Computational Statistics
https://doi.org/10.1007/s00180-025-01636-z

ORIGINAL PAPER

Maximum likelihood estimation of parameters for double poisson regression: a simulation study

Sebastian Appelbaum[1,2] · **Thomas Ostermann**[2] · **Uwe Konerding**[1,2]

Received: 3 April 2024 / Accepted: 27 April 2025

Abstract

Double Poisson Regression is specifically designed for regression of count variables and allows estimation of the parameters of a regression equation together with a dispersion parameter. Different computational procedures for obtaining maximum likelihood estimates of these parameters are possible. The objective of this contribution is to narrow down which of these computational procedures work best. Four different attributes of the computational procedures are investigated: (1) treatment of the normalisation factor in the Double Poisson with the two specifications: setting this factor equal to 1, and approximating this factor; (2) general estimation strategy with the two specifications: estimating the parameters of the regression equation and the dispersion parameters simultaneously, and estimating them sequentially; (3) starting value for the dispersion parameter with the two specifications: setting this value equal to 1, and computing it from data; and (4) algorithm with three variants of the Newton–Raphson algorithm, two variants of the BHHH algorithm and two variants of the BFGS algorithm as specifications. The four attributes of the computational procedures are investigated using simulation studies. The results of these studies show that the treatment of the normalisation factor very strongly affects parameter estimates and the quality of parameter estimation, whereas the other three attributes have no practically relevant effects. Moreover, the two treatments of the normalisation factor have opposite effects for different evaluation criteria. Therefore, neither treatment can be preferred. In data analyses, both treatments should be applied parallel to each other for sensitivity analysis.

Keywords Count variables · Double Poisson · Regression · Parameter estimation · Maximum likelihood

✉ Uwe Konerding
uwe.konerding@uni-bamberg.de

[1] Trimberg Research Academy, University of Bamberg, Bamberg, Germany

[2] Department of Psychology and Psychotherapy, University of Witten/Herdecke, Witten, Germany

Published online: 28 May 2025

1 Introduction

Double Poisson Regression (Efron 1986) is a statistical approach for identifying the determinants of count variables. Examples of count variables are number of migraine attacks (Silcocks et al. 2010), number of falls in old age (Luck et al. 2013), and number of days in hospital (Konerding et al. 2021). Double Poisson Regression has several advantages in comparison with other approaches also designed for count variables. For example, Double Poisson Regression is based on a model that describes the distribution of actual counts around true mean counts. This allows for the computing of proper likelihood functions and related statistics, and for the consideration of dispersion when estimating regression coefficients. This is not possible with approaches like Poisson Regression with robust standard errors, or Quasi-Poisson Regression (Cameron and Trivedi 2013; Hilbe 2014; Winkelmann 2008). Moreover, Double Poisson Regression allows for continuous modelling of the dispersion of the actual counts around the true mean count. This is not possible with many other model-based regressions for count data, such as classic Poisson Regression (Poisson 1837) and Negative Binomial Regression (Hilbe 2011). Furthermore, Double Poisson Regression provides regression coefficients that can be interpreted very well in terms of content. This is not given for Conway-Maxwell Regression (Conway and Maxwell 1962; Sellers and Premeaux 2020), which is also model-based and which also allows for continuous modelling.

In view of the features just described, Double Poisson Regression is a very promising approach for analysing the determinants of count variables. However, to tap the full potential of Double Poisson Regression, adequate procedures for obtaining maximum likelihood estimates of the corresponding parameters are needed. One approach that serves this purpose consists in formulating Double Poisson Regression as a special case of the generalised additive model for location, scale and shape (GAMLSS) (Rigby and Stasinopoulos 2005) and in using the corresponding R-package (Stasinopoulos and Rigby). Actually, this package already contains a function for Double Poisson Regression, and using it is certainly a viable option. However, computational tools that address a specific problem by treating it as a special case of a very general framework always run the risk of being unnecessarily cumbersome and less to the point regarding the needs of the specific problem.

With respect to parameter estimation for Double Poisson Regression, an approach specifically tailored to the problem consists in deriving, as far as possible, the procedure for parameter estimation directly from the regression model. However, the attempt to do so is connected with several problems. One problem arises because the Double Poisson distribution contains a normalisation factor that cannot be given in closed form (Efron 1986; Zou et al. 2013). This raises the question as to how this factor should be treated in parameter estimation. A second problem arises because two different kinds of parameters must be estimated. These are, on the one hand, parameters that determine the mean count, i.e., the additive parameter and the regression coefficients, and, on the other hand,

a parameter that models dispersion. Both estimations might interact with each other, resulting in impairment of parameter estimation (McNeish 2017). Hence, the question needs to be addressed as to how this problem should be handled. A third problem concerns the optimal starting value for the dispersion parameter. A fourth problem concerns the choice of the algorithm applied to determine the maximum of the likelihood function (Henningsen and Toomet 2011). Given this diversity of attributes according to which the computational procedure may vary, the question arises as to which combination of attribute specifications constitutes the optimal computational procedure.

The objective of the study presented here is to contribute to finding the optimal computational procedure. This contribution focusses on computational procedures that emerge through combinations of different solutions to the four problems just outlined. The whole argumentation is divided into five parts: (1) Double Poisson Regression is described in detail; (2) the different computational procedures to be evaluated are introduced; (3) the methodology for evaluating these procedures is presented; (4) the results of the evaluation are reported; and (5) the results are discussed.

2 Double Poisson Regression

Double Poisson Regression is based on a log-linear regression equation. The corresponding formula is

$$E\left[Y|x_i\right] = \exp\left(\beta_0 + x_i^T\beta\right) \tag{1}$$

with Y being the random variable of count values, and $\boldsymbol{x}_i$ being a vector composed of the predictor variable values of person i. $E\left[Y|\boldsymbol{x}_i\right]$ is the expected value of Y conditioned on $\boldsymbol{x}_i$, β_0 an additive parameter, and $\boldsymbol{\beta}$ the vector of the regression coefficients.

If μ_i is defined as

$$\mu_i = \exp\left(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}\right) \tag{2}$$

then the formula for Double Poisson Regression is

$$f\left(y_i|\mu_i,\alpha\right) = c\left(\mu_i,\alpha\right)\alpha^{1/2}e^{-\alpha\mu_i}\left(\frac{e^{-y_i}y_i^{y_i}}{y_i!}\right)\left(\frac{e\mu_i}{y_i}\right)^{y_i\alpha}, \alpha > 0 \tag{3}$$

with α being a parameter that models dispersion, and $c\left(\mu_i,\alpha\right)$ being the normalisation factor which guarantees that the sum of all probability mass function values is equal to one. This normalisation factor is

$$c\left(\mu_i,\alpha\right) = \frac{1}{\sum_{y=0}^{\infty}\alpha^{1/2}e^{-\alpha\mu_i}\left(\frac{e^{-y}y^y}{y!}\right)\left(\frac{e\mu_i}{y}\right)^{y\alpha}} \cong \frac{1}{1+\frac{1-\alpha}{12\alpha\mu_i}\left(1+\frac{1}{\alpha\mu_i}\right)} \tag{4}$$

According to Efron (1986), the approximation given in the right-hand term of this formula is usually close to one. The means conditioned on the predictor values and the dispersion parameter are

$$E\left[Y|\mu_i, \alpha\right] \cong \mu_i \tag{5}$$

and the corresponding conditioned variances

$$Var\left[Y|\mu_i, \alpha\right] \cong \frac{\mu_i}{\alpha}, \tag{6}$$

i.e., $0 < \alpha < 1$ signifies over-, $\alpha = 1$ equi-, and $\alpha > 1$ under-dispersion.

3 Computational procedures to be evaluated

As already stated in the introduction, there are different attributes according to which computational procedures for obtaining maximum likelihood estimates for the parameters of Double Poisson Regression can vary. These attributes include the treatment of the normalisation factor, the general strategy regarding the dispersion parameter, i.e., simultaneous versus sequential estimation of regression parameters and dispersion parameter, the starting value and the algorithm applied to determine the maximum of the likelihood. In the following chapter, firstly, these different attributes are described in more detail; subsequently, the complete study design is presented by means of which the computational procedures are evaluated that result from combining specifications of the different attributes; and, finally, some computational details of these procedures are discussed.

3.1 Attributes of the computational procedures to be evaluated

3.1.1 Normalisation factor

The first attribute of the computational procedures concerns the treatment of the normalisation factor. In the literature, three different approaches to the treatment of this factor are discussed: (1) setting the normalisation factor equal to one (Aragon et al. 2018); (2) approximating the normalisation factor using the approximation proposed by Efron (right term in Formula 4) (Toledo 2022); and (3) defining an upper limit for the originally infinite sum in the exact formulation (middle term in Formula 4) (Stasinopoulos and Rigby 2023). As the latter approach is the computationally most complicated, and as applying this approach necessitates further considerations and investigations regarding the optimal definition of the upper limit, the study presented here is restricted to the first two approaches (See Appendix for the log likelihood functions of the two resulting models).

The two different treatments of the normalisation factor investigated here may have a decisive impact on the estimates for the regression coefficients. If the normalisation factor is set equal to one, the gradient for the regression coefficient still contains the dispersion parameter (See Appendix, Formula A4). However, if the

gradient is used to determine an extremum and, for this reason, is set equal to zero, dividing the resulting equation by the dispersion parameter makes this parameter disappear. Consequently, with the normalisation factor set equal to one, the estimates of the regression coefficients do not depend on the dispersion. In fact, in this case, the same estimates result as for simple Poisson Regression. In contrast, if the normalisation factor is approximated, the dispersion parameter can no longer be removed after setting the gradient for the regression coefficients equal to zero (See Appendix, Formula A6). Consequently, in this case, the estimates for the regression coefficients might vary with the dispersion parameter.

3.1.2 Estimation strategy

The second attribute investigated concerns the strategy for estimating the dispersion parameter. Two different strategies are investigated: (1) simultaneous, and (2) sequential estimation. Simultaneous estimation consists in estimating the dispersion parameter together with the parameters of the regression equation. Sequential estimation consists in using a sequence of estimation processes comprising pairs of computational procedures. To be specific, one iteration cycle is divided into two separated cycles. In the first of these two separated cycles, the dispersion parameter is kept fixed, and the parameters of the regression equation are estimated. In the second of the two separated computational cycles, the parameters of the regression equation are kept fixed, and the dispersion parameter is estimated. At the very start of the sequential estimation process, it is the starting value of the dispersion parameter that is kept fixed. In each following cycle it is the result of the preceding cycle that is kept fixed. Sequential computation is also applied for parameter estimation in Negative Binomial Regression (Ripley et al. 2023).

3.1.3 Starting value for the dispersion parameter

The third attribute investigated concerns the determination of the starting value for the dispersion parameter. Two different approaches for determining these starting values are investigated. The first approach consists in applying 1 as the starting value, i.e., applying the value that describes equi-dispersion. The second approach consists in applying the maximum likelihood estimation of the dispersion parameter that results when the normalisation factor is set equal to 1. The corresponding equation is

$$\hat{\alpha} = \frac{1}{2 * \left(\frac{1}{n} \sum_{i=1}^{n} \left(y_i ln\left(y_i/\mu_i\right) - \left(y_i - \mu_i\right)\right)\right)} \tag{7}$$

If y_i is equal to zero, $y_i ln\left(y_i/\mu_i\right)$ is set equal to 0 because $\lim_{y \to 0} y \ln(y) = 0$ (Aragon et al. 2018). If sequential estimation is applied, this is the starting value for the cycle in which the additive parameter and the regression coefficients are kept fixed. Accordingly, the μ_i produced by the additive parameter and the regression coefficients estimated in the preceding cycle must be entered. If, additionally, the

normalisation factor is set equal to 1, Formula 7 directly provides the solution for the inner computation in which the additive parameter and the regressions coefficients are kept fixed. In other words, in this case no further iteration is necessary.

