragile families challenge. *Socius: Sociological Research for a Dynamic World*, *5*. https://doi.org/10.1177/2378023119825886

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480. https://doi.org/10.1093/biomet/asq017

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Duxbury Press.

Casella, G., & George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, *46*(3), 167–174. https://doi.org/10.2307/2685208

Chan, K. W., & Meng, X.-L. (2022). Multiple improvements of multiple imputation likelihood ratio tests. *Statistica Sinica*, *32*, 1489–1514. https://doi.org/10.5705/ss.202019.0314

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research*, *84*(4), 487–508. https://doi.org/10.3102/0034654314532697

Chen, Q., & Wang, S. (2013). Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in Medicine*, *32*(21), 3646–3659. https://doi.org/10.1002/sim.5783

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *11*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Y., Liu, J., Peng, L., Wu, Y., Xu, Y., & Zhang, Z. (2024). Auto-encoding variational bayes. *Cambridge Explorations in Arts and Sciences*, *2*(1). https://doi.org/10.61603/ceas.v2i1.33

Chen, Y., Ming, H., & Yang, H. (2024). Efficient variable selection for high-dimensional multiplicative models: A novel lpre-based approach. *Statistical Papers*, *65*(6), 3713–3737. https://doi.org/10.1007/s00362-024-01545-1

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*(432), 1313–1321. https://doi.org/10.1080/01621459.1995.10476635

Chib, S. (2001). Markov chain monte carlo methods: Computation and inference. In J. J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (pp. 3569–3649). Elsevier. https://doi.org/10.1016/s1573-4412(01)05010-3

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, *96*(453), 270–281. https://doi.org/10.1198/016214501750332848

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298. https://doi.org/10.1214/09-aoas285

Chipman, H. A., & McCulloch, R. E. (2024). *Bayestree: Bayesian additive regression trees* [R package version 0.3-1.5]. https://CRAN.R-project.org/package=BayesTree

Clarke, K. A. (2001). Testing nonnested models of international relations: Reevaluating realism. *American Journal of Political Science*, *45*(3), 724–744. https://doi.org/10.2307/2669248

Clyde, M., & George, E. I. (2004). Model uncertainty. *Statistical Science*, *19*(1), 81–94. https://doi.org/10.1214/088342304000000035

Cohen, J., Cohen, P., West, S., & Aiken, L. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge. https://doi.org/10.4324/9780203774441

Cox, D., & Hinkley, D. (1979). *Theoretical statistics*. Chapman; Hall/CRC. https://doi.org/10.1201/b14832

Crainiceanu, C. M., & Ruppert, D. (2003). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *66*(1), 165–185. https://doi.org/10.1111/j.1467-9868.2004.00438.x

Cranmer, S. J., & Desmarais, B. A. (2017). What can we learn from predictive modeling? *Political Analysis*, *25*(2), 145–166. https://doi.org/10.1017/pan.2017.3

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, *183*(3), 1447–1465. https://doi.org/10.1016/j.ejor.2006.09.100

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press.

David, H. A., & Nagaraja, H. N. (2003). *Order statistics*. Wiley. https://doi.org/10.1002/0471722162

Daziano, R. A., & Achtnicht, M. (2014). Forecasting adoption of ultra-low-emission vehicles using bayes estimates of a multinomial probit model and the ghk simulator. *Transportation Science*, *48*(4), 671–683. https://doi.org/10.1287/trsc.2013.0464

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Deng, Y., & Lumley, T. (2023). Multiple imputation through xgboost. *Journal of Computational and Graphical Statistics*, *33*(2), 352–363. https://doi.org/10.1080/10618600.2023.2252501

Dhal, P., & Azad, C. (2021). A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence*, *52*(4), 4543–4581. https://doi.org/10.1007/s10489-021-02550-9

Dobra, A. (2009). Variable selection and dependency networks for genomewide data. *Biostatistics*, *10*(4), 621–639. https://doi.org/10.1093/biostatistics/kxp018

Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1). https://doi.org/10.1186/2193-1801-2-222

Doove, L. L., van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, *72*, 92–104. https://doi.org/10.1016/j.csda.2013.10.025

Du, J., Boss, J., Han, P., Beesley, L. J., Kleinsasser, M., Goutman, S. A., Batterman, S., Feldman, E. L., & Mukherjee, B. (2022). Variable selection with multiply-imputed datasets: Choosing between stacked and grouped methods. *Journal of Computational and Graphical Statistics*, *31*(4), 1063–1075. https://doi.org/10.1080/10618600.2022.2035739

Eekhout, I., van de Wiel, M. A., & Heymans, M. W. (2017). Methods for significance testing of categorical covariates in logistic regression models after multiple imputation: Power and applicability analysis. *BMC Medical Research Methodology*, *17*(1). https://doi.org/10.1186/s12874-017-0404-7

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, *92*(438), 548–560. https://doi.org/10.2307/2965703

Enders, C. K. (2022). *Applied missing data analysis* (2nd ed.). The Guilford Press.

Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*, 101–148.

Freedman, D. (2009). *Statistical models: Theory and practice* (2nd ed.). Cambridge Univ. Press.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Frühwirth-Schnatter, S. (2010). *Finite mixture and markov switching models*. Springer.

Frühwirth-Schnatter, S., & Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, *26*(1), 78–89. https://doi.org/10.1198/073500107000000106

García-Laencina, P. J., Sancho-Gómez, J.-L., & Figueiras-Vidal, A. R. (2009). Pattern classification with missing data: A review. *Neural Computing and Applications*, *19*(2), 263–282. https://doi.org/10.1007/s00521-009-0295-6

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409. https://doi.org/10.1080/01621459.1990.10476213

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2023). *Bayesian data analysis*. Chapman; Hall/CRC. https://doi.org/10.1201/b16018

Gelman, A., & Shalizi, C. R. (2012). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741. https://doi.org/10.1109/tpami.1984.4767596

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2021). *mvtnorm: Multivariate normal and t distributions* [R package version 1.1-3]. https://CRAN.R-project.org/package=mvtnorm

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association*, *95*(452), 1304–1308. https://doi.org/10.1080/01621459.2000.10474336

George, E. I., & McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. https://doi.org/10.1080/01621459.1993.10476353

George, E. I., & McCulloch, R. E. (1997). Approaches to bayesian variable selection. *Statistica Sinica*, *7*(2), 339–373.

Geweke, J. (1989). Bayesian inference in econometric models using monte carlo integration. *Econometrica*, *57*(6), 1317–1339. https://doi.org/10.2307/1913710

Geweke, J. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Staff report (Federal Reserve Bank of Minneapolis. Research Department)*. https://doi.org/10.21034/sr.148

Geweke, J., & Keane, M. (2001). Chapter 56: Computationally intensive methods for integration in econometrics. In J. J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (pp. 3463–3568). Elsevier. https://doi.org/10.1016/S1573-4412(01)05009-7

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D., & Kirby, A. J. (1993). Modelling complexity: Applications of gibbs sampling in medicine. *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(1), 39–52. https://doi.org/10.1111/j.2517-6161.1993.tb01468.x

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, *106*(494), 746–762. https://doi.org/10.1198/jasa.2011.r10138

Godfrey-Smith, P. (2008). *Theory and reality: An introduction to the philosophy of science* (4th ed.). Univ. of Chicago Press.

Goldman, N., & Whelan, S. (2000). Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Molecular Biology and Evolution*, *17*(6), 975–978. https://doi.org/10.1093/oxfordjournals.molbev.a026378

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, *45*(3), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Greene, W. (2003). *Econometric analysis* (5th ed.). Prentice-Hall.

Gu, Y., Fiebig, D. G., Cripps, E., & Kohn, R. (2009). Bayesian estimation of a random effects heteroscedastic probit model. *Econometrics Journal*, *12*(2), 324–339. https://doi.org/10.1111/j.1368-423x.2009.00283.x

Guyon, I., & Elisseeff, A. (2003). An introduction of variable and feature selection. *Journal of Machine Learning Research*, *1*, 1157–1182. https://doi.org/10.1162/153244303322753616

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, *15*(3). https://doi.org/10.1214/19-ba1195

Hajivassiliou, V. A., & McFadden, D. L. (1998). The method of simulated scores for the estimation of ldv models. *Econometrica*, *66*(4), 863–896. https://doi.org/10.2307/2999576

Hajivassiliou, V. A., & Ruud, P. A. (1994). Chapter 40: Classical estimation methods for ldv models using simulation. In J. J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (pp. 2383–2441). Elsevier. https://doi.org/10.1016/S1573-4412(05)80009-1

Hans, C., Dobra, A., & West, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, *102*(478), 507–516. https://doi.org/10.1198/016214507000000121

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, *75*(4), 1175–1189. https://doi.org/10.1111/j.1468-0262.2007.00785.x

Hasan, M. K., Alam, M. A., Roy, S., Dutta, A., Jawad, M. T., & Das, S. (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked*, *27*, 100799. https://doi.org/10.1016/j.imu.2021.100799

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

Heckman, J. J., & Willis, R. J. (1976). Estimation of a stochastic model of reproduction: An econometric approach. In N. E. Terleckyj (Ed.), *Household production and consumption* (pp. 99–146). NBER Studies in Income and Wealth.

