Behavior of convergence in logistic regression models

Assessing the drop of the Kolmogorov distance between the sampling distribution and the asymptotic distribution of estimators and test statistics in logistic regression analysis

Dissertation

zur Erlangung des akademisches Grades
eines Doktors der Sozial- und Wirtschaftswissenschaften
(Dr. rer. pol.)
an der Fakultät der Sozial- und Wirtschaftswissenschaften
der Otto-Friedrich-Universität Bamberg

vorgelegt von
Dipl.-Stat. Mariana Saskia Nold
aus München

Bamberg, Januar 2014

To my children Marleen, Valentin and Oliver

Erstgutachterin: Prof. Dr. Susanne Rässler

Zweitgutachter: Prof. Dr. Georg Heinze

Tag der mündlichen Prüfung: 14.04.2014

Table of Contents

Acknowledgment

I would like to begin by thanking Professor Susanne Rässler whose comments and suggestions were of great value for my thesis. I appreciate her constructive feedback and have greatly benefited from her useful suggestions and meaningful face-to-face discussions. I am also grateful to Professor Georg Heinze who provided not only technical help and sincere encouragement but also very helpful and constructive advices and comments. Additionally, I extend my thanks to my family members for their moral support and warm encouragements.

I feel gratitude for the educational and professional opportunities granted to me during my academical career. I cannot thank all of my teachers and mentors here, but I do want to name Dr. Ulrich Kaczmar , Dr. habil. Ulrich Pötter, Professor Helmut Pruscha, and PD Dr. Christian Heumann for their highly valuable support.

Finally, I would like to thank the Free State of Bavaria for two scholarships, first the Step-by-Step-scholarship for the support of women, later a scholarship of the Bayerische Eliteförderung, for a grant that made it possible to write this doctoral thesis.

1 Outline

1.1 Abstract

Using classical inference, hypothesis tests and confidence intervals are often based on large-sample assumptions, which are said to hold if the sample size is large enough. The weakness of this approach is, that the researcher does not know what sample size is required for this purpose in a concrete situation.

A common problem that encounters in statistics is the procedure of modeling the relationship between explanatory variables and a binary response. Here logistic regression analysis often represents the appropriate method. This method is used to estimate the probability or odds of occurrence of the binary response in dependence of explanatory variables. But, what is the sample size to be large enough to base statistical conclusions on asymptotic properties?

The type of convergence, with which we are dealing here, is convergence in law, in the following denoted as $\mathcal{L}$-convergence. If the limiting distribution of a statistic is continuous, then $\mathcal{L}$-convergence is equivalent to convergence with respect to the Kolmogorov distance. Therefore, the Kolmogorov distance is an effective tool for discussing the behavior of $\mathcal{L}$-convergence.

The present work uses an autogenerated process that involves the classical theory of logistic regression analysis to explore the behavior of $\mathcal{L}$-convergence by means of the Kolmogorov distance. This autogenerated process is a special case of the autogenerated process given in McCullagh (2008). In that paper, McCullagh introduces a Cox process that is fully compatible with the standard logistic random effect model, while this work uses a Poisson process that corresponds to the standard logistic regression model.

Based on the Kolmogorov distance two methods are developed in order to investigate the behavior of $\mathcal{L}$-convergence and its impacts on statistical conclusions. The first serves to extend the spectrum of methods to discuss the impacts of the Firth-penalization, the second to use the classical inference as a more deliberate method with respect to asymptotic properties.

The first method consists of the distance-sample-size-diagram and the accuracy-diagram. The distance-sample-size-diagram represents the the mean approximate Kolmogorov distance as a function of the predefined sample size. The predefined sample size is displayed on the horizontal axis and the mean approximate Kolmogorov distance between the statistic of interest and its limiting distribution on the vertical axis. This is a fruitful graphical representation of the behavior of $\mathcal{L}$-convergence in dependence of the rate at which empirical information accrues. Finally the accuracy-diagram presents the actual accuracy function of a confidence interval and its reference derived from asymptotics. This diagram complements the distance-sample-size-diagram as a tool to study the impact of penalizations.

The second method, the p-value-uniform-diagram, shows the actual empirical cumulative distribution function of the p-values of a statical test and the cumulative distribution function of the uniform distribution as the reference of the former. A deviation from this reference indicates that $\mathcal{L}$-convergence is not reached.

1.2 Structure of this work

The second chapter introduces basic terms, definitions and concepts concerning logistic regression analysis. Section 2.1 reviews the classical theory of logistic regression analysis, defining the maximum likelihood estimator (MLE) and the deviance statistic and deriving their large-sample properties, following Pruscha (2000). Here, the logistic regression model is confined to deterministic explanatory variables. Subsequently, section 2.2 presents asymptotic confidence intervals and asymptotic hypothesis tests.

Chapter 3 introduces the autogenerated process, which is based on multivariate Poisson processes, as outlined in section 3.1. The central section of chapter 3, section 3.3, derives the large-sample properties of the regression coefficient estimator and the deviance statistic based on the autogenerated process.

Chapter 4 reveals the origin of $\mathcal{L}$-convergence problems and presents approaches to overcome them. Section 4.1 introduces separation, which is a condition that causes nonexistence of the covariate effect estimator. Section 4.2 informs about penalized likelihood estimation methods. Here, a penalty term is added to the estimating equation to ensure the existence of its root. In particular, the Firth-penalization is introduced.

Chapter 5 represents the mean approximate Kolmogorov distance as a measure to describe the behavior of $\mathcal{L}$-convergence in dependence of the sample size. Section 5.1 introduces the Kolmogorov distance. In general this metric cannot be computed exactly. Thus the approximate Kolmogorov distance is defined in section 5.2. In section 5.3 familiar convergence theorems, as for example the de Moivre-Laplace theorem, are used to discus of the behavior of $\mathcal{L}$-convergence by means of the distance-sample-size-diagram.

Thereafter, chapter 6 compares, by means of a fictive example, penalized and unpenalized likelihood estimation methods. This chapter demonstrates how the mean approximate Kolmogorov distance can be used to extend the spectrum of methods to explore the impacts of Firth-penalization.[1] The distance-sample-size-diagram is applied and provides a fruitful basis for a better understanding of the influence of the Firth-penalization on $\mathcal{L}$-convergence. Finally the accuracy-diagram is used to compare the actual accuracy of

[1] The effects of this penalization are for example discussed in Heinze (2006) and Heinze/Schemper (2002).

confidence intervals to the accuracy assumed in asymptotics, for both penalized and unpenalized likelihood estimation methods.

Chapter 7 demonstrates, by means of two real data sets, how the p-value-uniform-diagram is used to apply asymptotic methods taking into account encountering $\mathcal{L}$-convergence problems. Both examples serve for measuring the relationship between a binary response and three binary independent variables. The statistic of interest is here the residual deviance which is defined in section 7.1. In fact, this deviance is prone to a slow drop of the Kolmogorov distance with increasing sample size. The p-value-uniform-diagram is defined in section 7.2 and information is given also for the practical application.

The first example is taken from Altman (1990) and discussed in section 7.3. It deals with the potential influence of smoking, obesity and snoring on the hypertensive status in human patients. The second example, discussed in section 7.4, uses data from a study investigating the potential influence of three binary covariates on the occurrence or non-occurrence of infection following birth by Cesarian section. It is taken from Fahrmeir/Tutz (2001).

Finally section 8 reflects critically the usefulness of the developed methods. On the one hand the distance-sample-size-diagram and the accuracy-diagram proved beneficially for the comparison of penalized and unpenalized likelihood methods. On the other hand, the p-value-uniform-diagram seems to be appropriate to to raise the researcher's awareness for possible $\mathcal{L}$-convergence problems. The computation of this diagram in many statistical applications may reveal the diagnostic value of this approach in particular to be successful in making implicit assumptions explicit. In any case all three diagrams have a high instructive value.

2 Basic terms and concepts

2.1 Logistic regression analysis

The introduction of the logistic regression models follows Pruscha (2000) pages 278ff. The mathematical derivation of the asymptotic distribution of the estimator $\underline{\hat{\beta}}$ is based on Pruscha (2000) pages 187 ff.

2.1.1 Notation

We will use capital letters for random variates, and small letters for their realizations and for deterministic values. A vector is always denoted by underlining the (small or capital) letter. To describe a matrix the letter is underlined twice. The vector $\underline{v}$ is a column vector. The vector $\underline{v}'$ is a row vector.

The logistic regression model is a framework for modelling an univariate binary response Y in dependence of a set of m deterministic explanatory variables $x_1, \ldots, x_m$, $m \in \mathbb{N}^0$.[1] The explanatory variables are denoted covariates. In this work all covariates are binary. Nevertheless, the data is not grouped in this chapter. In fact, the introduction given here is also valid for continuous covariates.

In the following the index j is used to label the covariates, $j \in \{1, \ldots, m\}$. We assume data consisting of $n \in \mathbb{N}$ observations Y_i, $i \in \{1, \ldots, n\}$, of the response Y. The response vector is $\underline{Y}' := (Y_1, \ldots, Y_i, \ldots Y_n)'$. The values of the explanatory variables of unit or individual

[1] In the present work the set of all natural numbers $\mathbb{N}$ does not include zero, thus $\mathbb{N} := \{1, 2, 3, \ldots\}$.

i are collected in a row vector $\underline{x}_i^{'} := (1, x_{i,1}, \ldots, x_{i,j}, \ldots, x_{i,m})$. The design matrix $\underline{\underline{x}}$ is

$$\begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,j} & \ldots & x_{1,m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i,1} & \ldots & x_{i,j} & \ldots & x_{i,m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,j} & \ldots & x_{n,m} \end{pmatrix}.$$

An interaction effect is said to be present, if the effect of one covariate on the response, depends on the particular value of one or several other covariates. If interaction effects are considered, the design matrix is expanded by the corresponding products of the covariates. In general, the explanatory variables can be continuous real variables or binary variables. The binary variables will be coded by 0 and 1. The unknown covariate effects are given by the vector $\underline{\beta}$. If no interaction effect is taken into account $\underline{\beta}^{'} := (\beta_0, \beta_1, \ldots \beta_m)$. Otherwise the dimension of $\underline{\beta}$ is d, $d \in \mathbb{N}$. The component of $\underline{\beta}$ which corresponds to the interaction between x_{j_1} and x_{j_2} is β_{j_1,j_2}. The expectation $\mathbb{E}(Y_i)$ is denoted as μ_i. The present work uses the symbol β_j to term the jth component of the covariate effect vector.

2.1.2 Logistic regression model

The model is defined by two assumptions. (See Fahrmeir/Tutz (2001) pages 19ff.)

1. **Distributional assumption:** The response of individual i_1 is Bernoulli distributed with success probability μ_{i_1}, $Y_{i_1} \sim \text{Bin}(\mu_{i_1}, 1)$. Correspondingly the response of individual i_2, $i_1 \neq i_2$, is $\text{Bin}(\mu_{i_2}, 1)$ distributed and Y_{i_1} and Y_{i_2} are assumed independent.

2. **Structural assumption:** The expectation μ_i is related to the linear predictor $\eta_i := \underline{x}_i^{'}\underline{\beta}$ by the response function $\rho(\eta_i)$. This logistic response function is[2]

$$\mu_i = \rho(\eta_i) := \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}.$$

   The inverse of ρ is denoted as link function ρ^{-1}, and is for the logistic model given

[2] Other response functions, such as the complementary loglog or the probit function are sometimes used, but not considered here.

by

$$\eta_i = \rho^{-1}(\mu_i) := \log\left(\frac{\mu_i}{1-\mu_i}\right).$$

2.1.3 Parameter estimation

The estimation is based on the criterion-function $l(\underline{\underline{x}}, \underline{y}, \underline{\beta})$, which is defined as

$$\sum_{i=1}^{n} \{y_i \log[\rho(\eta_i)] + (1-y_i)\log[1-\rho(\eta_i)]\}. \tag{2.1}$$

This function is the log-likelihood function of the logistic regression model.

The estimating equation $\underline{U}(\underline{\underline{x}}, \underline{y}, \underline{\beta})$ is given by the partial derivatives of the criterion-function with respect to $\underline{\beta}$. (See Pruscha (2000) page 188.) The components of the estimating equation are

$$U_j(\underline{\underline{x}}, \underline{y}, \underline{\beta}) := \frac{\partial\, l(\underline{\underline{x}}, \underline{y}, \underline{\beta})}{\partial \beta_j} = \underline{x}_j^{'}[\underline{y} - \underline{\rho}],$$

with $\underline{\rho}^{'} := (\rho(\eta_1), \ldots, \rho(\eta_n))$. (See Pruscha (2000) page 282.) Here the index j runs from 0 to m. The vector $\underline{x}_j$ is the $(j+1)$th column vector of the design matrix $\underline{\underline{x}}$. The estimating equation

$$\underline{U}(\underline{\underline{x}}, \underline{y}, \underline{\beta}) := \left(U_0(\underline{\underline{x}}, \underline{y}, \underline{\beta}), \ldots, U_m(\underline{\underline{x}}, \underline{y}, \underline{\beta})\right)^{'} \tag{2.2}$$

is also called score function. To achieve simple and clear formulas $\underline{U}(\underline{\beta})$ is used instead of $\underline{U}(\underline{\underline{x}}, \underline{y}, \underline{\beta})$ in the following. The root of the estimating equation (2.2) is denoted $\hat{\underline{\beta}}$. It is known as maximum likelihood estimator (MLE) of the logistic regression model.

For $m = 1$, the score function has two components. The first component is

$$\sum_{i=1}^{n} (y_i - \rho(\eta_i)) \tag{2.3}$$

and the second is

$$\sum_{i=1}^{n} x_i (y_i - \rho(\eta_i)). \tag{2.4}$$

Let $\underline{\underline{W}}(\underline{\beta})$ be the (deterministic) derivative of the estimating equation with respect to $\underline{\beta}$.

3

To derive the asymptotic distribution of the MLE for increasing sample size n, the index n is used to label sequences of vectors, matrices or estimators. If the sequence $\underline{\underline{W}}_n(\underline{\beta})$ satisfies the following condition, the root $\hat{\underline{\beta}}_n$ has a Gaussian limiting distribution,

$$\underline{\underline{\Gamma}}_n \underline{\underline{W}}_n(\hat{\underline{\beta}}_n)\underline{\underline{\Gamma}}_n \xrightarrow{\mathbb{P}} -\underline{\underline{\Sigma}}(\underline{\beta}). \tag{2.5}$$

Here

$$\underline{\underline{\Gamma}}_n := \mathrm{Diag}_{(m+1)}\left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$$

is a $(m+1) \times (m+1)$-diagonal matrix. Each diagonal element of this matrix equals $\frac{1}{\sqrt{n}}$. The matrix $\underline{\underline{\Sigma}}(\underline{\beta})$ is a covariance matrix, in particular positive semidefinite and symmetric. (See Pruscha (2000) page 288 and page 192 remark 2.)

If condition (2.5) holds, then with increasing n, a suitably scaled sequence of estimators $\hat{\underline{\beta}}_n$ converges in distribution to a normal distribution with covariance matrix $\underline{\underline{\Sigma}}(\underline{\beta})^{-1}$.

More precisely

$$\underline{\underline{\Gamma}}_n^{-1}(\hat{\underline{\beta}}_n - \underline{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}_{m+1}(\underline{0}, \underline{\underline{\Sigma}}(\underline{\beta})^{-1}).$$

The limiting covariance matrix is denoted $\underline{\underline{V}}(\underline{\beta}) := \underline{\underline{\Sigma}}(\underline{\beta})^{-1}$.

In fact,

$$\underline{\underline{W}}_n(\underline{\beta}) = -\underline{\underline{x}}_n^{'}\mathrm{Diag}_n(\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n))\underline{\underline{x}}_n. \tag{2.6}$$

(See Pruscha (2000) page 286.) Note that, for each sequence of matrices $\underline{\underline{x}}_n$ the condition (2.5) needs to be proven. (See Pruscha (2000) page 292.)

2.1.4 Example

As an example consider a model with $m = 1$ binary deterministic covariate. Define

$$x_i := \begin{cases} 1, & \text{if the index } i \text{ is divisible by three;} \\ 0, & \text{otherwise.} \end{cases}$$

[3] As above $\underline{\underline{W}}(\underline{\beta})$ stands for $\underline{\underline{W}}(\underline{x}, \underline{y}, \underline{\beta})$

In this way, for each n, the deterministic set of covariates is defined. For $n = 1$ it is $\{0\}$, for $n = 2$ it is $\{0, 0\}$, for $n = 3$ it is $\{0, 0, 1\}$ and so on. Here the relative frequency of $x_i = 1$ converges to $p = \frac{1}{3}$.

For a given data set with sample size n, define $\tilde{p} := \frac{1}{n} \sum_{i=1}^{n} x_i$. Following equation (2.6)

$$-\frac{1}{n}\underline{\underline{W}}_n(\underline{\beta}) = \begin{pmatrix} (1-\tilde{p})\pi_1(1-\pi_1) + \tilde{p}\pi_2(1-\pi_2) & \tilde{p}\pi_2(1-\pi_2) \\ \tilde{p}\pi_2(1-\pi_2) & \tilde{p}\pi_2(1-\pi_2) \end{pmatrix}.$$

Obviously, if $\tilde{p} \to p$, for $n \to \infty$, equation (2.5) holds with

$$\underline{\underline{\Sigma}}(\underline{\beta}) = \begin{pmatrix} (1-p)\pi_1(1-\pi_1) + p\pi_2(1-\pi_2) & p\pi_2(1-\pi_2) \\ p\pi_2(1-\pi_2) & p\pi_2(1-\pi_2) \end{pmatrix}.$$

Here $\pi_1 := \rho(\beta_0)$ and $\pi_2 := \rho(\beta_0 + \beta_1)$. Thus for large n, the MLE $\underline{\hat{\beta}}_n = (\hat{\beta}_{0,n}, \hat{\beta}_{1,n})'$ has the following asymptotic distribution,

$$\sqrt{n(1-p)\pi_1(1-\pi_1)}(\hat{\beta}_{0,n} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \tag{2.7}$$

and

$$\sqrt{\frac{n(1-p)\pi_1(1-\pi_1)p\pi_2(1-\pi_2)}{(1-p)\pi_1(1-\pi_1) + p\pi_2(1-\pi_2)}}(\hat{\beta}_{1,n} - \beta_1) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1). \tag{2.8}$$

2.1.5 Profile likelihood and deviance

A likelihood ratio test is a statistical test, used to compare the fit of two models. Here one of these models, the restricted model, is a special case of the other. The null hypothesis assumes that the restricted model it true. Therefore this model is denoted null model. The unrestricted model is called alternative model. Twice the logarithm of the likelihood ratio of these models is called deviance. The reason is, that the likelihood ratio expresses how many times more likely the data are under one model than the other.

To define the deviance statistic, we first define the function h, which is used to set the last $d-c$ components of the d-dimensional regression coefficient vector of the alternative model

at certain values, $c \in \mathbb{N}$. The function h maps a c-dimensional vector $\underline{\vartheta}$ to a d-dimensional vector β. The null hypotheses is:

$$H_0 : \beta_{c+1} = b_{c+1}, \ldots \beta_d = b_d.$$

The vector $\underline{\vartheta} := (\vartheta_1, \ldots, \vartheta_c)'$. The corresponding function h is

$$h(\underline{\vartheta}) = (\vartheta_1, \ldots, \vartheta_c, b_{c+1}, \ldots, b_d)'.$$

Thus the regression coefficient estimator of the null model is computed by optimizing the h-restricted likelihood-function. The corresponding regression coefficient estimator, $\hat{\underline{\vartheta}} := \text{argmax}\{l(\underline{\underline{x}}, \underline{y}, h(\underline{\vartheta}))\}$ is the h-MLE of the restricted model. The deviance (with respect to h) is defined as

$$\text{dev}_h(\underline{\underline{x}}, \underline{Y}, \hat{\underline{\beta}}, \hat{\underline{\vartheta}}) := -2 \log \left[\frac{l(\underline{\underline{x}}, \underline{Y}, h(\hat{\underline{\vartheta}}))}{l(\underline{\underline{x}}, \underline{Y}, \hat{\underline{\beta}})} \right].$$

Under regularity conditions, if the null model is true, this statistic is asymptotically χ^2-distributed with $d^* := d - c$ degrees of freedom. (See Pruscha (2000) page 288 and page 255 example (c).) More generally, the deviance is χ^2-distributed with d^* degrees of freedom if d^* components of $\underline{\beta}$ are set at their true values. It is not necessary to choose the last d^* components.

Consider the function

$$h_d(\underline{\vartheta}) := (\vartheta_1, \ldots, \vartheta_{d-1}, b_d)'.$$

This function sets the last component of the vector of covariate effects at a certain value b_d. Here $c = d - 1$. Later the function $h_j(\underline{\vartheta})$ is of special interest. This function sets, per definition, the jth component of the covariate effects at a certain value b_j. The (normalized) profile likelihood (with respect to h_j) is defined as

$$\text{prl}_j(\underline{\underline{x}}, \underline{Y}, \hat{\underline{\beta}}, b_j) := \frac{l(\underline{\underline{x}}, \underline{Y}, h_j(\hat{\underline{\vartheta}}_j))}{l(\underline{\underline{x}}, \underline{Y}, \hat{\underline{\beta}})}.$$

This profile likelihood is considered as a function of the argument b_j and takes values between 0 and 1. For each value b_j the value $\hat{\underline{\vartheta}}_j$ needs to be computed by optimizing the corresponding h_j-restricted likelihood.

2.2 Tests and confidence intervals

The asymptotic normality of the estimator $\hat{\underline{\beta}}$ and the asymptotic χ^2-distribution of the deviance statistic are used as basis for the construction of hypothesis tests and confidence intervals. Generally one can construct a confidence interval for the scalar parameter β_j by inverting the hypothesis test with null hypothesis $H_0 : \beta_j = b_j$. The confidence interval with confidence level γ, $\gamma \in (0, 1)$, consists of all those values b_j for which the test of H_0 is not rejected at a significance level of $\alpha := 1 - \gamma$.

The test function $\varphi(S_{j,n})$, yields for the realization $s_{j,n}$ of the test statistic $S_{j,n}$ either one for rejecting H_0 or zero for not rejecting H_0,

$$\varphi(s_{j,n}) := \begin{cases} 1 & s_{j,n} \in \mathrm{K}; \\ 0 & s_{j,n} \notin \mathrm{K}. \end{cases}$$

K is called rejection region. (See Rüger (1999) pages 140ff.)

2.2.1 Wald method

The Wald test of $H_0 : \beta_j = b_j$ versus $H_1 : \beta_j \neq b_j$ is based on the corresponding Wald statistic

$$T_{j,n}^{\mathrm{Wald}} := \frac{\left(\hat{\beta}_j - b_j\right)^2}{\hat{w}_{j,n}}, \tag{2.9}$$

with $\hat{w}_{j,n}$ denoting the $(j+1)$th diagonal element of $\underline{\underline{W}}_n(\hat{\underline{\beta}})^{-1}$. $T_{j,n}^{\mathrm{Wald}}$ is asymptotically χ^2-distributed with one degree of freedom. In fact, $\sqrt{\hat{w}_{j,n}}$ is an estimate of the standard error of $\hat{\beta}_j$. (See Pruscha (2000) pages 196, 197, 252, 253.) For $m = 1$

$$\hat{w}_{1,n} = \frac{1}{n}\left(\frac{(1-\tilde{p})\hat{\pi}_1(1-\hat{\pi}_1) + \tilde{p}\hat{\pi}_2(1-\hat{\pi}_2)}{(1-\tilde{p})\hat{\pi}_1(1-\hat{\pi}_1)\tilde{p}\hat{\pi}_2(1-\hat{\pi}_2)}\right). \tag{2.10}$$

Here $\hat{\pi}_1 := \rho(\hat{\beta}_{0,n})$ and $\hat{\pi}_2 := \rho(\hat{\beta}_{0,n} + \hat{\beta}_{1,n})$. (See the example in section 2.1.4 on page 14.)

The (one-dimensional) two-sided Wald interval is

$$\hat{\beta}_j - \tau_{1-\frac{\alpha}{2}}\sqrt{\hat{w}_{j,n}} \leq \beta_j \leq \hat{\beta}_j + \tau_{1-\frac{\alpha}{2}}\sqrt{\hat{w}_{j,n}},$$

with $\tau_{1-\frac{\alpha}{2}}$ as $(1-\frac{\alpha}{2})$-fractile of the standard Gaussian distribution. The confidence level $\gamma = 1-\alpha$. The corresponding test is

$$\varphi(t_{j,n}^{\text{Wald}}) := \begin{cases} 1 & \sqrt{t_{j,n}^{\text{Wald}}} > \tau_{1-\frac{\alpha}{2}}; \\ 0 & \sqrt{t_{j,n}^{\text{Wald}}} \leq \tau_{1-\frac{\alpha}{2}}, \end{cases}$$

with $t_{j,n}^{\text{Wald}}$ denoting the realization of the Wald statistic defined in equation (2.9).

2.2.2 Deviance method

Denote with $T_{j,n}$ the deviance statistic which corresponds to the hypothesis test of

$$H_0 : \beta_j = b_j$$

versus

$$H_1 : \beta_j \neq b_j.$$

Thus

$$T_{j,n} := \text{dev}_{h_j}(\underline{\underline{x}}, \underline{Y}, \underline{\hat{\beta}}, \underline{\hat{\vartheta}}) = -2\log\left[\frac{l((\underline{\underline{x}}, \underline{Y}), h_j(\underline{\hat{\vartheta}}))}{l((\underline{\underline{x}}, \underline{Y}), \underline{\hat{\beta}})}\right].$$

The corresponding hypothesis test is

$$\varphi(t_{j,n}) := \begin{cases} 1 & t_{j,n} > \kappa_{1-\alpha}^1; \\ 0 & t_{j,n} \leq \kappa_{1-\alpha}^1. \end{cases}$$

where $\kappa_{1-\alpha}^1$ denotes the $(1-\alpha)$-fractile of of the χ^2-distribution with one degree of freedom. (See Pruscha (2000) pages 248ff.) For example the profile likelihood confidence interval for component β_j, denoted as profile interval for β_j in the present work, consists of all values b_j, so that

$$t_{j,n} \leq \kappa_{1-\alpha}^1.$$

2.2.3 Approximate accuracy function

The coverage probability of a confidence interval denotes the probability that the confidence interval contains the true parameter value. This probability could be made one, if the confidence interval is only made wide enough. However, then it will also cover many "false" parameter values. An ideal confidence interval would contain the true parameter with probability one and a false parameter with probability zero. The use of the coverage probability is thus only one aspect to asses how useful a confidence interval is. This section will make use of the accuracy function to enable a more sophisticated analysis of the properties of confidence intervals. The accuracy function informs about the probability, that a particular parameter value, the argument of the function, is included in the confidence interval. Hence the probability that a wrong parameter value is included can also be considered. The accuracy function may be used to assess whether a given confidence interval method comes close to this ideal situation.

The ability of a test to reject H_0, if it is false, is measured by the power (function). In the following we give the definition of the power of a significance test $\varphi(S_{j,n})$ with test statistic $S_{j,n}$. The null hypothesis is tested $H_0 : \beta_j = b_j$ versus $H_1 : \beta_j \neq b_j$.

The power of this test is defined as

$$Q_\varphi(\beta_j, b_j, n) := \mathbb{E}_{\beta_j}[\varphi(S_{j,n})].$$

Thus for given β_j, the power function is a function of b_j. It calculates the probability that the test with null hypothesis $H_0 : \beta_j = b_j$ will reject H_0, if β_j is the true parameter value. (See Rüger (1999) page 142.)

The type I error is defined as the probability to reject H_0 if it is true, and can be described, using the power function, by

$$\alpha(\varphi) := Q_\varphi(\beta_j, \beta_j, n).$$

The coverage is given by

$$\gamma(\varphi) := 1 - Q_\varphi(\beta_j, \beta_j, n) = 1 - \alpha(\varphi).$$

As stated above a significance test, based on the test statistic $S_{j,n}$ can be inverted to obtain a confidence interval $C(S_{j,n})$. The two-sided confidence interval with confidence level γ consists of all those values b_j for which the test of $H_0 : \beta_j = b_j$ is not rejected at a significance level of $\alpha = 1 - \gamma$. Therefore the accuracy function of a two-sided confidence interval is closely linked to the power function of the corresponding test. The accuracy function is defined as

$$A_C(\beta_j, b_j, n) := \mathbb{P}_{\beta_j}(b_j \in C(S_{j,n})).$$

It is the probability, that the parameter value b_j is included in the confidence interval C, if the jth covariate effect has the value β_j. (See RÜGER (1999) pages 138ff.) The link between the power and accuracy functions is

$$A_C(\beta_j, b_j, n) + Q_\varphi(\beta_j, b_j, n) = 1$$

or alternatively

$$A_C(\beta_j, b_j, n) = 1 - Q_\varphi(\beta_j, b_j, n).$$

This equation states that, the acceptance probability of $H_0 : \beta_1 = b_1$ equals the probability that b_1 belongs to the confidence interval if β_1 is the true value.