3.1.4 Algorithm

The fourth attribute concerns the computational algorithm applied. Seven different algorithms are investigated. They result as variants of three, more general algorithms: (1) the Newton–Raphson algorithm (Verbeke and Cools 1995; Ypma 1995); (2) the Berndt-Hall-Hall-Hausman (BHHH) algorithm (Berndt et al. 1974); and (3) the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Broyden 1970; Fletcher 1970; Goldfarb 1970; Shanno 1970). The variants of these algorithms result from choosing different approaches for determination of the gradients and the Hessian matrix. This can be done either numerically or algebraically. For the Newton–Raphson algorithm, three variants are considered: (1) both components are determined numerically; (2) the gradients are determined algebraically and the Hessian matrix is determined numerically; and (3) both components are determined algebraically (for the algebraic formulations of the gradients and the Hessian matrix, see Appendix). For the BHHH and the BFGS algorithm, the third variant does not make sense because these two algorithms are specially designed for numerical determination of the Hessian matrix. For this reason, for these two algorithms, only the first two variants are investigated. Generally, the variation between numerical and algebraical computation is considered because, although numerical computation might work well for simple models, there might be problems for more complex models (Henningsen and Toomet 2011).

3.2 Study design

In the study presented here, all combinations of the specifications of the attributes just described are investigated. This amounts to 56 different computational procedures to be compared (see Table 1).

3.3 Computational realisation

All computational procedures are implemented using the corresponding functions from the R-package maxLik (Toomet et al. 2022). The function maxNR is used for the Newton–Raphson algorithm; the function maxBHHH for the BHHH algorithm; and the function maxBFGS for the BFGS algorithm. To avoid negative values for the dispersion parameter α, the log likelihood function is maximised with regard to α' with Exp(α') equal to α. To ensure that log likelihood functions can also be computed for large y_i, $ln\left(y_i!\right)$ is computed using the function *lgamma* from R-base (Becker et al. 1988; R Core Team 2023) by computing $ln\left(\Gamma(y_i + 1)\right)$.

For all components of the procedures that are performed iteratively, the default termination criteria of maxNR are applied, i.e., the iteration terminates either when the absolute value of the difference between the log likelihood functions of the

Table 1 Computational procedures evaluated

Normalisation factor	Estimation strategy	Starting value	Algorithm[a]
1	Simultaneously	1	NR, Numerical
			NR, Mixed
			NR, Algebraic
			BHHH, Numerical
			BHHH, Mixed
			BFGS, Numerical
			BFGS, Mixed
		Computed via Formula 7	NR, Numerical
			NR, Mixed
			NR, Algebraic
			BHHH, Numerical
			BHHH, Mixed
			BFGS, Numerical
			BFGS, Mixed
	Sequentially	1	NR, Numerical
			NR, Mixed
			NR, Algebraic
			BHHH, Numerical
			BHHH, Mixed
			BFGS, Numerical
			BFGS, Mixed
		Computed via Formula 7	NR, Numerical
			NR, Mixed
			NR, Algebraic
			BHHH, Numerical
			BHHH, Mixed
			BFGS, Numerical
			BFGS, Mixed

Table 1 (continued)

Normalisation factor	Estimation strategy	Starting value	Algorithm[a]
Approximated via Formula 4	Simultaneously	1	NR, Numerical
			NR, Mixed
			NR, Algebraic
			BHHH, Numerical
			BHHH, Mixed
			BFGS, Numerical
			BFGS, Mixed
		Computed via Formula 7	NR, Numerical
			NR, Mixed
			NR, Algebraic
			BHHH, Numerical
			BHHH, Mixed
			BFGS, Numerical
			BFGS, Mixed
	Sequentially	1	NR, Numerical
			NR, Mixed
			NR, Algebraic
			BHHH, Numerical
			BHHH, Mixed
			BFGS, Numerical
			BFGS, Mixed
		Computed via Formula 7	NR, Numerical
			NR, Mixed
			NR, Algebraic
			BHHH, Numerical
			BHHH, Mixed
			BFGS, Numerical
			BFGS, Mixed

[a]*NR* Newton–Raphson, *BHHH* Berndt-Hall-Hall-Hausman, *BFGS* Broyden-Fletcher-Goldfarb-Shanno, *Numerical* gradient and Hessian matrix numerically determined, *Mixed* gradient algebraically and Hessian matrix numerically determined, *Algebraic* gradient and Hessian matrix algebraically determined

last and the preceding cycle is smaller than 10^{-8} or when the number of iterations exceeds 150. For the outer computation in the sequential procedures, the iteration terminates either when the absolute value of the difference between the log likelihood functions of the last and the preceding cycle is smaller than 10^{-8} or when the number of iterations exceeds 50.

In all cases, maximum likelihood estimators for simple Poisson Regression are taken as starting values for the additive parameter and the regression coefficients. These estimators can always be determined via the log likelihood function for the simple Poisson Regression (Cameron and Trivedi 2013, Chap. 3.2.1).

4 Methods for evaluation

The computational procedures just outlined are evaluated via simulation studies, i.e., they are applied to data sets for which the true parameters to be found are known. The whole investigation is restricted to data sets that could emerge in a randomised controlled trial (RCT). Accordingly, the super-ordinated criterion with regard to which the computational procedures are evaluated is the extent to which they help to correctly recognise whether the intervention has an effect. The whole methodology for reaching this goal consists of four different components: (1) the data sets for the simulation studies; (2) the regression equations for which the parameters are estimated; (3) the specific criteria according to which the computational procedures are evaluated; and (4) the analyses performed with the data obtained from the simulation study.

4.1 Data sets

Different data generating processes are applied to produce the data by means of which the computational procedures are tested. Each specific data generating process defines a specific scenario. All data generating processes are specifications of a general data generating model. This model consists of four major components:

(1) A log-linear regression equation that regresses the mean count on one dichotomous variable X_1 with $X_1 \in \{0, 1\}$, and on two continuous variables X_2 and X_3 with

$$X_2, X_3 \in [0, 1], \textit{i.e.},$$

$$E\left[Y|\boldsymbol{x}_i\right] = \exp\left(\beta_{g,0} + \beta_{g,1}x_1 + \beta_{g,2}x_2 + \beta_{g,3}x_3\right); \tag{8}$$

with $\beta_{g,j}$ being the parameters of the data generating model;

(2) a Double Poisson distribution, which is combined with the regression equation;
(3) the distributions of the predictor variables; and.
(4) the sample size.

The dichotomous variable corresponds to the study conditions, i.e. to the intervention and the control condition to be compared in the RCT. Accordingly, $\beta_{g,1}$ describes the effect of the intervention. The continuous variables correspond to covariates that might additionally affect the outcome, i.e. these variables are possible confounders. The coefficients $\beta_{g,2}$ and $\beta_{g,3}$ determine whether there is actually an effect and, if yes, how large this effect is.

All scenarios investigated here have three attributes in common: (1) the sample size is always equal to 600; (2) $\beta_{g,0} = \ln(3)$; and (3) all predictor variables are uncorrelated. The restriction that X_2 and X_3 are uncorrelated with X_1 follows directly from restricting the scope on data that emerge from an RCT. The crucial defining feature of an RCT is that the entities of investigation have been assigned randomly to the study conditions, i.e., to the two levels of X_1. This directly implies that the data generating process for the RCT is free of correlations between the study condition and any other variables. Non-zero correlations between X_2 and X_3 are excluded because otherwise the true effects of these variables on the count variable can, in principle, not be identified without including both variables in the regression equation. As a result, any failure in detecting the true effects would not be due to deficiencies of the estimation procedure but to a misspecification of the regression equation, and issues of misspecification do not belong to the problem addressed in this study.

Apart from the three attributes just outlined, the scenarios differ with respect to the other components of the general data generating model. To be specific:

(1) Two different values of $\beta_{g,1}$, i.e., $\beta_{g,1} = 0$, and $\beta_{g,1} = \ln\left(\frac{10}{9}\right)$, are considered. The first value means that the intervention has no effect on the outcome. The second value means that there is such an effect.

(2) Three different combinations of values of $\beta_{g,2}$ and $\beta_{g,3}$, i.e., $\beta_{g,2} = \beta_{g,3} = 0$, $\beta_{g,2} = 1$ and $\beta_{g,3} = 0$, and $\beta_{g,2} = \beta_{g,3} = 1$, are considered. The first combination means that neither of the two possible covariates actually has an effect on the outcome variable. The second combination means that only X_2 but not X_3 has an effect of the outcome. The third combination means that both, X_2 and X_3, have an effect of the outcome. Among other things, the latter two combinations are included to produce cases in which the empirical distributions of actual values conditioned on the predicted values are not Double Poisson distributions. In this way, not-well-behaved cases are produced.

(3) Four different values of the dispersion parameter α of the Double Poisson distribution are considered, i.e., $\alpha_g = 0.1$, $\alpha_g = 0.5$, $\alpha_g = 1$, and $\alpha_g = 2$, with $\alpha_g = 0.1$ and $\alpha_g = 0.5$ thereby signifying over-, $\alpha_g = 1$ equi-, and $\alpha_g = 2$ under-dispersion (see Formula 6). A dispersion parameter of 1 is selected because this parameter marks the state of equi-dispersion and because this state plays a central role in modelling count data. Dispersion parameters for over- and under-dispersion are chosen to cover both possible deviations from equi-dispersion. For over-dispersion, one instead of two parameters are chosen because over-dispersion is more frequent than under-dispersion. One of these parameters has been chosen to produce a very large over-dispersion because there is sometimes very large

over-dispersion in real data and because the normalisation factor differs most strongly from 1 in cases of very large dispersion.

(4) Two different distributions of X_1 are considered, i.e., $p(X_1 = 1) = 0.1$, and $p(X_1 = 1) = 0.5$. The second case is the most common in RCTs. The first case is included as pars pro toto of all possible deviations from the most common case.

(5) Two different distributions of X_2 are considered, i.e., a beta-distribution with a mean of 1/5 and a variance of 16/1100 (right-skewed), and a beta-distribution with a mean of 4/5 and a variance of 16/1100 (left-skewed). The two skewed beta-distributions are chosen because they resemble distributions of possible confounders. For example, age in a sample of students is usually right-skewed, whereas a health-utility index (Horsman et al. 2003; Rabin and de Charro 2001) in such a sample is usually left-skewed.

(6) Two different distributions of X_3 are considered, i.e., a beta-distribution with a mean of 1/5 and a variance of 16/1100 (right-skewed), and a beta-distribution with a mean of 4/5 and a variance of 16/1100 (left-skewed). The reason for choosing these distributions is the same as for the distributions of X_2.

All combinations of the different specifications of the model components are applied. This results in altogether 192 different scenarios. For each scenario, 2000 data sets are produced.

4.2 Investigated regression equations

The data sets just described are analysed using two different regression equations. The first equation contains only one predictor variable, i.e.,

$$E[Y|x_1] = \exp(\beta_{a,0} + \beta_{a,1}x_1) \tag{9}$$

with $\beta_{a,0}$ and $\beta_{a,1}$ being the parameters of the data analytical model. In the following text this equation will be referred to as the one-predictor equation. Such an equation could be used to investigate the effect of the intervention without controlling for any covariates. The second equation contains two predictor variables, i.e.,

$$E[Y|x_1, x_2] = \exp(\beta_{a,0} + \beta_{a,1}x_1 + \beta_{a,2}x_2) \tag{10}$$

with $\beta_{a,0}$, $\beta_{a,1}$ and $\beta_{a,2}$ being the parameters of the data analytical model. In the following text, this equation will be referred to as the two-predictor equation. Such an equation could be used for investigating the effect of the intervention with controlling for one covariate.