Held, L., & Ott, M. (2018). On p-values and bayes factors. *Annual Review of Statistics and Its Application*, *5*(1), 393–419. https://doi.org/10.1146/annurev-statistics-031017-100307

Heymans, M. W., & Twisk, J. W. (2022). Handling missing data in clinical research. *Journal of Clinical Epidemiology*, *151*, 185–188. https://doi.org/10.1016/j.jclinepi.2022.08.016

Heymans, M. W., van Buuren, S., Knol, D. L., van Mechelen, W., & de Vet, H. C. (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Medical Research Methodology*, *7*(33). https://doi.org/10.1186/1471-2288-7-33

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

Hocking, R. (1976). A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, *32*(1), 1–49. https://doi.org/10.2307/2529336

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–417. https://doi.org/10.1214/ss/1009212519

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The bayesian approach* (3rd ed.). Open Court.

Huber, P. J., & Ronchetti, E. M. (2011). *Robust statistics* (2nd ed.). Wiley.

Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2001). *Bayesian survival analysis*. Springer New York. https://doi.org/10.1007/978-1-4757-3447-8

Imai, K., & van Dyk, D. A. (2005). A bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics*, *124*(2), 311–334. https://doi.org/10.1016/j.jeconom.2004.02.002

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773. https://doi.org/10.1214/009053604000001147

Jackman, S. (2009). *Bayesian analysis for the social sciences*. Wiley & Sons, Incorporated, John.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r* (2nd ed.). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Jaynes, E. T. (2010). *Probability theory: The logic of science* (7th ed.). Cambridge Univ. Press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press, USA.

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, *50*(2), 105–115. https://doi.org/10.1016/j.artmed.2010.05.002

Joel, L. O., Doorsamy, W., & Paul, B. S. (2022). A review of missing data handling techniques for machine learning. *International Journal of Innovative Technology and Interdisciplinary Sciences*, *Vol. 5 No. 3*, Issue 3. https://doi.org/10.15157/IJITIS.2022.5.3.971-1005

Jolliffe, I. (2004). *Principal component analysis* (2nd ed.). Springer.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Junge, M. R. J., & Dettori, J. R. (2018). Roc solid: Receiver operator characteristic (roc) curves as a foundation for better diagnostic tests. *Global Spine Journal*, *8*(4), 424–429. https://doi.org/10.1177/2192568218778294

Kapelner, A., & Bleich, J. (2016). bartMachine: Machine learning with bayesian additive regression trees. *Journal of Statistical Software*, *70*(4), 1–40. https://doi.org/10.18637/jss.v070.i04

Kaplan, D. (2021). On the quantification of model uncertainty: A bayesian perspective. *Psychometrika*, *86*(1), 215–238. https://doi.org/10.1007/s11336-021-09754-5

Kaplan, D., & Chen, J. (2014). Bayesian model averaging for propensity score analysis. *Multivariate Behavioral Research*, *49*(6), 505–517. https://doi.org/10.1080/00273171.2014.928492

Kaplan, D., & Lee, C. (2018). Optimizing prediction using bayesian model averaging: Examples using large-scale educational assessments. *Evaluation Review*, *42*(4), 423–457. https://doi.org/10.1177/0193841x18761421

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Keane, M. P. (1994). A computationally practical simulation estimator for panel data. *Econometrica*, *62*(1), 95–116. https://doi.org/https://doi.org/10.2307/2951477

Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey research methods*, *13*, 73–93. https://doi.org/10.5684/soep.v32

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. https://doi.org/10.48550/ARXIV.1312.6114

Kleinert, C., Christoph, B., & Ruland, M. (2019). Experimental evidence on immediate and long-term consequences of test-induced respondent burden for panel attrition. *Sociological Methods & Research*, *50*(4), 1552–1583. https://doi.org/10.1177/0049124119826145