2.2.4 Example

As an example consider again a model with $m = 1$ binary deterministic covariate. In the following the approximate accuracy function which corresponds to the Wald interval

$$\hat{\beta}_1 - \tau_{1-\frac{\alpha}{2}}\sqrt{\hat{w}_{1,n}} \leq b_1 \leq \hat{\beta}_1 + \tau_{1-\frac{\alpha}{2}}\sqrt{\hat{w}_{1,n}}, \tag{2.11}$$

is derived. (The estimate $\hat{w}_{1,n}$ is given in equation (2.10) on page 17.) Remember the definition $\underline{\underline{V}}(\underline{\beta}) := \underline{\underline{\Sigma}}(\underline{\beta})^{-1}$. Let v_j denote the $(j+1)$th diagonal element of $\underline{\underline{V}}(\underline{\beta})$. Obviously

$$\frac{1}{\hat{w}_{1,n}} \xrightarrow{\mathbb{P}} \frac{n}{v_1},$$

with

$$v_1 = \frac{(1-p)\pi_1(1-\pi_1) + p\,\pi_2(1-\pi_2)}{(1-p)\pi_1(1-\pi_1)p\pi_2(1-\pi_2)}.$$

If $\mathcal{L}$-convergence is reached,

$$\sqrt{\frac{n}{v_1}}\,(\hat{\beta}_1 - \beta_1) \sim \mathcal{N}(0,1).$$

Per definition the approximate accuracy function is the accuracy function of the confidence interval

$$\hat{\beta}_1 - \tau_{1-\frac{\alpha}{2}}\sqrt{\frac{v_1}{n}} \leq b_1 \leq \hat{\beta}_1 + \tau_{1-\frac{\alpha}{2}}\sqrt{\frac{v_1}{n}}.$$

To derive this accuracy function we need the distribution of the statistic

$$\sqrt{\frac{n}{v_1}}\,(\hat{\beta}_1 - b_1).$$

In fact, it is asymptotically normal distributed with mean $\sqrt{\frac{n}{v_1}}(\beta_1 - b_1)$ and standard deviation one. Thus, the approximate accuracy function $A_1(\beta_1, b_1, n)$ is

$$\mathbb{P}_{\beta_1}\left(|\hat{\beta}_1 - b_1| \cdot \sqrt{\frac{n}{v_1}} \leq \tau_{1-\frac{\alpha}{2}}\right) = \Phi(\tau_{1-\frac{\alpha}{2}} - \delta) - \Phi(-\tau_{1-\frac{\alpha}{2}} - \delta), \tag{2.12}$$

where Φ is the cumulative distribution function of the standard normal distribution and

$$\delta := (\beta_1 - b_1)\sqrt{\frac{n}{v_1}}.$$

It is easy to simulate the actual accuracy function of the Wald interval of interest. To do this, one simulates n_{sim}, $n_{\text{sim}} \in \mathbb{N}$, data sets and computes the Wald interval (2.11), that is

$$\left[\hat{\beta}_1 - \tau_{1-\frac{\alpha}{2}}\sqrt{\hat{w}_{1,n}}, \hat{\beta}_1 + \tau_{1-\frac{\alpha}{2}}\sqrt{\hat{w}_{1,n}}\right].$$

The simulated accuracy function at b_1 is the relative frequency of the simulated Wald intervals containing b_1. If convergence is reached at a certain value b_1, the simulated accuracy function and $A_1(\beta_1, b_1, n)$ nearly coincide. If one bases statistical conclusions on large-sample assumptions one assumes implicitly that the accuracy of the Wald interval of β_1 equals (nearly) the approximate accuracy function $A_1(\beta_1, b_1, n)$. Thus the comparison of the approximate accuracy and the actual accuracy is informative. On the one hand, such a comparison may reveal whether the sample size is large enough to rely on asymptotic

methods. On the other hand, if the sample size is too small, it informs about the effect of the mistaken large-sample assumptions. For example, it can indicate that the probability that the Wald interval (2.11) contains wrong parameter values is significantly higher than the probability that it includes the true parameter value.

In fact, the comparison of the actual accuracy function of the Wald interval and the approximate accuracy function of the Wald interval is very interesting. Therefore it is desirable to derive the approximate accuracy function of the profile interval. Per definition this accuracy is the accuracy based on the assumption, that $T_{1,n}$ follows exactly its limiting distribution. Under H_0, this limiting distribution is the central χ^2-distributed with one degree of freedom. It turns out, that a closed form solution of this accuracy does not exist, but the approximate Wald-accuracy function and the approximate profile-accuracy function coincide near the true parameter value. This holds because, under the local alternative hypotheses $H_1 : b_1 = \beta_1 + \frac{r}{\sqrt{n}}$, $r \in \mathbb{R}$, the deviance $T_{1,n}$ is asymptotically χ^2-distributed with noncentrality parameter

$$\frac{n}{v_1}(\beta_1 - b_1)^2.$$

(See Pruscha (2000) page 250 and page 254.) Note that the approximate Wald-accuracy function given in equation (2.12) equals

$$\mathbb{P}_{\beta_1}\left(\sqrt{\frac{n}{v_1}}|\hat{\beta}_1 - b_1| \leq \tau_{1-\frac{\alpha}{2}}\right) = \mathbb{P}_{\beta_1}\left(\frac{n}{v_1}(\hat{\beta}_1 - b_1)^2 \leq \kappa^1_{1-\alpha}\right).$$

Section 6.5 will give an idea of the comparison of the approximate and the actual accuracy function.

3 Autogenerated process

3.1 Poisson process

The mathematic definition of the univariate and multivariate Poisson process follows Zocher (2005).

3.1.1 Counting process

Let $(\Omega, \mathcal{F}, P)$ be the underlying probability space of the stochastic process $\{N_\tau\}_{\tau \in \mathbb{R}_+}$. The process N_τ[1] is a counting process, if there exists a null set $M \in \mathcal{F}$ such that the following properties hold for every $\omega \in \Omega \setminus M$:

1. $N_0(\omega) = 0$
2. $N_\tau(\omega) \in \mathbb{N}_0$ for all $t > 0$,
3. $N_\tau(\omega) = \inf_{s \in (\tau, \infty)} N_s(\omega)$ for all $\tau > 0$,
4. $\sup_{s \in [\tau, \infty)} N_s(\omega) \leq N_\tau(\omega) \leq \sup_{s \in [\tau, \infty)} N_s(\omega) + 1$ for all $\tau > 0$ and
5. $\sup_{\tau \in \mathbb{R}} N_\tau(\omega) = \infty$.

A counting process N_τ is usually understood as the number of events occurring in an observation interval $(0, \tau]$. For example, the number of childbirths in a hospital in a given time interval $(0, \tau]$ is a counting process.

[1] N_τ is used for as an abbreviation of $\{N_\tau\}_{\tau \in \mathbb{R}_+}$,

3.1.2 Multivariate Poisson process

A multivariate counting process in k dimensions counts k different types of events, $k \in \mathbb{N}$. For instance, one could count the number of births of boys and girls in a hospital in a given time interval $(0, \tau]$. Here $k = 2$. Per definition, a multivariate stochastic process N_τ, in k dimensions is called a multivariate counting process if every coordinate $N_\tau^{(l)}$, $l \in \{1, \ldots, k\}$, is a counting process and moreover the sum $\sum_{i=l}^{k} N_\tau^{(l)}$ of all coordinates is a counting process.

If N_τ is a multivariate counting process the following conditions are equivalent:

1. N_τ is a multivariate Poisson process

2. N_τ has stationary and independent increments and

$$\mathbb{P}(N_\tau^{(1)} = n_1, \ldots, N_\tau^{(k)} = n_k) = \exp\Big(\sum_{l=1}^{k} \lambda_l \tau\Big) \frac{\prod_{l=1}^{k} (\lambda_l \tau)^{n_l}}{\prod_{l=1}^{k} n_l!},$$

with $\lambda_l \in \mathbb{R}_+$. The parameter λ_l is called the intensity of the process $N_\tau^{(l)}$. The total intensity is denoted $\Lambda := \sum_{l=1}^{k} \lambda_l$.

3.1.3 Limit theorem for Poisson processes

A Poisson process is a recurrent renewal process. Thus

1.
$$\lim_{\tau \to \infty} \frac{N_\tau}{\tau} = \Lambda.$$

2.
$$\lim_{\tau \to \infty} \mathbb{P}\Big(\frac{N_\tau - \Lambda\, \tau}{\Lambda\, \tau} \leq s\Big) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{s} \exp\big(-\frac{r^2}{2}\big) d\, r,$$

with $s, r \in \mathbb{R}$. (See Lange (1982) pages 140 and 142ff.) Correspondingly for each component

3.
$$\lim_{\tau \to \infty} \frac{N_\tau^{(l)}}{\tau} = \lambda_l. \tag{3.1}$$

and

4.
$$\lim_{\tau\to\infty} \mathbb{P}\Big(\frac{N_\tau^{(l)} - \lambda_l\,\tau}{\lambda_l\,\tau} \leq s\Big) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{s} \exp\big(-\frac{r^2}{2}\big) d\,r, \tag{3.2}$$

holds.

3.2 Autogenerated process

In section 2.1 the logistic regression model is defined as a framework for modelling a binary response Y conditional on a set of m deterministic explanatory variables $x_1, \ldots, x_j, \ldots, x_m$. In this section the autogenerated process is introduced as an idealized stochastic model for an event that occurs randomly in the time interval $(0, \tau]$. Here, the question is not whether an event occurs, but rather a distinction is made between two possible event types. The event type is described by the binary response variable y of the logistic regression model. If, for example, the event is voting, the event type can be the vote for a certain party,

$$y := \begin{cases} 1, & \text{vote for the party of interest;} \\ 0, & \text{vote for an other party.} \end{cases}$$

The event type is assumed to depend on a linear combination of m binary explanatory variables, possibly also including their interactions.

3.2.1 Grouped data

To provide a better basis for comprehension, the data is now grouped. Given that some of the rows of the design matrix $\underline{\underline{x}}$ in section 2.1 have identical covariate values, grouping is possible. After relabeling the index i, this index denotes group i instead of individual i. So only rows $\underline{x}_i^{'}$ with different combinations of covariate values appear in the design matrix $\underline{\underline{x}}$. For each row $\underline{x}_i^{'}$ the numbers of both event types are given. Here the group index i runs from 1 to $g = 2^m$. The covariates are denoted $x_{(j)}$ instead of x_j. [2] The elements of $\underline{x}_i^{'}$ are $(x_{i,0}, \ldots, x_{i,j}, \ldots x_{i,g})$. The index j runs from 0 to m. Per definition $x_{(0)} = 1$. Hence the dimension of the design matrix is $2^m \times (m+1)$, if no interaction is included.

The pattern, defining which combination of covariate values corresponds to a certain index i, still remains to be specified, depending on m. For $m = 2$

$$\underline{\underline{x}} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix},$$

while for $m = 3$

$$\underline{\underline{x}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

This pattern defines the covariate values which correspond to index i. If for example $m = 3$ covariates are included in the model specification, the fifth row of $\underline{\underline{x}}$ is $(1, 0, 0, 1)$. Thus, in this case, $\underline{x}_5^{'} = (1, 0, 0, 1)$. If $m \geq 2$, it is possible to include interaction effects in the

[2] For grouped data vector $\underline{x}_i$ is the combination of covariate values corresponding to group index i. For $m = 1$, $\underline{x}_1^{'} = (0, 0)$. Indeed, x_1 is the realization of X_1, $x_1 \in \{0, 1\}$. However $\underline{x}_1^{'}$ and x_1 could easily be mistaken. Thus to avoid misunderstandings the index j is placed in parentheses.

model specification. The definition of design matrices given above can easily be expanded to include additional columns corresponding to products of covariate values, by which interaction effects can be modeled.

For $m = 3$ the linear predictor including all pairwise and higher-order interactions is

$$\eta(\underline{x}_i, \underline{\beta}) := \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_{1,2} x_{i,1} x_{i,2} + \beta_{1,3} x_{i,1} x_{i,3} + \beta_{2,3} x_{i,2} x_{i,3} + \beta_{1,2,3} x_{i,1} x_{i,2} x_{i,3}.$$

The corresponding design matrix is

$$\underline{\underline{x}} = \left(\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{array}\right). \tag{3.3}$$

3.2.2 Definition of the autogenerated process

The autogenerated process is based on a multi-dimensional Poisson process. In the following the intensity of this time-homogeneous process is given in dependence of the binary response y and the combination of covariate values $\underline{x}_i'$. The number of observations of $(\underline{x}_i', y)$ in the time interval $(0, \tau]$ is denoted as $N_\tau^{(i,y)}$. The index τ will not be used if it is not necessary for comprehension, which might be the case, e. g. to distinguish different observation periods. Per definition of the autogenerated process $N^{(i,y)}$ is Poisson distributed with expectation

$$\mathbb{E}\big(N^{(i,y)}\big) = \lambda_y(i) \cdot \tau.$$

Here $\lambda_y(i)$ is the intensity of the Poisson process which depends on the covariate group i and the response y. It is defined below in equation (3.4). The random variable

$$N^{(i,\,.\,)} := \sum_{y=0}^{1} N^{(i,y)}$$

is denoted by $N^{(i)}$. Its realization is $n^{(i)}$. The sum

$$\sum_{i=1}^{g} n^{(i)}$$

is the observed sample size n. This process can, for example, model the electoral behavior of a population of students. In this fictive example students, who enter the cafeteria in the time interval $(0, \tau]$ are asked: "Would you vote for the *party of improvement* (PI) if federal elections were held this sunday? Here

$$x_{(1)} := \begin{cases} 1, & \text{student of natural sciences;} \\ 0, & \text{other student} \end{cases}$$

and

$$x_{(2)} := \begin{cases} 1, & \text{female student;} \\ 0, & \text{male student.} \end{cases}$$

The response

$$y := \begin{cases} 1, & \text{student would vote for PI;} \\ 0, & \text{student would not vote for PI.} \end{cases}$$

Here $N^{(2,1)}$ is the random number of males studying natural sciences who enter the cafeteria in $(0, \tau]$ and would vote for PI. Correspondingly $N^{(2)}$ is the random number of male students of natural sciences entering the cafeteria in $(0, \tau]$. Its realization is $n^{(2)}$.

The following Poisson process defines the autogenerated process. It consists of $2^{(m+1)}$ stochastically independent time-homogeneous Poisson processes with intensities

$$\lambda_y(i, \underline{\beta}, \underline{\theta}) := \begin{cases} \theta_i & \text{y} = 0; \\ \theta_i \exp(\eta(\underline{x}_i, \underline{\beta})) & \text{y} = 1. \end{cases} \tag{3.4}$$

The vector $\underline{\theta}$ contains a weight θ_i for each covariate group i. Thus it has length $g = 2^m$. Per definition of $\underline{\theta}$

$$\sum_{i=1}^{g} \theta_i = 1$$

holds. To achieve simple and clear formulas $\lambda_y(i)$ is used instead of $\lambda_y(i, \underline{\beta}, \underline{\theta})$. Respectively Λ corresponds to $\Lambda(\underline{\beta}, \underline{\theta})$. Recall $\Lambda := \sum_{i=1}^{g} \sum_{y=0}^{1} \lambda_y(i)$. Note that equation (3.4) contains a restriction on the intensities, in fact $\lambda_0(.) = 1$. Hence equation (3.4) is a bijective function which maps the covariate effects $\underline{\beta}$ and the weight parameter vector $\underline{\theta}$ on the intensities $\lambda_y(i)$. Consider for example $m = 1$. The corresponding Poisson process has three intensities $\lambda_0(1)$, $\lambda_1(1)$, and $\lambda_1(2)$. The fourth parameter is then $\lambda_0(2) = 1 - \lambda_0(1)$.

The inverse map is

$$\theta_1 = \lambda_0(1).$$

The covariate effect $\underline{\beta} := (\beta_0, \beta_1)'$ is

$$\left(\log\left(\frac{\lambda_1(1)}{\lambda_0(1)}\right), \log\left(\frac{\lambda_0(1)}{\lambda_1(1)} \cdot \frac{\lambda_1(2)}{\lambda_0(2)}\right)\right).$$

Within the autogenerated process the waiting time between the lth and the $(l+1)$th event is exponentially distributed with rate Λ. The probability that the waiting time between the lth and the $(l+1)$th event is equal to or less than τ and that the $(l+1)$th event is of

type $(\underline{x}_i, y)$ is

$$\frac{\lambda_y(i)}{\Lambda}(1 - \exp[-\Lambda\,\tau]).$$

This model actually gives rise to several interrelated random processes, in particular to the sequence of arrival times $\mathcal{T} \equiv \mathcal{T}_n$ and a counting process $N \equiv N_\tau$ [3]. The arrival time process $\mathcal{T}$ is here the random duration until the nth event occurs. The counting process

$$N := \sum_{y=0}^{1} \sum_{i=1}^{g} N^{(i,y)}$$

counts the events until t. The arrival time process $\mathcal{T}_n$ and the counting process N_τ are inversions of one another. This means the following:

- $\mathcal{T}_n \leq \tau$ if and only if $N_\tau \geq n$. This means that there are at least n arrivals in $(0, \tau]$.
- $N_\tau = n$ if and only if $\mathcal{T}_n \leq \tau < \mathcal{T}_{n+1}$. This means that there are exactly n arrivals in $(0, \tau]$.

The autogenerated process is used to derive two sampling protocols. In the first protocol, the sample size n is predefined and thus the duration of the fictive study $\mathcal{T}_n$ is random. In the second sampling scheme the sample size N_τ is random while the observation time τ is predefined. The former protocol is denoted fixed sample size protocol, and the latter fixed time protocol. Both sampling schemes represent two types of observational studies with two different criteria to stop data collection.

For fixed sample size n, let $Z_n^{(i,y)}$ denote the number of observations of $(\underline{x}_i', y)$. Thus $Z_n^{(i,y)}$ corresponds to $N_\tau^{(i,y)}$. Again, the index n is not used if it is not necessary. The observed units can be stored in a data set in temporal order. For the fixed sample size protocol the data set is given in table 3.1.

Per definition

$$\sum_{y=0}^{1} \sum_{i=1}^{g} z^{(i,y)} = n.$$

Consider the simple case in which only one covariate is included in the model specification. Here $g = 2^1 = 2$. Thus $i \in \{1, 2\}$ and $\underline{x}_1' = (0, 0)$ and $\underline{x}_2' = (0, 1)$. $N^{(i,y)}$ is the number of

[3] The index n is not used, if it is not necessary for comprehension for example to distinguish between $\mathcal{T}_{n_1}$ and $\mathcal{T}_{n_2}$, $n_1 \neq n_2$, $n_1, n_2 \in \mathbb{N}$.

Data set			
Variables	$\underline{X}_i^{'}$	$Z^{(i,1)}$	$Z^{(i,0)}$
Observation	$\underline{x}_1^{'}$	$z^{(1,1)}$	$z^{(1,0)}$
	$\underline{x}_2^{'}$	$z^{(2,1)}$	$z^{(2,0)}$
	$\vdots$	$\vdots$	$\vdots$
	$\underline{x}_i^{'}$	$z^{(i,1)}$	$z^{(i,0)}$
	$\vdots$	$\vdots$	$\vdots$
	$\underline{x}_g^{'}$	$z^{(g,1)}$	$z^{(g,0)}$

Table 3.1: Data set observed under the fixed sample size protocol.

observations of mark $(\underline{x}_i^{'}, y)$ in the observation period $(0, \tau]$. By definition of the autogenerated process, $N^{(i,y)}$ follows a Poisson distribution with expectation $\lambda_y(i) \cdot t$. The data observed in the period $(0, \tau]$ can be arranged in a two-by-two table:

		y = 1	y = 0	
x	1	$N^{(1,1)}$	$N^{(1,0)}$	$N^{(1)}$
	2	$N^{(2,1)}$	$N^{(2,0)}$	$N^{(2)}$
		$N^{(\cdot,1)}$	$N^{(\cdot,0)}$	N

The data observed until the nth event occurs can be, in the same way, arranged in a two-by-two table:

		y = 1	y = 0	
x	1	$Z^{(1,1)}$	$Z^{(1,0)}$	$Z^{(1)}$
	2	$Z^{(2,1)}$	$Z^{(2,0)}$	$Z^{(2)}$
		$Z^{(\cdot,1)}$	$Z^{(\cdot,0)}$	n

Here $Z^{(i)} := Z^{(i,\cdot)}$. The data matrix

$$\underline{\underline{Z}} := \begin{pmatrix} Z_n^{(1,1)} & Z_n^{(1,0)} \\ Z_n^{(2,1)} & Z_n^{(2,0)} \end{pmatrix},$$

follows a multinomial distribution with parameters n and probability matrix

$$\underline{\underline{\mathfrak{p}}} := \frac{1}{\Lambda} \begin{pmatrix} \lambda_1(1) & \lambda_0(1) \\ \lambda_1(2) & \lambda_0(2) \end{pmatrix}.$$

The random variable $Z^{(i)}$ is binomially distributed with number of successes n and success probability $\frac{\lambda_{\cdot}(i)}{\Lambda}$. The conditional distribution $Z^{(i,1)}$ given that $Z^{(i)} = z^{(i)}$ is binomially distributed with number of successes $z^{(i)}$ and success probability[4]

$$\frac{\lambda_1(i)}{\lambda_{\cdot}(i)} = \frac{\exp(\eta_i)}{1+\exp(\eta_i)} = \pi_i.$$

with $\eta_i := \underline{x}_i'\underline{\beta}$. Here $\underline{\beta} := (\beta_0, \beta_1)$.

The two sampling protocols defined above are stochastically proportional. Given that

$$\mathbb{E}(N) = \Lambda \cdot \tau,$$

the data sets observed under the fixed sample size protocol with predefined sample size n and the fixed time protocol with predefined time $\tau = \frac{\Lambda}{n}$ are expected to be similar.

3.3 Large sample properties

In chapter 2 the parameter estimation and the derivation of large-sample properties are based on properties of the likelihood and the score function i.e. the derivative of the logarithm of the likelihood function with respect to the covariate effect vector $\underline{\beta}$. To derive the large-sample properties of the covariate effect estimator and the deviance within the autogenerated process a more general approach is needed. Indeed, the estimating equation approach extends the likelihood approach. Within the likelihood approach the covariate effect estimate is, per definition, the root of the score function. By contrast, the estimating equation approach bases the parameter estimation directly on an estimating equation . This estimating equation is not necessarily the derivative of the log-likelihood function, but the derivative of a criterion-function. The criterion-function is a function which involves the random number of observations in the respective covariate groups, $N^{(i,y)}$ or $Z^{(i,y)}$, and the parameter vector $\underline{\beta}$. The dimension of the estimating equation is $m+1$, if no interaction effect is included in the specification of the linear predictor. If this equation satisfies certain conditions, its root has a normal limiting distribution (following the theorem 3.4 in Pruscha (2000) on page 194).

[4] This corresponds to the definition of π_i in the example of section 2.1.4 on page 14.

In the following the asymptotic distribution of the root of the estimating equation is derived, first for the fixed time protocol for $\tau \to \infty$. This root is called the covariate effects estimator $\underline{\hat{\beta}}_\tau$. To begin with, the asymptotic distribution of $\underline{\hat{\beta}}_\tau$ is derived for $m = 1$. Then, the derivation is generalized to the case of $m = 3$ covariates. Following this, the asymptotic distribution of the deviance statistic $T_{j,\tau}$ is derived. Afterwards, the same proof is given for the fixed sample size protocol. Here the estimator depends on the sample size n and is denoted $\underline{\hat{\beta}}_n$. Recall that in section 2.2.2 the parameter estimation is based on the binomial distributed response. Within the fixed sample size protocol the observed data is multinomial distributed. However in both cases the sample size n is deterministic, so that the estimator is denoted $\underline{\hat{\beta}}_n$. Indeed, these two estimators are roots of different estimating equations. In section 2.2.2 the likelihood approach is used wheres this section uses the estimating equation-approach. Thus, in this chapter $\underline{\hat{\beta}}_n$ is always the root of the estimating equation corresponding to the fixed sample size protocol.

The statistic $T_{j,\tau}$ is defined in accordance with $T_{j,n}$ (defined in section 2.2.2).

$$T_{j,\tau} := \mathrm{dev}_{h_j}(\underline{\underline{x}}, \underline{y}, \underline{\hat{\beta}}_\tau, \underline{\hat{\vartheta}}_\tau).$$

3.3.1 Fixed time protocol with one covariate

In the following, the asymptotic distribution of $\underline{\hat{\beta}}_\tau$ is derived for the fixed time protocol with $m = 1$ for $\tau \to \infty$. For $m = 1$ the autogenerated process corresponds to the following four-dimensional Poisson process:

$$\lambda_y(i, \underline{\beta}, \underline{\theta}) := \begin{cases} \theta_i & \mathrm{y} = 0; \\ \theta_i \exp(\eta_i) & \mathrm{y} = 1, \end{cases}$$

with $\eta_i := \underline{x}_i' \underline{\beta}$ and $i \in \{1, 2\}$. Here $\underline{x}_1' = (1, 0)$ and $\underline{x}_2' = (1, 1)$. The covariate effect vector is $\underline{\beta} = (\beta_0, \beta_1)'$ and the weight parameter vector is $\underline{\theta} = (\theta_1, \theta_2)'$ with $\theta_i \in (0, 1)$, $\theta_1 + \theta_2 = 1$.

To derive the asymptotic distribution of $\underline{\hat{\beta}}_\tau$, we show that the conditions of theorem 3.4 on page 194 in Pruscha (2000) hold. This theorem states, that under certain regularity conditions, a $\underline{\underline{\Gamma}}_\tau^{-1}$-consistent zero-estimator is asymptotic normal distributed. A $\underline{\underline{\Gamma}}_\tau^{-1}$-consistent zero-estimator $\underline{\hat{\beta}}_\tau$ of the parameter $\underline{\beta}$ is a sequence of random variables fulfilling

(see

Pruscha (2000) page 189)

1.
$$\lim_{\tau\to\infty} \mathbb{P}(\underline{U}_\tau(\hat{\underline{\beta}}_\tau) = \underline{0}) = 1$$

and

2.
$$\underline{\underline{\Gamma}}_\tau^{-1}(\hat{\underline{\beta}}_\tau - \underline{\beta}) \tag{3.5}$$

is $\mathbb{P}$-stochastically bounded.

The scaling matrix $\underline{\underline{\Gamma}}_\tau$ is by definition

$$\underline{\underline{\Gamma}}_\tau := \begin{pmatrix} \frac{1}{\sqrt{\tau}} & 0 \\ 0 & \frac{1}{\sqrt{\tau}} \end{pmatrix}.$$

Let $\tilde{\underline{\beta}}_\tau$ be an estimator fulfilling the condition (3.5)

$$\underline{\underline{\Gamma}}_\tau^{-1}(\tilde{\underline{\beta}}_\tau - \underline{\beta}) \tag{B^*}$$

is $\mathbb{P}$-stochastically bounded. In particular, the estimator $\tilde{\underline{\beta}}_\tau$ converges $\mathbb{P}$-stochastically to $\underline{\beta}$.

First, we show that condition U^*, that is

$$\underline{\underline{\Gamma}}_\tau \underline{U}_\tau(\underline{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}_2(\underline{0}, \underline{\underline{\Xi}}(\underline{\beta})), \tag{U^*}$$

holds. Here $\underline{\underline{\Xi}}(\underline{\beta})$ is a positive semidefinite and symmetric 2×2-matrix. The estimating equation

$$\underline{U}_\tau(\underline{\beta}) := \big(U_{0,\tau}(\underline{\beta}), U_{1,\tau}(\underline{\beta})\big)'$$

is defined in accordance with the score function of the logistic regression model, see equations (2.3) and (2.4) of section 2.1.3. As stated above it is not necessary that this estimating equation is the derivative of the log-likelihood function. If the estimating equation $\underline{U}_\tau(\underline{\beta})$

satisfies the required regularity conditions U^* and W^*, [5] a $\underline{\underline{\Gamma}}_{\tau}^{-1}$-consistent zero-estimator $\underline{\hat{\beta}}_{\tau}$ is asymptotically normal distributed. The first component of $\underline{U}_{\tau}(\underline{\beta})$ is

$$U_{0,\tau}(\underline{\beta}) := -N^{(1,0)}\pi_1 + N^{(1,1)}(1-\pi_1) - N^{(2,0)}\pi_2 + N^{(2,1)}(1-\pi_2)$$

and the second component is

$$U_{1,\tau}(\underline{\beta}) := -N^{(2,0)}\pi_2 + N^{(2,1)}(1-\pi_2).$$

Here $\pi_1 := \rho(\beta_0)$ and $\pi_2 := \rho(\beta_0 + \beta_1)$. First, we calculate the expected value and the covariance matrix of the estimating equation $U_{\tau}(\underline{\beta})$ in dependence of τ. Then we use the central limit theorem for Poisson processes in combination with the Continuous Mapping Theorem (CMT) (see Pruscha (2000) p. 346) to proof U^*. The expected value $\mathbb{E}\{U_{0,\tau}(\underline{\beta})\} = 0$. The same is true for $\mathbb{E}\{U_{1,\tau}(\underline{\beta})\}$.

The expectation of the first component is

$$\begin{aligned}\mathbb{E}\{U_{0,\tau}(\underline{\beta})\} = \mathbb{E}\Big\{ &- N^{(1,0)}\frac{\exp(\beta_0)}{1+\exp(\beta_0)} + N^{(1,1)}\frac{1}{1+\exp(\beta_0)} \\ &- N^{(2,0)}\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} + N^{(2,1)}\frac{1}{1+\exp(\beta_0+\beta_1)}\Big\} = \\ = \tau\Big[&- (1-\theta_2)\frac{\exp(\beta_0)}{1+\exp(\beta_0)} + (1-\theta_2)\exp(\beta_0)\frac{1}{1+\exp(\beta_0)} \\ -\theta_2&\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} + \theta_2\exp(\beta_0+\beta_1)\frac{1}{1+\exp(\beta_0+\beta_1)}\Big] = 0.\end{aligned} \tag{3.6}$$

In the same way

$$\mathbb{E}\{U_{1,\tau}(\underline{\beta})\} = 0. \tag{3.7}$$

The covariance of the estimating equation is

$$\underline{\underline{\Xi}}_{\tau}(\underline{\beta}) := \tau \begin{pmatrix} \pi_1\,(1-\theta_2) + \pi_2\,\theta_2 & \pi_2\,\theta_2 \\ \pi_2\,\theta_2 & \pi_2\,\theta_2 \end{pmatrix}.$$

Given that

$$\mathbb{V}\{\pi_i\, N^{(i,y)}\} = \pi_i^2\,\lambda_y(i)\,\tau,$$

[5] defined on page 36

$$\frac{1}{\tau}\mathbb{V}\{U_{1,\tau}(\underline{\beta})\} = \pi_2^2\lambda_0(2) + \left(1-\pi_2\right)^2\lambda_1(2) =$$
$$\pi_2^2\lambda_\cdot(2) + (1-2\pi_2)\lambda_1(2) =$$
$$\pi_2^2\lambda_\cdot(2) + (\pi_2 - 2\pi_2^2)\lambda_\cdot(2) =$$
$$\pi_2(1-\pi_2)\lambda_\cdot(2) = \theta_2\pi_2.$$

Moreover, obviously

$$\mathbb{C}ov\{U_{0,\tau}, U_{1,\tau}\} = \mathbb{V}\{U_{1,\tau}(\underline{\beta})\}$$

holds.