If the one-predictor equation is applied for analysis of the data, this model only matches the data generating model if $\beta_{g,2} = \beta_{g,3} = 0$. If the two-predictor equation is applied, the data analysing model also matches the data generating model if $\beta_{g,2} = 1$ and $\beta_{g,3} = 0$. If the data generating and the data analysing models do not match, the distributions of actual values conditioned on the predicted values are no longer

well-behaved. Such cases have been constructed to test how well Double Poisson Regression works as a universal model for count data.

4.3 Evaluation criteria

The computational procedures are evaluated with respect to (1) the feasibility of the computation, and (2) the quality of the parameter estimates. The criterion for assessing feasibility is the number of non-converging computations, i.e., of convergence failures. The criteria for assessing the quality of parameter estimates are (1) bias, (2) root mean square error (RMSE) and (3) coverage (Morris et al. 2019). All criteria regarding the quality of parameter estimates are determined separately for each combination of scenario, regression equation and computational procedure. Bias is thereby determined as the mean difference between the estimated and the true parameter across the data sets for each of these combinations; RMSE as the root of the mean squared difference between the estimated and the true parameter across these data sets; and coverage as the percentage of data sets for which the true parameter lies within the estimated 95%-confidence interval across these data sets. Bias and RMSE are computed for the additive parameter, the regression coefficients and the logarithmised dispersion parameter; coverage only for the additive parameter and the regression coefficients.

Bias, RMSE and coverage for β_0 and β_2 are not directly relevant for correctly identifying the effect of an intervention. They are nevertheless investigated because this provides more insights into the quality of parameter estimation. Moreover, RMSE differs from bias and coverage in that there is a monotonous relationship between this statistic and the quality of parameter estimation. To be specific, the lower the RMSE, the higher the quality. In contrast, the optimal value for bias is zero, with values lower and higher than zero indicating losses in quality. In a similar way, the optimal value of coverage with regard to a 95%-confidence interval is 95, with values lower and higher than 95 indicating losses in quality. For these reasons, RMSEs for β_1 and for $ln(\alpha)$, i.e., the logarithmised dispersion parameter, play the most central role in evaluating quality of parameter estimation.

Bias, RMSE and coverage can only be computed if the values of the true parameters are given. The additive parameter, i.e. to $\beta_{a,0}$, must be computed. The formula for computing this parameter depends upon the data analytical model. If this is the one-predictor equation, the true additive parameter $\beta_{a,0;true}$ is

$$\beta_{a,0;true} = \beta_{g,0} + \beta_{g,2}E(X_2) + \beta_{g,3}E(X_3), \tag{11}$$

with $\beta_{g,0}$, $\beta_{g,2}$, and $\beta_{g,3}$ being the parameters from the data generating process, $E(X_2)$ the expected value of X_2 as presupposed in the data generating process, and $E(X_3)$ the corresponding expected value of X_3. If the two-predictor equation is applied as the data analytical model, the true additive parameter $\beta_{a,0;true}$ is

$$\beta_{a,0;true} = \beta_{g,0} + \beta_{g,3}E(X_3). \tag{12}$$

The true values of the regression coefficients, i.e., of $\beta_{a,1}$ and $\beta_{a,2}$, are identical with the corresponding values of the data generating models.

In contrast to the additive parameter and the regression coefficients, the dispersion parameter is only defined as a component of the Double Poisson Regression. Therefore, in a strict sense, a true meaning of this parameter only exists if the distribution of the actual values conditioned on the predicted values is a Double Poisson distribution. In the setting considered here, this is this case if the model applied for analysis of the data matches the model for generating these data. This, in turn, is the case for both data analytical regression equations if they are applied to scenarios with $\beta_{g,2}$ and $\beta_{g,3}$ equal to zero and, additionally, for the two-predictor equation if this equation is applied to scenarios with $\beta_{g,2}$ equal to one and $\beta_{g,3}$ equal to zero. In these cases, the dispersion parameters of the data generating models constitute the true parameters. The analyses presented here are restricted to these cases.

4.4 Analyses

The major objective of the study presented here is to find the optimal computational procedure that can be constructed by combining the different specifications of the four investigated attributes of the computational procedures. However, to put this adequately into context, some other aspects must be considered. To be specific, with the treatment of the core question included, six specific research questions emerge:

(1) How do the four different attributes of the computational procedures affect the parameter estimates?
(2) How do the four different attributes of the computational procedures affect the feasibility and the quality of parameter estimation?
(3) To what extent do the effects on quality of parameter estimation vary with the scenarios?
(4) What is the optimal computational procedure or, respectively, what are the optimal computational procedures?
(5) What are the characteristics of the optimal computational procedure or, respectively, the optimal computational procedures?
(6) How does the optimal computational procedure or, respectively, do the optimal computationalprocedures perform in a case study?

The first research question calls for information regarding the extent to which attributes of the computational procedures can lead to differences in the quality of parameter estimation. If there are no effects on the parameter estimates, there can be no effects on the quality of parameter estimation. The second research question calls for an understanding of the function of the individual computational procedure attributes. Moreover, the answers to the second question provide the information needed to answer the third question. The third research question calls for information as to how well the results for the investigated scenarios can be generalised to scenarios not investigated. The fourth research question is the central research question of the whole study. The fifth research question calls for information regarding the performance of the optimal procedure or, respectively, optimal procedures. The sixth and final research question calls for additional information regarding performance with real data.

4.4.1 Effects of computational procedure attributes on parameter estimates

The effects of the computational procedure attributes on the parameter estimates are investigated separately for the one- and for the two-predictor equation. For the one-predictor equation, only the parameters β_0, β_1, and $ln(\alpha)$ are considered; for the two-predictor equation additionally the parameter β_2. The effect of a computational procedure attribute on the parameters is operationalised via the agreement between estimates produced by computational procedures that only differ with regard to this attribute. The lower the agreement is, the larger is the effect of this attribute on the parameter estimates, and the more the quality of the estimates produced by different specialisations of this attribute may differ. In the case of absolute agreement, the computational procedure attribute cannot have any effect at all on quality of parameter estimates.

The agreement between parameter estimates is determined using intraclass correlations (ICCs) for absolute values (Shrout and Fleiss 1979). These computations are performed separately for the two investigated regression equations, for each regression equation for all 192 scenarios and within each scenario for each combination of estimated parameter and investigated computational procedure attribute. The ICCs are aggregated across all scenarios by computing means and standard deviations. For those computational procedure attributes for which the ICCs deviate from zero to a noteworthy extent, the effects of scenario attributes on ICCs are also investigated. For binary attributes, this is performed using t-tests for independent samples; for attributes with more than two specifications using analyses of variance (ANOVAs) with the respective attribute as a between-scenario factor.

4.4.2 Effects of computational procedure attributes on feasibility and quality of parameter estimation

Just like the effects of the computational procedure attributes on the parameter estimates, the effects of these attributes on the evaluation criteria are also investigated separately for the one- and the two-predictor equation. In both cases, the same statistical approaches are applied. The effects on convergence failures are tested on the level of the individual data sets using goodness-of-fit chi-square tests for deviation from equal distribution. The effects on criteria regarding the quality of parameter estimation are tested on the level of individual scenarios using ANOVAs with the four attributes of the computational procedures as within-scenario factors. For criteria referring to β_0, β_1, and β_2, all scenarios are included. For criteria referring to $ln(\alpha)$, the selection of included scenarios depends upon the regression equation applied. If the one-predictor equation is applied, only scenarios with $\beta_{g,2} = \beta_{g,3} = 0$ are included. If the two-predictor equation is applied, these are only scenarios with $\beta_{g,3} = 0$. The restrictions regarding $ln(\alpha)$ are necessary because the true values for this parameter can only be determined for these scenarios.

4.4.3 Variability of results across scenarios

To assess the likelihood with which the results of the investigated scenarios can be generalised to different scenarios, the extent to which the effects of the computational procedure attributes on quality of parameter estimation vary with the scenarios is investigated. The more these effects vary with the investigated scenarios, the less likely it is that the effects found for the total sample of scenarios will be found for different scenarios. As the variability across scenarios is only relevant for those attributes that affect the parameter estimates, only these attributes are considered in this context. As before, these investigations are performed separately for the two regression equations. All evaluation criteria regarding the quality of parameter estimation are considered. For criteria referring to β_0, β_1, and β_2, all scenarios, i.e., 192 scenarios, are included. For criteria referring to $ln(\alpha)$, the selection of included scenarios depends upon the regression equation applied. If the one-predictor equation is applied, only scenarios with $\beta_{g,2} = \beta_{g,3} = 0$, i.e., 64 scenarios, are included. If the two-predictor equation is applied, these are only scenarios with $\beta_{g,3} = 0$, i.e., 128 scenarios.To assess the extent to which the included computational attributes affect the evaluation criteria in the same way for different scenarios, the ordinal relations in the result patterns for the individual scenarios are compared with the ordinal relations in the result pattern for all scenarios together.

As several computational procedures are applied in the study, several result patterns belong to one and the same scenario. For a binary computational procedure attribute, there are 2*2*7 = 28 different patterns for one scenario. For the attribute 'algorithm', there are 2*2*2 = 8 different patterns. In the analyses, all these result patterns are treated as discrete analytical units. Accordingly, 192*28 = 5,376 different analytical units are considered if the effect of a binary attribute on an evaluation criterion referring to β_0, β_1, or β_2 is investigated. In contrast, if the effect of the attribute 'algorithm' on a criterion referring to $ln(\alpha)$ is investigated, and if the one-predictor equation is applied, the number of investigated analytical units is only 64*8 = 512.

For binary computational procedure attributes, the agreement of the patterns for individual analytical units with that for all scenarios together is investigated by determining the percentages for three different constellations: (1) concordant results, i.e., the constellation that both effects are in the same direction with ties on both sides counting as effects in the same direction; (2) ambiguous results, i.e., the constellation that there is a tie on one side but none on the other; and (3) discordant results, i.e., the constellation that both effects are in opposite directions. For the attribute 'algorithm', a slightly more complicated procedure must be applied. This procedure consists of three steps: (1) on the level of analytical units, each pair of specifications of this attribute is categorised according to the three constellations just described; (2) for each analytical unit the percentages of constellations across all pairs is determined; and (3) the means of percentages for the three different constellations are computed across all analytical units. For especially relevant combinations of computational procedure attributes and

evaluation criteria for which the percentage of concordant results is low, the distributions of scenario attribute specifications within the three constellations are analysed.

4.4.4 Identification of the optimal computational procedure or, respectively, procedures

The optimal computational procedure is determined via a sequence of examinations, decisions and actions. The first examination addresses the question as to which of the investigated computational procedure attributes has a noteworthy effect on the parameter estimates. For those attributes that have no such effect, the selection of the optimal attribute specification is based, firstly, on convergence failures and, secondly, on simplicity in programming. For those attributes that have such an effect, the specification is based on the quality of parameter estimation and on the extent to which the results regarding the quality of parameter estimation seem stable.

As the superordinate criterion for evaluating the computational procedures is their capability to support identification of the effect of an intervention, and as only the RMSEs have a monotonous relationship to quality of parameter estimation, the crucial criteria for judging the quality of parameter estimation are the RMSEs for β_1 and $ln(\alpha)$. Accordingly, those specifications are selected that have the lowest RMSEs for both parameters. If the effects are neither stable across scenarios nor very large, the decision is based, firstly, on convergence failures and, secondly, on simplicity in programming. For computational procedure attributes that affect the RMSEs for β_1 and $ln(\alpha)$ distinctly in opposite directions, no decision is made. In this case, no single optimal computational procedure, but only a set of potentially optimal computational procedures, can be identified. In data analyses, these optimal computational procedures might then be applied parallel to each other for cross-validation.