Kohn, R., Smith, M., & Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, *11*(4), 313–322. https://doi.org/10.1023/a:1011916902934

Koop, G., León-González, R., & Strachan, R. W. (2010). Efficient posterior simulation for cointegrated models with priors on the cointegration space. *Econometric Reviews*, *29*(2), 224–242. https://doi.org/10.1080/07474930903382208

Korobilis, D., & Shimizu, K. (2022). Bayesian approaches to shrinkage and sparse estimation. *Foundations and Trends in Econometrics*, *11*(4), 230–354. https://doi.org/10.1561/0800000041

Kotsiantis, S. B. (2011). Feature selection for machine learning classification problems: A recent overview. *Artificial Intelligence Review*, *42*(1), 157–157. https://doi.org/10.1007/s10462-011-9230-1

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, *5*(2), 369–411. https://doi.org/10.1214/10-ba607

Lamnisos, D., Griffin, J. E., & Steel, M. F. J. (2009). Transdimensional sampling algorithms for bayesian variable selection in classification problems with many more variables than

observations. *Journal of Computational and Graphical Statistics*, *18*(3), 592–612. https://doi.org/10.1198/jcgs.2009.08027

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., & Mallick, B. K. (2003). Gene selection: A bayesian variable selection approach. *Bioinformatics*, *19*(1), 90–97. https://doi.org/10.1093/bioinformatics/19.1.90

Lehmann, R., & Lösler, M. (2016). Multiple outlier detection: Hypothesis tests versus model selection by information criteria. *Journal of Surveying Engineering*, *142*(4). https://doi.org/10.1061/(asce)su.1943-5428.0000189

Lewis, F., Butler, A., & Gilbert, L. (2010). A unified approach to model selection using the likelihood ratio test. *Methods in Ecology and Evolution*, *2*(2), 155–162. https://doi.org/10.1111/j.2041-210x.2010.00063.x

Li, K. H. (1988). Imputation using markov chains. *Journal of Statistical Computation and Simulation*, *30*(1), 57–79. https://doi.org/10.1080/00949658808811085

Li, P., & Redden, D. T. (2015). Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *BMC Medical Research Methodology*, *15*(1). https://doi.org/10.1186/s12874-015-0026-x

Liesenfeld, R., & Richard, J.-F. (2010). Efficient estimation of probit models with correlated errors. *Journal of Econometrics*, *156*(2), 367–376. https://doi.org/10.1016/j.jeconom.2009.11.006

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, *113*(522), 626–636. https://doi.org/10.1080/01621459.2016.1264957

Little, R., & Rubin, D. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley. https://doi.org/10.1002/9781119013563

Liu, Y., Wang, Y., Feng, Y., & Wall, M. M. (2016). Variable selection and prediction with incomplete high-dimensional data. *The Annals of Applied Statistics*, *10*(1), 418–450. https://doi.org/10.1214/15-AOAS899

Llorente, F., Martino, L., Delgado, D., & Lopez-Santiago, J. (2020). Marginal likelihood computation for model selection and hypothesis testing: An extensive review. *SIAM Review, 2022*, *65*(1), 3–58. https://doi.org/10.1137/20m1310849

Lucchetti, R., & Pedini, L. (2023). The spherical parametrisation for correlation matrices and its computational advantages. *Computational Economics*, *64*(2), 1023–1046. https://doi.org/10.1007/s10614-023-10467-3

Lupu, N., & Michelitch, K. (2018). Advances in survey methods for the developing world. *Annual Review of Political Science*, *21*(1), 195–214. https://doi.org/10.1146/annurev-polisci-052115-021432

Lütkepohl, H. (1996). *Handbook of matrices* (1st ed.). Wiley.