Thus, due to the central limit theorem (3.2) in combination with the CMT, condition U^* holds with

$$\underline{\underline{\Xi}}(\underline{\beta}) := \frac{1}{\tau}\underline{\underline{\Xi}}_\tau(\underline{\beta}).$$

The condition W^* is

$$\underline{\underline{\Gamma}}_\tau\underline{\underline{W}}_\tau(\underline{\tilde{\beta}}_\tau)\underline{\underline{\Gamma}}_\tau \xrightarrow{\mathbb{P}} -\underline{\underline{\tilde{\Xi}}}. \qquad (W^*)$$

Here $\underline{\underline{\tilde{\Xi}}}$ is a positive semidefinite and symmetric 2×2-matrix and

$$\underline{\underline{W}}_\tau(\underline{\beta}) := \frac{\partial \underline{U}_\tau(\underline{\beta})}{\partial \underline{\beta}}(\underline{\beta}).$$

Thus

$$\underline{\underline{\Gamma}}_\tau\underline{\underline{W}}_\tau(\underline{\beta})\underline{\underline{\Gamma}}_\tau = \frac{1}{\tau}\underline{\underline{W}}_\tau(\underline{\beta}) =$$

$$-\frac{1}{\tau}\begin{pmatrix} N^{(1)}\pi_1(1-\pi_1) + N^{(2)}\pi_2(1-\pi_2) & N^{(2)}\pi_2(1-\pi_2) \\ N^{(2)}\pi_2(1-\pi_2) & N^{(2)}\pi_2(1-\pi_2) \end{pmatrix}.$$

Following the law of large numbers (3.1)

$$\underline{\underline{\Gamma}}_\tau\underline{\underline{W}}_\tau(\underline{\beta})\underline{\underline{\Gamma}}_\tau \xrightarrow{\mathbb{P}} -\underline{\underline{\Xi}}(\underline{\beta}).$$

The CMT is used to derive the limit of

$$\underline{\underline{\Gamma}}_\tau\underline{\underline{W}}_\tau(\underline{\tilde{\beta}}_\tau)\underline{\underline{\Gamma}}_\tau = \frac{1}{\tau}\underline{\underline{W}}_\tau(\underline{\tilde{\beta}}_\tau).$$

$$\frac{1}{\tau}\underline{\underline{W}}_\tau(\underline{\tilde{\beta}}_\tau) = -\begin{pmatrix} \left[\frac{N^{(1)}\exp(\tilde{\beta}_{0,\tau})}{(1+\exp(\tilde{\beta}_{0,\tau}))^2} + \frac{N^{(2)}\exp(\tilde{\beta}_{0,\tau}+\tilde{\beta}_{1,\tau})}{(1+\exp(\tilde{\beta}_{0,\tau}+\tilde{\beta}_{1,\tau}))^2}\right] & \frac{N^{(2)}\exp(\tilde{\beta}_{0,\tau}+\tilde{\beta}_{1,\tau})}{(1+\exp(\tilde{\beta}_{0,\tau}+\tilde{\beta}_{1,\tau}))^2} \\ \frac{N^{(2)}\exp(\tilde{\beta}_{0,\tau}+\tilde{\beta}_{1,\tau})}{(1+\exp(\tilde{\beta}_{0,\tau}+\tilde{\beta}_{1,\tau}))^2} & \frac{N^{(2)}\exp(\tilde{\beta}_{0,\tau}+\tilde{\beta}_{1,\tau})}{(1+\exp(\tilde{\beta}_{0,\tau}+\tilde{\beta}_{1,\tau}))^2} \end{pmatrix}.$$

Following the CMT, $\frac{\exp(\tilde{\beta}_{0,\tau})}{(1+\exp(\tilde{\beta}_{0,\tau}))^2}$ converges in probability to $\frac{\exp(\beta_0)}{(1+\exp(\beta_0))^2} = \pi_1(1-\pi_1)$. Hence, due to the Cramér-Slutzky theorem (see Pruscha (2000) p. 346),

$$\frac{N^{(1)}}{\tau}\frac{\exp(\tilde{\beta}_{0,\tau})}{(1+\exp(\tilde{\beta}_{0,\tau}))^2} \xrightarrow{\mathbb{P}} \lambda_\cdot(1)\pi_1(1-\pi_1) = (1-\theta_2)\,\pi_1.$$

So

$$\underline{\underline{\Gamma}}_\tau\underline{\underline{W}}_\tau(\underline{\tilde{\beta}}_\tau)\underline{\underline{\Gamma}}_\tau \xrightarrow{\mathbb{P}} -\underline{\underline{\Xi}}(\underline{\beta}).$$

Thus W^* holds with $\underline{\underline{\tilde{\Xi}}} = \underline{\underline{\Xi}}$.

Given that $\underline{\hat{\beta}}_\tau$ is the unique root of the estimating equation, following remark 2 on page 192 in Pruscha (2000),

$$\underline{\underline{\Gamma}}_\tau^{-1}(\underline{\hat{\beta}}_\tau - \underline{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}_2(\underline{0}, \underline{\underline{\Xi}}(\underline{\beta})^{-1}).$$

3.3.2 Fixed time protocol with three covariates

The derivation of the asymptotic distribution of the root of the estimating equation for $m = 3$ covariates is completely analogous to the proof given above. The dimension of the estimating equation is $m + 1 = 4$ if the model specification does not contain interaction effects. The dimension equals eight if all possible interactions are taken into account. As above the estimating equation equals formally the score function of the logistic regression model with the same model specification. To begin with, we give this estimating equation if all pairwise and higher-order interactions are taken into account. In this case, the linear predictor η_i is

$$\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_{1,2} x_{i,1} x_{i,2} + \beta_{1,3} x_{i,1} x_{i,3} + \beta_{2,3} x_{i,2} x_{i,3} + \beta_{1,2,3} x_{i,1} x_{i,2} x_{i,3}.$$

Again, $\pi_i = \frac{\exp(\eta_i)}{\exp(1+\eta_i)}$. The corresponding design matrix is given in equation (3.3).

The estimating equation is a vector $U_\tau(\underline{\beta})$ with length eight. The first component is

$$U_{0,\tau}(\underline{\beta}) = \sum_{i=1}^{8} -N^{(i,0)}\,\pi_i + N^{(i,1)}\,(1-\pi_i).$$

The second component of the estimating equation is

$$U_{1,\tau}(\underline{\beta}) = \sum_{i\in\{2,4,6,8\}} -N^{(i,0)}\,\pi_i + N^{(i,1)}\,(1-\pi_i).$$

Here the index set corresponds to the rows of the design matrix with $x_{(1)} = 1$. The index set of the third component contains all indices with $x_{(2)} = 1$. It is

$$U_{2,\tau}(\underline{\beta}) = \sum_{i\in\{3,4,7,8\}} -N^{(i,0)}\,\pi_i + N^{(i,1)}\,(1-\pi_i).$$

In the same way

$$U_{3,\tau}(\underline{\beta}) = \sum_{i\in\{5,6,7,8\}} -N^{(i,0)}\,\pi_i + N^{(i,1)}\,(1-\pi_i).$$

The next component corresponds to the interaction effect $\beta_{1,2}$. Thus it is denoted $U_{(1,2),\tau}$. It is

$$U_{(1,2),\tau}(\underline{\beta}) = \sum_{i\in\{4,8\}} -N^{(i,0)}\,\pi_i + N^{(i,1)}\,(1-\pi_i).$$

Correspondingly

$$U_{(1,3),\tau}(\underline{\beta}) = \sum_{i\in\{6,8\}} -N^{(i,0)}\,\pi_i + N^{(i,1)}\,(1-\pi_i),$$

$$U_{(2,3),\tau}(\underline{\beta}) = \sum_{i\in\{7,8\}} -N^{(i,0)}\,\pi_i + N^{(i,1)}\,(1-\pi_i)$$

and

$$U_{(1,2,3),\tau}(\underline{\beta}) = -N^{(8,0)}\,\pi_8 + N^{(8,1)}\,(1-\pi_8).$$

In fact, as described above, for this estimating equation the corresponding conditions U^* and W^* hold[6].

The estimator $\hat{\underline{\beta}}_\tau$ is asymptotic normally distributed. In fact,

$$\underline{\underline{\Gamma}}_\tau^{-1}\left(\hat{\underline{\beta}}_\tau - \underline{\beta}\right) \xrightarrow{\mathcal{L}} \mathcal{N}_8\left(\underline{0}, \underline{\underline{\Xi}}^{-1}\right)$$

[6] These conditions are defined on pages 34 and 34 respectively.

holds, for $t \to \infty$. Due to the proposition 4.2 in PRUSCHA (2000) on page 252,

$$\sqrt{\tau}(\hat{\beta}_{j,\tau} - \beta_j) \xrightarrow{\mathcal{L}} \mathcal{N}_1(0, \mathrm{Diag}_j(\underline{\underline{\Xi}}^{-1}))$$

also holds, for $t \to \infty$. Here $\mathrm{Diag}_j(\underline{\underline{\Xi}}^{-1})$ means the $(j+1)$ diagonal element of $\underline{\underline{\Xi}}^{-1}$. The limiting matrix $\underline{\underline{\Xi}}$ has dimension 8×8, is symmetric and given by

$$\begin{pmatrix}
\sum_{i=1}^{8} \nu_i & \sum_{i\in\{5,6,7,8\}} \nu_i & \sum_{i\in\{3,4,7,8\}} \nu_i & \sum_{i\in\{2,4,6,8\}} \nu_i & \sum_{i\in\{7,8\}} \nu_i & \sum_{i\in\{6,8\}} \nu_i & \sum_{i\in\{4,8\}} \nu_i & \nu_8 \\
\dots & \sum_{i\in\{5,6,7,8\}} \nu_i & \sum_{i\in\{7,8\}} \nu_i & \sum_{i\in\{6,8\}} \nu_i & \sum_{i\in\{7,8\}} \nu_i & \sum_{i\in\{6,8\}} \nu_i & \nu_8 & \nu_8 \\
\dots & \dots & \sum_{i\in\{3,4,7,8\}} \nu_i & \sum_{i\in\{4,8\}} \nu_i & \sum_{i\in\{7,8\}} \nu_i & \nu_8 & \nu_8 & \nu_8 \\
\dots & \dots & \dots & \sum_{i\in\{2,4,6,8\}} \nu_i & \nu_8 & \nu_8 & \nu_8 & \nu_8 \\
\dots & \dots & \dots & \dots & \sum_{i\in\{7,8\}} \nu_i & \nu_8 & \nu_8 & \nu_8 \\
\dots & \dots & \dots & \dots & \dots & \sum_{i\in\{6,8\}} \nu_i & \nu_8 & \nu_8 \\
\dots & \dots & \dots & \dots & \dots & \dots & \sum_{i\in\{4,8\}} \nu_i & \nu_8 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \nu_8
\end{pmatrix}, \quad (3.8)$$

with $\nu_i := \pi_i \theta_i$. Of course, following the same principle, one can derive the asymptotic distribution of the root of the estimating equation for an arbitrary model specification.

3.3.3 Deviance statistic within the fixed time protocol

In this section the asymptotic distribution of the deviance statistic $T_{j,\tau}$ is derived. As stated above, the statistic $T_{j,\tau}$ is defined in accordance with $T_{j,n}$ (defined in section 2.2.2). $T_{j,\tau}$ is

$$\mathrm{dev}_{h_j}(\underline{\underline{x}}, \underline{y}, \underline{\hat{\beta}}_\tau, \underline{\hat{\vartheta}}_\tau) = -2 \log \left[\frac{l((\underline{\underline{x}}, \underline{y}), h_j(\underline{\hat{\vartheta}}_\tau))}{l((\underline{\underline{x}}, \underline{y}), \underline{\hat{\beta}}_\tau)} \right].$$

The criterion-function is given in equation (2.1) on page 13. For grouped data this function is rewritten in dependence of the random variables $N^{(i,y)}$. Denote with $\underline{N}^{\,y}$ the vector $\left(N^{(1,y)}, \dots, N^{(i,y)}, \dots, N^{(2^m,y)}\right)'$. The criterion-function is

$$l((\underline{N}^{\,0}, \underline{N}^{\,1}, \underline{\beta}) = \sum_{i=1}^{2^m} N^{(i,1)}\, \eta_i - N^{(i,\,\cdot\,)} \log\left(1 + \exp(\eta_i)\right). \quad (3.9)$$

Denote with $\underline{\beta}^{(0,j)} := h_j(\underline{\vartheta})$. With this notation,

$$T_{j,\tau} = -2 \log \left[\frac{l(\underline{N}^0, \underline{N}^1, \hat{\underline{\beta}}_\tau^{(0,j)})}{l(\underline{N}^0, \underline{N}^1, \hat{\underline{\beta}}_\tau)} \right].$$

Define the random variate T_τ as

$$T_\tau := -2 \log \left[\frac{l(\underline{N}^0, \underline{N}^1, \underline{\beta})}{l(\underline{N}^0, \underline{N}^1, \hat{\underline{\beta}}_\tau)} \right].$$

Following the theorem 4.1 in Pruscha (2000) on page 249:

$$T_\tau \xrightarrow{\mathcal{L}} \chi^2_d,$$

for $\tau \to \infty$, where d is the dimension of the parameter vector. Correspondingly, due to theorem 4.3 Pruscha (2000) on page 253, $T_{j,\tau}$ is asymptotically χ^2-distributed with one degree of freedom. As in section 2.1.5 the deviance corresponding to the null hypotheses where $d^* = d - c$ parameters are set to their true values, is asymptotically χ^2-distributed with d^* degrees of freedom.

3.3.4 Fixed sample size protocol with one covariate

The asymptotic distribution given above is derived according to the following pattern: The conditions U^* and W^* play a key role. If these conditions hold, a $\underline{\underline{\Gamma}}_\tau^{-1}$-consistent zero-estimator $\hat{\underline{\beta}}_\tau$ is asymptotically normal.[7] Due to the uniqueness of the root of the estimating equation, this root equals the $\underline{\underline{\Gamma}}_\tau^{-1}$-consistent zero-estimator $\hat{\underline{\beta}}_\tau$.[8] Given that the functions h_j satisfy certain regularity conditions, the asymptotic normality of $\hat{\underline{\beta}}_\tau$ holds component-wisely.[9] One possibility to proof the condition U^* is to compute the expectation and the variance of the estimating equation. Due to the CLT for Poisson processes $U_\tau(\underline{\beta})$ is asymptotic normal. The condition W^* holds due to the law of large numbers for Poisson processes in combination with the CMT and the theorem of Cramér-Slutzky.

[7] See theorem 3.4 in Pruscha (2000) on page 194.
[8] See remark 2 on page 192 in combination with the proposition on page 194.
[9] See proposition 4.2 in Pruscha (2000) on page 252. The condition H^* is satisfied as shown in example (c) on page 255.

The intention of this section is to follow the same pattern within the fixed sample size protocol. To begin with, we show for $m = 1$, that for $n \to \infty$ the root of the estimating equation

$$\underline{U}_n(\underline{\beta}) := \left(U_{0,n}(\underline{\beta}), U_{1,n}(\underline{\beta})\right)'$$

is asymptotically normally distributed. The first component of $\underline{U}_n(\underline{\beta})$ is

$$U_{0,n}(\underline{\beta}) := -Z^{(1,0)}\pi_1 + Z^{(1,1)}(1-\pi_1) - Z^{(2,0)}\pi_2 + Z^{(2,1)}(1-\pi_2)$$

and the second component is

$$U_{1,n}(\underline{\beta}) := -Z^{(2,0)}\pi_2 + Z^{(2,1)}(1-\pi_2).$$

The condition U^* is here

$$\underline{\underline{\Gamma}}_n \underline{U}_n(\underline{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}_2(\underline{0}, \underline{\underline{\Sigma}}(\underline{\beta})), \tag{U_n^*}$$

with

$$\underline{\underline{\Gamma}}_n := \begin{pmatrix} \frac{1}{\sqrt{n}} & 0 \\ 0 & \frac{1}{\sqrt{n}} \end{pmatrix}.$$

The matrix $\underline{\underline{\Sigma}}(\underline{\beta})$ is again a positive semidefinite and symmetric matrix. To proof this condition we compute the expectation and the variance of the estimating equation.

As explained above, the data matrix

$$\underline{\underline{Z}} := \begin{pmatrix} Z^{(1,1)} & Z^{(1,0)} \\ Z^{(2,1)} & Z^{(2,0)} \end{pmatrix},$$

follows a multinomial distribution with parameters n and probability matrix

$$\underline{\underline{\mathfrak{p}}} := \frac{1}{\Lambda} \begin{pmatrix} \lambda_1(1) & \lambda_0(1) \\ \lambda_1(2) & \lambda_0(2) \end{pmatrix}.$$

The expectation $\mathbb{E}\{U_{0,n}(\underline{\beta})\}$ is computed as in equation (3.6) with $\frac{n}{\Lambda}$ instead of t. In fact,

$$\mathbb{E}\{\underline{U}_n(\underline{\beta})\} = (0,0)'$$

holds. To derive the variance matrix $\mathbb{V}\{\underline{U}_n(\underline{\beta})\}$, we make use of the Fisher-regularity of the multinomial distribution.(See RÜGER (1999) pages 103ff.) Due to this regularity

$$\mathbb{V}\{\underline{U}_n(\underline{\beta})\} = -\mathbb{E}\{\underline{\underline{W}}_n(\underline{\beta})\} \tag{3.10}$$

holds. (See Lemma 1 and Lemma 2 in PRUSCHA (2000) on page 175.) The matrix $\underline{\underline{W}}_n(\underline{\beta})$ is again $\frac{\partial \underline{U}_n(\underline{\beta})}{\partial \underline{\beta}}(\underline{\beta})$. Thus

$$\underline{\underline{W}}_n(\underline{\beta}) = -\begin{pmatrix} Z^{(1)}\pi_1(1-\pi_1) + Z^{(2)}\pi_2(1-\pi_2) & Z^{(2)}\pi_2(1-\pi_2) \\ Z^{(2)}\pi_2(1-\pi_2) & Z^{(2)}\pi_2(1-\pi_2) \end{pmatrix}.$$

Recall, that $Z^{(i)} := Z^{(i,\,\cdot\,)}$. The random variable $Z^{(i)}$ is binomially distributed with number of successes n and success probability $\frac{\lambda_{\,\cdot\,(i)}}{\Lambda}$. Thus,

$$\mathbb{E}\{\underline{\underline{W}}_n(\underline{\beta})\} = -\frac{n}{\Lambda}\begin{pmatrix} \pi_1\theta_1 + \pi_2\theta_2 & \pi_2\theta_2 \\ \pi_2\theta_2 & \pi_2\theta_2 \end{pmatrix}.$$

Due to equation (3.10),

$$\mathbb{V}\{\underline{U}_n(\underline{\beta})\} = \frac{n}{\Lambda}\begin{pmatrix} \pi_1\theta_1 + \pi_2\theta_2 & \pi_2\theta_2 \\ \pi_2\theta_2 & \pi_2\theta_2 \end{pmatrix}.$$

Thus by virtue of of the CLT for multinomially distributed random variables, U_n^* holds with

$$\underline{\underline{\Sigma}}(\underline{\beta}) := \frac{1}{\Lambda}\underline{\underline{\Xi}}(\underline{\beta}).$$

The condition W_n^* is

$$\underline{\underline{\Gamma}}_n\underline{\underline{W}}_n(\underline{\tilde{\beta}}_n)\underline{\underline{\Gamma}}_n \xrightarrow{\mathbb{P}} -\underline{\underline{\tilde{\Sigma}}}(\underline{\beta}). \tag{W_n^*}$$

In the same way as above for the fixed time protocol one shows that W_n^* holds with $\underline{\underline{\tilde{\Sigma}}}(\underline{\beta}) = \underline{\underline{\Sigma}}(\underline{\beta})$. In fact, $\underline{\hat{\beta}}_n$ is asymptotically normally distributed:

$$\underline{\underline{\Gamma}}_n^{-1}(\underline{\hat{\beta}}_n - \underline{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}_2(\underline{0}, \underline{\underline{\Sigma}}^{-1})$$

3.3.5 Fixed sample size protocol with three covariates

Here

$$\underline{\underline{\Gamma}}_n^{-1}(\hat{\underline{\beta}}_n - \underline{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}_8(\underline{0}, \Lambda \cdot \underline{\underline{\Xi}}^{-1})$$

holds with $\underline{\underline{\Xi}}^{-1}$ being the inverse of $\underline{\underline{\Xi}}$ given in equation (3.8). The matrix $\underline{\underline{\Gamma}}_n$ is the 8×8 diagonal matrix with diagonal elements $\frac{1}{\sqrt{n}}$. The component-wise asymptotic normality holds correspondingly. The derivation of these results follows exactly the same pattern as in section 3.3.2.

3.3.6 Deviance statistic within the fixed sample size protocol

Denote with $\underline{Z}^{\,y}$ the vector $\left(Z^{(1,y)}, \ldots, Z^{(i,y)}, \ldots, Z^{(2^m,y)}\right)'$. Thus, the data matrix $\underline{\underline{Z}} = (\underline{Z}^{\,1}, \underline{Z}^{\,0})$. For grouped data the criterion-function function is rewritten in dependence of the random variables $Z^{(i,y)}$. In analogy with equation (3.9) it is

$$l(\underline{Z}^{\,0}, \underline{Z}^{\,1}, \underline{\beta}) = \sum_{i=1}^{2^m} Z^{(i,1)}\, \eta_i - Z^{(i)} \log\left(1 + \exp(\eta_i)\right). \tag{3.11}$$

Define $\underline{\beta}^{(0,j)} := h_j(\underline{\vartheta})$.

With this notation,

$$T_{j,n} := -2 \log\left[\frac{l(\underline{Z}^{\,0}, \underline{Z}^{\,1}, \hat{\underline{\beta}}_n^{(0,j)})}{l(\underline{Z}^{\,0}, \underline{Z}^{\,1}, \hat{\underline{\beta}}_n)}\right].$$

Define the random variate T_n as

$$-2 \log\left[\frac{l(\underline{Z}^{\,0}, \underline{Z}^{\,1}, \underline{\beta})}{l(\underline{Z}^{\,0}, \underline{Z}^{\,1}, \hat{\underline{\beta}}_n)}\right].$$

As above in section 3.3.3, following the theorem 4.1 in Pruscha (2000) on page 249:

$$T_n \xrightarrow{\mathcal{L}} \chi^2_d,$$

for $n \to \infty$.

Correspondingly, due to the theorem 4.3 in Pruscha (2000) on page 253, the deviance $T_{j,n}$ is asymptotically χ^2-distributed with one degree of freedom. Finally, as in section

2.1.5 the deviance which corresponds to the null hypothesis where $d^* = d - c$ parameters are set to their true values, is asymptotically χ^2-distributed with d^* degrees of freedom.

4 Separation and Penalization

The following considerations are based on the logistic regression model with deterministic binary covariates defined in section 2. To derive the definition of separation, we need to identify conditions causing the criterion-function to be monotone. In this case the root of estimating equation does not exist.

4.1 Separation

Consider the case $m = 2$. The criterion-function is given in equation (3.11), it is

$$l(\underline{z}^{\,0}, \underline{z}^{\,1}, \underline{\beta}) = \sum_{i=1}^{4} z^{(i,1)}\, \eta_i - z^{(i)} \log\left(1 + \exp(\eta_i)\right).$$

Note that in case of deterministic binary covariates this criterion-function is the log-likelihood function. The derivation of summand i at point $\underline{\beta}$ in direction $\underline{b}$ is

$$\frac{z^{(i,1)}\, \eta(\underline{b}, \underline{x}_i) - z^{(i,0)}\, \eta(\underline{b}, \underline{x}_i)\, \exp(\eta_i)}{1 + \exp(\eta_i)}. \tag{4.1}$$

Here the denominator is strictly positive and does not depend on the data. Specializing the numerator of equation (4.1) for the cases $z^{(i,0)} = 0$ or alternatively $z^{(i,1)} = 0$ leads to

$$\begin{cases} -z^{(i,0)}\, \eta(\underline{b}, \underline{x}_i)\, \exp(\eta_i) & \text{if } z^{(i,1)} = 0 \\ z^{(i,1)}\, \eta(\underline{b}, \underline{x}_i). & \text{if } z^{(i,0)} = 0 \end{cases}$$

The criterion-function is strictly increasing in direction $\underline{b}$ if and only if

$$\begin{aligned} \eta(\underline{b}, \underline{x}_i) &< 0 \quad \text{if } z^{(i,1)} = 0 \\ \eta(\underline{b}, \underline{x}_i) &> 0 \quad \text{if } z^{(i,0)} = 0 \end{aligned} \tag{4.2}$$

holds for $i \in \{1, 2, 3, 4\}$.

If there exists a direction $\underline{b}$ such that equation (4.2) holds, at least one component of the root of the estimating equation is not finite. This situation is often called separation.

In general, consider a model with m covariates. The data set, with $g = 2^m$ covariate groups,

Data set			
Variables	$\underline{X}_i'$	$z^{(i,1)}$	$z^{(i,0)}$
Observation	$\underline{x}_1'$	$z^{(1,1)}$	$z^{(1,0)}$
	$\underline{x}_2'$	$z^{(2,1)}$	$z^{(2,0)}$
	$\vdots$	$\vdots$	$\vdots$
	$\underline{x}_i'$	$z^{(i,1)}$	$z^{(i,0)}$
	$\vdots$	$\vdots$	$\vdots$
	$\underline{x}_g'$	$z^{(g,1)}$	$z^{(g,0)}$

is separated, if there exists a direction $\underline{b}$ such that equation (4.2) holds for all $i \in \{1, \ldots, g\}$. Moreover, per definition, a data set is quasi-complete separated, if there exists a vector $\underline{b}$, so that

$$\begin{aligned} \eta(\underline{b}, \underline{x}_i) &\leq 0 \quad \text{if } z^{(i,1)} = 0 \\ \eta(\underline{b}, \underline{x}_i) &\geq 0 \quad \text{if } z^{(i,0)} = 0 \end{aligned}$$

holds for $i \in \{1, \ldots, g\}$. In this situation, if $\eta(\underline{b}, \underline{x}_i) = 0$ holds, in the covariate group i the response is allowed to be zero or one. Separation and quasi-complete separation were defined in Albert/Anderson (1984). A fundamental difference between separation and quasi-complete separation is the limiting value of the criterion-function in direction $\underline{b}$. It is zero in the case of complete separation, whereas it is less than zero in the case of quasi-complete separation.

4.2 Penalization

If a data set is separated at least one component of the root of the estimating equation does not exist. One way to modify the estimation procedure so that the estimator is finite, is to add an appropriate penalty term to the criterion-function. The penalized estimating equation is then the derivation of the penalized criterion-function. In the following section the Firth-penalization is introduced. Thereafter, the corresponding criterion-function is derived for the fixed sample size protocol with $m = 2$ covariates. The Firth-penalty term was suggested by Firth (1993) and is strongly recommended by Heinze.[1] From a Bayesian point of view, the Firth-penalization corresponds to Jeffreys prior. It leads to an unique and finite root of the estimating equation. (See Firth (1993).)

4.2.1 Penalized estimating equation

To derive the penalized criterion-function a penalty term is added to the criterion-function. This penalty term depends on the number of covariates m. If m covariates are included in the model, the number of covariate groups is $g = 2^m$. Remember

$$\pi_i \equiv \pi(\underline{x}_i, \underline{\beta}) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}.$$

The linear predictor may involve interaction effects.

The Firth-penalty term for a logistic regression log-likelihood function with m binary covariates is

$$\begin{aligned} \text{pen}(\underline{\beta}) := 0.5 \log \big[\prod_{i=1}^{g} \{\pi_i(1-\pi_i)\} \big] = \\ 0.5 \sum_{i=1}^{g} \big[\log(\pi_i) + \log(1-\pi_i) \big], \end{aligned} \tag{4.3}$$

if all pairwise and higher-order interactions are taken into account. To understand this, consider the logistic regression model with one deterministic binary covariate x and a binomial response Y as described in section 2.1.4. The likelihood function is

$$\prod_{i=1}^{2} \pi_i^{z^{(i,1)}} (1-\pi_i)^{z^{(i)} - z^{(i,1)}}.$$

[1] See for example Heinze (2006) and Heinze/Schemper (2002) and the references given there.

The Fisher-information is

$$\begin{pmatrix} z^{(1)}\pi_1(1-\pi_1) + z^{(2)}\pi_2(1-\pi_2) & z^{(2)}\pi_2(1-\pi_2) \\ z^{(2)}\pi_2(1-\pi_2) & z^{(2)}\pi_2(1-\pi_2) \end{pmatrix}.$$

Following Jeffreys rule the prior distribution is proportional to the root of the determinant of this Fisher information matrix. (See RÜGER (1999) p. 228.) This determinant is

$$z^{(1)}\pi_1(1-\pi_1)z^{(2)}\pi_2(1-\pi_2).$$

Thus the prior is proportional to[2]

$$\sqrt{\pi_1(1-\pi_1)\pi_2(1-\pi_2)}.$$

The penalized likelihood is consequently

$$\prod_{i=1}^{2} \pi_i^{z^{(i,1)}+0.5}(1-\pi_i)^{z^{(i)}-z^{(i,1)}+0.5}.$$

In the same way one can proof that for m covariates, if all pairwise and higher-order interactions are taken into account, Jeffreys prior corresponds to equation (4.3). In this case the penalized criterion-function, which is defined as the criterion-function (3.11) plus this penalty term, is

$$l^{\text{pen}}(\underline{z}^{\,0}, \underline{z}^{\,1}, \underline{\beta}) := \sum_{i=1}^{g} \left(z^{(i,1)} + 0.5\right) \eta_i - \left(z^{(i)} + 1\right) \log\left(1 + \exp(\eta_i)\right). \tag{4.4}$$

This function corresponds to the unpenalized criterion-function of the following modified data set:

[2] Recall that here $z^{(1)}$ and $z^{(2)}$ are deterministic.