4.4.5 Characteristics of the optimal computational procedure or, respectively, procedures

To identify the optimal computational procedure or, respectively, the potentially optimal computational procedures, means and standard deviations as well as minima and maxima are determined for the evaluation criteria regarding quality of parameter estimation. This is, again, conducted separately for both regression equations. Moreover, if the standard deviations are large, additional analyses are conducted regarding the dependencies on scenario attributes.

4.4.6 Case study

The case study is performed with data from an evaluation study regarding an intervention designed to reduce exacerbations in chronic obstructive pulmonary disease (COPD) (Storgaard et al. 2018). In this study, 100 persons had been randomised to the intervention and 100 to the control group. The study outcomes were the number of exacerbations of COPD and the number of days in hospital. Both variables are applied as the outcome variables in the case study. Both the one- and the

two-predictor equation are employed. In the latter case, number of exacerbations or, respectively, number of days in hospital in the year before the study are taken as the covariate. When analysing the data, different approaches are compared. On the one hand, Negative Binomial Regression and Poisson Regression with robust standard errors are applied; on the other hand, Double Poisson Regression, either with the particular computational procedure that has been identified as optimal, or with those computational procedures that have been identified as potentially optimal.

4.4.7 Computational realisation

The analyses performed on the level of the individual data sets are performed with R. All other analyses are performed using SPSS Version 29.

5 Results

5.1 Effects of computational procedure attributes on parameter estimates

The attributes of the computational procedures affect the parameter estimates to a different extent. The treatment of the normalisation factor has a large effect on the parameters, the estimation strategy a moderate effect, and starting value and algorithm hardly any effect (see Table 2). The effect of the normalisation factor shows for all four parameters, with $ln(\alpha)$ being affected very strongly, β_0 affected moderately, and the other two parameters affected slightly. The effect of estimation strategy only shows for those two parameters that are mostly affected by the treatment of the normalisation factor, i.e., for β_0 and $ln(\alpha)$ (see Table 2). The agreement between

Table 2 ICCs for parameter estimates with computational procedure attribute listed in the first column varying

Computational procedure attribute	Parameters of Double Poisson regression[a]			
	β_0	β_1	β_2	$ln(\alpha)$
One-predictor equation				
Normalisation factor	.711 (.461)	.992 (.014)	–	.192 (.730)
Estimation strategy	.999 (.001)	1.000 (< .001)	–	.997 (.006)
Starting value	1.000 (< .001)	1.000 (< .001)	–	1.000 (< .001)
Algorithm	1.000 (< .001)	1.000 (< .001)	–	1.000 (< .001)
Two-predictor equation				
Normalisation factor	.864 (.278)	.993 (.013)	.980 (.035)	.202 (.737)
Estimation strategy	1.000 (.001)	1.000 (< .001)	1.000 (< .001)	.996 (.006)
Starting value	1.000 (< .001)	1.000 (< .001)	1.000 (< .001)	1.000 (< .001)
Algorithm	1.000 (< .001)	1.000 (< .001)	1.000 (< .001)	1.000 (< .001)

[a] β_0, β_1, β_2, and $ln(\alpha)$ are the generic terms of the parameters to be estimated. Cell entries are mean (standard deviation). All statistics are based on 192 scenarios

parameter estimates produced by different treatments of the normalisation factor varies by far to the greatest degree with dispersion (see Table 3). The agreement seems to be largest for dispersion close to the point of equi-dispersion and decreases to both sides of this point. The differences in agreement are largest for the estimation of the logarithmised dispersion parameter. For dispersion parameters of 0.1 and of 2.0, i.e., for very strong over- and very strong under-dispersion, the ICCs even become negative (see Table 3). This means that the mean squares between the two parameter estimates produced by different methods for the same data set are larger than the mean squares between the parameter estimates produced by the same method for different data sets.

5.2 Effects of computational procedure attributes on evaluation criteria

Each of the four attributes has a high statistically significant effect on feasibility as operationalised by failures of convergence. This holds true, both if the one-predictor and if the two-predictor equation is applied (see Tables 4 and 5). The results are very distinct for the attribute 'algorithm'. There are only failures in the case of the Newton–Raphson and the BHHH algorithm, whereas there are no failures at all for the BFGS algorithm (see Tables 4 and 5). However, in all cases the relative frequencies of failures are extremely low. They vary from exactly zero for all applications of the BFGS algorithm

Table 3 Relation between dispersion and ICCs regarding different treatments of the normalisation factor

Dispersion parameter of data generating model	Parameters of Double Poisson regression[a]			
	β_0	β_1	β_2	$ln(\alpha)$
One-predictor equation				
0.1	–.032	.971	–	–..796
0.5	.995	1.000	–	.931
1.0	1.000	1.000	–	.658
2.0	.883	.998	–	–..024
Standard deviation[b]	.499	.014	–	.771
Statistical test[c]	p < .001	p < .001	–	p < .001
Two-predictor equation				
0.1	.497	.972	.926	–..797
0.5	.998	1.000	1.000	.939
1.0	1.000	1.000	1.000	.675
2.0	.959	.998	.995	–..007
Standard deviation[b]	.245	.014	.036	.777
Statistical test[c]	p < .001	p < .001	p < .001	p < .001

[a] β_0, β_1, β_2, and $ln(\alpha)$ are the generic terms of the parameters to be estimated. Except for rows denominated with standard deviation or statistical test, cell entries are means. Each mean is based on 48 scenarios

[b] Standard deviation between means

[c] Test for difference between means

Table 4 Effects of the computational procedure attributes: One-predictor equation

Computational procedure attribute	Feasibility[a]	Quality of parameter estimation[b]							
	Failures	β_0 [c]			β_1 [c] l			$n(\alpha)$[d]	
		$Bias_{100}$[e]	$RMSE_{100}$[e]	Coverage	$Bias_{100}$[e]	$RMSE_{100}$[e]	Coverage	$Bias_{100}$[e]	$RMSE_{100}$[e]
Normalisation factor									
1	13	6.9831	8.7074	73.34%	−..5950	6.9479	95.25%	10.8116	20.2591
Approximation	35	9.6868	11.0505	68.95%	−..7392	6.4004	94.23%	24.3650	27.3631
Standard deviation	15.556	1.9118	1.6568	3.10	.1020	.3871	.72	9.5837	5.0233
Statistical test	p < .01	p < .001	p < .001	p < .001	p < .001	p < .001	p < .001	p < .001	p < .001
Estimation strategy									
Simultaneous	43	8.3804	9.9228	71.18%	−..6691	6.6666	94.80%	17.3028	23.5312
Sequential	5	8.2895	9.8351	71.11%	−..6651	6.6817	94.67%	17.8738	24.0910
Standard deviation	26.870	.0643	.0620	.03	.0028	.0107	.10	.4038	.3958
Statistical test	p < .001	p < .001	p < .001	p < .001	p < .001	p < .001	p < .001	p < .001	p < .001
Starting value									
1	19	8.3348	9.8788	71.15%	−..6672	6.6741	94.74%	17.5887	23.8114
Computed	29	8.3351	9.8792	71.15%	−..6671	6.6741	94.74%	17.5880	23.8109
Standard deviation	7.071	.0002	.0003	.00	.0001	.0000	.00	.0005	.0004
Statistical test	p = .194	p < .001	p < .001	p = .856	p < .001	p < .01[f]	p = .426	p < .001	p < .001
Algorithm[g]									
NR; Numerical	6	8.3350	9.8790	71.07%	−..6671	6.6741	94.70%	17.5882	23.8111
NR; Mixed	9	8.3350	9.8790	71.07%	−..6671	6.6741	94.70%	17.5882	23.8111
NR; Algebraic	15	8.3350	9.8790	71.07%	−..6671	6.6741	94.70%	17.5882	23.8111
BHHH; Numerical	9	8.3349	9.8789	71.35%	−..6671	6.6741	94.84%	17.5885	23.8113
BHHH; Mixed	9	8.3349	9.8789	71.35%	−..6671	6.6741	94.84%	17.5885	23.8113
BFGS; Numerical	0	8.3350	9.8790	71.07%	−..6671	6.6741	94.70%	17.5883	23.8111

Table 4 (continued)

Computational procedure attribute	Feasibility[a]	Quality of parameter estimation[b]							
	Failures	β_0 [c]			β_1 [c] l			$n(\alpha)$[d]	
		$Bias_{100}$[e]	$RMSE_{100}$[e]	Coverage	$Bias_{100}$[e]	$RMSE_{100}$[e]	Coverage	$Bias_{100}$[e]	$RMSE_{100}$[e]
BFGS; Mixed	0	8.3350	9.8790	71.06%	–..6671	6.6741	94.70%	17.5883	23.8110
Standard deviation	5.398	<.0001	<.0001	.09	.0000	.0000	.04	.0001	.0001
Statistical test	p<.001	p<.01	p<.01	p<.001	p<.001[f]	p=.138	p<.001	p<.01	p<.001

[a]Except for rows denominated with 'Standard deviation' or 'Statistical test' cell entries are absolute frequencies. Entries in the rows denominated with 'Standard deviation' are standard deviations between these frequencies, and entries in the rows denominated with 'Statistical test' are tests regarding differences between these frequencies. For binary attributes, the number of computations for each specification of the attribute is 192 (scenarios) *28 (specification combinations of the other computational procedure attributes) *2000 (data sets) = 10,752,000; for the attribute algorithm, the number of computations for each specification is 192 (scenarios) *8 (combinations of the other computational procedure attributes) * 2000 (data sets) = 3,072,000

[b] β_0, β_1, and α are the generic terms for the parameters to be estimated. Except for rows denominated with 'Standard deviation' or 'Statistical test' cell entries are means. Entries in the rows denominated with 'Standard deviation' are standard deviations between these means, and entries in the rows denominated with 'Statistical test' are tests regarding differences between these means

[c]For binary attributes, means are based on 192 (scenarios) *28 (specification combinations of the other computational procedure attributes) = 5.376 values; for the attribute 'algorithm, means are based on 192 (scenarios) *8 (specification combinations of the other computational procedure attributes) = 1.536 values

[d]Only scenarios with $\beta_2 = \beta_3 = 0$ are included. Accordingly, for binary attributes, means are based on 64 (scenarios) *28 (specification combinations of the other computational procedure attributes) = 1,792 values; for the attribute 'algorithm', means are based on 64 (scenarios) * (specification combinations of the other computational procedure attributes) = 512 values

[e]Bias and RMSE have been multiplied by 100 to increase readability of the table entries

[f]Difference between the descriptive values only shows after the fourth digit following the decimal point

[g]*NR* Newton–Raphson, *BHHH* Berndt-Hall-Hall-Hausman, *BFGS* Broyden-Fletcher-Goldfarb-Shanno, *Numerical* gradient and Hessian matrix numerically determined, *Mixed* gradient algebraically and Hessian matrix numerically determined, *Algebraic* gradient and Hessian matrix algebraically determined

Table 5 Effects of the computational procedure attributes: Two-predictor equation