Mallows, C. L. (1973). Some comments on cp. *Technometrics*, *15*(4), 661–675. https://doi.org/10.1080/00401706.1973.10489103

Marill, T., & Green, D. (1963). On the effectiveness of receptors in recognition systems. *IEEE Transactions on Information Theory*, *9*(1), 11–17. https://doi.org/10.1109/tit.1963.1057810

Marshall, A., Altman, D. G., Holder, R. L., & Royston, P. (2009). Combining estimates of interest in prognostic modelling studies after multiple imputation: Current practice and guidelines. *BMC Medical Research Methodology*, *9*(1). https://doi.org/10.1186/1471-2288-9-57

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, *42*(9), 1–21. https://doi.org/10.18637/jss.v042.i09

Meng, X.-L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, *79*(1), 103–111. https://doi.org/10.2307/2337151

Mersmann, O., Trautmann, H., Steuer, D., & Bornkamp, B. (2023). *Truncnorm: Truncated normal distribution* [R package version 1.0-9]. https://CRAN.R-project.org/package=truncnorm

Miller, A. (2019). *Subset selection in regression* (2nd ed.). Taylor & Francis Group.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032. https://doi.org/10.1080/01621459.1988.10478694

Mittelhammer, R. C. (2013). *Mathematical statistics for economics and business* (2nd ed.). Springer.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics* (3rd ed.). McGraw-Hill.

Mortensen, D. T., & Pissarides, C. A. (1994). Job creation and job destruction in the theory of unemployment. *The Review of Economic Studies*, *61*(3), 397–415. https://doi.org/10.2307/2297896

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective* (1st ed.). MIT press.

Nasrollahzadeh, S. (2007). The analysis of bayesian probit regression of binary and polychotomous response data. *IJE Transactions B: Applications*, (20), 237–248.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*. https://doi.org/10.3389/fnbot.2013.00021

NEPS, National Educational Panel Study. (2021). Neps-startkohorte 4: Klasse 9 (sc4 12.0.0). https://doi.org/10.5157/NEPS:SC4:12.0.0

NEPS Network. (2024a). National educational panel study, scientific use file of starting cohort grade 9. https://doi.org/10.5157/NEPS:SC4:14.0.0

NEPS Network. (2024b). Scientific use file of neps starting cohort 6: Adults. https://doi.org/10.5157/NEPS:SC6:15.0.0

Newton, M. A., & Raftery, A. E. (1994). Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, *56*(1), 3–26. https://doi.org/10.1111/j.2517-6161.1994.tb01956.x

O'Hara, R. B., & Sillanpää, M. J. (2009). A review of bayesian variable selection methods: What, how and which. *Bayesian Analysis*, *4*(1), 85–117. https://doi.org/10.1214/09-BA403

Pacifico, A., & Pilone, D. (2024). Penalized bayesian approach-based variable selection for economic forecasting. *Journal of Risk and Financial Management*, *17*(2). https://doi.org/10.3390/jrfm17020084

Pajor, A. (2017). Estimating the marginal likelihood using the artihmetic mean identiy. *Bayesian Analysis*, *12*(1), 261–287. https://doi.org/10.1214/16-BA1001

Panken, A. M., & Heymans, M. W. (2022). A simple pooling method for variable selection in multiply imputed datasets outperformed complex methods. *BMC Medical Research Methodology*, *22*(1). https://doi.org/10.1186/s12874-022-01693-8

Park, M. Y., & Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(4), 659–677. https://doi.org/10.1111/j.1467-9868.2007.00607.x

Pearl, J. (2016). *Causal inference in statistics: A primer* (1st ed.). John Wiley Sons, Incorporated.

Perrakis, K., Ntzoufras, I., & Tsionas, E. G. (2014). On the use of marginal posteriors in marginal likelihood estimation via importance sampling. *Computational Statistics & Data Analysis*, *77*, 54–69. https://doi.org/10.1016/j.csda.2014.03.004

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*(4), 525–556. https://doi.org/10.3102/00346543074004525

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, *73*(1), 74–97. https://doi.org/10.1093/poq/nfp014

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The American Economic Review*, *62*(4), 659–661.

Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation*, *7*(4), 353–383. https://doi.org/10.1076/edre.7.4.353.8937

Piironen, J., & Vehtari, A. (2016). Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735. https://doi.org/10.1007/s11222-016-9649-y

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in s and s-plus* (3rd ed.). Springer.