Data set			
Variables	$\underline{X}_i'$	$z^{(i,1)}$	$z^{(i,0)}$
Observation	$\underline{x}_1'$	$z^{(1,1)} + 0.5$	$z^{(1,0)} + 0.5$
	$\underline{x}_2'$	$z^{(2,1)} + 0.5$	$z^{(2,0)} + 0.5$
	$\vdots$	$\vdots$	$\vdots$
	$\underline{x}_i'$	$z^{(i,1)} + 0.5$	$z^{(i,0)} + 0.5$
	$\vdots$	$\vdots$	$\vdots$
	$\underline{x}_g'$	$z^{(g,1)} + 0.5$	$z^{(g,0)} + 0.5$

If the linear predictor does not contain all pairwise and higher order interactions, the Firth-penalty term is

$$\text{pen}(\underline{\beta}) = \frac{1}{2} \log \Big(\text{Det} [\underline{\underline{x}}' \text{Diag}(z^{(i)} \cdot \pi_i \cdot (1 - \pi_i)) \underline{\underline{x}}] \Big). \tag{4.5}$$

(See Heinze/Schemper (2002) or Rüger (1999) page 228 and Pruscha (2000) page 286.) The penalized criterion-function, which is defined as the criterion-function plus this penalty term, is univariate. The parameter value maximizing the penalty term (4.5) is $\underline{0}$. (See Chen et al. (2008).) Therefore the penalized estimator $\hat{\underline{\beta}}^{\text{pen}}$ is, compared to its unpenalized version, shifted towards $\underline{0}$. Note that the penalty (4.3) is a special case of penalty (4.5).

The penalized estimating equation $\underline{U}^{\text{pen}}(\underline{\beta})$ is given by the partial derivatives of the penalized criterion function with respect to $\underline{\beta}$. Its root is $\hat{\beta}^{\text{pen}}$. The derivation of the penalized criterion-function (4.4) with respect to the $(j+1)$ component of $\underline{\beta}$ is

$$\Big[\underline{z} + 0.5 \cdot \underline{1} - \underline{\pi}(\underline{\beta}) \Big] \underline{x}_j. \tag{4.6}$$

Here the vector $\underline{x}_j$ is the $(j+1)$th column vector of the design matrix $\underline{\underline{x}}$. The vector $\underline{z} = (z^{(1)}, \ldots, z^{(i)}, \ldots, z^{(g)})'$, correspondingly $\underline{\pi}(\underline{\beta}) = (\pi_1, \ldots, \pi_i, \ldots, \pi_g)'$ and the unitary vector $\underline{1}$ of length g.

In general, if no cell of the data set is empty, the root of the estimating equation is finite. Thus another way to penalize the estimating equation with the intention to ensure the existence of its root, is to add a constant $c^{(i,y)}$, $c^{(i,y)} \in \mathbb{R}^+$ to the the observation $z^{(i,y)}$. The

penalty weights $c^{(i,y)}$, can be stored in a matrix $\underline{\underline{c}}$. The corresponding penalized criterion-function is denoted $l^{\underline{\underline{c}}}(\underline{z}^{\,0}, \underline{z}^{\,1}, \underline{\beta})$. If each entry of the matrix $\underline{\underline{c}}$ has the same value, say c, one can define the c-penalized criterion-function as

$$l^{\mathrm{c}}(\underline{z}^{\,0}, \underline{z}^{\,1}, \underline{\beta}) := \sum_{i=1}^{g} \left(z^{(i,1)} + c\right) \eta_i - \left(z^{(i)} + 2\,c\right) \log\left(1 + \exp(\eta_i)\right),$$

with $c \in \mathbb{R}^+$. The root of the corresponding estimating equation is denoted $\underline{\hat{\beta}}^{c}_{n}$. As above, in section 2.1.5, the root of the h-restricted criterion-function is $h(\underline{\hat{\vartheta}}^{c}_{n})$. Recall that in classical inference it is necessary to define $c^{(i,y)}$ before viewing the data.

The Firth-penalization is based on a heuristic argument given in Firth (1993). If an estimating equation U is linear in the parameter of interest ζ

$$\mathbb{E}\{U(\hat{\zeta})\} = 0 \;\Leftrightarrow\; \mathbb{E}\{\hat{\zeta}\} = \zeta.$$

But, if $U(\zeta)$ is convex

$$\mathbb{E}\{U(\hat{\zeta})\} = 0 \;\Leftrightarrow\; \mathbb{E}\{\hat{\zeta}\} > \zeta.$$

Firth (1993) suggests introducing a small bias in the estimating equation to reduce the bias in its root $\hat{\zeta}$. The correction of the estimating equation is based on its derivation. In this way it is guaranteed that, the estimating equation is shifted in the right direction and that the amount of shifting is appropriate.

4.2.2 Example

Consider, for example, the model with $m = 2$ covariates and one interaction term. The corresponding design matrix is

$$\underline{\underline{x}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

The linear predictor η_i is

$$\eta_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_{1,2} x_{i,1} x_{i,1}.$$

The Firth-penalized criterion-function $l^{\text{pen}}(\underline{z}^0, \underline{z}^1, \underline{\beta})$ is

$$\begin{aligned}\sum_{i=1}^{4} z^{(i,1)}\,\eta_i - z^{(i)} \log\left(1+\exp(\eta_i)\right) + 0.5 \sum_{i=1}^{4} \{\log(\pi_i) + \log(1-\pi_i)\} = \\ = \sum_{i=1}^{4} (z^{(i,1)} + 0.5)\,\eta_i - (z^{(i)} + 1) \log\left(1+\exp(\eta_i)\right).\end{aligned}$$

The Firth-penalized estimating equation $\underline{U}^{\text{pen}}(\beta) = \left(U_0(\underline{\beta}), U_1(\underline{\beta}), U_1(\underline{\beta}), U_{(1,2)}(\underline{\beta})\right)'$ is

$$\underline{\underline{x}}' \left[\underline{z} + 0.5 \cdot \underline{1} - \underline{\pi}(\underline{\beta})\right].$$

If the pairwise interaction $\beta_{1,2}$ is set to zero, the penalty term is

$$\text{pen}(\underline{\beta}) = \upsilon_1 \upsilon_3 \upsilon_4 + \upsilon_2 \upsilon_3 \upsilon_4 - \upsilon_2^2 (\upsilon_3 + \upsilon_4),$$

with

$$\upsilon_i := z^{(i)} \cdot \pi_i \cdot (1 - \pi_i),$$

in accordance with equation (4.5).

5 Approximate Kolmogorov distance

5.1 Convergence with respect to the Kolmogorov distance

This work deals with the $\mathcal{L}$-convergence of estimators or test statistics to their limiting distributions. $\mathcal{L}$-convergence means convergence in distribution. The definition of convergence in distribution is:

A sequence $S_1, S_2, \ldots$ of random variables is said to converge in distribution to a random variable S if

$$\lim_{n\to\infty} G_n(s) = G(s), n \in \mathbb{N} \tag{5.1}$$

for every number $s \in \mathbb{R}$ at which G is continuous. Here G_n and G are the cumulative distribution functions of random variables S_n and S correspondingly. (See RÜGER (2002) pages 41ff.)

The metric, which corresponds exactly to this form of convergence is the Lévy-Prohorov metric. In other words: Convergence with respect to the Lévy-Prohorov metric is equivalent to convergence in distribution. (See RÜGER (2002) pages 48ff.) If the limiting distribution is continuous, convergence with respect to the Lévy-Prohorov metric and convergence with respect to the Kolmogorov distance are equivalent. (See RÜGER (2002) page 51.)

The Kolmogorov distance d_K between two (one-dimensional) distributions with corresponding cumulative distribution functions G_n and G is defined as

$$d_K(G, G_n) := \sup_{s\in\mathbb{R}} |G(s) - G_n(s)|. \tag{5.2}$$

Convergence with respect to the Kolmogorov distance is equivalent to uniform convergence in distribution. That means in (5.1), the cumulative distribution functions $G_n(s)$, converges uniformly to $G(s)$ for all s in $\mathbb{R}$.

5.2 Mean approximate Kolmogorov distance

We want to measure the Kolmogorov distance d_K between the actual distribution $G_n(s)$ of an one-dimensional estimator S_n and the corresponding limiting distribution $G(s)$. In fact, the actual distribution $G_n(s)$ is not given analytically. Thus it is necessary to use an estimator $\hat{G}_n(s)$. The distance d_K between $G_n(s)$ and $\hat{G}_n(s)$ should be insignificant compared with the distance of interest $d_K(G, G_n)$.

An obvious estimator $\hat{G}_n(s)$ is the empirical cumulative distribution function. To compute the empirical cumulative distribution function of an estimator S_n, one simulates n_{sim} data sets and each time computes the realization of the estimator S_n, which is denoted $s_n^{(i)}$, $i \in \{1, \dots n_{\text{sim}}\}$. The empirical cumulative distribution function $\hat{G}_n^{(n_{\text{sim}})}(s)$ is defined as

$$\hat{G}_n^{(n_{\text{sim}})}(s) := \frac{1}{n_{\text{sim}}} \#[s_n^{(i)} \leq s], \tag{5.3}$$

with # as symbol for "number of." In fact,

$$n_{\text{sim}} \cdot \hat{G}_n^{(n_{\text{sim}})}(s) \sim \text{Bin}\big(n_{\text{sim}}, G_n(s)\big).$$

By definition the Kolmogorov distance $d_K(G_n, \hat{G}_n^{(n_{\text{sim}})})$ is

$$\sup_{i \in \{1 \dots n_{\text{sim}}\}} \big| G_n(s_n^{(i)}) - \hat{G}_n^{(n_{\text{sim}})}(s_n^{(i)}) \big|. \tag{5.4}$$

It should be noted that, this Kolmogorov distance is a random variable. Following the theorem of Glivenko-Cantelli, the empirical cumulative distribution function $\hat{G}_n$ converges uniformly to G_n, that is

$$d_K(G_n, \hat{G}_n^{(n_{\text{sim}})}) \to 0 \tag{5.5}$$

almost surely.

In fact, we want to measure $d_K(G, G_n)$, but we need to approximate G_n by $\hat{G}_n^{(n_{\text{sim}})}$. Thus, we measure $d_K(G, \hat{G}_n^{(n_{\text{sim}})})$. The distance $d_K(G, \hat{G}_n^{(n_{\text{sim}})})$ is a random variate and is denoted the approximate Kolmogorov distance. The remaining question is, how large needs n_{sim} to be, so that, a realization of $d_K(G, \hat{G}_n^{(n_{\text{sim}})})$ is a good approximation of $d_K(G, G_n)$.

Indeed,

$$\lim_{n_{\text{sim}} \to \infty} \mathbb{P}\big(\sqrt{n_{\text{sim}}} \cdot d_K(G_n, \hat{G}_n^{(n_{\text{sim}})}) \leq s\big) = K(s),$$

with

$$K(s) := \begin{cases} 1 - 2\sum_{i=0}^{\infty} (-1)^{i-1} \exp(-2i^2 s^2), & s > 0; \\ 0. & s \leq 0. \end{cases}$$

$K(s)$ is denoted Kolmogorov distribution. The approximation

$$\mathbb{P}\big(\sqrt{n_{\text{sim}}} \cdot d_K(G_n, \hat{G}_n^{(n_{\text{sim}})}) \leq s\big) \approx K(s), \tag{5.6}$$

works for $n_{\text{sim}} \geq 40$. For $s \geq 1$,

$$K(s) \approx 1 - 2\exp(-2s^2) \tag{5.7}$$

holds. (See Pruscha (2000) pages 156ff and Rüger (2002) page 194.) Define s_γ, so that,

$$K(s_\gamma) = \gamma,$$

with $\gamma \in (0, 1)$.

With equation (5.7)

$$s_\gamma \approx \sqrt{-0.5 \log\left(\frac{1-\gamma}{2}\right)}.$$

Using this approximation, $s_{0.95} \approx \sqrt{-0.5 \log(0.025)} \approx 1.358$. Thus, for $n_{\text{sim}} := 1.5 \cdot 10^4$,

$$\mathbb{P}\Big(d_K(G_n, \hat{G}_n^{(n_{\text{sim}})}) \leq \frac{s_{0.95}}{n_{\text{sim}}}\Big) = \mathbb{P}\big(d_K(G_n, \hat{G}_n^{(n_{\text{sim}})}) \leq 9.05 \cdot 10^{-4}\big) \approx 0.95.$$

Hence for $n_{\text{sim}} = 1.5 \cdot 10^4$, it is very probable that the empirical cumulative distribution function $\hat{G}_n^{(n_{\text{sim}})}$ is sufficiently approximating G_n. In fact, the random variate $d_K(G, \hat{G}_n^{(1500)})$ is assumed to have a very low variance. To verify this assumption in the following section the mean approximate will be represented graphically. The standard deviation can also be included in the graphic. The name of this graphic is distance-sample-size-diagram.

The random variate $d_K(G, \hat{G}_n^{(n_{\text{sim}})})$ is the approximate Kolmogorov distance. The definition of the mean approximate Kolmogorov distance with respect to a particular $\mathcal{L}$-convergence theorem in form of equation (5.1) is: the mean of a sequence $d_K^l(G, \hat{G}_n^{(n_{\text{sim}})})$, $l \in \{1, \ldots, k\}$, where each random variate has the same distribution as the approximate Kolmogorov distance $d_K(G, \hat{G}_n^{(n_{\text{sim}})})$. Moreover these random variates are mutually independent. Thus $\{d_K^l(G, \hat{G}_n^{(n_{\text{sim}})})\}_{l \in \{1,\ldots,k\}}$ is an independently and identically distributed sequence. The mean approximate Kolmogorov distance is defined as

$$A_n^{(n_{\text{sim}},k)} \equiv makd(n) := \frac{1}{k} \sum_{l=1}^{k} d_K^l(G, \hat{G}_n^{(n_{\text{sim}})}). \tag{5.8}$$

Obviously for $k = 1$ the mean approximate Kolmogorov distance equals the approximate Kolmogorov distance.

Now consider the sequence of the mean approximate Kolmogorov distance for increasing sample size n. For a given value of n the mean approximate Kolmogorov distance is a random variate denoted as $A_n^{(n_{\text{sim}},k)}$.[1] Per definition the sequence $\{A_n^{(n_{\text{sim}},k)}\}_{n\in\mathbb{N}}$ is a family of dependent random variables indexed by n. More precisely $\{A_n^{(n_{\text{sim}},k)}\}_{n\in\mathbb{N}}$ is a continuous stochastic process with discrete state space. A stochastic process with discrete state space is also denoted a random sequence. The stochastic dependence of the process $\{A_n^{(n_{\text{sim}},k)}\}_{n\in\mathbb{N}}$ is due to following definition:

To realize two values $A_{n_1}^{(n_{\text{sim}},k)}$ and $A_{n_2}^{(n_{\text{sim}},k)}$, $n_1 < n_2$, belonging to the same trajectory of $\{A_n^{(n_{\text{sim}},k)}\}_{n\in\mathbb{N}}$ one simulates n_2 data sets to compute the realization of $A_{n_2}^{(n_{\text{sim}},k)}$. The first n_1 data sets serve then as a basis for calculating the realization of $A_{n_1}^{(n_{\text{sim}},k)}$. In the same way one can compute a sample path of the random sequence for a given index set.[2]

5.3 Distance-sample-size-diagram

This section introduces the distance-sample-size-diagram. This diagram represents the behavior of $\mathcal{L}$-convergence defined as the mean approximate Kolmogorov distance as a function of the predefined sample size. The predefined sample size is displayed on the horizontal axis and the mean approximate Kolmogorov distance between the statistic of interest and its limiting distribution on the vertical axis. Indeed the mean approximate Kolmogorov distance is random. Its variation depends on the number of simulations n_{sim}. It is possible to take into account the variation of the approximate Kolmogorov distance by plotting error bars along the curve of the mean approximate Kolmogorov distance.

5.3.1 Fixed sample size protocol-model without covariates

It is commonly known that the binomial distribution can be approximated by the normal distribution and that the goodness-of-fit depends on both parameters of this distribution, thus of the number of trails and the success probability. (See for example Nagaev/Cheb-

[1] This random variate is also denoted with $makd(n)$ as the abbreviation of mean approximate Kolmogorov distance. In this abbreviation the number of simulation n_{sim} and the number of realizations of the approximate Kolmogorov distance k are not included. The abbreviation $makd(n)$ will be used in the distance-sample-size-diagram.

[2] Using the programming language R it might be advantageous to use the R-function set.seed for the simulation of a sample path.

OTAREV (2011).) In this section, this common example of $\mathcal{L}$-convergence is used to make the measurement of the behavior of $\mathcal{L}$-convergence by means of the approximate Kolmogorov distance accessible and understandable.

For $m = 0$ covariates, the only regression coefficient is β_0. Within the fixed sample size protocol without covariates the data is binomial distributed with success probability $\pi = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$. The number of trails equals the predefined sample size n. This simple model is suitable for discussing the behavior of $\mathcal{L}$-convergence. Let $Z_n^{(\cdot,1)}$ denote the number of successes, that is to say the number of observations with $Y = 1$. In fact, the MLE $\hat{\pi}_n$ is $\frac{Z_n^{(\cdot,1)}}{n}$ and correspondingly $\hat{\beta}_{0,n} = \log\left(\frac{Z_n^{(\cdot,1)}}{n-Z_n^{(\cdot,1)}}\right)$. In the same way as in section 3.3.4 for one covariate, one shows for $m = 0$ that

$$\sqrt{n} \cdot \sqrt{\frac{\exp(\beta_0)}{(1+\exp(\beta_0))^2}}(\hat{\beta}_{0,n} - b_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathfrak{m}, 1), \tag{5.9}$$

holds, with

$$\mathfrak{m} := \sqrt{n} \cdot \sqrt{\frac{\exp(\beta_0)}{(1+\exp(\beta_0))^2}}(\beta_0 - b_0).$$

and $b_0 \in \mathbb{R}$.

For $b_0 = \beta_0$ this corresponds to the de Moivra-Laplace theorem. (See FOATA/FUCHS (1999) page 286.) In fact, it is known that the behavior of $\mathcal{L}$-convergence depends strongly on β_0. Moreover, it is likely that b_0 has little influence on the drop of the (mean) Kolmogorov distance with increasing sample size.

Graphic 5.1 illustrates the behavior of $\mathcal{L}$-convergence of equation (5.9) for $\beta_0 = 4$ and for different values of b_0. The cumulative distribution function $G_{0,n}$ denotes the cumulative distribution function of the suitably scaled version of the $\hat{\beta}_{0,n}$ given in equation (5.9). The function $\hat{G}_{0,n}^{(n_{\text{sim}})}$ is the corresponding empirical cumulative distribution function based on n_{sim} simulations. The limiting normal distribution is denoted with $\Phi(\mathfrak{m})$ in the following. The approximate Kolmogorov distance is $d_K(\Phi(\mathfrak{m}), \hat{G}_{0,n}^{(n_{\text{sim}})})$. The mean approximate Kolmogorov distance is plotted in figure 5.1 as a function of the predefined sample size n for sample sizes from 20 to 400 with an increment of 20. Here the number of simulations $n_{\text{sim}} = 1500$. The mean approximate Kolmogorov distance for a given predefined sample size n is per definition

$$\frac{1}{k}\sum_{l=1}^{k} d_K^l(\Phi(\mathfrak{m}), \hat{G}_{0,n}^{(n_{\text{sim}})}). \tag{5.10}$$

In the following $k = 20$. Indeed, to compute a realization of $d_K(\Phi(\mathfrak{m}), \hat{G}_{0,n}^{(1500)})$ for the predefined sample size n, $n_{\text{sim}} = 1500$ data sets are simulated with $\beta_0 = 4$. For each data set the regression coefficient is estimated. Based on these estimators the empirical cumulative distribution function $\hat{G}_{0,n}^{(1500)}$ is computed. Finally, the Kolmogorov distance between $\hat{G}_{0,n}^{(1500)}$ and $\Phi(\mathfrak{m})$ is measured. For each predefined sample size n the number of n_{sim} data sets are simulated 20 times in order to compute the mean approximate Kolmogorov distance given in equation (5.10) using the programming language R. The R-function set.seed() is used to define a set of 20 seeds used for each predefined sample size n. In this way a trajectory of the family of dependent random variables $\{A_n^{(n_{\text{sim}},k)}\}_{n\in\mathbb{N}}$ is computed according to the definition of this process given above.

The three lines in figure 5.1 represent three different values of b_0. The error bars give the standard deviations of the approximate Kolmogorov distance for each predefined sample size. It turns out, that, in fact, the choice of b_0 has little influence on the approximate Kolmogorov distance between the sampling distribution of the estimator $\hat{\beta}_{0,n}$ and its limiting distribution. The situation is very different if one looks at the mean approximate Kolmogorov distance in dependence of different values of the true parameter β_0. This parameter is crucial for the behavior of $\mathcal{L}$-convergence. It is a bijective function of the successes probability π. The behavior of $\mathcal{L}$-convergence of the binomial distribution to the

corresponding normal distribution depends on this parameter.

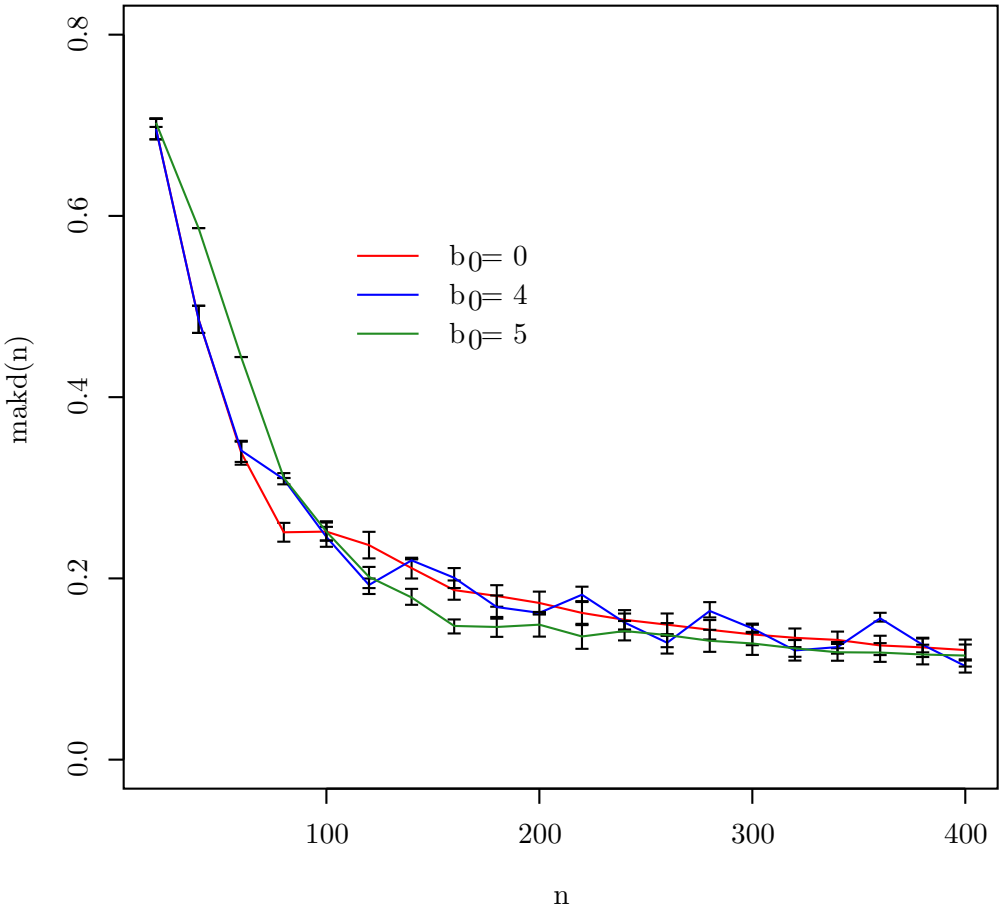

Figure 5.1: The graphic is the distance-sample-size-diagram of equation (5.9) for different values of b_0 in dependence of the predefined sample size n. The true parameter $\beta_0 = 4$.

Figure 5.2 illustrates the dependence of the behavior of $\mathcal{L}$-convergence for different success probabilities. Here $b_0 = 0$. The mean approximate Kolmogorov distance is plotted as a function of the predefined sample size n for sample sizes from 20 to 150 with an increment of 10 and the number of simulations $n_{\text{sim}} = 1500$. The true regression $\beta_0 \in \{0, 2.5, 4\}$. This corresponds to success probabilities $\pi \in \{0.5, 0.9241, 0.982\}$. The Fisher information of β_0 is $n\,\pi\,(1-\pi)$. This function is maximal for $\pi = 0.5$. This corresponds to $\beta_0 = 0$. Thus, it is to be expected, that the drop of the mean approximate Kolmogorov distance with increasing sample size becomes slower as $|\beta_0|$ increases. It is remarkable that , if β_0 is changed from 2.5 to 4, the sample size necessary to reach an approximate Kolmogorov distance less than 0.1 increases from 140 to 560.

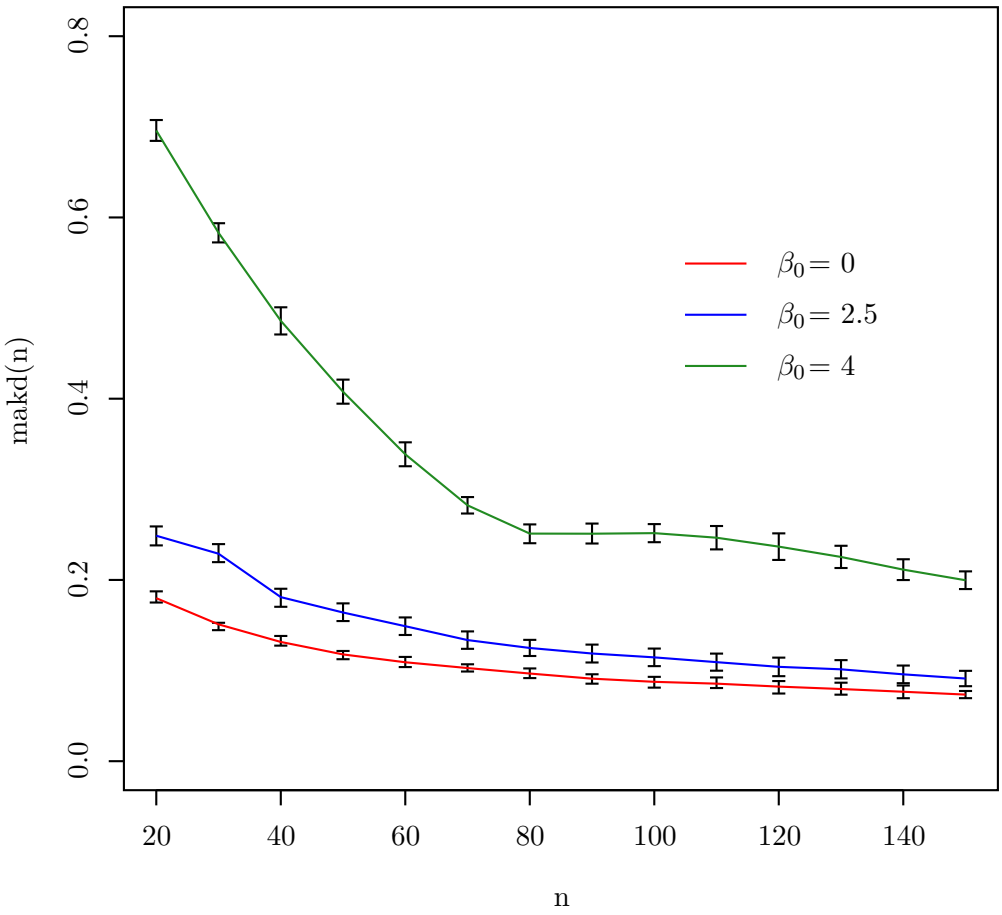

Figure 5.2: The graphic gives the distance-sample-size-diagram of equation (5.9) for different values of β_0 in dependence of the predefined sample size n. Here $b_0 = 0$.

5.3.2 Fixed sample size protocol-model with one covariate

Recall, for $m = 1$ the data observed until the nth event occurs can be arranged in a two-by-two table:

		Y: 1	Y: 0	
X	1	$Z_n^{(1,1)}$	$Z_n^{(1,0)}$	$Z_n^{(1)}$
	2	$Z_n^{(2,1)}$	$Z_n^{(2,0)}$	$Z_n^{(2)}$
		$Z_n^{(\cdot,1)}$	$Z_n^{(\cdot,0)}$	n

Here $Z_n^{(i)} := Z_n^{(i,\cdot)}$.

The data matrix

$$\underline{\underline{Z}}_n := \begin{pmatrix} Z_n^{(1,1)} & Z_n^{(1,0)} \\ Z_n^{(2,1)} & Z_n^{(2,0)} \end{pmatrix},$$

follows a multinomial distribution with parameters n and probability matrix

$$\underline{\underline{\mathfrak{p}}} := \frac{1}{\Lambda} \begin{pmatrix} \lambda_1(1) & \lambda_0(1) \\ \lambda_1(2) & \lambda_0(2) \end{pmatrix}.$$

Here

$$\lambda_y(i, \underline{\beta}, \underline{\theta}) := \begin{cases} \theta_i & \text{y} = 0; \\ \theta_i \exp(\eta(\underline{x}_i, \underline{\beta})) & \text{y} = 1. \end{cases}$$

The vector $\underline{\theta}$ contains a weight θ_i for each covariate group i, with $i \in \{1, 2\}$ and $\underline{x}_1^{'} = (0, 0)$ and $\underline{x}_2^{'} = (0, 1)$. Thus it has length $g = 2^1 = 2$.