Computational procedure attribute	Feasibility[a]	Quality of parameter estimation[b]					
	Failures	β_0 [c]			β_1 [c]		
		$Bias_{100}$[e]	$RMSE_{100}$[e]	Coverage	$Bias_{100}$[e]	$RMSE_{100}$[e]	Coverage
Normalisation factor							
1	10	8.2723	16.2868	82.84%	–..5919	6.8462	95.22%
Approximation	7	12.1234	18.8133	76.87%	–..7333	6.3059	94.23%
Stand. deviation	2.121	2.7231	1.7865	4.22	.1000	.3820	.70
Statistical test	p = .629	p < .001	p < .001	p < .001	p < .001	p < .001	p < .001
Estimation strategy							
Simultaneous	12	10.2566	17.5990	79.86%	–..6642	6.5684	94.79%
Sequential	5	10.1391	17.5011	79.85%	–..6605	6.5837	94.67%
Stand. deviation	4.950	.0831	.0692	.03	.0026	.0108	.09
Statistical test	p = .096	p < .001	p < .001	p = .267	p < .001	p < .001	p < .001
Starting value							
1	15	10.1976	17.5499	79.86%	–..6626	6.5760	94.73%
Computed	2	10.1981	17.5503	79.86%	–..6626	6.5761	94.73%
Stand. deviation	9.192	.0004	.0003	.00	.0000	.0001	.00
Statistical test	p < .01	p < .001	p < .001	p = .388	p < .001[f]	p < .001	p = .575
Algorithm[g]							
NR; Num	5	10.1978	17.5501	79.74%	–..6626	6.5761	94.66%
NR; Mixed	0	10.1978	17.5501	79.74%	–..6626	6.5761	94.66%
NR; Algebraic	0	10.1978	17.5501	79.74%	–..6626	6.5761	94.66%
BHHH; Num	6	10.1978	17.5501	80.13%	–..6626	6.5760	94.89%
BHHH; Mixed	6	10.1978	17.5501	80.13%	–..6626	6.5760	94.89%
BFGS; Num	0	10.1979	17.5500	79.75%	–..6626	6.5760	94.66%
BFGS; Mixed	0	10.1979	17.5500	79.74%	–..6626	6.5761	94.66%
Stand. deviation	3.047	< .0001	< .0001	.11	.0000	.0001	.10
Statistical test	p < .001	p = .845	p = .189	p < .001	p < .01[f]	p < .05	p < .001

Computational procedure attribute	Feasibility[a]	Quality of parameter estimation[b]				
	Failures	β_2 [c]			$ln(\alpha)$[d]	
		$Bias_{100}$[e]	$RMSE_{100}$[e]	Coverage	$Bias_{100}$[e]	$RMSE_{100}$[e]
Normalisation factor						
1	10	–.4.2466	22.6216	94.08%	9.0516	17.9827
Approximation	7	–.6.6124	22.4712	90.52%	20.9710	24.2020
Stand. deviation	2.121	1.6729	.1063	2.52	8.4283	4.3977
Statistical test	p = .629	p < .001	p = .528	p < .001	p < .001	p < .001
Estimation strategy						
Simultaneous	12	–.5.4580	22.5459	92.35%	14.7627	20.8508
Sequential	5	–.5.4011	22.5468	92.26%	15.2599	21.3359
Stand. deviation	4.950	.0402	.0006	.04	.3516	.3430
Statistical test	p = .096	p < .001	p = .916	p < .001	p < .001	p < .001

Table 5 (continued)

Computational procedure attribute	Feasibility[a]	Quality of parameter estimation[b]				
	Failures	β_2 [c]			*ln(α)*[d]	
		$Bias_{100}$[e]	$RMSE_{100}$[e]	Coverage	$Bias_{100}$[e]	$RMSE_{100}$[e]
Starting value						
1	15	−.5.4293	22.5463	92.30%	15.0117	21.0935
Computed	2	−.5.4297	22.5464	92.30%	15.0109	21.0932
Stand. deviation	9.192	.0003	.0001	.00	.0006	.0002
Statistical test	p < .01	p < .001	p < .05	p = .947	p < .001	p < .001
Algorithm[g]						
NR; Num	5	−.5.4296	22.5464	92.27%	15.0114	21.0934
NR; Mixed	0	−.5.4296	22.5464	92.27%	15.0114	21.0934
NR; Algebraic	0	−.5.4296	22.5464	92.27%	15.0114	21.0934
BHHH; Num	6	−.5.4298	22.5465	92.39%	15.0142	21.0936
BHHH; Mixed	6	−.5.4298	22.5465	92.39%	15.0138	21.0936
BFGS; Num	0	−.5.4291	22.5461	92.27%	15.0142	21.0932
BFGS; Mixed	0	−.5.4292	22.5462	92.27%	15.0147	21.0931
Stand. deviation	3.047	.0002	.0002	.03	.0015	.0002
Statistical test	p < .001	p < .05	p < .05	p < .05	p < .001	p < .001

[a]Except for rows denominated with 'Stand. deviation' or 'Statistical test' cell entries are absolute frequencies. Entries in the rows denominated with 'Stand. deviation' are standard deviations between these frequencies, and entries in the rows denominated with 'Statistical test' are tests regarding differences between these frequencies. For binary attributes, the number of computations for each specification of the attribute is 192 (scenarios) *28 (specification combinations of the other computational procedure attributes) *2000 (data sets) = 10,752,000; for the attribute algorithm, the number of computations for each specification is 192 (scenarios) *8 (combinations of the other computational procedure attributes) *2000 (data sets) = 3,072,000

[b] β_0. β_1. β_2. and α are the generic terms for the parameters to be estimated. Except for rows denominated with 'Stand. deviation' or 'Statistical test' cell entries are means. Entries in the rows denominated with 'Stand. deviation' are standard deviations between these means, and entries in the rows denominated with 'Statistical test' are tests regarding differences between these means

[c]For binary attributes, means are based on 192 (scenarios) *28 (specification combinations of the other computational procedure attributes) = 5,376 values; for the attribute 'algorithm, means are based on 192 (scenarios) *8 (specification combinations of the other computational procedure attributes) = 1,536 values

[d]Only scenarios with $\beta_3 = 0$ are included. Accordingly, for binary attributes, means are based on 128 (scenarios) *28 (specification combinations of the other computational procedure attributes) = 3,584 values; for the attribute 'algorithm', means are based on 128 (scenarios) *8 (specification combinations of the other computational procedure attributes) = 1,024 values

[e]Bias and RMSE have been multiplied by 100 to increase readability of the table entries

[f]Difference between the descriptive values only shows after the fourth digit following the decimal point

[g]*NR* Newton–Raphson, *BHHH* Berndt-Hall-Hall-Hausman, *BFGS* Broyden-Fletcher-Goldfarb-Shanno, *Num* gradient and Hessian matrix numerically determined, *Mixed* gradient algebraically and Hessian matrix numerically determined, *Algebraic* gradient and Hessian matrix algebraically determined

to 0.0000014 for the completely algebraic variant of the Newton–Raphson-algorithm, if it is applied for the one-predictor equation (see Tables 4 and 5).

The four computational procedure attributes affect the criteria regarding the quality of parameter estimation to very different degrees. The standard deviations between the means of the different specifications of the same computational procedure attribute are very large for the treatment of the normalisation factor, moderate for the estimation strategy, and hardly existent for the remaining two attributes (see Tables 4 and 5). Admittedly, there are also several statistically significant effects for these remaining two attributes. However, these effects mainly result because there is hardly any error variance for these two attributes. The differences just outlined between the four computational procedure attributes are completely in line with the results regarding the effects of the computational procedure attributes on the parameter estimates (see Table 2).

5.3 Variability of results across scenarios

As starting value and algorithm have hardly any effect on the parameter estimates, variability of results regarding the effects on quality of parameter estimation are only investigated for the treatment of the normalisation factor and the general estimation strategy. With a few exceptions, the percentage of concordant analytical units is at best moderate (see Table 6). For the attribute 'estimation strategy' there are, with regard to coverage, even quite considerable percentages of ambiguous analytical units. Altogether, these results indicate that the computational procedure attributes have different effects for different scenarios. Accordingly, generalising the results obtained from the scenarios obtained in this study to different scenarios is highly questionable.

As the treatment of the normalisation factor has by far the largest effect on quality of parameter estimation and as the RMSEs for β_1 and for $ln(\alpha)$ are most important, closer inspection is given to the way in which scenario attributes affect the effect of the normalisation factor on these two evaluation criteria. The main result is that the dispersion parameter in the data generating model affects these effects to the largest extent. For a dispersion parameter of 0, i.e., for very strong over-dispersion, all results are concordant for both parameters. For a dispersion parameter of 0.5., i.e., for moderate over-dispersion, nearly all results for β_1 and absolutely all results for $ln(\alpha)$ are discordant. For a dispersion parameter of 1, i.e., for equi-dispersion, the results for both parameters are oppositional. For β_1, nearly all results are concordant and for $ln(\alpha)$ absolutely all results are discordant. For a dispersion parameter of 1, i.e., for moderate under-dispersion, all results are concordant except for $ln(\alpha)$, if the two-predictor equation is applied. In the latter case, only about two thirds of the results are concordant and the remaining results are discordant (see Table 7).

5.4 Identification of the optimal computational procedure or, respectively, procedures

As starting value and algorithm have virtually no effect on the parameter estimates, the selection of the optimal specification of these two attributes is solely based on convergence failures and on simplicity of programming. The attribute

Table 6 Percentages of concordance between results for individual analytical units and for all data[a]

Computational procedure attribute	Quality of parameter estimation[b]										
	β_0 [c]			β_1 [c]			β_2 [c]			$ln(\alpha)$[d]	
	Bias	RMSE	Coverage	Bias	RMSE	Coverage	Bias	RMSE	Coverage	Bias	RMSE
One-predictor equation											
Normalisation factor											
Concordant[e]	75.0%	56.3%	65.0%	55.8%	75.0%	95.8%				100.0%	50.0%
Ambiguous[f]	–	–	21.9%	–	–	0.3%				–	–-
Discordant[g]	25.0%	43.8%	13.2%	44.2%	25.0%	3.9%				0.0%	50.0%
Estimation strategy											
Concordant[e]	55.3%	53.3%	27.3%	57.8%	59.7%	36.1%				54.0%	64.2%
Ambiguous[f]	–	–	68.1%	–	–	61.2%				–	–-
Discordant[g]	44.7%	46.7%	4.7%	42.2%	40.3%	2.7%				46.0%	35.8%
Two-predictor equation											
Normalisation factor											
Concordant[e]	75.0%	60.7%	99.0%	56.0%	75.0%	95.8%	48.0%	75.0%	90.7%	100.0%	42.2%
Ambiguous[f]	–	–	0.9%	–	–	0.9%	–	–	1.7%	–	–-
Discordant[g]	25.0%	39.3%	0.1%	44.0%	25.0%	3.3%	52.0%	25.0%	7.5%	0.0%	57.8%
Estimation strategy											
Concordant[e]	57.6%	50.0%	30.7%	58.4%	63.7%	35.7%	58.5%	54.2%	34.4%	52.1%	64.5%
Ambiguous[f]	0.1%	0.5%	58.0%	–	0.8%	61.8%	–	0.8%	55.2%	–	–-
Discordant[g]	42.4%	49.5%	11.3%	41.6%	35.4%	2.5%	41.5%	45.0%	10.4%	47.9%	35.5%

Table 6 (continued)

[a]The analytical units are combinations of scenarios with specification combinations of the other computational procedure attributes. For the two attributes included in the table these are always 28 specification combinations

[b]β_0, β_1, β_2, and α are the generic terms for the parameters to be estimated

[c]All 192 scenarios are included

[d]Only scenarios with $\beta_{g,2} = \beta_{g,3} = 0$, i.e., 64 scenarios are included

[e]Percentages of analytical units with both effects are in the same direction with ties on both sides counting as effects in the same direction

[f]Percentages of analytical units one tie on one side but none on the other

[e]Percentages of analytical units with both effects are in opposite directions

'algorithm' has a very distinct effect on convergence failures. For two specifications of this attribute, i.e., for the two variants of the BFGS algorithm, there are exactly no convergence failures at all. For the other five, there are, at least, very few convergence failures. Accordingly, these other five algorithms are discarded. With these five algorithms discarded, there are no longer any convergence failures. Therefore, the final selection of the specifications for the attributes 'algorithm' and 'starting value' can be based solely on simplicity of programming. As it is easier to set the starting value equal to 1 than to compute it, the first alternative is chosen. Moreover, as it is easier, with the BFGS Algorithm at hand, to have the gradients determined numerically rather than algebraically, the completely numerical version is chosen.