Popper, K. R. (2002). *The logic of scientific discovery* (2nd ed.). Routledge Classics,

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. https://doi.org/10.2307/271063

Raftery, A. E., & Zheng, Y. (2003). Discussion: Performance of bayesian model averaging. *Journal of the American Statistical Association*, *98*(464), 931–938. https://doi.org/10.1198/016214503000000891

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-56706-x

Reichl, J. (2020). Estimating marginal likelihoods from the posterior draws through a geometric identity. *Monte Carlo Methods and Applications*, *26*(3), 205–221. https://doi.org/10.1515/mcma-2020-2068

Robert, C. P., & Casella, G. (2004). *Monte carlo statistical methods* (2nd ed.). Springer New York. https://doi.org/10.1007/978-1-4757-4145-2

Roberts, G., & Smith, A. (1994). Simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. *Stochastic Processes and their Applications*, *49*(2), 207–216. https://doi.org/10.1016/0304-4149(94)90134-1

Ročková, V., & George, E. I. (2018). The spike-and-slab LASSO. *Journal of the American Statistical Association*, *113*(521), 431–444. https://doi.org/10.1080/01621459.2016.1260469

Ročková, V., & van der Pas, S. (2020). Posterior concentration for bayesian regression trees and forests. *The Annals of Statistics*, *48*(4), 2108–2131. https://doi.org/10.1214/19-aos1879

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1981). The bayesian bootstrap. *The Annals of Statistics*, *9*(1), 130–134. https://doi.org/10.1214/aos/1176345338

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*(4), 1151–1172. https://doi.org/10.1214/aos/1176346785

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (1st ed.). Wiley. https://doi.org/10.1002/9780470316696

Rusdah, D. A., & Murfi, H. (2020). Xgboost in handling missing values for life insurance risk prediction. *SN Applied Sciences*, *2*(1336). https://doi.org/10.1007/s42452-020-3128-y

Russu, A., Malovini, A., Puca, A. A., & Bellazzi, R. (2012). Stochastic model search with binary outcomes for genome-wide association studies. *Journal of the American Medical Informatics Association*, *19*(e1), e13–e20. https://doi.org/10.1136/amiajnl-2011-000741

Sabbe, N., Thas, O., & Ottoy, J.-P. (2013). EMLasso: Logistic lasso with missing data. *Statistics in Medicine*, *32*(18), 3143–3157. https://doi.org/10.1002/sim.5760

Schafer, J. L. (1997). *Analysis of incomplete multivariate data* (1st ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9781439821862

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989x.7.2.147

Scheipl, F., Kneib, T., & Fahrmeir, L. (2013). Penalized likelihood and bayesian function selection in regression models. *AStA Advances in Statistical Analysis*, *97*(4), 349–385. https://doi.org/10.1007/s10182-013-0211-3

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*(398), 605–610. https://doi.org/10.1080/01621459.1987.10478472

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-sts330

Sixt, M., & Aßmann, C. (2020). The influence of regional school infrastructure and labor market conditions on the transition process to secondary schooling in germany.

Journal for Educational Research Online, *12*(2), 36–66. https://www.waxmann.com/artikelART104168

Smith, J. Q. (2010). *Bayesian decision analysis*. Cambridge University Press.

Sparapani, R., Spanbauer, C., & McCulloch, R. (2021). Nonparametric machine learning and efficient computation with bayesian additive regression trees: The bart r package. *Journal of Statistical Software*, *97*(1), 1–66. https://doi.org/10.18637/jss.v097.i01

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Stan Development Team. (2025). RStan: The R interface to Stan [R package version 2.32.7]. https://mc-stan.org/

Steinhauer, H. W., & Aßmann, C. (2018). *Modelling nonresponse in educational multi-informant studies: A multilevel approach using bivariate probit models* (LIfBi WorkingPaper No. 74). Bamberg. https://www.lifbi.de/Portals/2/Working%20Papers/WP_LXXIV.pdf

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, *50*(4), 1171–1177. https://doi.org/10.2307/2533455

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1). https://doi.org/10.1186/1471-2105-8-25

Tanner, M. A., & Wong, W. H. (2010). From EM to data augmentation: The emergence of MCMC bayesian computation in the 1980s. *Statistical Science*, *25*(4), 506–516. https://doi.org/10.1214/10-sts341

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528–540. https://doi.org/10.1080/01621459.1987.10478458

Therneau, T., & Atkinson, B. (2018). *Rpart: Recursive partitioning and regression trees, [computer software manual].* Version (R package version 4.1-13). https://CRAN.R-project.org/package=rpart

Thulasiraman, K., & Swamy, M. N. S. (1992). *Graphs: Theory and algorithms* (1st ed.). Wiley.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *73*(3), 273–282. https://doi.org/10.1111/j.1467-9868.2011.00771.x

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91–108. https://doi.org/10.1111/j.1467-9868.2005.00490.x

Train, K. (2009). *Discrete choice methods with simulation* (2nd ed.). Cambridge University Press.