Per definition of $\underline{\theta}$

$$\sum_{i=1}^{2} \theta_i = 1 \tag{5.11}$$

holds. Hence, the vector of regression coefficients $\underline{\beta} = (\beta_0, \beta_1)'$, the weight parameter θ_2 and the predefined sample size n define the model. The behavior of $\mathcal{L}$-convergence as a function of n must depend on $\underline{\beta}$ and θ_2. In the following $\theta_2 \equiv \theta$.[3]

In fact, due to section 3.3.4

$$\sqrt{\frac{\pi_1(1-\theta)}{\Lambda}}\,\sqrt{n}\,(\hat{\beta}_{0,n} - b_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathfrak{m}_1, 1) \tag{5.12}$$

with

$$\mathfrak{m}_1 = \sqrt{\frac{\pi_1(1-\theta)}{\Lambda}}\,\sqrt{n}\,(\beta_0 - b_0).$$

Moreover

$$\sqrt{\frac{\pi_1(1-\theta)\cdot\pi_2\theta}{\pi_1(1-\theta)+\pi_2\theta}}\sqrt{n}\,(\hat{\beta}_{1,n} - b_1) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathfrak{m}_2, 1). \tag{5.13}$$

with

$$\mathfrak{m}_2 = \sqrt{\frac{\pi_1(1-\theta)\cdot\pi_2\theta}{\pi_1(1-\theta)+\pi_2\theta}}\sqrt{n}\,(\beta_1 - b_1).$$

Furthermore,

$$n(\underline{\hat{\beta}}_n - \underline{b})' \underline{\underline{\Sigma}}(\underline{\beta})(\underline{\hat{\beta}}_n - \underline{b}) \xrightarrow{\mathcal{L}} \chi^2_2(\tilde{\mathfrak{m}}), \tag{5.14}$$

with

$$\underline{\underline{\Sigma}}(\underline{\beta}) = \Lambda \begin{pmatrix} \pi_1(1-\theta)+\pi_2\theta & \pi_1(1-\theta) \\ \pi_1(1-\theta) & \pi_1(1-\theta) \end{pmatrix}$$

and noncentrality parameter

$$\tilde{\mathfrak{m}} = n(\underline{\beta} - \underline{b})' \underline{\underline{\Sigma}}(\underline{\beta})(\underline{\beta} - \underline{b}).$$

For all three equations (5.12), (5.13) and (5.14) one could measure the behavior of $\mathcal{L}$-convergence as a function of the predefined sample size n in dependence of the true regression coefficient $\underline{\beta}$ and the true weight parameter θ. In fact, the assumption is obvious that the drop of the mean approximate Kolmogorov distance is the fastest at $\underline{\beta} = (0, 0)'$

[3] This is possible because of equation (5.11).

and $\theta = 0.5$.

The following graphics 5.3, 5.4 and 5.5 illustrate the distance-sample-size-diagrams of equation (5.14). Again, $n_{\text{sim}} = 1500$ and the computation of the mean is based on 20 independently and identically distributed realizations of the mean approximate Kolmogorov distance of interest.

As above, the values b_0 and b_1 do little influence the behavior of $\mathcal{L}$-convergence. Figure 5.3 shows the mean approximate Kolmogorov distance for different values of $\underline{b}$.

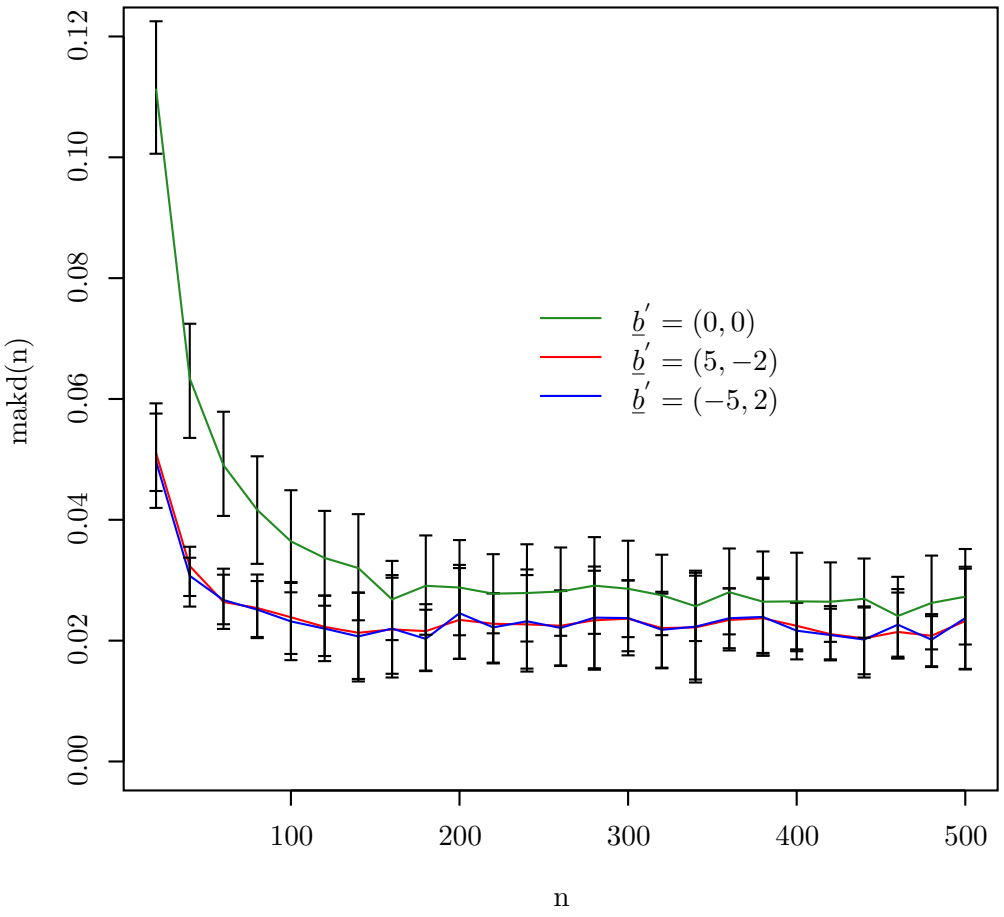

Figure 5.3: The distance-sample-size-diagram which corresponds to equation (5.14) is plotted for different values of $\underline{b}$. Here $\underline{\beta} = (0,0)'$ and $\theta = 0.5$.

The following graphic 5.4 confirms that the drop of the mean approximate Kolmogorov distance with increasing sample size becomes slower as $|\theta - 0.5|$ increases.

It is more difficult to understand how the true regression coefficients β_0 and β_1 influence the behavior of $\mathcal{L}$-convergence of equation (5.14). As was to be expected, $\mathcal{L}$-convergence is very fast for $\underline{\beta} = (0,0)$, as shown in figure 5.5.

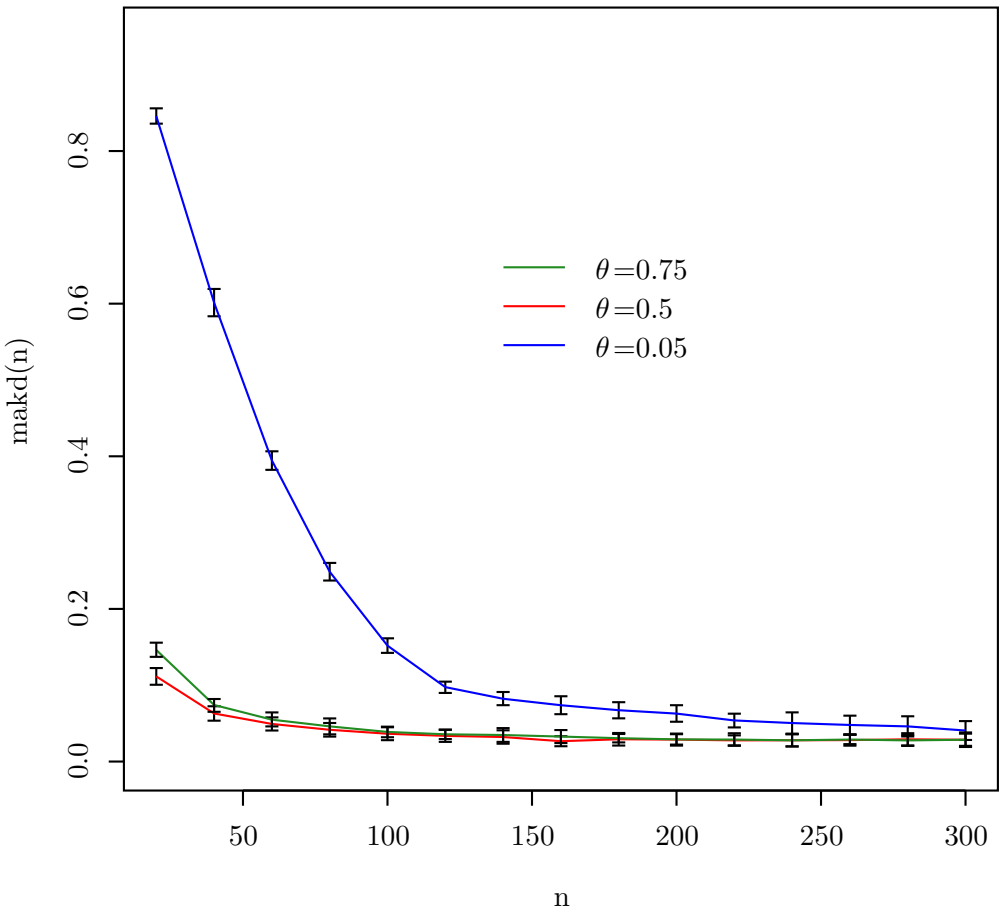

Figure 5.4: The distance-sample-size-diagram which corresponds to equation (5.14) is plotted for different values of θ. Here $\underline{b} = \underline{\beta} = (0,0)'$.

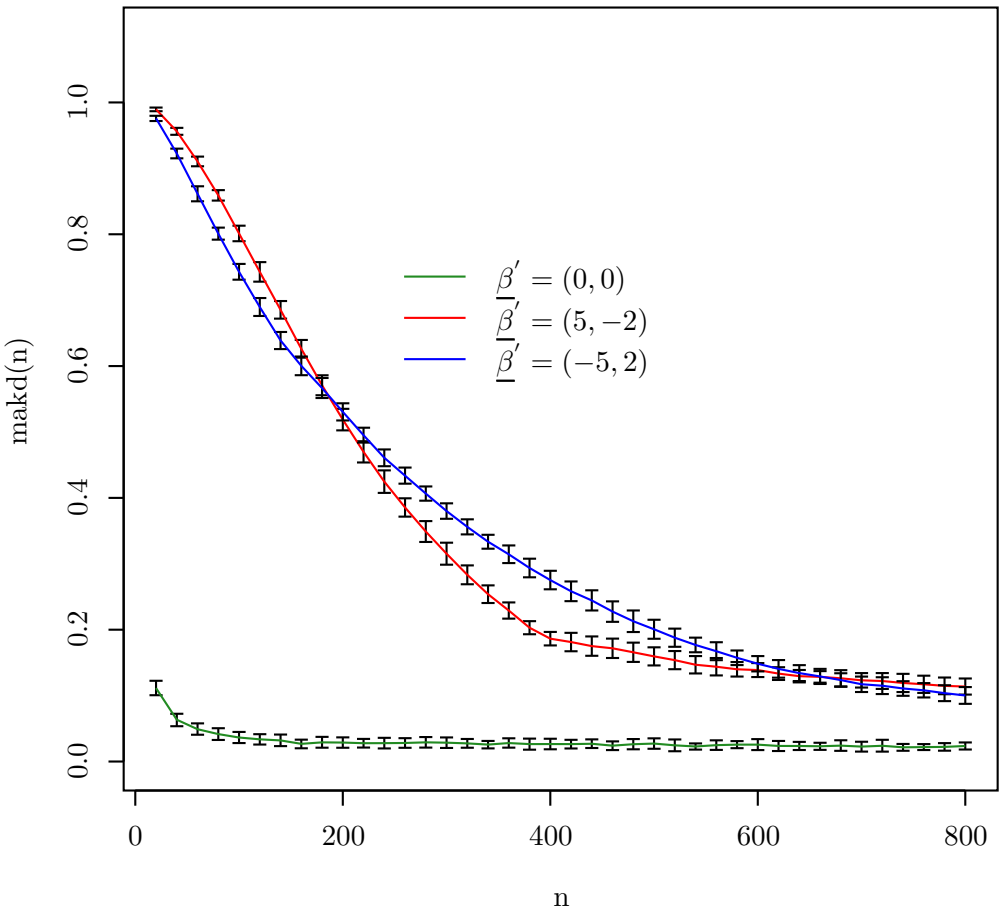

Figure 5.5: The distance-sample-size-diagram which corresponds to equation (5.14) is plotted for different values of $\underline{\beta}$. Here $\underline{b} = (0, 0)'$ and $\theta = 0.5$.

The next chapter explores the behavior of $\mathcal{L}$-convergence of the logistic regression model with $m = 1$ covariate for both Firth-penalized likelihood and unpenalized likelihood methods.

6 Comparison of of Firth-penalized and unpenalized likelihood methods

The fixed sample size protocol[1] can, for example, model the electoral behavior of a student population. The aim of the fictive study is to find out, how a binary factor $X_{(1)}$ affects the result of an election. Thus the regression coefficient of interest is β_1. In this fictive example students, who enter the cafeteria are asked: "Would you vote for the *party of improvement* (PI) if federal elections were held this sunday?

The response is

$$Y := \begin{cases} 1, & \text{student would vote for PI;} \\ 0, & \text{student would not vote for PI} \end{cases}$$

and the covariate of interest is

$$X_{(1)} := \begin{cases} 1, & \text{student of natural sciences;} \\ 0, & \text{other student.} \end{cases}$$

The sample size n is defined in advance. Thus the fixed sample size protocol is used to collect the data. This sampling protocol is used in Heinze (2001) and Heinze/Schemper (2002). The spacial case of deterministic covariates is used in Heinze (2006). For reasons of simplicity X is used instead of $X_{(1)}$. Here $\underline{x}_1 = (1, 0)$ and $\underline{x}_2 = (1, 1)$. The data, observed until the nth event occurs, can be arranged in the two-by-two table, given above in section 5.3.2.

[1] defined section 3.2.2

6.1 Convergence theorems of interest

Equation (5.8) gives the definition of the mean approximate Kolmogorov distance with respect to a particular $\mathcal{L}$-convergence theorem.[2] This section informs about the $\mathcal{L}$-convergence theorems we are interest in. For these theorems the distance-sample-size-diagrams will be given in this chapter. In our example, we assume the following true covariate effects:

$$\beta_0 = -3,$$

$$\beta_1 = 7$$

and

$$\theta = 0.05.$$

Thus

$$\pi_1 = 0.0474,$$

$$\pi_2 = 0.9820$$

and

$$\underline{\underline{\mathfrak{p}}} = \frac{1}{\Lambda} \begin{pmatrix} \lambda_1(1) & \lambda_0(1) \\ \lambda_1(2) & \lambda_0(2) \end{pmatrix} = \begin{pmatrix} 0.013 & 0.252 \\ 0.723 & 0.013 \end{pmatrix}.$$

Hence, in this model the decision to vote for the PI depends strongly on factor x. The observation of $(x = 1, y = 0)$ and $(x = 0, y = 1)$ are rare phenomenons in this fictive example. The MLE $\underline{\hat{\beta}}_n := (\hat{\beta}_{0,n}, \hat{\beta}_{1,n})'$ is

$$\left(\log\left(\frac{Z^{(1,1)}}{Z^{(1,0)}} \right), \log\left(\frac{Z^{(1,0)}}{Z^{(1,1)}} \cdot \frac{Z^{(2,1)}}{Z^{(2,0)}} \right) \right)'. \tag{6.1}$$

The limiting distribution of $\underline{\hat{\beta}}_n$ for $n \to \infty$ is, in accordance with section 3.3.4,

[2] See page 56.

$$\sqrt{n}(\underline{\hat{\beta}}_n - \underline{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}_2(\underline{0}, \underline{\underline{\Sigma}}(\underline{\beta})^{-1}), \tag{6.2}$$

with

$$\underline{\underline{\Sigma}}(\underline{\beta})^{-1} = \begin{pmatrix} \frac{1}{\pi_1(1-\pi_1)\frac{\lambda_{\cdot}(1)}{\Lambda}} & -\frac{1}{\pi_1(1-\pi_1)\frac{\lambda_{\cdot}(1)}{\Lambda}} \\ -\frac{1}{\pi_1(1-\pi_1)\frac{\lambda_{\cdot}(1)}{\Lambda}} & \frac{\pi_1(1-\pi_1)\frac{\lambda_{\cdot}(1)}{\Lambda}+\pi_2(1-\pi_2)\frac{\lambda_{\cdot}(2)}{\Lambda}}{\pi_1(1-\pi_1)\frac{\lambda_{\cdot}(1)}{\Lambda}\,\pi_2(1-\pi_2)\frac{\lambda_{\cdot}(2)}{\Lambda}} \end{pmatrix} =$$

$$= \Lambda \begin{pmatrix} \frac{1}{\pi_1(1-\theta)} & -\frac{1}{\pi_1(1-\theta)} \\ -\frac{1}{\pi_1(1-\theta)} & \frac{\pi_1(1-\theta)+\pi_2\theta}{\pi_1(1-\theta)\cdot\pi_2\theta} \end{pmatrix} = 3.777 \begin{pmatrix} 22.20 & -22.20 \\ -22.20 & 42.56 \end{pmatrix}.$$

Hence

$$\begin{aligned} \sqrt{\pi_1(1-\pi_1)\frac{\lambda_{\cdot}(1)}{\Lambda}}\,\sqrt{n}\,(\hat{\beta}_{0,n} - \beta_0) = \\ = \sqrt{\frac{\pi_1(1-\theta)}{\Lambda}}\,\sqrt{n}\,(\hat{\beta}_{0,n} - \beta_0) = \\ = 0.1092\,\sqrt{n}\,(\hat{\beta}_{0,n} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \end{aligned} \tag{6.3}$$

and

$$\begin{aligned} \sqrt{\frac{\pi_1(1-\pi_1)\frac{\lambda_{\cdot}(1)}{\Lambda}\,\pi_2(1-\pi_2)\frac{\lambda_{\cdot}(2)}{\Lambda}}{\pi_1(1-\pi_1)\frac{\lambda_{\cdot}(1)}{\Lambda}+\pi_2(1-\pi_2)\frac{\lambda_{\cdot}(2)}{\Lambda}}}\,\sqrt{n}\,(\hat{\beta}_{1,n} - \beta_1) = \\ = \sqrt{\frac{\pi_1(1-\theta)\cdot\pi_2\theta}{\pi_1(1-\theta)+\pi_2\theta}}\sqrt{n}\,(\hat{\beta}_{1,n} - \beta_1) = \\ = 0.0245\,\sqrt{n}\,(\hat{\beta}_{1,n} - \beta_1) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \end{aligned} \tag{6.4}$$

Equation (6.3) corresponds to equation (2.7) with $p = \frac{\lambda_{\cdot}(2)}{\Lambda}$. Analogous equation (6.4) corresponds to equation (2.8).[3] Due to equation (6.2),

$$\tilde{T}_n^{\text{Wald}} := n \cdot (\underline{\hat{\beta}}_n - \underline{\beta})' \cdot \underline{\underline{\Sigma}}(\underline{\beta}) \cdot (\underline{\hat{\beta}}_n - \underline{\beta}) \tag{6.5}$$

converges to the central χ^2-distribution with two degrees of freedom. Note that the behavior of $\mathcal{L}$-convergence of $\tilde{T}_n^{\text{Wald}}$ correspond exactly to the behavior of $\mathcal{L}$-convergence of the suitably scaled version of the estimator $\underline{\hat{\beta}}$ to its limiting bivariate standard normal distribution. In this section we aim at exploring this convergence. Therefore, the focus is to discuss the behavior of $\mathcal{L}$-converge of this statistic.

[3] See page 15.

The equations (6.3), (6.4) and (6.5) are statements about convergence in distribution. In equation (6.3) and equation (6.4) the limiting distribution is the standard normal distribution, in equation (6.5) it is the central χ^2-distribution with two degrees of freedom. In equations (6.3) and (6.4) a suitably scaled version of the $\hat{\beta}_{0,n}$ or alternatively $\hat{\beta}_{1,n}$ converges to the standard normal distribution. The cumulative distribution function $G_{j,n}$ denotes the suitably scaled version of the $\hat{\beta}_{j,n}$ and $\hat{G}_{j,n}^{(n_{\text{sim}})}$ the corresponding empirical cumulative distribution function based on n_{sim} simulations. The limiting standard normal distribution is denoted with Φ. The approximate Kolmogorov distance is $d_K(\Phi, \hat{G}_{j,n}^{(n_{\text{sim}})})$. The approximate Kolmogorov distance, which corresponds to equation (6.5), is denoted $d_K(\mathcal{H}_2, \hat{G}_n^{(n_{\text{sim}})})$. Here $\mathcal{H}_d$ is the cumulative distribution function of the central χ^2-distribution with d degrees of freedom and $\hat{G}_n^{(n_{\text{sim}})}$ is the empirical cumulative distribution function of $\tilde{T}_n^{\text{Wald}}$.

In the following we compute the distance-sample-size-diagram for the mean approximate Kolmogorov distances with respect to the statistics defined in the equations (6.3), (6.4) and (6.5) respectively. The approximate Kolmogorov distance with respect to equations (6.3) or (6.4) is

$$D_j(n) := d_K(\Phi, \hat{G}_{j,n}^{(n_{\text{sim}})}), \tag{6.6}$$

$j \in \{0, 1\}$ and the approximate Kolmogorov distance with respect to equation (6.5) is

$$D_{0,1}(n) := d_K(\mathcal{H}_2, \hat{G}_n^{(n_{\text{sim}})}). \tag{6.7}$$

To compute the approximate Kolmogorov distance means to draw a realization of the random variable $D_j(n)$ or alternatively $D_{0,1}(n)$.

6.2 Truncated multinomial distribution

To define the Kolmogorov distances $D_0(n)$, $D_1(n)$ and $D_{0,1}(n)$ the following conditions need to be satisfied: The corresponding empirical cumulative distribution functions $\hat{G}_{0,n}^{(n_{\text{sim}})}$, $\hat{G}_{1,n}^{(n_{\text{sim}})}$ and $\hat{G}_n^{(n_{\text{sim}})}$ must exist. If not, the approximate Kolmogorov distance is not defined. The definitions of these empirical cumulative distribution functions work if one component or both components of the estimator diverge.[4] In such a case the corresponding empirical cumulative distribution function is improper. Whereas, if one or both component of $\underline{\hat{\beta}}_n$ are undefined at least one of these empirical cumulative distribution functions does not exist. Thus to ensure that all three approximate Kolmogorov distance are defined, data sets causing undefined estimators are excluded in the following.

[4] Recall that the matrix $\underline{\underline{\Sigma}}(\underline{\beta})$ in the definition of the statistic $\tilde{T}_n^{\text{Wald}}$ is positive semidefinite.

Consider for example the following two-by-two tables: In the two-by-two table given in table 6.1 both components of $\hat{\underline{\beta}}_n$ diverge, the same applies for the two-by-two table given in table 6.2. In the two-by-two table given in table 6.3 $\hat{\beta}_{1,n}$ diverges. It is not unlikely to observe one of these combinations for the assumed regression coefficients.

		Y		
		1	0	
X	1	0	$z^{(1,0)}$	$z^{(1)}$
	2	$z^{(2,1)}$	0	$z^{(2)}$
		$z^{(\,.\,,1)}$	$z^{(\,.\,,0)}$	n

Table 6.1: $\hat{\beta}_{0,n} = -\infty$ and $\hat{\beta}_{1,n} = \infty$.

		Y		
		1	0	
X	1	0	$z^{(1,0)}$	$z^{(1)}$
	2	$z^{(2,1)}$	$z^{(2,0)}$	$z^{(2)}$
		$z^{(\,.\,,1)}$	$z^{(\,.\,,0)}$	n

Table 6.2: $\hat{\beta}_{0,n} = -\infty$ and $\hat{\beta}_{1,n} = \infty$.

		Y		
		1	0	
X	1	$z^{(1,1)}$	$z^{(1,0)}$	$z^{(1)}$
	2	$z^{(2,1)}$	0	$z^{(2)}$
		$z^{(\,.\,,1)}$	$z^{(\,.\,,0)}$	n

Table 6.3: $\hat{\beta}_{1,n} = \infty$.

		Y		
		1	0	
X	1	$z^{(1,1)}$	0	$z^{(1)}$
	2	$z^{(2,1)}$	0	$z^{(2)}$
		$z^{(\,.\,,1)}$	0	n

Table 6.4: $\hat{\beta}_{0,n} = \infty$ and $\hat{\beta}_{1,n} =$NaN.

However, in table 6.4 the estimator $\hat{\beta}_{1,n}$ is undefined. So the corresponding data set is excluded. This strategy means to truncate the distribution of the data matrix $\underline{\underline{Z}}$ so that invalid data sets are excluded from its domain.

Given that the matrix $\underline{\underline{Z}}$ follows a multinomial distribution, the probability to observe a two-by-two table that results in an undefined estimator decreases as the sample size increases. For the true parameter values assumed here, this probability is vanishingly small for $n > 30$, and still less than 0.001 for $n = 25$.

Because of the truncation of the domain of $\underline{\underline{Z}}$, in fact the distribution of this data matrix is a truncated multinomial distribution. Let $q(n)$ denote the probability to observe an invalid two-by-two table. The truncated multinomial distribution has then the density

$$\frac{1}{1-q(n)} \cdot \binom{n}{z^{(1,0)} z^{(1,1)} z^{(2,0)} z^{(2,1)}} \prod_{i=1}^{2} \prod_{y=0}^{1} \left(\frac{\lambda_y(i)}{\Lambda} \right)^{z^{(i,y)}} \cdot \mathbf{1}_{\mathcal{D}}, \tag{6.8}$$

where $\mathcal{D}$ is the set of valid two-by-two tables.

6.3 Convergence of the covariate effect estimator

It is interesting to study the behavior of $\mathcal{L}$-convergence with the help of the mean approximate Kolmogorov distance. To do this, we compute the mean approximate Kolmogorov distance in dependence of the predefined sample size n for both Firth-penalized and unpenalized likelihood methods.

In section 4.2.1 the penalized estimator $\underline{\hat{\beta}}_n^{\text{pen}}$ is defined. The limiting distribution of this estimator is the same as for $\underline{\hat{\beta}}_n$, because the penalty term is asymptotically negligible. (See Firth (1993).) Thus it is very interesting to compare the behavior of $\mathcal{L}$-convergence of the penalized and the unpenalized estimator. Let $P_0(n)$, $P_1(n)$ and $P_{0,1}(n)$ denote the corresponding approximate Kolmogorov distances for the penalized estimator. The vector $\underline{D}(n) := \big(D_0(n), D_1(n), D_{0,1}(n)\big)'$ and $\underline{P}(n) := \big(P_0(n), P_1(n), P_{0,1}(n)\big)'$.

To compare the behavior of $\mathcal{L}$-convergence for a predefined sample size n taking into account the precision, we simulate both $\underline{D}(n)$ and $\underline{P}(n)$ 20 times and compute the means $\bar{D}_0(n)$, $\bar{D}_1(n)$, $\bar{D}_{0,1}(n)$, $\bar{P}_0(n)$, $\bar{P}_1(n)$, $\bar{P}_{0,1}(n)$, and their standard deviations. The mean approximate Kolmogorov distance is, according to its definition given in equation (5.8),

$$\bar{D}_0(n) := \frac{1}{20}\sum_{l=1}^{20} D_0^{(l)}(n), \tag{6.9}$$

with $D_0^{(l)}(n)$ as the lth independent and identically distributed realization of $D_0(n)$.

The source code of the simulation study is written in the programming language R. The R-package logistf[5] is used to compute both the penalized and unpenalized statistics. It is possible that the unpenalized estimator diverges. The criterion-function is implemented in the source code in such a way so that the limiting value is returned if the argument is not finite. The estimator of the free component $\hat{\vartheta}$ of the deviance statistic, considered in section 6.4,

$$T_{j,n} = -2\log\left[\frac{l\big((\underline{\underline{X}}, \underline{Y}), h_j(\hat{\vartheta})\big)}{l\big((\underline{\underline{X}}, \underline{Y}), \underline{\hat{\beta}}\big)}\right]$$

is, for $j \in \{1, 2\}$, computed analytically for the unpenalized likelihood estimation methods.

[5] version 1.21, developed by Georg Heinze, Meinhard Ploner, Daniela Dunkler (former versions), Harry Southworth (former versions)

For the penalized likelihood estimation methods the R-function logistftest is used. The confidence intervals, needed in section 6.5, are computed using the R-method logistf. This function can compute the penalized and unpenalized Wald interval as well as the penalized and unpenalized profile interval.

The following graphics 6.1, 6.2 and 6.3 illustrate how the the mean approximate Kolmogorov distance between the actual distribution of a certain estimator and the corresponding limiting distribution can be used to study the behavior of $\mathcal{L}$-convergence. The distance-sample-size-diagram given in figure 6.1 compares $\bar{D}_0(n)$ with $\bar{P}_0(n)$.