For the remaining two computational procedure attributes, the decision must be based on the effects regarding quality of parameter estimation. For the attribute 'estimation strategy', simultaneous estimation of the regression coefficients and the dispersion parameter leads to lower RMEs for β_1 and $ln(\alpha)$ for both regression equations. However, this result is not very stable across scenarios (see Table 6). On the other hand, the effects of the estimation strategy are generally not very large (see Tables 4 and 5). Consequently, choosing the specification for the estimation strategy on the grounds of simplicity of programming seems wise. This decision rule also leads to simultaneous estimation. Therefore, this specification is chosen. For the remaining attribute, i.e., for the treatment of the normalisation factor, no clear result emerges. For both regression equations, the RMSE for β_1 is higher if the normalisation factor is set equal to 1, whereas, in this case, the RMSE for $ln(\alpha)$ is lower. Moreover, the results vary strongly with the scenario, specifically with the dispersion (see Tables 6 and 7). Thus, no decision can be made as to whether the normalisation factor should be set equal to 1 or approximated via the formula proposed by Efron (see formula 4).

Because of the results regarding the normalisation factor, no unique optimal computational procedure, but only a set of potentially optimal computational procedures, can be identified. This set consists of two procedures. These two procedures have three components in common, i.e., (1) the regression coefficients and the dispersion parameter are estimated simultaneously; (2) the starting value for the dispersion parameter is 1; and (3) the completely numerical variant of the BFGS algorithm is used. The two potentially optimal procedures differ with regard to the treatment of the normalisation factor. In one of these procedures, this factor is set equal to 1; in the other it is approximated using the formula of Efron (see Formula 4).

5.5 Characteristics of the optimal computational procedure or, respectively, procedures

The two potentially optimal computational procedures have nearly the same values for the criteria regarding quality of parameter estimation as do the two attribute specifications with respect to which the procedures differ, i.e. the two treatments of the normalisation factor (see Tables 4, 5, and 8). However, there is a large variation across scenarios. This applies especially to the criteria referring to β_0. In some

Table 7 Distributions of concordant and discordant analytical units regarding normalisation factor separated for dispersion levels

	Concordant analytical units	Discordant analytical units
One-predictor equation		
RMSE: β_1 [a]		
$\alpha_g = 0.1$	100.0%	–
$\alpha_g = 0.5$	0.1%	99.9%
$\alpha_g = 1$	100.0%	–
$\alpha_g = 2$	100.0%	–
RMSE: $ln(\alpha)$[b]		
$\alpha_g = 0.1$	100.0%	–
$\alpha_g = 0.5$	–	100.0%
$\alpha_g = 1$	–	100.0%
$\alpha_g = 2$	100.0%	–
Two-predictor equation		
RMSE: β_1 [a]		
$\alpha_g = 0.1$	100.0%	–
$\alpha_g = 0.5$	0.1%	99.9%
$\alpha_g = 1$	96.9%	3.1%
$\alpha_g = 2$	100.0%	–
RMSE: $ln(\alpha)$[c]		
$\alpha_g = 0.1$	100.0%	–
$\alpha_g = 0.5$	–	100.0%
$\alpha_g = 1$	–	100.0%
$\alpha_g = 2$	68.8%	31.3%

[a]Number of analytical units per row is 48 (scenarios) * 28 (result patterns per scenario) = 1,344

[b]Only scenarios with $\beta_{g,2} = \beta_{g,3} = 0$ are included. Accordingly, the number of analytical units per row is 16 (scenarios) * 28 (result patterns per scenario) = 448

[c]Only scenarios with $\beta_{g,3} = 0$ are included. Accordingly, the number of analytical units per row is 32 (scenarios) *28 (result patterns per scenario) = 896

scenarios, the parameter estimation fails completely. There are even scenarios with zero coverages. These scenarios are exclusively scenarios with a dispersion parameter of 0.1, i.e., scenarios with very high over-dispersion. However, for practical applications, β_0, i.e., the additive parameter, is usually not important. In contrast, with regard to practical applications, the RMSEs for β_1 and $ln(\alpha)$ are crucial. Therefore, closer inspection is conducted regarding the way in which the scenario attributes determine these variables. It is, again, the dispersion that has the largest effect and it is, again, the dispersion level of 0.1. i.e., high over-dispersion, for which the results are by far the worst (see Table 9). However, the effects of dispersion on the estimation of β_1 are by far less detrimental than the effects on the estimation of β_0. The worst coverage obtained for β_1 is 85.20% (see Table 8).

5.6 Case study

If the two potentially optimal computational procedures are applied to the data of the case study (Storgaard et al. 2018), these procedures yield, with the exception of the significance level, the same results for all statistical tests regarding the parameters (see Table 10). With one exception, the two approaches used as comparators, i.e., Poisson Regression with robust standard errors and Negative Binomial Regression, provide the same test results, in so far as these tests can also be performed with these two comparators. The exception is due to the Negative Binomial Regression combined with the two-predictor equation. In contrast to all other approaches, this approach does not provide a statistically significant result for β_1 for the outcome 'days in hospital' (see Table 10).

For the outcome 'number of exacerbations', all approaches yield nearly the same estimates for β_0. For the outcome 'days in hospital', this also applies if the one-predictor equation is employed. The results produced by the two-predictor equation are different. Moreover, for this equation, the result of Negative Binomial Regression differs from the results of the other three procedures. The parameter estimates of β_1 are only close to each other if the one-predictor equation is applied to 'number of exacerbations'. With the two-predictor equation applied to this outcome, and with 'days in hospital' as outcome, the estimates of β_1 differ. For the same outcome, the estimates of β_2 are close together. In contrast, the estimates of $ln(\alpha)$ obtained with the same equation for the same outcome differ slightly from each other (see Table 10).

6 Discussion

Four different attributes of computational procedure for estimating parameters for the Double Poisson Regression have been investigated here. The investigations have provided important information for all four attributes. However, there are qualitative differences between the kinds of information provided for the four different attributes. For some attributes, this information mainly has direct implications for data analysis; for other attributes, this information has more implications for future research. For the treatment of the normalisation factor, the extent to which the information has implications for future research is largest, and, for the algorithm applied to find the parameters, it is smallest. For this reason, the four attributes are discussed in reverse order in comparison with the preceding text, i.e., (1) algorithm, (2) starting value, (3) estimation strategy, and (4) treatment of the normalisation factor.

6.1 Algorithm

The most important result regarding the algorithms and their different variants is that all of them provide virtually the same results. This result can be interpreted as a kind of cross-validation of the three algorithms and their different variants. This cross-validation implies that the three algorithms and their different variants all work almost perfectly. Each algorithm and each variant can be applied with hardly any problems. There might only be some very small problems with convergence failures for the

Table 8 Evaluation criteria values for potentially optimal computational procedures[a]

Specification of normalisation factor	Quality of parameter estimation[b]										
	β_0 [c]			β_1 [c]			β_2 [c]			$ln(\alpha)$[d]	
	Bias[e]	RMSE[e]	Coverage	Bias[e]	RMSE[e]	Coverage	Bias[e]	RMSE[e]	Coverage	Bias[e]	RMSE[e]
One-predictor equation											
Factor = 1											
Mean	6.9831	8.7075	73.17%	− 0.5950	6.9480	95.15%				10.8120	20.2587
Standard deviation	12.9346	12.3576	36.05	1.2122	3.7557	0.96				27.4378	22.0145
Minimum	− 0.2607	1.4365	0.00%	− 5.5001	1.9517	90.90%				− 8.6189	5.9072
Maximum	40.9234	41.2326	96.85%	0.3675	16.9903	97.05%				58.4202	58.6411
Factor approximated											
Mean	9.7779	11.2384	68.87%	− 0.7434	6.3852	94.33%				23.7938	26.8034
Standard deviation	17.0132	16.4874	38.25	1.5688	2.9417	1.84				35.4601	33.5827
Minimum	− 0.2314	1.4111	0.00%	− 6.2694	1.9517	85.20%				− 1.3900	5.2516
Maximum	53.1049	53.4172	96.10%	0.2845	13.8503	96.40%				84.9517	85.0795
Comparison											
Standard deviation	1.9762	1.7896	3.04	0.1049	0.3980	0.58				9.1795	4.6278
Statistical test	p < .001	p < .001	p < .001	p < .001	p < .001	p < .001				p < .001	p < .001
Two-predictor equation											
Factor = 1											
Mean	8.2719	16.2863	82.69%	− 0.5919	6.8461	95.10%	− 4.2442	22.6208	93.96%	9.0531	17.9815
Standard deviation	15.1961	14.9190	28.72	1.2131	3.8075	0.98	10.1031	12.2864	3.34	24.1130	19.1975
Minimum	− 1.9342	2.3583	0.55%	− 5.4545	1.7369	90.95%	− 36.9145	7.5085	81.45%	− 8.4759	5.6262
Maximum	41.0709	52.0138	96.90%	0.3536	16.9927	96.90%	2.1161	52.6186	96.70%	58.5962	58.8176
Factor approximated											
Mean	12.2407	18.9107	76.76%	− 0.7373	6.2906	94.31%	− 6.6666	22.4686	90.61%	20.4739	23.7188
Standard deviation	20.8879	19.3292	33.43	1.5686	2.9983	1.82	14.6439	13.0526	10.19	32.0846	30.1806
Minimum	− 0.7829	2.3475	0.00%	− 6.2330	1.7373	85.50%	− 50.0074	7.5104	55.75%	− 1.2482	5.1095

Table 8 (continued)

Specification of normalisation factor	Quality of parameter estimation[b]										
	β_0 [c]			β_1 [c]			β_2 [c]			$ln(\alpha)$[d]	
	Bias[e]	RMSE[e]	Coverage	Bias[e]	RMSE[e]	Coverage	Bias[e]	RMSE[e]	Coverage	Bias[e]	RMSE[e]
Maximum	54.8976	59.2004	96.35%	0.2719	13.8507	96.35%	1.9504	58.2900	96.00%	85.1224	85.2507
Comparison											
Standard deviation	2.8064	1.8557	4.19	0.1028	0.3928	0.56	1.7129	0.1076	2.37	8.0757	4.0569
Statistical test	p < .001	p < .001	p < .001	p < .001	p < .001	p < .001	p < .001	p = .538	p < .001	p < .001	p < .001

[a]Only the treatment of the normalisation factor varies. The specifications for the other three investigated factors are simultaneous estimation strategy, starting value for dispersion parameter = 1, and BFGS algorithm with gradients numerically determined

[b] $\beta_0, \beta_1, \beta_2$, and α are the generic terms for the parameters to be estimated

[c]All 192 scenarios are included

[d]For the one-predictor equation, only scenarios with $\beta_{g,2} = \beta_{g,3} = 0$, i.e., 64 scenarios, are included. For the two-predictor equation these are only scenarios with $\beta_{g,3} = 0$, i.e., 128 scenarios

[e]Bias and RMSE have been multiplied by 100 to increase readability of the table entries

Table 9 Mean RMSEs for β_1 and *ln(α)* separated for dispersion levels[a]

	Computational procedure	
	Normalisation factor = 1	Approximated normalisation factor
One-predictor equation		
100*RMSE: β_1 [b]		
$\alpha_g = 0.1$	11.7170	9.7275
$\alpha_g = 0.5$	3.7655	3.7789
$\alpha_g = 1$	5.1521	5.1473
$\alpha_g = 2$	7.1575	6.8870
100*RMSE: *ln(α)*[c]		
$\alpha_g = 0.1$	57.9876	84.3466
$\alpha_g = 0.5$	6.0126	5.7873
$\alpha_g = 1$	10.1814	5.3934
$\alpha_g = 2$	6.8533	11.6862
Two-predictor equation		
100*RMSE: β_1 [b]		
$\alpha_g = 0.1$	11.6939	9.7206
$\alpha_g = 0.5$	3.5937	3.6074
$\alpha_g = 1$	5.0319	5.0296
$\alpha_g = 2$	7.0649	6.8047
100:RMSE: *ln(α)*[d]		
$\alpha_g = 0.1$	49.5112	74.3490
$\alpha_g = 0.5$	5.9360	5.7942
$\alpha_g = 1$	8.7889	5.4893
$\alpha_g = 2$	7.6899	9.2426

[a]All differences related to dispersion are statistically significant with $p < 0.001$

[b]Number of scenarios per cell is 48

[c]Only scenarios with $\beta_{g,2} = \beta_{g,3} = 0$ are included. Accordingly, the number of scenarios per cell is 16

[d]Only scenarios with $\beta_{g,3} = 0$ are included. Accordingly, the number of scenarios per cell is 32

Newton–Raphson and the BHHH algorithm. This makes the BFGS algorithm the safest choice. Both variants of this algorithm are fine. However, having the gradients determined numerically causes the least problems in programming. For this reason, this variant seems to be optimal.