Trinkaus, O., & Kauermann, G. (2023). Can machine learning algorithms deliver superior models for rental guides? *AStA Wirtschafts- und Sozialstatistisches Archiv*, *17*(3), 305–330. https://doi.org/10.1007/s11943-023-00333-x

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, *27*(11), 1134–1142. https://doi.org/10.1145/1968.1972

van Buuren, S. (2018). *Flexible imputation of missing data, second edition*. Chapman; Hall/CRC. https://doi.org/10.1201/9780429492259

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

van Dyk, D., & Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, *10*(1), 1–50. https://doi.org/10.1198/10618600152418584

Venables, W. N., & Ripley, B. (2002). *Modern applied statistics with s* (4th ed.). Springer.

Vergouwe, Y., Royston, P., Moons, K. G., & Altman, D. G. (2010). Development and validation of a prediction model with missing predictor data: A practical approach. *Journal of Clinical Epidemiology*, *63*(2), 205–214. https://doi.org/10.1016/j.jclinepi.2009.03.017

von Collani, G., & Herzberg, P. Y. (2003). Eine revidierte fassung der deutschsprachigen skala zum selbstwertgefühl von rosenberg. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *24*(1), 3–7. https://doi.org/10.1024//0170-1789.24.1.3

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*(2), 307–333. https://doi.org/10.2307/1912557

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/bf03194105

Wasserman, L. (2008). *All of statistics: A concise course in statistical inference* (5th ed.). Springer.

West, M. (1993). Approximating posterior distributions by mixtures. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *55*(2), 409–422. https://doi.org/10.1111/j.2517-6161.1993.tb01911.x

Wilson, P. (2015). The misuse of the vuong test for non-nested models to test for zero-inflation. *Economics Letters*, *127*, 51–53. https://doi.org/10.1016/j.econlet.2014.12.029

Witten, D. M., & Tibshirani, R. (2011). Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(5), 753–772. https://doi.org/10.1111/j.1467-9868.2011.00783.x

Wood, A. M., White, I. R., & Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, *27*(17), 3227–3246. https://doi.org/10.1002/sim.3177

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). MIT Press.

Xu, D., Daniels, M. J., & Winterstein, A. G. (2016). Sequential bart for imputation of missing covariates. *Biostatistics*, *17*(3), 589–602. https://doi.org/10.1093/biostatistics/kxw009

Xu, Z. (2022). Handling ignorable and non-ignorable missing data through bayesian methods in jags. *Journal of Behavioral Data Science*, *2*(2), 1–28. https://doi.org/10.35566/jbds/v2n2/xu

Yan, T., & Williams, D. (2022). Response burden – review and conceptual framework. *Journal of Official Statistics*, *38*(4), 939–961. https://doi.org/10.2478/jos-2022-0041

Yang, X., Belin, T. R., & Boscardin, W. J. (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics*, *61*(2), 498–506. https://doi.org/10.1111/j.1541-0420.2005.00317.x

Yang, Y. (2005). Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika*, *92*(4), 937–950. https://doi.org/10.1093/biomet/92.4.937

Zhang, X., Yan, C., Gao, C., Malin, B. A., & Chen, Y. (2020). Predicting missing values in medical data via xgboost regression. *Journal of Healthcare Informatics Research*, *4*(4), 383–394. https://doi.org/10.1007/s41666-020-00077-1

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*(1), 95–112. https://doi.org/10.1016/j.jeconom.2015.02.006

Zinn, S., & Gnambs, T. (2020). Analyzing nonresponse in longitudinal surveys using bayesian additive regression trees: A nonparametric event history analysis. *Social Science Computer Review*, *40*(3), 678–699. https://doi.org/10.1177/0894439320928242

Zinn, S., Würbach, A., Steinhauer, H. W., & Hammon, A. (2018). Attrition and selectivity of the neps starting cohorts: An overview of the past 8 years. *NEPS Survey Papers*. https://doi.org/10.5157/NEPS:SP34:1.0

Zinn, S., Würbach, A., Steinhauer, H. W., & Hammon, A. (2020). Attrition and selectivity of the neps starting cohorts: An overview of the past 8 years. *AStA Wirtschafts- und Sozialstatistisches Archiv*, *14*(2), 163–206. https://doi.org/10.1007/s11943-020-00268-7

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x