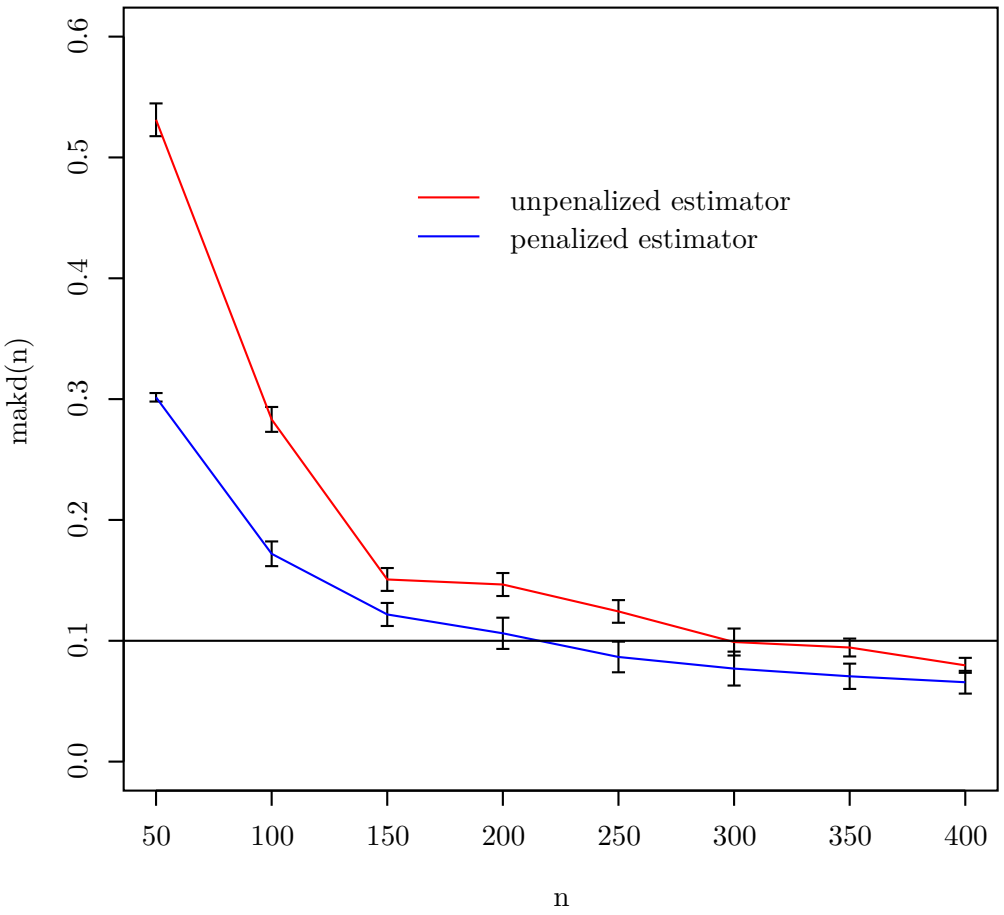

Figure 6.1: The distance-sample-size-diagram which corresponds to equation (6.3) is plotted for $\underline{b} = \underline{\beta} = (-3, 7)'$ and $\theta = 0.05$. The red line corresponds to $\bar{D}_0(n)$, the blue line to $\bar{P}_0(n)$.

Figure 6.2 compares $\bar{D}_1(n)$ and $\bar{P}_1(n)$. Finally, figure 6.3 compares $\bar{D}_{0,1}(n)$ and $\bar{P}_{0,1}(n)$. The mean approximate Kolmogorov distance is given for predefined sample sizes from 50 to 400 with an increment of 50. Both curves are computed with the help of the smooth.spline function of the R-package stats. The error bars along the curves give the standard deviations of the corresponding approximate Kolmogorov distance. The number of simulation

$n_{\text{sim}} = 1.5 \cdot 10^4$.

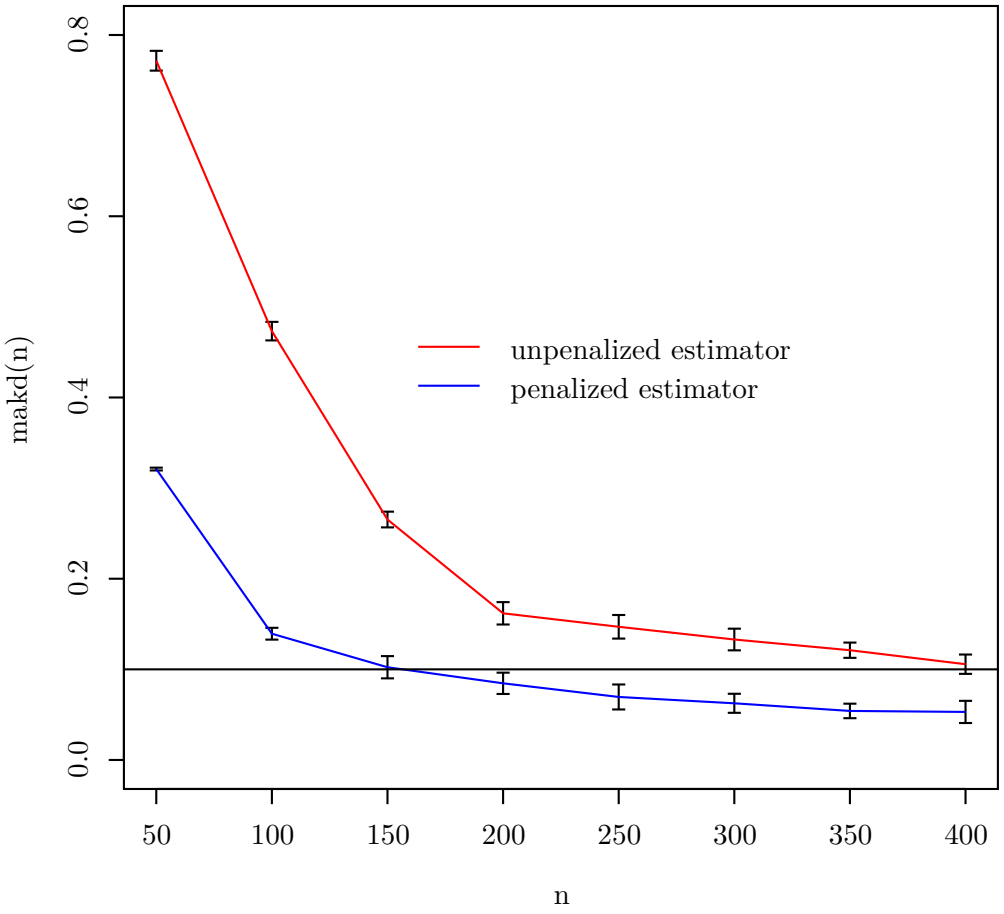

Figure 6.2: The distance-sample-size-diagram which corresponds to equation (6.4) is plotted for $\underline{b} = \underline{\beta} = (-3, 7)'$ and $\theta = 0.05$. The red line corresponds to $\bar{D}_1(n)$, the blue line to $\bar{P}_1(n)$.

In figure 6.1 the drop of the mean approximate Kolmogorov distance with increasing sample size is clearly faster for the penalized estimator. In particular for sample sizes smaller than 150 the penalized estimator performs much better. The distance-sample-size-diagram given in figure 6.2 illustrates that, for covariate effect β_1 the advantages of the penalization are even greater than for β_0. In particular for sample sizes smaller than 200 the penalized estimator performs much better. As expected, the distance-sample-size-diagram of equation (6.5) given in figure 6.3 provides the evidence: the drop of the mean approximate Kolmogorov distance with increasing sample size is clearly faster for the penalized estimator in this example.

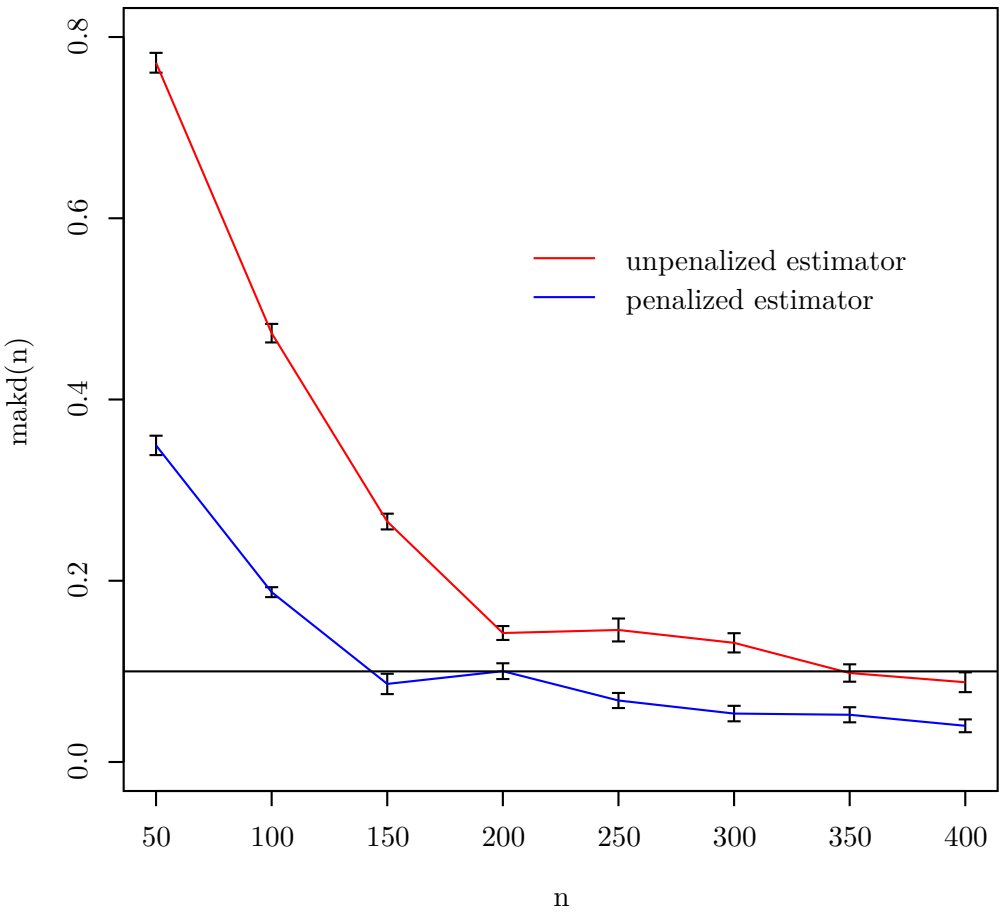

Figure 6.3: The distance-sample-size-diagram which corresponds to equation (6.5) is plotted for $\underline{b} = \underline{\beta} = (-3, 7)'$ and $\theta = 0.05$. The red line corresponds to the $\bar{D}_{0,1}(n)$, the blue line to $\bar{P}_{0,1}(n)$.

n	200	250	300	350	400
$\bar{D}_0(n)$	0.1466	0.1243	0.0990	0.0944	0.0797
$\bar{P}_0(n)$	0.1062	0866	0.0770	0.0706	0.0657
$\bar{D}_1(n)$	0.1618	0.1470	0.1329	0.1211	0.1057
$\bar{P}_1(n)$	0.0846	0.0695	0.0625	0.0541	0.0530
$\bar{D}_{0,1}(n)$	0.1423	0.1457	0.1315	0.0982	0.0880
$\bar{P}_{0,1}(n)$	0.1002	0.0679	0.0535	0.0521	0.0399

Table 6.5: The realization of the mean approximate Kolmogorov distances in dependence of the sample size n.

Table 6.5 describes the mean approximate Kolmogorov distances plotted in figures 6.1, 6.2 and 6.3, starting from a sample size of 200. Per definition the sample size required for convergence is the sample size, which ensures that the Kolmogorov distance is less than $c := 0.1$. As can be seen from this table, the sample size, which is required to reach convergence is significantly larger for the unpenalized likelihood estimation methods. The penalized likelihood estimation methods reach convergence for sample sizes marginally larger than 200. However a sample size larger than 400 is required to in order to achieve convergence without penalization. The corresponding standard deviations can be found in the subsequent table 6.6. It makes obvious, that the random variables given in table 6.5 have a very low variation. Thus they are nearly deterministic. Due to the low standard deviations one can conclude that the number of simulations $n_{\text{sim}} = 1.5 \cdot 10^4$ is large enough to estimate the Kolmogorov distance with sufficient precision.

n	200	250	300	350	400
$\mathsf{sd}(\bar{D}_0(n))$	0.0095	0.0094	0.0112	0.0074	0.0062
$\mathsf{sd}(\bar{P}_0(n))$	0.0129	0.0127	0.0140	0.0105	0.0094
$\mathsf{sd}(\bar{D}_1(n))$	0.0123	0.0130	0.0119	0.0084	0.0107
$\mathsf{sd}(\bar{P}_1(n))$	0.0118	0.0138	0.0105	0.0080	0.0122
$\mathsf{sd}(\bar{D}_{0,1}(n))$	0.0077	0.0126	0.0106	0.0096	0.0109
$\mathsf{sd}(\bar{P}_{0,1}(n))$	0.0087	0.0083	0.0086	0.0083	0.0071

Table 6.6: The estimated standard deviations of the means given in table 6.5.

6.4 Convergence of the deviance statistic

So far we explored the behavior of $\mathcal{L}$-convergence of the covariate effect estimator $\underline{\hat{\beta}}_n$. It is also interesting to study the behavior of $\mathcal{L}$-convergence of the deviance statistic $T_{j,n}$ defined in section 3.3.6 as

$$T_{j,n} = -2\log\left[\frac{l(\underline{Z}^0, \underline{Z}^1, \underline{h}_j(\hat{\vartheta}_n))}{l(\underline{Z}^0, \underline{Z}^1, \underline{\hat{\beta}}_n)}\right].$$

This statistic is, under H_0, asymptotically χ^2-distributed with one degree of freedom. The deviance

$$T_n = -2\log\left[\frac{l(\underline{Z}^0, \underline{Z}^1, \underline{\beta})}{l(\underline{Z}^0, \underline{Z}^1, \underline{\hat{\beta}}_n)}\right] \tag{6.10}$$

is asymptotically χ^2-distributed with two degrees of freedom. $\hat{F}_{j,n}^{(n_{\text{sim}})}$ denotes the empirical cumulative distribution function, which corresponds to $T_{j,n}$. Analogous, $\hat{F}_n^{(n_{\text{sim}})}$ is the empirical cumulative distribution function used to estimate the cumulative distribution function of T_n. In the following

$$K_j(n) := d_K(\mathcal{H}_1, \hat{F}_{j,n}^{(n_{\text{sim}})}) \tag{6.11}$$

$$K_{0,1}(n) := d_K(\mathcal{H}_2, \hat{F}_n^{(n_{\text{sim}})}). \tag{6.12}$$

Just as before the behavior of $\mathcal{L}$-convergence of the unpenalized and the penalized estimator will be compared. The penalized versions of the deviance are

$$T_{j,n}^{\text{pen}} = -2\log\left[\frac{l^{\text{pen}}(\underline{Z}^0, \underline{Z}^1, \underline{h}_j(\hat{\vartheta}_n^{\text{pen}}))}{l^{\text{pen}}(\underline{Z}^0, \underline{Z}^1, \underline{\hat{\beta}}_n^{\text{pen}})}\right]$$

and

$$T_n^{\text{pen}} = -2\log\left[\frac{l^{\text{pen}}(\underline{Z}^0, \underline{Z}^1, \underline{\beta})}{l^{\text{pen}}(\underline{Z}^0, \underline{Z}^1, \underline{\hat{\beta}}_n^{\text{pen}})}\right].$$

The corresponding approximate Kolmogorov distances are denoted by $J_0(n)$, $J_1(n)$ and $J_{0,1}(n)$. The following table 6.7 describes the behavior of $\mathcal{L}$-convergence in dependence of n. The predefined sample size n runs from 100 to 400 with an increment of 50. Obviously the unpenalized likelihood methods reach convergence for $n \approx 200$, whereas the penalized likelihood methods converge for $n \approx 100$. In comparison with table 6.5, the sample size required for convergence is halved. Thus the deviance converges faster as the estimator itself. Moreover the Firth-penalization improves the behavior of $\mathcal{L}$-convergence significantly.

n	100	150	200	250	300	350	400
$\bar{K}_0(n)$	0.2131	0.1302	0.0736	0.0460	0.0365	0.0316	0.0276
$\bar{J}_0(n)$	0.0959	0.0648	0.0487	0.0344	0.0301	0.0296	0.0234
$\bar{K}_1(n)$	0.2624	0.1515	0.0852	0.0564	0.0475	0.0403	0.0359
$\bar{J}_1(n)$	0.0990	0.0514	0.0377	0.0311	0.0303	0.0262	0.0233
$\bar{K}_{0,1}(n)$	0.2736	0.1642	0.0940	0.0566	0.0424	0.0362	0.0364
$\bar{J}_{0,1}(n)$	0.1024	0.0724	0.0462	0.0492	0.0396	0.0333	0.0279

Table 6.7: The realization of the mean approximate Kolmogorov distances in dependence of the sample size n.

n	100	150	200	250	300	350	400
$\mathsf{sd}(\bar{K}_0(n))$	0.0092	0.0071	0.0048	0.0069	0.0097	0.0105	0.0075
$\mathsf{sd}(\bar{J}_0(n))$	0.0110	0.0097	0.0066	0.0068	0.0094	0.0067	0.0039
$\mathsf{sd}(\bar{K}_1(n))$	0.0104	0.0077	0.0063	0.0076	0.0102	0.0113	0.0090
$\mathsf{sd}(\bar{J}_1(n))$	0.0127	0.0063	0.0089	0.0086	0.0076	0.0092	0.0059
$\mathsf{sd}(\bar{K}_{0,1}(n))$	0.0124	0.0082	0.0085	0.0099	0.0100	0.0097	0.0080
$\mathsf{sd}(\bar{J}_{0,1}(n))$	0.0134	0.0103	0.0085	0.0096	0.0045	0.0094	0.0048

Table 6.8: The estimated standard deviations of the means given in table 6.7.

After all, the estimation of the Kolmogorov distance with $n_{\text{sim}} = 1.5 \cdot 10^4$ shows a very good precision. The standard deviation is low. Thus the number of simulations seems to be high enough to estimate the Kolmogorov distance with satisfying accuracy. For the parameter values assumed in this example, the Firth-penalization performs clearly better. In particular the penalized deviance statistic has very good convergence properties. It reaches convergence for $n = 100$.

If β_1 is decreased from 7 to 4, the advantage of penalization diminishes, but still the penalized likelihood estimation methods perform significantly better than the unpenalized likelihood methods. The assumed true regression coefficients are now:

$$\beta_0 = -3,$$

$$\beta_1 = 4,$$

$$\theta = 0.05,$$

thus

$$\pi_1 = 0.0474,$$

$$\pi_2 = 0.7311.$$

Table 6.9 describes $\underline{\bar{D}}(n)$ and $\underline{\bar{P}}(n)$. The subsequent table 6.10 informs about $\underline{\bar{K}}(n)$ and $\underline{\bar{J}}(n)$. The predefined sample size n is in $\{45, 55, 60, 70, 80, 90, 100, 125, 150\}$. The tables show that penalization is very effective to improve the $\mathcal{L}$-convergence of the convergence theorems given in equations (6.4) and (6.5).

n	45	55	60	70	80	90	100	125	150
$\bar{D}_0(n)$	0.1846	0.1876	0.1789	0.1564	0.1364	0.1222	0.1226	0.0964	0.0863
$\bar{P}_0(n)$	0.1752	0.1573	0.1542	0.1372	0.1184	0.1179	0.0998	0.0907	0.0787
$\bar{D}_1(n)$	0.2795	0.1945	0.1535	0.1254	0.1177	0.1125	0.1024	0.0874	0.0803
$\bar{P}_1(n)$	0.0806	0.0654	0.0656	0.0619	0.0551	0.0544	0.0488	0.0463	0.0405
$\bar{D}_{0,1}(n)$	0.2827	0.1955	0.1536	0.1213	0.1301	0.1226	0.0992	0.0784	0.0651
$\bar{P}_{0,1}(n)$	0.0782	0.0826	0.0869	0.0746	0.0547	0.0482	0.0529	0.0415	0.0342

Table 6.9: The realization of the mean approximate Kolmogorov distances in dependence of the sample size n for the covariate effect estimator and $\tilde{T}_n^{\text{Wald}}$.

n	25	35	45	55	60	70	80	90	100
$\bar{K}_0(n)$	0.2574	0.1905	0.1368	0.1028	0.1182	0.0888	0.0725	0.0776	0.0865
$\bar{J}_0(n)$	0.2345	0.1992	0.1206	0.0767	0.1295	0.0908	0.0617	0.0624	0.0801
$\bar{K}_1(n)$	0.3015	0.2285	0.1544	0.1123	0.0886	0.0647	0.0498	0.0471	0.0391
$\bar{J}_1(n)$	0.1148	0.0684	0.0516	0.0378	0.0374	0.0325	0.0345	0.0316	0.0318
$\bar{K}_{0,1}(n)$	0.2752	0.2046	0.1460	0.1048	0.0826	0.0597	0.0505	0.0448	0.0361
$\bar{J}_{0,1}(n)$	0.1308	0.0848	0.0634	0.0537	0.0539	0.0466	0.0422	0.0335	0.0383

Table 6.10: The realization of the mean approximate Kolmogorov distances in dependence of the sample size n for the deviance.

Note that the behavior of $\mathcal{L}$-convergence of $\tilde{T}_n^{\text{Wald}}$ corresponds exactly to the behavior of $\mathcal{L}$-convergence of the suitably scaled version of the estimator $\underline{\hat{\beta}}$ to its limiting standard normal distribution. One could also explore the behavior of $\mathcal{L}$-convergence of the Wald statistic

$$T_n^{\text{Wald}} := (\underline{\hat{\beta}}_n - \underline{\beta})' \cdot \underline{\underline{W}}_n(\underline{\hat{\beta}}_n) \cdot (\underline{\hat{\beta}}_n - \underline{\beta}).$$

Indeed, that is not precisely the same as to measure the behavior of $\mathcal{L}$-convergence of the root of the estimating equation. However, it is to be expected that the behavior of $\mathcal{L}$-convergence of T_n^{Wald} and $\tilde{T}_n^{\text{Wald}}$ are similar. Both statistics converge, due to the Cramér-Slutzky device, to the same limiting distribution.

For the changed assumed covariate effects the probability to observe a two-by-two table, which results in an undefined estimator, increases. For $n > 80$ it is vanishingly small. For $n = 60$, it is still less than 10^{-4}. For $n = 35$ it is approximately 0.005.

The exploration of the behavior of $\mathcal{L}$-convergence done so far suggests that the Firth-penalization improves the behavior of $\mathcal{L}$-convergence in general. The key advantage of the measurement of the behavior of $\mathcal{L}$-convergence with the help of the Kolmogorov distance is that it allows a sophisticated and clear comparison of penalized and unpenalized likelihood methods. The quality of an estimation method is often measured with help of the mean squared error, in particular of the bias. The distance-sample-size-diagram extends the spectrum of methods to explore the impacts of the Firth-penalization and to measure its quality.

6.5 Approximate accuracy function

The approximate accuracy function is introduced in section 2.2.3. It is a univariate function with argument b_j. Recall, the accuracy function

$$A_C(\beta_j, b_j, n) := \mathbb{P}_{\beta_j}(b_j \in C)$$

is defined as the probability that the value b_j is included in the confidence interval C. The approximate accuracy function $A_j(\beta_j, b_j, n)$ is the accuracy function based on the assumption that

$$\sqrt{\frac{n}{v_j}}\,(\hat{\beta}_{j,n} - b_j) \sim \mathcal{N}\Big(\sqrt{\frac{n}{v_j}}\,(\beta_j - b_j), 1\Big)$$

holds exactly. Recall that

$$\sqrt{\frac{1}{\hat{w}_{j,n}}}\,(\hat{\beta}_{j,n} - \beta_j) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

holds for $n \to \infty$. In the following we illustrate the approximate accuracy function again for the model with $m = 1$ covariate. In this case

$$\hat{w}_{1,n} := \frac{1}{n}\left(\frac{(1-\tilde{p})\hat{\pi}_1(1-\hat{\pi}_1) + \tilde{p}\hat{\pi}_2(1-\hat{\pi}_2)}{(1-\tilde{p})\hat{\pi}_1(1-\hat{\pi}_1)\tilde{p}\hat{\pi}_2(1-\hat{\pi}_2)}\right),$$

with $\hat{\pi}_1 := \frac{\exp(\hat{\beta}_{n,0})}{1+\exp(\hat{\beta}_{n,0})}$, $\hat{\pi}_2 := \frac{\exp(\hat{\beta}_{n,0}+\hat{\beta}_{n,1})}{1+\exp(\hat{\beta}_{n,0}+\hat{\beta}_{n,1})}$ and $\tilde{p} := \frac{z^{(2)}}{n}$. The corresponding Wald interval is

$$\hat{\beta}_1 - \tau_{1-\frac{\alpha}{2}}\sqrt{\hat{w}_{1,n}} \le b_1 \le \hat{\beta}_1 + \tau_{1-\frac{\alpha}{2}}\sqrt{\hat{w}_{1,n}}.$$

The penalized confidence interval is easily derived by replacing the unpenalized estimator by the penalized estimator. Given that the penalty term of the Firth-penalization is asymptotically negligible, the approximate accuracy function of the penalized estimator and the unpenalized estimator are the same. Thus it is interesting to compare the actual accuracy functions of the penalized and the unpenalized estimator with the approximate accuracy function. Following table 6.9 and table 6.10, for the true regression coefficients $\beta_0 = -3$, $\beta_1 = 4$ and $\theta = 0.05$, the penalized estimation methods outperform the unpenalized likelihood estimation methods, in particular for the covariate effect β_1. The performance of the corresponding confidence intervals is explored in this section.

The approximate accuracy function is

$$A_1(\beta_1, b_1, n) = \Phi(\tau_{\frac{\alpha}{2}} - \delta) - \Phi(-\tau_{\frac{\alpha}{2}} - \delta), \tag{6.13}$$

where Φ is the cumulative distribution function of the standard normal distribution and

$$\delta = (\beta_1 - b_1)\sqrt{\frac{n}{v_1}},$$

with

$$v_1 = \left(\frac{(1-p)\pi_1(1-\pi_1) + p\,\pi_2(1-\pi_2)}{(1-p)\pi_1(1-\pi_1)p\pi_2(1-\pi_2)}\right).$$

In this section $p := \frac{\lambda_{\cdot}(2)}{\Lambda}$. Obviously $\mathbb{E}(Z^{(2)}) = p \cdot n$, thus $\mathbb{E}(\frac{Z^{(2)}}{n}) = p$. In the following we give the actual accuracy function for the 95%-Wald interval, the penalized 95%-Wald interval, the 95%-profile interval and the penalized 95%-profile interval for β_1. These functions are denoted by $A_1^{\text{Wald}}(\beta_1, b_1, n)$, $A_1^{\text{PenWald}}(\beta_1, b_1, n)$, $A_1^{\text{Prof}}(\beta_1, b_1, n)$ and $A_1^{\text{PenProf}}(\beta_1, b_1, n)$, respectively [6].

The approximate accuracy function could also be denoted approximate Wald-accuracy function to distinguish it from the approximate profile-accuracy function. This function is the accuracy function of the profile interval based on the assumption that convergence is reached. The approximate Wald-accuracy function is easily derived, because the distribution of the Wald statistic is known as well under H_0 as under H_1. However the distribution of the deviance under H_1 cannot be expressed in an analytical form. At least near β_1 the distribution of the deviance statistic and the Wald statistic are similar. Thus the approximate Wald-accuracy function and the approximate profile-accuracy function coincide near the true regression coefficient β_1. (See section 2.2.4.) Given that the impact of the penalization vanishes with growing sample size n, the approximate accuracy function $A_1(\beta_1, b_1, n)$ should, close to β_1, nearly coincide with $A_1^{\text{Wald}}(\beta_1, b_1, n)$, $A_1^{\text{PenWald}}(\beta_1, b_1, n)$, $A_1^{\text{Prof}}(\beta_1, b_1, n)$ and $A_1^{\text{PenProf}}(\beta_1, b_1, n)$, if n is large enough.

For the assumed covariate effects $\beta_0 = -3$, $\beta_1 = 4$ and $\theta = 0.05$, the drop of the mean approximate Kolmogorov distance $\bar{D}_1(n)$ with increasing sample size is significantly slower than the drop of $\bar{P}_1(n)$. The same is true for $\bar{K}_1(n)$ and $\bar{J}_1(n)$. Table 6.11 summarizes

[6] The confidence level is 0.95 for all accuracy functions given in this work. It is omitted in the following.

the mean approximate Kolmogorov distance for the predefined sample sizes $n = 60$ and $n = 100$. For $n = 100$ convergence is (nearly) reached. Thus, one could expect that for $n = 100$, the accuracy functions of the penalized Wald interval, the unpenalized Wald interval, the profile interval and the unpenalized profile interval are very close to the corresponding approximate accuracy function. For $n = 60$ at least the unpenalized Wald interval is expected to perform badly.

Figure 6.4 shows the accuracy-diagrams for the predefined sample size $n = 100$. The accuracy functions $A_1^{\text{Wald}}(4, b_1, 100)$ and $A_1^{\text{PenWald}}(4, b_1, 100)$ are presented. Moreover this figure gives the accuracy functions $A_1^{\text{Prof}}(4, b_1, 100)$ and $A_1^{\text{PenProf}}(4, b_1, 100)$.

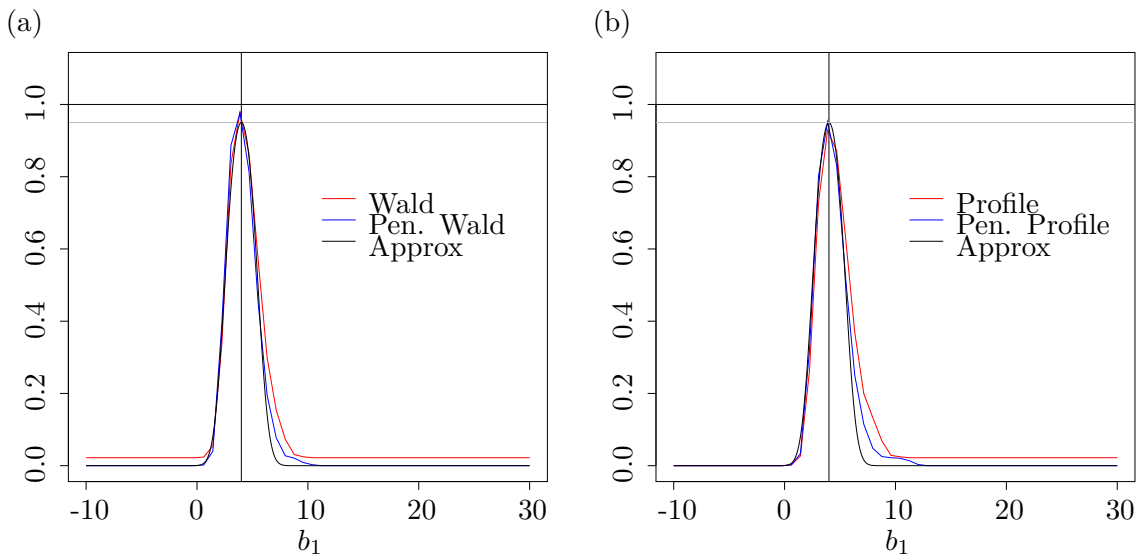

Figure 6.4: In both accuracy-diagrams, (a) and (b), the approximate accuracy function $A_1(4, b_1, 100)$ of the 95% Wald interval (black line) as given in equation (6.13) is plotted. The horizontal gray line marks the 95% confidence level. The vertical black line marks the true parameter value $\beta_1 = 4$.
In (a) the blue line depicts $A_1^{\text{PenWald}}(4, b_1, 100)$. The red line corresponds to $A_1^{\text{Wald}}(4, b_1, 100)$. In figure (b) the blue line is $A_1^{\text{PenProf}}(4, b_1, 100)$. The red line marks the function $A_1^{\text{Prof}}(4, b_1, 100)$.

n	60	100
$\bar{D}_1(n)$	0.1535	0.1024
$\bar{P}_1(n)$	0.0656	0.0488
$\bar{K}_1(n)$	0.0886	0.0391
$\bar{J}_1(n)$	0.0374	0.0318

Table 6.11: The mean approximate Kolmogorov distance for the predefined sample sizes $n = 60$ and $n = 100$.