6.2 Starting value

The results for the starting value are similar to the results for the algorithm. Both starting values investigated here produce virtually the same parameter estimates.

Table 10 Case study[a]

Analytical approach	Parameters[b]			
	β_0	β_1	β_2	$ln(\alpha)$
Number of exacerbations				
One-predictor equation				
Double Poisson regression; normalisation factor = 1	1.49 (0.08); [1.33; 1.66]; p < .001	− 0.43 (0.13); [− 0.69; -0.17]; p < .01	–-	− 1.15 (0.10); [− 1.35; − 0.95]; p < .001
Double Poisson regression; approximated normalisation factor	1.52 (0.07); [1.38; 1.67]; p < .001	− 0.38 (0.11); [− 0.60; − 0.15]; p < .001	–-	− 0.95 (0.08); [− 1.11; − 0.79]; p < .001
Poisson regression with robust standard errors	1.49 (0.08); [1.34; 1.65]; p < .001	− 0.43 (0.13); [− 0.68; − 0.18]; p < .01	–-	–-
Negative binomial regression	1.49 (0.10); [1.30; 1.68]; p < .001	− 0.43 (0.26); [− 0.71; -0.15]; p < .01	–-	–-
Two-predictor equation				
Double Poisson regression; normalisation factor = 1	1.52 (0.09); [1.34; 2.70]; p < .001	− 0.74 (0.15); [− 1.02; − 0.45]; p < .001	0.09 (0.02); [0.05, 0.14]; p < .001	− 0.82 (0.09); [− 1.00; − 0.63]; p < .001
Double Poisson regression; approximated normalisation factor	1.49 (0.11); [1.27; 1.70]; p < .001	− 0.92 (0.19); [− 1.30; − 0.55]; p < .001	0.11 (0.03); [0.06; 0.17]; p < .001	− 1.08 (0.15); [− 1.37; − 0.79]; p < .001
Poisson regression with robust standard errors	1.49 (0.09); [1.31; 1.68]; p < .001	− 0.83 (0.16); [− 1.15; − 0.51]; p < .001	0.10 (0.02); [0.06; 0.15]; p < .001	–-
Negative binomial regression	1.48 (0.12); [1.24; 1.72]; p < .001	− 0.86 (0.18); [− 1.22; − 0.52]; p < .001	0.12 (0.03); [0.06; 0.18]; p < .001	–-
Days in hospital				
One-predictor equation				
Double Poisson regression; normalisation factor = 1	2.43 (0.12); [2.20; 2.66]; p < .001	− 0.61 (0.18); [− 1.00; − 0.22]; p < .001	–-	− 2.74 (0.10); [− 2.94; − 2.55]; p < .001
Double Poisson regression; approximated normalisation factor	2.42 (0.09); [2.40; 2.74]; p < .001	− 0.36 (0.13); [− 0.74; − 0.26]; p < .01	–-	− 2.42 (0.08); [− 2.57; − 2.26]; p < .001

Table 10 (continued)

Analytical approach	Parameters[b]			
	β_0	β_1	β_2	$ln(\alpha)$
Poisson regression with robust standard errors	2.43 (0.13); [2.18; 2.68]; p < .001	− 0.61 (0.23); [− 1.06; − 0.16]; p < .01	--	--
Negative binomial regression	2.43 (0.19); [2.08; 2.81]; p < .001	− 0.61 (0.26); [− 1.13; − 0.09]; p < .05	--	--
Two-predictor equation				
Double Poisson regression; normalisation factor = 1	2.12 (0.12); [1.89; 2.35]; p < .001	− 0.79 (0.18); [− 1.14; -0.42]; p < .001	0.04 (0.005); [0.03; 0.05]; p < .001	− 2.54 (0.10); [− 2.74; − 2.35]; p < .001
Double Poisson regression; approximated normalisation factor	2.31 (0.09); [2.13; 2.49]; p < .001	− 0.50 (0.12); [− 0.74; − 0.26]; p < .001	0.03 (0.004); [0.02, 0.04]; p < .001	− 2.22 (0.08); [− 2.38; − 2.06]; p < .001
Poisson regression with robust standard errors	2.12 (0.11); [1.90; 2.34]; p < .001	− 0.78 (0.23); [− 1.23; − 0.34]; p < .001	0.04 (0.004); [0.03; 0.05]; p < .001	--
Negative binomial regression	1.96 (0.19); [1.56; 2.40]; p < .001	− 0.49 (0.25); [− 0.99; 0.02]; p = 0.056	0.04 (0.01); [0.02; 0.08]; p < .001	--

[a]Cell entries are: parameter estimate (standard error); [95%-confidence interval]; p-value for test regarding deviation from zero

[b] $\beta_0, \beta_1, \beta_2$, and α are the generic terms for the parameters to be estimated

This indicates that both work fine. This, in turn, provides justification for choosing the simplest approach, i.e., for taking 1 as a starting value.

6.3 Estimation strategy

The results for the estimation strategy are not as unambiguous as the results for the algorithm and the starting value. The two estimation strategies do not produce exactly the same parameter estimates. Moreover, there are also slight differences with respect to the evaluation criteria regarding quality of parameter estimation. To the extent that these differences exist, they indicate an advantage for simultaneous in contrast to sequential estimation. However, these differences are not very stable across scenarios. This provides some justification for putting the decision for simultaneous estimation in question. On the other hand, the effects of the estimation strategy on quality of parameter estimation are, as a whole, very small. Moreover, simultaneous estimation is computationally simpler than sequential estimation. This, again, is an argument for simultaneous estimation. Nevertheless, the fact that there are some unclear effects of estimation strategy on the parameter estimates and quality of parameter estimation could be seen as a starting point for research aimed at identifying the conditions that determine which of the two strategies is better. However, considering the small size of the effects, this research will not be of much practical relevance.

6.4 Normalisation factor

The results regarding the treatment of the normalisation factor differ substantially from the results for the other computational procedure attributes. The treatment of the normalisation factor strongly affects the parameter estimates and quality of parameter estimation. However, the results provide no hint as to which of the two treatments is better. Consequently, based on the results presented here, the best and only recommendation that can be given for data analysis is to apply both treatments parallel to each other for sensitivity analysis. If both analyses provide the same answer for the research question addressed, the answer that should be given is clear. Otherwise, there are some interpretational problems.

This uncertainty regarding the treatment of the normalisation factor defines new research questions that are definitely of high practical relevance. One of these questions is what the conditions are that determine which of the two investigated treatments is better. A further question is how treatment of the normalisation factor could be improved. The next candidate for a reasonable answer to this question is the treatment applied in the R-package for GAMLSS (Stasinopoulos and Rigby 2023). This consists in defining a limit for the infinite sum in the mathematically correct formulation of the normalisation factor and computing the sum until this limit. The central problem of this approach is to find reasonable rules for defining the limit. Setting the limit too low will render the computation inaccurate due to missing information. Setting the limit to high will render the computation inaccurate due to accumulation

of computational inaccuracy. A further approach for a better treatment of the normalisation factor might consist in a better approximation than that proposed by Efron. All in all, there are several avenues for promising research in this context.

Appendix

Log likelihood functions, gradients, and Hessian matrices.

1. General remarks

As the log likelihood functions are maximised with respect to $\alpha\prime$ with $Exp(\alpha\prime) = \alpha$, all equations are given in two presentations, one using $Exp(\alpha\prime)$, the other using α.

2. Log likelihood functions

2.1. Normalisation factor set equal to 1

If the normalisation factor is set equal to 1, the log likelihood function is

$$\begin{aligned} LL(core) &= \sum_{i=1}^{n} \left[\alpha'/2 - \mu_i Exp(\alpha') - y_i + y_i(\ln y_i - ln(\Gamma(y_i+1)) + y_i Exp(\alpha')(1 + \ln \mu_i - \ln y_i)\right] \\ &= \sum_{i=1}^{n} \left[\ln(\alpha)/2 - \mu_i \alpha - y_i + y_i(\ln y_i - ln(\Gamma(y_i+1)) + y_i \alpha(1 + \ln \mu_i - \ln y_i)\right]. \end{aligned} \tag{A1}$$

with n the number of cases and μ_i defined as in Eq. 2 in the main text, and Γ the gamma function.

2.2. Approximated normalisation factor

If the normalisation factor is approximated, the log likelihood function is

$$\begin{aligned} LL(complete) &= LL(core) - \sum_{i=1}^{n} ln\left[1 + \frac{1 - Exp(\alpha')}{12 Exp(\alpha')\mu_i}\left(1 + \frac{1}{Exp(\alpha')\mu_i}\right)\right] \\ &= \sum_{i=1}^{n} \left[\ln(\alpha)/2 - \mu_i \alpha - y_i + y_i(\ln y_i - ln(\Gamma(y_i+1)) + y_i \alpha(1 + \ln \mu_i - \ln y_i) - ln\left[1 + \frac{1-\alpha}{12\alpha\mu_i}\left(1 + \frac{1}{\alpha\mu_i}\right)\right]\right]. \end{aligned} \tag{A2}$$

3. Gradients

3.1. Normalisation factor set equal to 1

If the normalisation factor is set equal to 1, the gradients are

$$\begin{aligned}\frac{\partial LL(core)}{\partial \alpha'} &= \sum_{i=1}^{n}\left[\frac{1}{2} - \mu_i Exp(\alpha') + y_i Exp(\alpha')\left(1 + \ln \mu_i - \ln y_i\right)\right] \\ &= \sum_{i=1}^{n}\left[\frac{1}{2} - \mu_i \alpha + y_i \alpha\left(1 + \ln \mu_i - \ln y_i\right)\right]\end{aligned} \tag{A3}$$

and

$$\frac{\partial LL(core)}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} Exp(\alpha')\left(y_i - \mu_i\right) = \sum_{i=1}^{n} x_{ij}\alpha\left(y_i - \mu_i\right). \tag{A4}$$

with x_{ij} being the value of case i on variable j.