According to table 6.11 convergence is reached for this sample size. Thus it is to be expected that all four accuracy functions are close to the approximate accuracy function $A_1(4, b_1, 100)$. Figure 6.4 confirms this expectation.

For $n = 60$ the mean approximate Kolmogorov distance $\bar{D}_1(n)$ has not converged. In fact, figure 6.5 shows that only the penalized confidence intervals deliver good results.

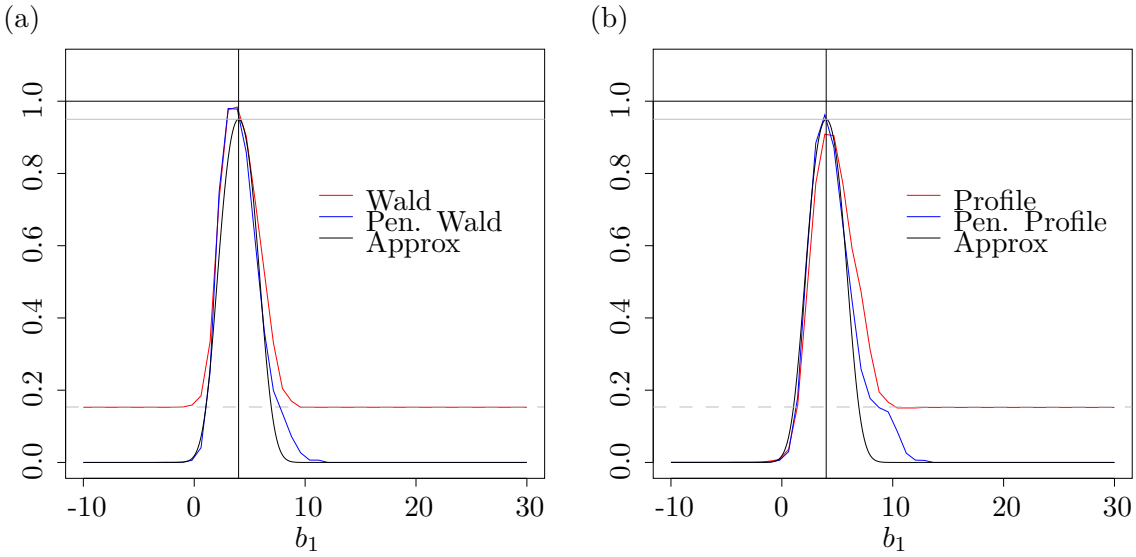

Figure 6.5: In both accuracy-diagrams, (a) and (b), the approximate accuracy function $A_1(4, b_1, 60)$ of the 95% Wald interval (black line) as given in equation (6.13) is plotted. The horizontal gray line marks the 95% confidence level. The vertical black line marks the true parameter value $\beta_1 = 4$. The horizontal gray, dashed line marks $\bar{D}_1(60)$. Here $\bar{D}_1(60)$ corresponds to the proportion of all simulations in which $\hat{\beta}_1$ diverges.
In (a) the blue line depicts $A_1^{\text{PenWald}}(4, b_1, 60)$. The red line corresponds to $A_1^{\text{Wald}}(4, b_1, 60)$. In figure (b) the blue line is $A_1^{\text{PenProf}}(4, b_1, 60)$. The red line marks the function $A_1^{\text{Prof}}(4, b_1, 60)$.

Indeed, the fast $\mathcal{L}$-convergence of $\bar{K}_1(n)$ does not imply that the corresponding confidence interval works well. The reason is that $\bar{K}_1(n) = 0.0886$, while $\hat{\beta}_1 = \infty$ in 15.35% of all simulations. If $\hat{\beta}_1 = \infty$ the limiting value of $T_{1,n}$ (under H_0) exists. Thus it is possible to compute the improper empirical cumulative distribution function of this statistic. Following table 6.11 the mean approximate Kolmogorov distance $\bar{D}_1(60) = 0.1535$. In fact, $\hat{\beta}_1 = \infty$ in 15.35% of all simulations. Figure 6.5 (a) represents this fact in form of

an accuracy function $A_1^{\text{Wald}}(4, b_1, 60)$ with "heavy tails."As expected, the actual accuracy functions of the unpenalized estimator performs badly, whereas $A_1^{\text{PenWald}}(4, b_1, 60)$ nearly coincides with the approximate accuracy function. In figure 6.5 (b), as expected, the actual accuracy functions of the penalized estimator $A_1^{\text{PenProf}}(4, b_1, 60)$ performs well. The function $A_1^{\text{Prof}}(4, b_1, 60)$ reflects the fact that $\hat{\beta}_1$ diverges to $+\infty$ in $\bar{D}_1(60) = 0.1535$ of all simulations.

Note the R-function logistf is used to compute the confidence intervals of interest. For computional reasons this R-function returns finite intervals even if the corresponding estimator is infinite. In fact, these confidence intervals are very wide.

7 Two applications of the p-value-uniform-diagram

This chapter uses two real data examples to show how the p-value-uniform-diagram can be useful to discover potential $\mathcal{L}$-convergence problems and to identify additional assumptions, which are necessary for basing statistical conclusions on the validity of the large-sample properties of the residual deviance. Such assumptions can concern the weight parameter $\underline{\theta}$ of the autogenerated process. If it should appear that it is questionable if the large-sample properties of the residual deviance hold, it is advisable to carefully consider if one can collect further data or to add special knowledge. In this way one can possibly justify to set one or more interaction parameters to zero and then use classical frequentist inference or to specify a prior distribution in order to analyze the data using Bayesian methods.

Both real data examples serve for measuring the relationship between a binary response Y and three binary independent variables stored in a vector $\underline{X} := (1, X_{(1)}, X_{(2)}, X_{(3)})$ by means of logistic regression analysis. For $m = 3$ the linear predictor η_i including all pairwise and higher-order interactions is

$$\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_{1,2} x_{i,1} x_{i,2} + \beta_{1,3} x_{i,1} x_{i,3} + \beta_{2,3} x_{i,2} x_{i,3} + \beta_{1,2,3} x_{i,1} x_{i,2} x_{i,3}.$$

There are $g = 2^m = 8$ covariate groups. The corresponding design matrix contains the covariate groups as row vectors.[1] In this section we assume that the fixed sample size protocol was used to collect the data. Even when, actually the data was collected in a time interval $(0, \tau]$ following the fixed time protocol, it may be more appropriate to base the exploration of the behavior of $\mathcal{L}$-convergence on the fixed sample size protocol. Indeed, the fixed sample size protocol is not based on the assumption of a homogenous Poisson

[1] It is given in equation (3.3) on page 27. The data set gives for each covariate group i the number of observations $z^{(i,y)}$ and can be found on page 31.

process. The fixed sample size protocol can perform well if the homogeneity assumption does not hold. This sampling protocol is based on the multinomial distribution with parameters n and probability matrix

$$\underline{\underline{\mathfrak{p}}} := \begin{pmatrix} \mathfrak{p}_{1,1} & \mathfrak{p}_{1,0} \\ \mathfrak{p}_{2,1} & \mathfrak{p}_{2,0} \\ \vdots & \vdots \\ \mathfrak{p}_{8,1} & \mathfrak{p}_{8,0} \end{pmatrix}.$$

This matrix can be parametrized in a way aiming at embedding the covariate effects $\underline{\beta}$ of the logistic regression model. The weight parameter $\underline{\theta}$ can be defined as a nuisance parameter of the bijective function mapping $\underline{\underline{\mathfrak{p}}}$ to the matrix

$$\frac{1}{\Lambda} \begin{pmatrix} \theta_1 \exp(\eta_1) & \theta_1 \\ \theta_2 \exp(\eta_2) & \theta_2 \\ \vdots & \vdots \\ \theta_8 \exp(\eta_8) & \theta_8, \end{pmatrix}$$

with

$$\sum_{i=1}^{8} \theta_i = 1,$$

$\lambda_y(i) := (\theta_i)^{1-y}(\theta_i \exp(\eta_i))^y$ and $\Lambda = \sum_{i=1}^{8} \sum_{y=0}^{1} \lambda_y(x)$.

7.1 Residual deviance

$\mathcal{L}$-convergence problems are more likely to occur if the linear predictor includes all pairwise and higher-order interactions. From section 4 we know that separation can not occur if no cell in the data set is empty. But on the other hand, an empty cell does not necessarily cause separation. If the vector of covariate effects is restricted, the root of the estimating equation can be finite, even if empty cells are present in the data set.

If the data set is sparse, it is likely to be nearly separated. In this case the estimator of covariate effects exists, but its asymptotic properties may be far from convergence.

In fact, if (nearly) separation occurs, the criterion-function is asymmetric, and thus the approximation by its Taylor polynomial close to the true covariate effect does not work well. However, this approximation is needed to proof the convergence of the deviance statistic. (See Pruscha (2000) page 249.) That is the reason why, a slow convergence of the deviance may be expected in scenarios where nearly separation is likely to occur.

The pairwise and higher order interactions are described by the vector

$$\underline{\beta}_{\text{inter}} := \begin{pmatrix} \beta_{1,2} \\ \beta_{1,3} \\ \beta_{2,3} \\ \beta_{1,2,3} \end{pmatrix}.$$

To decide, for example, whether the null hypothesis

$$\underline{\beta}_{\text{inter}} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

is rejected or not, one uses a significance test based an the asymptotic χ^2-distribution of the deviance statistic with respect to the corresponding function h. Here h is[2]

$$h : \underline{\vartheta} = \begin{pmatrix} \vartheta_0 \\ \vartheta_1 \\ \vartheta_2 \\ \vartheta_3 \end{pmatrix} \mapsto \begin{pmatrix} \vartheta_0 \\ \vartheta_1 \\ \vartheta_2 \\ \vartheta_3 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

The null model contains here four free covariate effects, while the saturated model has eight free covariate effects.

[2] For the definition of the deviance statistic see section 3.3.6 on page 43.

If one uses the glm-function of the programming language R (in the base-package), R automatically compares the fit of the null model with the fit of the saturated model. This procedure will be denoted in the future as the deviance test of the null model against the alternative model. The programming language R refers to the difference between the number of covariate effects of the saturated model and the null model as residual degrees of freedom. The corresponding deviance statistic is denoted as residual deviance. This statistic compares the fit of the h-null model with the fit of the saturated model. It is

$$T_n^{\text{ResDev}} = -2 \log \left[\frac{l(\underline{z}^0, \underline{z}^1; h(\hat{\underline{\vartheta}}_n))}{l(\underline{z}^0, \underline{z}^1; \hat{\underline{\beta}}_n)} \right].$$

The vector $\underline{z}^y$ is defined in section 3.3.6 as $\left(z^{(1,y)}, \ldots, z^{(i,y)}, \ldots, z^{(2^m,y)}\right)'$. The estimator, $\hat{\underline{\vartheta}} := \text{argmax}\{l(\underline{x}, \underline{y}; h(\underline{\vartheta}))\}$ is the parameter vector $\underline{\vartheta}$ maximizing the h-restricted criterion-function. As shown in section 3.3.6, under H_0, this statistic is asymptotically χ^2-distributed with d^* degrees of freedom. Here d^* is the number of tested regression coefficients. Thus in this example $d^* = 4$.

The corresponding p-value is

$$\mathrm{p} := 1 - \mathcal{H}_4(T_n^{\mathrm{ResDev}}), \tag{7.1}$$

where $\mathcal{H}_4$ is the cumulative distribution function of the χ^2-distribution with four degrees of freedom. If the residual deviance greatly exceeds the residual degrees of freedom, the null model is rejected. But, in fact, the Kolmogorov distance between the residual deviance and its reference distribution is suspected to drop slowly with increasing sample size.

The hypothesis test with significance level α, based on the residual deviance, is

$$\varphi(t_n^{\mathrm{ResDev}}) := \begin{cases} 1 & \mathrm{p} \leq \alpha; \\ 0 & \mathrm{p} > \alpha. \end{cases} \tag{7.2}$$

Generally, to introduce the test, based on p-values, has both advantages and disadvantages. On the one hand it defines a general scale for tests based on different test statistics. (See Rüger (2002) page 36ff). On the other hand it encourages the data-analyst to define the significance level α after computing the p-value. This practice is incompatible with principles of the classical inference. As a rule, the failure to reject the null hypothesis does not mean to accept the null hypothesis. But, large p-values correspond to low deviance values.

If the null hypothesis is true, then the p-value is uniformly distributed. This holds because of equation (7.1). It needs to be emphasized that under H_0, the p-value is uniformly distributed if $\mathcal{L}$-convergence is reached. In fact, a deviation from this reference (that is the uniform distribution) indicates either, that H_0 is false or that $\mathcal{L}$-convergence is not reached or both. Thus, if we simulate data under H_0 a deviation from the uniform suggests $\mathcal{L}$-convergence problems.

Now suppose that the sample size is large enough to ensure that the residual deviance reaches convergence. Then, if H_1 is true, it becomes more probable to observe small p-values. In this case the cumulative distribution function of the p-values is concave downwards. Thus, under H_0, a concave empirical cumulative distribution function indicates that the null hypothesis is rejected too often. In this case the statistical test is said to be anti-conservative. On the other hand a convex function indicates that it is rejected

too seldom. One speaks here of a conservative test.

Hence, to discover potential $\mathcal{L}$-convergence problems, the use of the p-value-scale is highly advantageous. Indeed, this scale reveals not only if $\mathcal{L}$-convergence problems arise, but also allows to identify if, as a consequence, the null hypothesis is rejected too often or too seldom. It would appear that it is unproblematic to reject H_0 too seldom if H_0 is true. But, the point is, that a very conservative test reveals a very low power to identify a wrong null hypothesis.

7.2 p-Value-uniform-diagram

Let G_n denote the actual cumulative distribution function of the residual deviance T_n^{ResDev} in dependance of the predefined sample size n. This predefined sample size is a measure of the rate at which empirical information accrues. Obviously, the Kolmogorov distance between G_n and its limiting distribution $\mathcal{H}_4$ and the Kolmogorov distance between the cumulative distribution function of the corresponding p-values given in equation (7.1), denoted as F_n, and the uniform cumulative distribution function $\mathcal{U}$ are equal. This Kolmogorov distance is denoted as $\tilde{R}(n) := d_K(\mathcal{H}_4, G_n)$. Given that G_n is unknown, $\tilde{R}(n)$ is approximated by $d_K(\mathcal{H}_4, \hat{G}_n^{(n_{\text{sim}})})$. Here $\hat{G}_n^{(n_{\text{sim}})}$ is the corresponding empirical cumulative distribution function.[3] This approximate Kolmogorov distance is denoted $R(n)$. It should be borne in mind that $\tilde{R}(n)$ is a deterministic value whereas $R(n)$ is a random variate. In fact,

$$R(n) := d_K(\mathcal{H}_4; \hat{G}_n) = d_K(\mathcal{U}; \hat{F}_n).$$

The following procedure assumes that H_0 is true and then simulates $\hat{F}_n$ and the corresponding realization of $R(n)$ to explore the behavior of $\mathcal{L}$-convergence under H_0. Here n equals the observed sample size. To discover whether the residual deviance suffers from $\mathcal{L}$-convergence problems, one simulates n_{sim} data sets. For each data set the residual deviance is computed. In this way the empirical cumulative distribution function of the residual deviance $\hat{G}_n$ is simulated. Due to equation (7.1) one can convert this empirical cumulative distribution function to the corresponding empirical cumulative distribution

[3] To achieve simple and clear formulas, In the following, the index n_{sim} indicating the number of simulations, used to estimate $\hat{G}_n^{(n_{\text{sim}})}$, will be left out.

function of the p-values $\hat{F}_n$. The realization of $R(n)$, that is $r(n)$, will be calculated based on this empirical cumulative distribution function.

First of all, it is of interest if $\mathcal{L}$-convergence is reached, in other words, if $r(n)$ is smaller than or equal to 0.1. Otherwise, one should study the course of the empirical cumulative distribution function $\hat{F}_n$. The p-value-uniform-diagram shows the course of this empirical cumulative distribution function and the cumulative distribution function of the uniform distribution $\mathcal{U}$ as its reference. A deviation from this reference indicates that $\mathcal{L}$-convergence is not reached. Let $\underline{\mathrm{p}} := (\mathrm{p}_1, \ldots, \mathrm{p}_l, \ldots, \mathrm{p}_{n_{\mathrm{sim}}})'$ denote the vector storing the simulated p-values. In fact, the vector $(F_n(\mathrm{p}_1), \ldots, F_n(\mathrm{p}_l), \ldots, F_n(\mathrm{p}_{n_{\mathrm{sim}}}))'$ is then a vector of uniformly distributed random numbers. Therefore the p-value-uniform-diagram can be used to define the corrected p-value as $\hat{F}_n(\mathrm{p})$, with p denoting the observed p-value.

7.3 Hypertension example

The example described in this section, demonstrates how the p-value-uniform-diagram can be used. It is taken from Altman (1990) and deals with the potential influence of the covariates smoking, obesity and snoring on the hypertensive status in human patients. The authors analyze the data using logistic regression and conclude, that obesity and snoring increase the risk of developing high blood pressure.

A central question is, whether the model specification is adequate. In fact, the authors did not include interactions between these variables in their model specification. In this section we verify whether the deviance test of the null model against the alternative model may suffer from $\mathcal{L}$-convergence problems.

The definition of the binary risk factor $X_{(j)}$, $j \in \{1, 2, 3\}$ is

$$X_{(j)} := \begin{cases} 0, & \text{risk factor } j \text{ is is present ;} \\ 1, & \text{risk factor } j \text{ is not is present.} \end{cases}$$

The definition of the binary response Y is

$$Y := \begin{cases} 0, & \text{no hypertension;} \\ 1, & \text{hypertension.} \end{cases}$$

7.3.1 Separated data

The data is stored in an R data.frame. The three binary covariates define eight groups. The R-vector hyp corresponds to $\underline{z}^1$. It contains the number of patients suffering from hypertension per group. The number of healthy patients is stored in not.hyp (which corresponds to $\underline{z}^0$). Remark that the group of smoking, obese, non-snoring patients ($i = 4$) has group size two. In this group both patients have normal blood pressure. The group of non-smoking, obese, non-snoring patients ($i = 3$) consists of eight patients out of which one has hypertension. In the group of smoking, non-obese, non-snoring patients ($i = 2$) only two suffer from high blood pressure, while the other 15 are not affected. All other components of the R-vectors hyp and not.hyp contain at least five patients.

smoking	obesity	snoring	hyp	not.hyp	total
No	No	No	5	55	60
Yes	No	No	2	15	17
No	Yes	No	1	7	8
Yes	Yes	No	0	2	2
No	No	Yes	35	152	187
Yes	No	Yes	13	72	85
No	Yes	Yes	15	36	51
Yes	Yes	Yes	8	15	23

The output of the R-function glm of the model including the main effects is:

```
Call:  glm(formula = data ~ smoking + obesity + snoring,
           family = binomial)

Coefficients:
  (Intercept)       smoking       obesity       snoring
   -2.37766      -0.06777       0.69531       0.87194
```

	R	true
$\hat{\beta}_0$	-2.398	-2.398
$\hat{\beta}_1$	0.383	0.383
$\hat{\beta}_2$	0.452	0.452
$\hat{\beta}_3$	0.929	0.929
$\hat{\beta}_{1,2}$	-21.340	$-\infty$
$\hat{\beta}_{1,3}$	-0.626	-0.626
$\hat{\beta}_{2,3}$	0.141	0.141
$\hat{\beta}_{1,2,3}$	21.830	∞

Table 7.1: The table gives the the true estimator and the R-estimator. All components are equal except $\hat{\beta}_{1,2}$ and $\hat{\beta}_{1,2,3}$.

```
Degrees of Freedom: 7 Total (i.e. Null);  4 Residual
Null Deviance:      14.13
Residual Deviance:  1.618       AIC: 34.54
```

Obviously the residual deviance is 1.618. The corresponding p-value is 0.8056.

```
pchisq(1.618,df=4,log.p=F,lower.tail=F)
[1] 0.8055534
```

The null hypothesis is not rejected. In fact, to compute the deviance the covariate effect estimator of the saturated model is needed. The data is separated because the group of smoking, obese, non-snoring patients ($i = 4$) consists of two healthy patients only, and there are no other ill patients, with the consequence $z^{(4,1)} = 0$. Thus the estimator of the saturated model must at least have one diverging component. However, the covariate effect estimator for the saturated model, computed by R, is finite. This is a consequence of the fact, that convergence is declared if the log-likelihood changes for less than a certain value between two iterations of the estimation algorithm. Indeed, two components of the covariate effect estimator diverge. To compute residual deviance, one needs to calculate the corresponding limiting value. It turns out, that this limiting value is 0.8056. Both estimates, the true estimator and the R-estimator, are given in the table 7.1.

7.3.2 Approximate Kolmogorov distance of the residual deviance

The following procedure can either confirm that, under H_0, $\mathcal{L}$-convergence is reached, or point out that this assumption is problematic.

We assume a true covariate effect of

$$\underline{\beta}_{H_0} := h(\underline{\hat{\vartheta}}) = \begin{pmatrix} -2.3777 \\ -0.0678 \\ 0.6953 \\ 0.87194 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

and a suitable vector $\underline{\theta}$ to simulate data following the fixed sample size protocol. The sample size used in the simulation equals the true sample size $n = 433$.

The vector $\underline{\theta}$ used in the simulation is

$$\begin{pmatrix} 0.1554 \\ 0.0424 \\ 0.0198 \\ 0.0056 \\ 0.4294 \\ 0.2034 \\ 0.1017 \\ 0.0424 \end{pmatrix}.$$

It is defined, so that the group sizes of the eight covariate groups correspond to the observed group sizes. In fact

$$\theta_i := \frac{z^{(i,0)}}{\sum_{i=1}^{8} z^{(i,0)}}. \tag{7.3}$$

Following the fixed sample size protocol the data matrix

$$\underline{\underline{Z}} = \begin{pmatrix} Z_n^{(1,1)} & Z_n^{(1,0)} \\ Z_n^{(2,1)} & Z_n^{(2,0)} \\ Z_n^{(3,1)} & Z_n^{(3,0)} \\ Z_n^{(4,1)} & Z_n^{(4,0)} \\ Z_n^{(5,1)} & Z_n^{(5,0)} \\ Z_n^{(6,1)} & Z_n^{(6,0)} \\ Z_n^{(6,1)} & Z_n^{(6,0)} \\ Z_n^{(8,1)} & Z_n^{(8,0)} \end{pmatrix},$$

follows a multinomial distribution with parameters n and probability matrix

$$\underline{\underline{\mathfrak{p}}} := \frac{1}{\Lambda} \begin{pmatrix} \lambda_1(1) & \lambda_0(1) \\ \lambda_1(2) & \lambda_0(2) \\ \lambda_1(3) & \lambda_0(3) \\ \lambda_1(4) & \lambda_0(4) \\ \lambda_1(5) & \lambda_0(5) \\ \lambda_1(6) & \lambda_0(6) \\ \lambda_1(7) & \lambda_0(7) \\ \lambda_1(8) & \lambda_0(8) \end{pmatrix}.$$

The matrix of expected values, thus $n \cdot \underline{\underline{\mathfrak{p}}}$, is given in the following R data.frame

smoking	obesity	snoring	hyp	not.hyp	total
No	No	No	5.094	54.906	60
Yes	No	No	1.356	15.644	17
No	Yes	No	1.254	6.746	8
Yes	Yes	No	0.296	1.704	2
No	No	Yes	33.954	153.046	187
Yes	No	Yes	14.596	70.404	85
No	Yes	Yes	15.698	35.302	51
Yes	Yes	Yes	6.752	16.248	23

Penalization	no	Firth
$r(433)$	0.06363	0.1311

Table 7.2: This table contains the approximate Kolmogorov distance $r(433)$. The first value corresponds to the unpenalized criterion-function. The Firth-penalization was used to compute the second value.

The approximate Kolmogorov distance $R(n)$ is simulated with $n_{\text{sim}} = 10^4$. In fact, $R(n)$ takes values between 0 and 1, with small values indicating that $\mathcal{L}$-convergence is reached.

First we compute the realization of the approximate Kolmogorov distance $R(n)$ between the empirical cumulative distribution function $\hat{F}_{433}$ and the uniform distribution. Thereafter we compute the same value based on the Firth-penalized criterion-function. The corresponding estimates are given in table 7.2 (on page 100). Obviously $\mathcal{L}$-convergence is reached for the unpenalized case. In case of the Firth-penalization $\mathcal{L}$-convergence is nearly reached. Denote with $\hat{F}^{\text{F}}_{433}$ the empirical cumulative distribution function of p-values computed using the Firth-penalization. This empirical cumulative distribution function is convex downwards. Thus, the corresponding test is a bit too conservative. The measurement of the approximate Kolmogorov distance does not indicate $\mathcal{L}$-convergence problems in this example. Thus from this point of view, there is no reasonable doubt regarding the deviance test of the null model against the saturated model.

7.4 Cesarean section example

This second example uses data from a study investigating the potential influence of three binary covariates on the occurrence or non-occurrence of infection following birth by cesarean section. It is taken from Fahrmeir/Tutz (2001). The authors analyze the data using logistic regression and conclude, in a first step of their study, that the prophylactic application of antibiotics affects positively the protection of infections. Moreover, the presence of risk factors, as for example diabetes or excessive weight, increase the risk of infection. This effect is higher, if the cesarean section was not planned. A central question is, whether the model specification is adequate. In this section, as in the previous section, we verify if the residual deviance may suffer from $\mathcal{L}$-convergence problems.

7.4.1 Covariate effect estimator

The definition of the binary covariate $X_{(j)}$, $j \in \{1,2,3\}$ is as follows: $X_{(1)}$ answers the question whether or not the cesarean section was planned,

$$X_{(1)} := \begin{cases} 0, & \text{the cesarean section was planned;} \\ 1, & \text{the cesarean section was not planned.} \end{cases}$$

This covariate is denoted NOPLAN by Fahrmeir/Tutz (2001). The second covariate informs about the presence of one or more risk factors, such as diabetes, excessive weight, early labor and others. Thus, $X_{(2)}$ answers the question whether or not such a risk factor was present,

$$X_{(2)} := \begin{cases} 0, & \text{there were no risk factors present;} \\ 1, & \text{there were risk factors present.} \end{cases}$$

This covariate is denoted FACTOR by Fahrmeir/Tutz (2001). Finally, the third covariate $X_{(3)}$ states whether antibiotics were given as prophylaxis. It is

$$X_{(3)} := \begin{cases} 0, & \text{there were no antibiotics given;} \\ 1, & \text{there were antibiotics given.} \end{cases}$$

This covariate is denoted ANTIB by Fahrmeir/Tutz (2001). The definition of the binary response Y is

$$Y := \begin{cases} 0, & \text{no infection;} \\ 1, & \text{infection.} \end{cases}$$

The data is stored in a R data.frame. The R-vector INF corresponds to $\underline{z}^1$. It contains the number of infections per group. The number of births without infection is stored in NO.INF.