3.2. Approximated normalisation factor

If the normalisation factor is approximated, the gradients are

$$\begin{aligned}\frac{\partial LL(complete)}{\partial \alpha'} &= \frac{\partial LL(core)}{\partial \alpha'} - \sum_{i=1}^{n} \frac{Exp(\alpha') - \mu_i Exp(\alpha') - 2}{12\mu_i^2 Exp(\alpha')^2 + (1 - Exp(\alpha'))\left(\mu_i Exp(\alpha') + 1\right)} \\ &= \sum_{i=1}^{n}\left[\frac{1}{2} - \mu_i \alpha + y_i \alpha\left(1 + \ln \mu_i - \ln y_i\right) - \frac{\alpha - \mu_i \alpha - 2}{12\mu_i^2 \alpha^2 + (1 - \alpha)\left(\mu_i \alpha + 1\right)}\right]\end{aligned} \tag{A5}$$

and

$$\begin{aligned}\frac{\partial LL(complete)}{\partial \beta_j} &= \frac{\partial LL(core)}{\partial \beta_j} + \sum_{i=1}^{n} x_{ij} \frac{\left(1 - Exp(\alpha')\right)\left(\mu_i Exp(\alpha') + 2\right)}{12\mu_i^2 Exp(\alpha')^2 + (1 - Exp(\alpha'))\left(\mu_i Exp(\alpha') + 1\right)} \\ &= \sum_{i=1}^{n} x_{ij}\left[\alpha\left(y_i - \mu_i\right) + \frac{(1 - \alpha)\left(\mu_i \alpha + 2\right)}{12\mu_i^2 \alpha^2 + (1 - \alpha)\left(\mu_i \alpha + 1\right)}\right]\end{aligned} \tag{A6}$$

4. Hessian matrices

4.1. Normalisation factor set equal to 1

If the normalisation factor is set equal to 1, the entries of the Hessian matrix are

$$\frac{\partial^2 LL(core)}{\partial \alpha' \partial \alpha'} = \sum_{i=1}^{n} Exp(\alpha')\left[y_i\left(1 + \ln \mu_i - \ln y_i\right) - \mu_i\right] = \sum_{i=1}^{n} \alpha\left[y_i\left(1 + \ln \mu_i - \ln y_i\right) - \mu_i\right]. \tag{A7}$$

$$\frac{\partial^2 LL(core)}{\partial\alpha'\partial\beta_j} = \frac{\partial^2 LL(core)}{\partial\beta_j\partial\alpha'} = \sum_{i=1}^{n} x_{ij} Exp\left(\alpha'\right)\left(y_i - \mu_i\right) = \sum_{i=1}^{n} x_{ij}\alpha\left(y_i - \mu_i\right). \quad \text{(A8)}$$

and

$$\frac{\partial^2 LL(core)}{\partial\beta_j\partial\beta_k} = -\mu_i Exp\left(\alpha'\right) \sum_{i=1}^{n} x_{ij}x_{ik} = -\mu_i\alpha \sum_{i=1}^{n} x_{ij}x_{ik}. \quad \text{(A9)}$$

4.2. Approximated normalisation factor

If the normalisation factor is approximated, the entries of the Hessian matrix are

$$\begin{aligned}\frac{\partial^2 LL(complete)}{\partial\alpha'\partial\alpha'} &= \frac{\partial^2 LL(core)}{\partial\alpha'\partial\alpha'} - \sum_{i=1}^{n} Exp\left(\alpha'\right) \frac{Exp\left(\alpha'\right)^2\left[12\mu_i^3 - 12\mu_i^2 - \mu_i^2 + \mu_i\right] + Exp\left(\alpha'\right)\left[48\mu_i^2 - 4\mu_i\right] + \mu_i - 1}{\left(12\mu_i^2 Exp(\alpha')^2 + (1 - Exp(\alpha'))\left(\mu_i Exp(\alpha') + 1\right)\right)^2} \\ &= \alpha \sum_{i=1}^{n}\left[y_i\left(1 + \ln\mu_i - \ln y_i\right) - \mu_i - \frac{\alpha^2\left[12\mu_i^3 - 12\mu_i^2 - \mu_i^2 + \mu_i\right] + \alpha\left[48\mu_i^2 - 4\mu_i\right] + \mu_i - 1}{\left(12\mu_i^2\alpha^2 + (1-\alpha)\left(\mu_i\alpha + 1\right)\right)^2}\right]\end{aligned} \quad \text{(A10)}$$

$$\begin{aligned}&\frac{\partial^2 LL(complete)}{\partial\alpha'\partial\beta_j} = \frac{\partial^2 LL(complete)}{\partial\beta_j\partial\alpha\prime} = \frac{\partial^2 LL(core)}{\partial\alpha'\partial\beta_j} \\ &+ \sum_{i=1}^{n} x_{ij}\mu e^{\alpha} \frac{-12\mu_i^2 Exp\left(\alpha'\right)^2 + 24\mu_i Exp\left(\alpha'\right)^2 - Exp\left(\alpha'\right)^2 - 48\mu_i Exp\left(\alpha'\right) + 2Exp\left(\alpha'\right) - 1}{\left(12\mu_i^2 Exp(\alpha')^2 + (1 - Exp(\alpha'))\left(\mu_i Exp(\alpha') + 1\right)\right)^2} \\ &= \alpha \sum_{i=1}^{n} x_{ij}\left[y_i - \mu_i + \mu_i \frac{-12\mu_i^2\alpha^2 + 24\mu_i\alpha^2 - \alpha^2 - 48\mu_i\alpha + 2\alpha - 1}{\left(12\mu_i^2\alpha^2 + (1-\alpha)\left(\mu_i\alpha + 1\right)\right)^2}\right]\end{aligned} \quad \text{(A11)}$$

and

$$\begin{aligned}\frac{\partial^2 LL(complete)}{\partial\beta_j\partial\beta_k} &= \frac{\partial^2 LL(core)}{\partial\beta_j\partial\beta_k} + \sum_{i=1}^{n} x_{ij}x_{ik}\mu_i Exp\left(\alpha'\right)\left(Exp\left(\alpha'\right) - 1\right) \frac{Exp\left(\alpha'\right)^2 12\mu_i^2 + Exp\left(\alpha'\right)\left(48\mu_i - 1\right) + 1}{\left(12\mu_i^2 Exp(\alpha')^2 + (1 - Exp(\alpha'))\left(\mu_i Exp(\alpha') + 1\right)\right)^2} \\ &= \alpha \sum_{i=1}^{n} x_{ij}x_{ik}\mu_i\left[(\alpha - 1)\frac{12\mu_i^2\alpha^2 + 48\mu_i\alpha - \alpha + 1}{\left(12\mu_i^2\alpha^2 + (1-\alpha)\left(\mu_i\alpha + 1\right)\right)^2} - 1\right].\end{aligned} \quad \text{(A12)}$$

Acknowledgements We would like to thank Pete Bereza for his linguistic advice. The work for this manuscript was funded by the Innovation Fund of the Federal Joint Committee (Germany) under the funding code 01NVF18033.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Aragon DC, Achcar JA, Martinez EZ (2018) Maximum likelihood and Bayesian estimators for the double Poisson distribution. J Statist Theory Pract 12:886–911. https://doi.org/10.1080/15598608.2018.1489919

Becker RA, Chambers JM, Wilks AR (1988) The New S Language. Cole, USA: Wadsworth & Brooks

Berndt E, Hall B, Hall R, Hausman J (1974) Estimation and inference in nonlinear structural models. Annals Soc Measur 3:653–665

Broyden CG (1970) The Convergence of a class of double-rank minimization algorithms. J Institute of Math Appl 6:76–90

Cameron AC, Trivedi PK (2013) Regression analysis of count data, 2nd edn. Cambridge University Press, New York, USA

Conway RW, Maxwell WL (1962) A queuing model with state dependent service rates. J Ind Eng 12:132–136

Efron B (1986) Double exponential families and their use in generalized linear regression. J Am Stat Assoc 81(395):709–721

Fletcher R (1970) A new approach to variable metric algorithms. Computer J 13:317–322

Goldfarb D (1970) A family of variable metric updates derived by variational means. Math Comput 24:23–26

Henningsen A, Toomet O (2011) maxLik: A package for maximum likelihood estimation in R. Comput Stat 26:443–458. https://doi.org/10.1007/s00180-010-0217-1

Hilbe JM (2011) Negative binomial regression, 2nd edn. Cambridge University Press, New York, USA

Hilbe JM (2014) Modelling count data. Cambridge University Press, New York, USA

Horsman J, Furlong W, Feeny D, Torrance G (2003) The Health Utilities Index (HUI): concepts, measurement properties and applications. Health Qual Life Outcomes 16(1):54. https://doi.org/10.1186/1477-7525-1-54

Konerding U, Redaèlli M, Ackermann K, Altin S, Appelbaum S, Biallas B et al (2021) A pragmatic randomised controlled trial referring to a Personalised Self-management SUPport Programme (P-SUP) for persons enrolled in a disease management programme for type 2 diabetes mellitus and/or for coronary heart disease. Trials 22:659. https://doi.org/10.1186/s13063-021-05636-4

Luck T, Motzek T, Luppa M, Matschinger H, Fleischer S, Sesselmann Y, Roling G, Beutner K, König HH, Behrens J, Riedel-Heller SG (2013) Effectiveness of preventive home visits in reducing the risk of falls in old age: a randomized controlled trial. Clin Interv Aging 8:697–702. https://doi.org/10.2147/CIA.S43284

McNeish D (2017) Small sample methods for multilevel modeling: a colloquial elucidation of REML and the Kenward-Roger correction. Multivar Behav Res 52:661–670. https://doi.org/10.1080/00273171.2017.1344538

Morris TP, White IR, Crowther MJ (2019) Using simulation studies to evaluate statistical methods. Stat Med 38:2074–2102. https://doi.org/10.1002/sim.8086

Poisson SD (1837) Recherches sur la probabilité des jugements en matière criminelle et en matière civile. Bachelier, Paris, France

R Core Team (2023) A Language and Environment for Statistical Computing Reference Index Version 4.3.1. 2023: https://cran.r-project.org/ (Accessed 2023/08/24)

Rabin R, de Charro F (2001) EQ-5D: a measure of health status from the EuroQol Group. Ann Med 33(5):337–343. https://doi.org/10.3109/07853890109002087

Rigby RA, Stasinopoulos DM (2005) Generalized additive models for location, scale and shape. Appl Statist 54(3):507–554

Ripley B, Venables B, Bates DM, Hornik K, Gebhardt A Firth D (2023). Support Functions and Datasets for Venables and Ripley's MASS. https://cran.r-project.org/web/packages/MASS/MASS.pdf (Accessed 2023/05/22)

Sellers KF, Premeaux B (2020) Conway–Maxwell–Poisson regression models for dispersed count data. Wires Comput Statist. https://doi.org/10.1002/wics.1533

Shanno DF (1970) Conditioning of Quasi-Newton methods for function minimization. Math Comput 24:647–656

Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86:420–428

Silcocks P, Whitham D, Whitehouse WP (2010) P3MC: a double blind parallel group randomised placebo controlled trial of Propranolol and Pizotifen in preventing migraine in children. Trials 11:71. https://doi.org/10.1186/1745-6215-11-71

Stasinopoulos M, Rigby R (2023). gamlss.dist: Distributions for generalized additive models for location scale and shape. R package version 6.1–1, https://CRAN.R-project.org/package=gamlss.dist

Storgaard LH, Hockey H, Laursen BS, Weinreich UM (2018) Long-term effects of oxygen-enriched nasal high flow treatment in COPD with chronic hypoxemic respiratory failure. Int J Chron Obs Pulmon Dis. https://doi.org/10.2147/COPD.S159666

Toledo D, Umetsu CA, Camargo AFM, Idemauro ARL (2022) Flexible models for non-equidispersed count data: comparative performance of parametric models to deal with underdispersion. Adv Stat Anal. https://doi.org/10.1007/s10182-021-00432-6

Toomet O, Henningsen A, Graves S, Croissant Y, Hugh-Jones D, Scrucca L (2022) maxLik: Maximum Likelihood Estimation and Related Tools. https://cran.r-project.org/web/packages/maxLik/maxLik.pdf (Accessed 2023/05/11)

Verbeke J, Cools R (1995) The Newton-Raphson method. Int J Math Educ Sci Technol 26:177–193

Winkelmann R (2008) Econometric Analysis of Count Data. Fifth Edition. Berlin Heidelberg, Germany: Springer Verlag.Ypma TJ (1995) Historical development of the Newton-Raphson method. SIAM Review. 37:531–551

Zou Y, Geedipally SR, Lord D (2013) Evaluating the double Poisson generalized linear model. Accid Anal Prev 59:497–505. https://doi.org/10.1016/j.aap.2013.07.017