	NOPLAN	FACTOR	ANTIB	INF	NO.INF	total
1	No	No	No	8	32	40
2	Yes	No	No	0	9	9
3	No	Yes	No	28	30	58
4	Yes	Yes	No	23	3	26
5	No	No	Yes	0	2	2

6	Yes	No	Yes	0	0	0
7	No	Yes	Yes	1	17	18
8	Yes	Yes	Yes	11	87	98

Mind that the group $i = 6$ is empty. Moreover, in the groups with $i = 2$ and $i = 5$ the number of infections $z^{(i,1)}$ is zero. Thus, the covariate effect estimator of the saturated model is undefined. Indeed for the model including all pairwise and higher order interactions this estimator is given by a closed form solution. The covariate effect estimator

$$\hat{\beta}_{1,3} = \log\left(\frac{Z_n^{(1,1)}}{Z_n^{(1,0)}}\right) - \log\left(\frac{Z_n^{(2,1)}}{Z_n^{(2,0)}}\right) - \log\left(\frac{Z_n^{(5,1)}}{Z_n^{(5,0)}}\right) + \log\left(\frac{Z_n^{(6,1)}}{Z_n^{(6,0)}}\right).$$

The lack of information in the data concerns mainly this interaction effect and the third order interaction $\beta_{1,2,3}$. Hence, these parameters are suspected to be very sensitive to the kind of penalization. The other two pairwise interaction estimators are in case of the saturated model

$$\hat{\beta}_{1,2} = \log\left(\frac{Z_n^{(1,1)}}{Z_n^{(1,0)}}\right) - \log\left(\frac{Z_n^{(2,1)}}{Z_n^{(2,0)}}\right) - \log\left(\frac{Z_n^{(3,1)}}{Z_n^{(3,0)}}\right) + \log\left(\frac{Z_n^{(4,1)}}{Z_n^{(4,0)}}\right)$$

and

$$\hat{\beta}_{2,3} = \log\left(\frac{Z_n^{(1,1)}}{Z_n^{(1,0)}}\right) - \log\left(\frac{Z_n^{(3,1)}}{Z_n^{(3,0)}}\right) - \log\left(\frac{Z_n^{(5,1)}}{Z_n^{(5,0)}}\right) + \log\left(\frac{Z_n^{(7,1)}}{Z_n^{(7,0)}}\right).$$

In Fahrmeir/Tutz (2001), the authors have included one interaction term in their analysis, namely $\beta_{1,2}$ the interaction between the covariates NOPLAN and FACTOR. The authors assume that the interaction effects $\beta_{1,3}, \beta_{2,3}$ and $\beta_{1,2,3}$ are zero. Furthermore, they penalized the criterion-function by adding a fictitious datum to the group with $i = 2$. More precisely: 0.5 is added to $z^{(2,0)}$ and to $z^{(2,1)}$. The corresponding estimator is given in the following R-output.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3880     0.3952  -3.512 0.000444 ***
NOPLAN       -1.5565     1.5038  -1.035 0.300661
FACTOR        1.3622     0.4729   2.881 0.003968 **
ANTIB        -3.8294     0.6025  -6.355 2.08e-10 ***
```

```
NOPLAN*FACTOR    3.4052        1.6138    2.110  0.034861 *
```

```
Residual deviance:  0.95499  on 2  degrees of freedom
```

Obviously the residual deviance is 0.95499 with two degrees of freedom. The corresponding p-value is 0.6203. This p-value does not suggest rejecting the null model. Mind that R does not compare the fit of the null model with the fit of the saturated model. Indeed, the covariate effect estimator of the saturated model does not exist, even after penalization. Therefore the alternative model used to compute the residual deviance is the model based on the assumption $\beta_{1,2,3} = 0$. In the following, this model will be denoted nearly saturated model. Section 7.4.2 will illustrate the effects of different penalizations.

7.4.2 Three different penalizations

In section 4.2 the c-penalized criterion-function is defined as

$$l^c(\underline{z}^0, \underline{z}^1; \underline{\beta}) := \sum_{i=1}^{g} \left(z^{(i,1)} + c\right) \eta(\underline{\beta}, \underline{x}_i) - \left(z^{(i)} + 2\,c\right) \log\left(1 + \exp(\eta(\underline{\beta}, \underline{x}_i))\right),$$

with $c \in \mathbb{R}^+$. As outlined in section 4.2 the c-penalized criterion-function $l^c(\underline{z}^0, \underline{z}^1; \underline{\beta})$ ensures that the covariate effect estimator exists. In the following the approximate Kolmogorov distance is denoted $R(c, n) := d_K(\mathcal{H}_2; \hat{G}_n(c))$, where $\hat{G}_n(c)$ is the empirical cumulative distribution function of

$$T^{\text{ResDev}}_{(c,n)} := -2 \log\left[\frac{l^c(\underline{z}^0, \underline{z}^1; h(\underline{\hat{\vartheta}}^c_n))}{l^c(\underline{z}^0, \underline{z}^1; \underline{\hat{\beta}}^c_n)}\right].$$

Accordingly $\hat{F}_n(c)$ denotes the empirical cumulative distribution function of the p-values

$$\text{p} := 1 - \mathcal{H}_2(T^{\text{ResDev}}_{(c,n)}).$$

Thus

$$R(c, n) = d_K(\mathcal{H}_2; \hat{G}_n(c)) = d_K(\mathcal{U}; \hat{F}_n(c)).$$

Again, $\mathcal{U}$ denotes the cumulative distribution function of the uniform distribution, $\mathcal{H}_2$ is the cumulative distribution function of the χ^2-distribution with two degrees of freedom. Correspondingly the approximate Kolmogorov distance based on the penalization criterion-function $l^{\underline{\underline{c}}}(\underline{z}^0, \underline{z}^1; \underline{\beta})$ is denoted $R(\underline{\underline{c}}, n) = d_K(\mathcal{H}_2; \hat{G}_n(\underline{\underline{c}})) = d_K(\mathcal{U}; \hat{F}_n(\underline{\underline{c}}))$. The following matrix $\underline{\underline{c}}$ corresponds to penalization implied by Fahrmeir/Tutz (2001),

$$\underline{\underline{c}}_{\mathrm{FT}} := \begin{pmatrix} 0 & 0 \\ 0.5 & 0.5 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

The following table gives the covariate effect estimators of the nearly saturated model based on three different penalizations. The estimators are computed by the R-function glm. The first penalization is implemented by adding the value 0.5 to $z^{(i,0)}$ and to $z^{(i,1)}$, for $i \in \{1, \ldots, 8\}$. The second is implemented by adding the value 0.5 to both $z^{(i,0)}$ and to $z^{(i,1)}$ for $i \in \{2, 5\}$. In the third case only for $i = 2$ the value of 0.5 is added. The third penalization is applied by Fahrmeir/Tutz (2001). The significant estimators are written in italic.[4] Note that, the estimators $\hat{\beta}_0$, $\hat{\beta}_2$ and $\hat{\beta}_{1,2}$ are significantly different from zero for all three penalizations. In the first case $\hat{\beta}_{2,3}$ is also significant. The comparison of the three estimators shows that the assessment of the preventive treatment with antibiotics depends on penalization. The data analysis suggests for all three penalizations, that the presence of risk factors increases the risk of infection. This effect is even stronger if the cesarean section was unplanned. It is also worthwhile noting that in the nearly saturated model, the interaction effect estimator $\hat{\beta}_{2,3}$ is very sensitive with respect to the kind of penalization. Presumably the effect of setting the parameter $\beta_{1,2,3}$ to zero is that the estimator of the pairwise interaction $\beta_{1,3}$ becomes less sensitive with respect to the penalization.

In the case of the first penalization it is also possible to estimate the covariate effects

[4] The significance level is 0.1.

	$i \in \{1,\ldots,8\}$	$i \in \{2,5\}$	$i = 2$
$\hat{\beta}_0$	*-1.407*	*-1.386*	*-1.386*
$\hat{\beta}_1$	-0.867	-1.558	-1.558
$\hat{\beta}_2$	*1.368*	*1.317*	*1.317*
$\hat{\beta}_3$	0.611	-0.223	-21.516
$\hat{\beta}_{1,2}$	*2.675*	*3.664*	*3.664*
$\hat{\beta}_{1,3}$	-0.973	-1.341	-1.341
$\hat{\beta}_{2,3}$	*-3.395*	-2.541	18.752
$\beta_{1,2,3}$	**0**	**0**	**0**

Table 7.3: Covariate effect estimators of the nearly saturated model.

for the saturated model. The corresponding residual deviance is 2.2609 with one degree of freedom. The p-value is thus 0.1327. This value does not suggest rejecting the nearly saturated model.

7.4.3 Approximate Kolmogorov distance of the residual deviance

In this section the approximate Kolmogorov distance of the residual deviance which tests the null hypothesis

$$H_0 : \beta_{1,3} = 0, \beta_{2,3} = 0, \beta_{1,2,3} = 0$$

versus the alternative hypothesis

$$H_1 : \beta_{1,2,3} = 0,$$

is estimated. The intention is, to examine the $\mathcal{L}$-convergence properties of the deviance test of the null model of Fahrmeir/Tutz (2001) against the nearly saturated model.

Thus, in accordance with the R-output given in section 7.4.1, the null hypothesis is

$$\underline{\beta}_{H_0} := h(\underline{\hat{\vartheta}}) = \begin{pmatrix} \hat{\vartheta}_0 \\ \hat{\vartheta}_1 \\ \hat{\vartheta}_2 \\ \hat{\vartheta}_3 \\ \hat{\vartheta}_4 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -1.39 \\ -1.56 \\ 1.36 \\ -3.83 \\ 3.41 \\ 0 \\ 0 \end{pmatrix}.$$

Note that the covariate effect vector has length seven. The reason is, that $\beta_{1,2,3} = 0$ is not tested. The alternative model has seven free covariate effects. Under H_0, the last two are set to zero. The residual deviance based on the $\underline{c}$-penalized criterion-function is

$$T^{\text{ResDev}}_{(\underline{c},n)} = -2 \log \left[\frac{l^{\underline{c}}(\underline{z}^{\,0}, \underline{z}^{\,1}; h(\underline{\hat{\vartheta}}^{\underline{c}}_{n}))}{l^{\underline{c}}(\underline{z}^{\,0}, \underline{z}^{\,1}; \underline{\hat{\beta}}^{\underline{c}}_{n})} \right].$$

Recall, under H_0, and if $\mathcal{L}$-convergence is reached, the statistic

$$\text{p}(T^{\text{ResDev}}_{(\underline{c},n)}) = 1 - \mathcal{H}_2(T^{\text{ResDev}}_{(\underline{c},n)})$$

is uniformly distributed. This section will explore the behavior of $\mathcal{L}$-convergence of $\text{p}(T^{\text{ResDev}}_{(\underline{c},n)})$. To do this, we simulate $n_{\text{sim}} = 10^4$ data sets under H_0 and compute for each data set the realization $\text{p}(t^{\text{ResDev}}_{(\underline{c},n)})$. The corresponding empirical cumulative distribution function is $\hat{F}_n(\underline{c})$. To simulated data under H_0, first the weight vector $\underline{\theta}$ needs to be specified. Above, in section 7.3, it was defined, so that the expected group sizes of the eight covariate groups equals the observed group sizes. For the cesarean section data, the group size of the covariate group with $i = 6$ is zero. In the simulation it should be possible to observe data in this group. Hence, the expected group size $\mathbb{E}(Z^{(6)})$ is set to one. In accordance with equation (7.3) the weight vector $\underline{\theta}_1$ which corresponds to these expected

frequencies, thus to $\mathsf{total} = (40, 9, 58, 26, 2, 1, 18, 97)'$, is

$$\underline{\theta}_1 := \begin{pmatrix} 0.1781 \\ 0.0476 \\ 0.1635 \\ 0.0201 \\ 0.0111 \\ 0.0056 \\ 0.0981 \\ 0.4759 \end{pmatrix}.$$

To study the influence of the weight parameter $\underline{\theta}$, the p-value-uniform-diagram will be computed for another weight parameter $\underline{\theta}_2$. This parameter corresponds to the expected frequencies $\mathsf{total} = (30, 9, 38, 26, 32, 51, 18, 47)'$. It is

$$\underline{\theta}_2 := \begin{pmatrix} 0.1217 \\ 0.0433 \\ 0.0976 \\ 0.0183 \\ 0.1614 \\ 0.2583 \\ 0.0894 \\ 0.2100 \end{pmatrix}.$$

The approximate Kolmogorov distance is calculated for three different penalizations. The

first is based on the 0.01-penalized criterion-function. Hence the penalizing matrix is

$$\underline{c}_{\leqq \text{light}} := \begin{pmatrix} 0.01 & 0.01 \\ 0.01 & 0.01 \\ 0.01 & 0.01 \\ 0.01 & 0.01 \\ 0.01 & 0.01 \\ 0.01 & 0.01 \\ 0.01 & 0.01 \\ 0.01 & 0.01 \end{pmatrix}.$$

This penalization is used to guarantee the existence of the estimator. It means only a minimal modification of the data set. The second penalization is a slightly modified version of the penalization implemented by Fahrmeir/Tutz (2001). The penalizing matrix is

$$\underline{c}_{\leqq \text{FTmod}} := \begin{pmatrix} 0.01 & 0.01 \\ 0.5 & 0.5 \\ 0.01 & 0.01 \\ 0.01 & 0.01 \\ 0.01 & 0.01 \\ 0.01 & 0.01 \\ 0.01 & 0.01 \\ 0.01 & 0.01 \end{pmatrix}.$$

Finally, the third penalization is the Firth-penalization. The following table 7.4 gives the corresponding approximate Kolmogorov distances. In fact, for $\underline{\theta}_1$, all three values indicate that convergence problems might arise. Therefore one should consider the p-value-uniform-diagrams showing the empirical cumulative distribution functions $\hat{F}_{251}(\underline{c}_{\leqq \text{light}})$, $\hat{F}_{251}(\underline{c}_{\leqq \text{FTmod}})$ and $\hat{F}_{251}^{\text{Firth}}$ given in figure 7.1 (on page 110). In the following $\hat{F}_{251}(\underline{c}_{\leqq \text{light}})$ is denoted $\hat{F}_{251}^{a}$ and $\hat{F}_{251}(\underline{c}_{\leqq \text{FTmod}})$ is denoted $\hat{F}_{251}^{b}$. Actually, in figure 7.1 the functions $\hat{F}_{251}^{a}$ and $\hat{F}_{251}^{b}$ are convex, indicating that the test based on the residual deviance is too conservative.[5] Indeed, the function $\hat{F}_{251}^{\text{Firth}}$ is not convex. For p-values less than or equal to

[5] One might think that it is unproblematic to reject H_0 too seldom, if H_0 is true. But, a very conservative test reveals that the power to identify a wrong null hypothesis is very low.

Penalization	$\underline{\underline{c}}_{\text{light}}$	$\underline{\underline{c}}_{\text{FTmod}}$	Firth
approximate Kolmogorov distance for $\underline{\theta}_1$	0.31919	0.32743	0.43672
approximate Kolmogorov distance for $\underline{\theta}_2$	0.12103	0.13098	0.18675

Table 7.4: This table contains the approximate Kolmogorov distances in dependence of the weight parameter $\underline{\theta}$. The predefined sample size n equals the observed sample size 251.

0.1, the function $\hat{F}_{251}^{\text{Firth}}$ has nearly no deviation from the reference. Because of the definition of the hypothesis test with significance level α, given in equation (7.2), the range $(0, \alpha]$ is crucial for the test decision. Common values for α are smaller or equal to 0.1. Hence the function $\hat{F}_{251}^{\text{Firth}}$ performs well in the crucial region. Therefore it is interesting to compute the p-value of the residual deviance under the Firth-penalization. Above, in section 7.4.1, the observed p-value of the residual deviance was computed using the penalization matrix $\underline{\underline{c}}_{FT}$. Is is 0.6203. If one recomputes this p-value using Firth-penalization it is 0.1495. This p-value also does not suggest rejecting H_0, indeed, the decision is less clear now. For the weight parameter $\underline{\theta}_2$ all three empirical cumulative distribution functions are convex. The corresponding p-value-uniform-diagram is given in figure 7.2.

Note that, if the value 0.01 is changed to a value within the interval $[0.005, 0.1]$, the empirical cumulative distribution functions of interest remain essentially unchanged.

The procedure described here certainly provides only limited usefulness in assessing whether statistical conclusions can be based on asymptotic properties. (See discussion below.) Nevertheless, for this example it is advisable to reflect on the plausibility of the null model carefully, and these considerations should be based on profound medical knowledge.

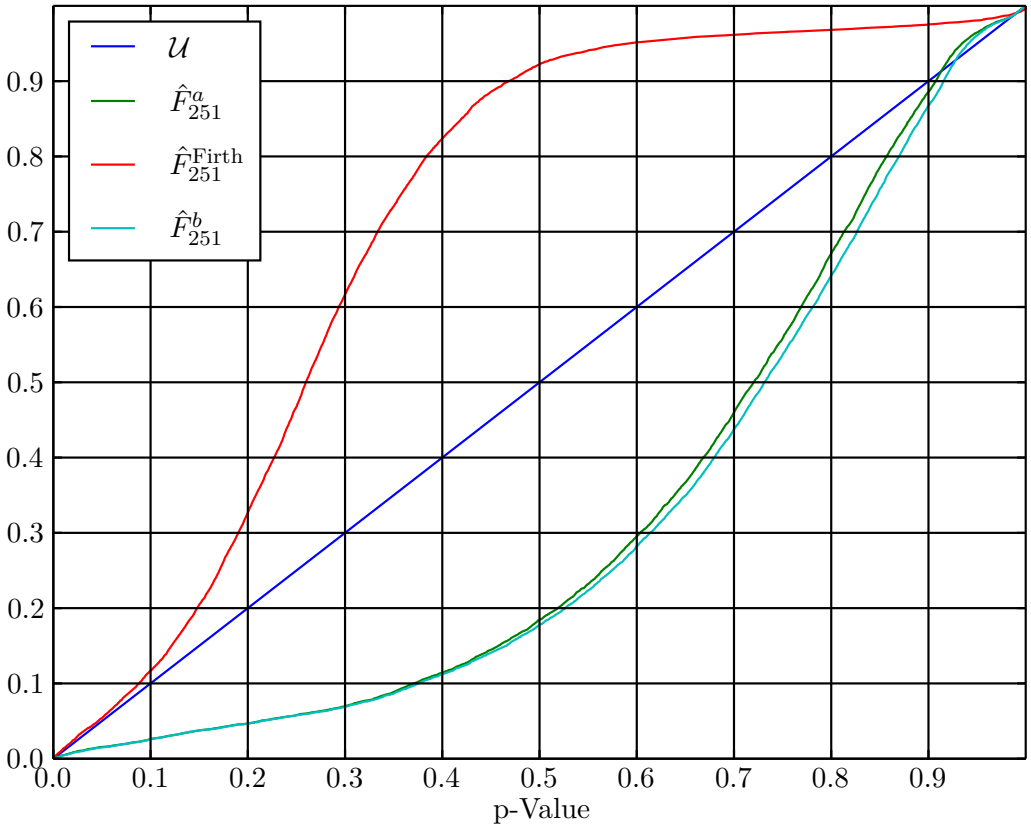

Figure 7.1: p-Value-uniform-diagram of $\hat{F}^{a}_{251}$, $\hat{F}^{\text{Firth}}_{251}$, $\hat{F}^{b}_{251}$ and the cumulative distribution function of the uniform distribution $\mathcal{U}$ for weight parameter θ_1.

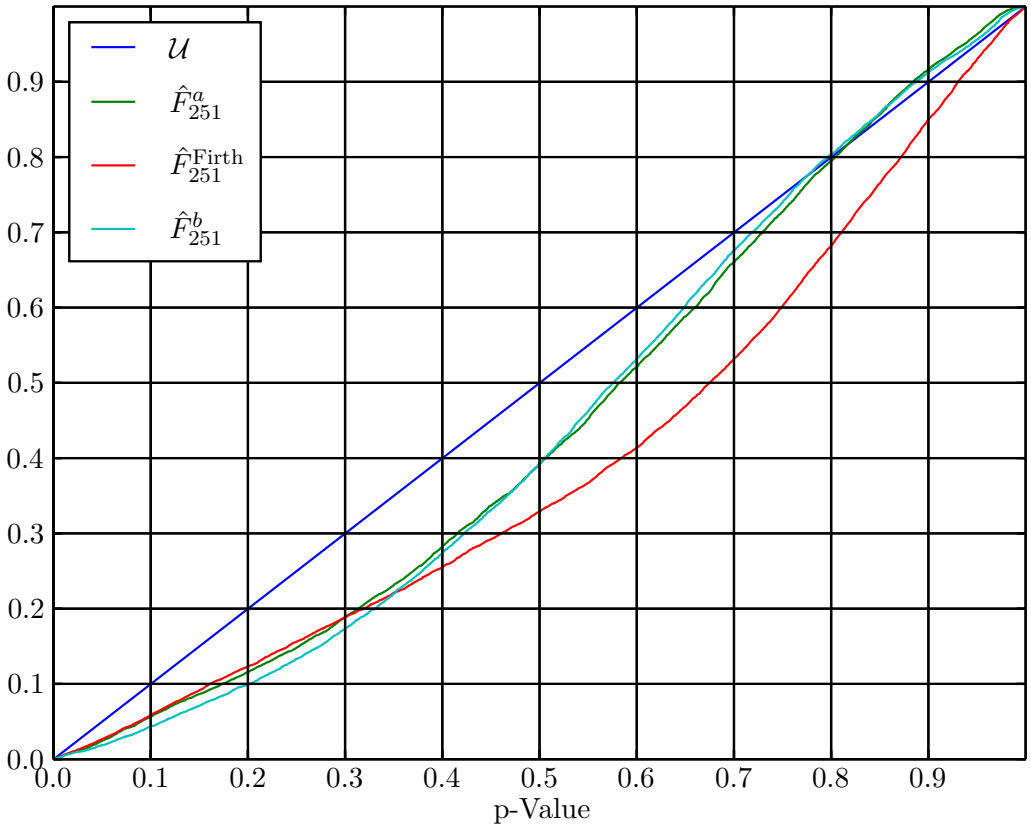

Figure 7.2: p-Value-uniform-diagram of the functions $\hat{F}^{a}_{251}$, $\hat{F}^{\text{Firth}}_{251}$, $\hat{F}^{b}_{251}$ and the cumulative distribution function of the uniform distribution $\mathcal{U}$ for weight parameter θ_2.

8 Concluding remarks and discussion

The present work reveals that the behavior of $\mathcal{L}$-convergence within logistic regression analysis is a complex phenomena. In fact, one has to assume that, within the logistic regression model, convergence problems often arise. As pointed out in sections 5.3, even the logistic regression model without covariates or with one covariate converge slowly for moderate covariate effect values. The fact that, with increasing covariate effect values, convergence becomes extremely slow is of particular interest. It is remarkable that, within the logistic regression model without covariates, if β_0 is changed from 2.5 to 4, the sample size necessary to reach an approximate Kolmogorov distance less than 0.1 goes from approximately 140 to approximately 560. If the number of covariates increases, the situation becomes more complicated.

In general, the data generating process, the specific definition of a (test) statistic and the estimation procedure determine not only the limiting distribution of this statistic but the whole process of convergence. The behavior of $\mathcal{L}$-convergence, more precisely the efficiency of the translation of the rate, at which empirical information accrues, into a drop of the Kolmogorov distance, is difficult to understand. The asymptotic frequentist interpretation of tests and confidence intervals is a very powerful tool, but it relies on large-sample assumptions. Consequently, the use of asymptotic methods in logistic regression analysis may result in misleading statical conclusions. In order to reveal the extend of this problem in applied statistics additional research is needed.

Indeed, if statistical conclusions are based on logistic regression, it is important that the data analyst has expert knowledge about classical asymptotic methods. Certainly these very powerful methods may lead to wrong statistical conclusions if large-sample assumptions do not hold. Due to this work it is possible to asses the drop of the Kolmogorov distance between the sampling distribution and the asymptotic distribution of (covari-

ate effect) estimators and test statistics in logistic regression analysis. In particular, the present work demonstrates how statistical tests, based on asymptotic properties, can be misleading if $\mathcal{L}$-convergence is not reached. The graphical representation of the behavior of $\mathcal{L}$-convergence gives an in-depth look at the validity of large-sample assumptions, at least within the autogenerated process.

The autogenerated process, which is fully compatible with the logistic regression model, is an interesting example of a data generating process. It is very instructive to use limit theorems for Poisson processes like the central limit theorem and the law of large numbers to proof key conditions for asymptotics. As a matter of fact, it is highly advisable to use this process in university lectures to explain asymptotic properties of statistical tests, estimators or confidence intervals within the logistic regression model. The autogenerated process provides a valuable opportunity to develop a clearer idea of how convergence in distribution works. Both the fixed time protocol and the fixed sample size protocol serve to illustrate how the sample distribution of the statistic of interest converges to its limiting distribution with increasing empirical information.

Of course, it is possible to generalize the autogenerated process as defined in chapter 3. For example the definition can involve continuous covariates.(See Diggle et al. (2010).) In this case the definition of the intensity process $\lambda_y(\underline{x}, \underline{\beta})$ is based on a weight function $\theta(\underline{x})$ instead of the weight parameter $\underline{\theta}$. For the fixed time protocol, it turns out that the variance matrix of the corresponding estimating equation is the expectation of the classical variance matrix. Note that Campbell's theorem is needed for the computation of this expectation. Hence, the methods presented in this work can be extended to logistic regression models with both binary and continuous independent variables. Using this approach one can, for example, study the impact of convergence problems on statistical conclusions based on large-sample properties.

The p-value-uniform-diagram, as discussed in chapter 7, illustrates how an asymptotic statistical test becomes delusive if large-sample assumptions do not hold. Actually, in the cesarean section example it is doubtful if the model specification can be based on the likelihood ratio test. Nevertheless the model specification is crucial for the statistical results. In Fahrmeir/Tutz (2001) the only interaction effect included in the model specification is the interaction between NOPLAN and FACTOR. If we include also the

interaction between ANTIB and FACTOR the conclusion that the prophylactic application of antibiotics affects positively the protection of infections becomes invalid. This example is also interesting because the statistical results depend on the penalization. The p-value-uniform-diagram shows how convergence properties of the likelihood ratio test depend on the penalization.

However, the p-value-uniform-diagram needs to be tested in applied statistics to reveal its diagnostic value. Moreover, the autogenerated process as it is used in this work should not be interpreted as an approach for modelling the real data generating process, but as a model reflecting some important aspects of this process. In any case, the use of this method can improve the discussion about the possibilities and limitations of inference concepts like classical inference, Bayesian inference or likelihood inference.

It is known that shrinkage methods, in particular the Firth-penalization improve parameter estimations in many situations.[1] The quality of an estimation method is often measured with help of the mean squared error, in particular of the bias. The graphical methods introduced in the present work add a new perspective. The distance-sample-size-diagram extends the spectrum of methods to explore the impacts of the Firth-penalization. As shown in chapter 6, the distance-sample-size-diagram and the accuracy-diagram are useful in this context. The distance-sample-size-diagram is a very sophisticated method to compare the behavior of convergence of penalized and unpenalized estimators. It reveals what sample size is necessary to reduce the approximate Kolmogorov distance between the sampling distribution of a statistic of interest and its limiting distribution to a certain value. It is interesting to contrast these sample sizes for the penalized and the unpenalized version of a statistic respectively. The accuracy-diagram usefully supplement the distance-sample-size-diagram. It is helpful to compare penalized and unpenalized confidence intervals. An important point to note is, that the unpenalized profile interval can be very wide despite the small Kolmogorov distance between the corresponding deviance and its asymptotic reference.

It is found that the Firth-penalization can improve the behavior of $\mathcal{L}$-convergence significantly. However, the present work is not meant to give a complete answer to this question. Instead, it provides valuable, beneficial methods to extend the ongoing discussion in the

[1] See for example Heinze (2006) and Heinze/Schemper (2002) and references given there.

scientific community.

Finally, the approaches developed in the present work invite to discuss the behavior of $\mathcal{L}$-convergence of penalized and unpenalized estimation methods in applied statistics. It will be interesting to review studies based on logistic regression analysis using these approaches. Particularly, It will be very instructive to compare the statistical conclusions derived from classical asymptotic inference to those derived from Bayesian inference.

Bibliography

Albert, A./Anderson, J. A. (1984): On the existence of maximum likelihood estimates in logistic regression models, in: Biometrika, 71(1), S. 1–10. 46

Altman, D. G. (1990): Practical Statistics for Medical Research (Chapman & Hall/CRC Texts in Statistical Science), 1. Aufl., Chapman and Hall/CRC. 10, 95

Chen, M.-H. H./Ibrahim, J. G./Kim, S. (2008): Properties and Implementation of Jeffreys's Prior in Binomial Regression Models., in: Journal of the American Statistical Association, 103(484), S. 1659–1664. 49

Diggle, P. J./Menezes, R./Su, T.-l. (2010): Geostatistical inference under preferential sampling, in: Journal of the Royal Statistical Society: Series C (Applied Statistics), 59(2), S. 191–232. 112

Fahrmeir, L./Tutz, G. (2001): Multivariate Statistical Modelling Based on Generalized Linear Models (Springer Series in Statistics), 2. Aufl., Springer. 10, 12, 100, 101, 102, 104, 105, 108, 112

Firth, D. (1993): Bias reduction of maximum likelihood estimates, in: Biometrika, 80(1), S. 27–38. 47, 50, 75

Foata, D./Fuchs, A. (1999): Wahrscheinlichkeitsrechnung (Grundstudium Mathematik) (German Edition), 1. Aufl., Birkhäuser Basel. 58

Heinze, G. (2001): The application of Firth's procedure to Cox and logistic regression, Techn. Ber., University of Vienna. 69

Heinze, G. (2006): A comparative investigation of methods for logistic regression with separated or nearly separated data., in: Statistics in medicine, 25(24), S. 4216–4226. 9,

47, 69, 113

Heinze, G./Schemper, M. (2002): A solution to the problem of separation in logistic regression., in: Statistics in medicine, 21(16), S. 2409–2419. 9, 47, 49, 69, 113

Lange, C. (1982): Fahrmeir, L./Kaufmann, H./Ost, F., Stochastische Prozesse. Eine Einführung in Theorie und Anwendungen. München-Wien, Carl Hanser Verlag 1981. XII, 366 S., 44 Abb., DM 44,âĂŤ. ISBN 3-446-13411-5, in: Z. angew. Math. Mech., 62(8), S. 426. 24

McCullagh, P. (2008): Sampling bias and logistic models, in: , 70(4), S. 643–677. 7

Nagaev, S. V./Chebotarev, V. I. (2011): On the bound of proximity of the binomial distribution to the normal one, in: Doklady Mathematics, 83(1), S. 19–21. 57

Pruscha, H. (2000): Vorlesungen über Mathematische Statistik (German Edition), 2000. Aufl., Vieweg+Teubner Verlag. 9, 11, 13, 14, 16, 17, 18, 22, 32, 33, 34, 35, 37, 39, 40, 42, 43, 49, 55, 91

Rüger, B. (1999): Testtheorie und Schätztheorie, Bd.1, Grundlagen, Oldenbourg. 17, 19, 20, 42, 48, 49

Rüger, B. (2002): Test- und Schätztheorie 2. Statistische Tests., Oldenbourg. 53, 55, 93

Zocher, M. (2005): Multivariate Mixed Poisson Processes, Dissertation, Technischen Universität Dresden,. 23

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorgelegte Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt und verfasst habe, insbesondere ohne die Hilfe einer Promotionsberaterin oder eines Promotionsberaters. Alle Hilfsmittel sind im Literaturverzeichnis genannt und alle Stellen, die ich wörtlich oder dem Sinne nach aus anderen Veröffentlichungen entnommen habe, sind kenntlich gemacht worden. Ich versichere außerdem, dass diese Dissertation in der vorgelegten oder einer ähnlichen Fassung noch nicht zu einem früheren Zeitpunkt an der Otto-Friedrich-Universität Bamberg oder einer anderen in- oder ausländischen Hochschule als Dissertation eingereicht worden ist.
Die vorliegende Arbeit ist bisher noch nicht publiziert worden.

Jena, den 21.01.2014

_____

Dipl.-Stat. Mariana Nold