HEALTH PSYCHOLOGY AND BEHAVIORAL MEDICINE
2023, VOL. 11, NO. 1, 2222164
https://doi.org/10.1080/21642850.2023.2222164

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Evaluation of frequentist test statistics using constrained statistical inference in the context of the generalized linear model

Caroline Keck[a], Axel Mayer[b] and Yves Rosseel[a]

[a]Department of Data Analysis, Faculty of Psychology and Educational Sciences, Ghent University, Ghent, Belgium; [b]Psychological Methods and Evaluation, Bielefeld University, Bielefeld, Germany

ABSTRACT

When faced with a binary or count outcome, informative hypotheses can be tested in the generalized linear model using the distance statistic as well as modified versions of the Wald, the Score and the likelihood-ratio test (LRT). In contrast to classical null hypothesis testing, informative hypotheses allow to directly examine the direction or the order of the regression coefficients. Since knowledge about the practical performance of informative test statistics is missing in the theoretically oriented literature, we aim at closing this gap using simulation studies in the context of logistic and Poisson regression. We examine the effect of the number of constraints as well as the sample size on type I error rates when the hypothesis of interest can be expressed as a linear function of the regression parameters. The LRT shows the best performance in general, followed by the Score test. Furthermore, both the sample size and especially the number of constraints impact the type I error rates considerably more in logistic compared to Poisson regression. We provide an empirical data example together with R code that can be easily adapted by applied researchers. Moreover, we discuss informative hypothesis testing about effects of interest, which are a non-linear function of the regression parameters. We demonstrate this by means of a second empirical data example.

ARTICLE HISTORY
Received 12 October 2022
Accepted 30 May 2023

KEYWORDS
Informative hypothesis testing; constrained statistical inference; type I error rates; generalized linear model; treatment effects

1. Introduction

A researcher wants to examine the relevance of five health indicators for the ability to live self-sufficiently, as opposed to living in a nursing home, after the age of 80. The health indicators are continuous variables, which are assessed by means of questionnaires and include 'access to health care' (x_1), 'use of preventive services' (x_2), 'mental health' (x_3), 'physical activity' (x_4) and 'nutritional status' (x_5). These indicators, amongst

CONTACT Caroline Keck Caroline.Keck@UGent.be Department of Data Analysis, Faculty of Psychology and Educational Sciences, Ghent University, Henri Dunantlaan 1, 9000 Ghent, Belgium

Supplemental data for this article can be accessed online at http://dx.doi.org/10.1080/21642850.2023.2222164.

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

others, have been described as the leading health indicators by the Centers for Disease Control and Prevention (CDC, 2020). The outcome Y, being able to live alone after the age of 80, is binary, where 1 represents success and 0 represents failure. This scenario will be used as a running example throughout the paper. A common technique to analyze data with a binary outcome is logistic regression (Hosmer et al., 2013). If the outcome is a count variable, for example the number of days spent in an intensive care unit in a hospital, Poisson regression can be used (Agresti, 2003).

In situations like this, researchers typically test a variety of hypotheses related to the regression coefficients. For example, the researcher can test whether ‘access to healthcare’ (x_1) has a significant effect on the outcome Y after controlling for the other variables. Or the researcher can test if a subset of regression coefficients, for example ‘physical activity’ (x_4) and ‘nutritional status’ (x_5), have a significant effect on the outcome Y after controlling for the other variables. In other words, the researcher can assess whether regression parameters of interest are significant in the model. For that, standard test statistics like the Wald, the Score or the LRT (see, e.g., Buse, 1982) can be used.

However, in some situations, a certain ordering or certain signs of the regression coefficients can be expected. Considering our exemplary predictors $x_1, \ldots, x_5$ and the binary outcome Y, being able to live self-sufficiently after the age of 80, the researcher assumes that all regression coefficients will be positive. Better access to health care or greater physical activity will lead to a higher probability of success. In the case of regular null hypothesis testing, the researcher can test the null hypothesis that the regression coefficients of all predictors are zero:

$$H_0 : \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0, \beta_5 = 0, \tag{1}$$

against the alternative hypothesis that at least one of the regression coefficients is nonzero in the model:

$$H_a : \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0, \beta_4 \neq 0, \beta_5 \neq 0. \tag{2}$$

If H_0 can be rejected in favor of H_a, the researcher can assess whether the regression coefficients are greater or smaller than zero via post-hoc tests. This procedure is somewhat unfortunate, since the researcher assumed a positive sign for all the regression coefficients right from the start.

In contrast to regular null hypothesis testing, constrained statistical inference (Hoijtink, 2012; Silvapulle & Sen, 2005) allows the researcher to take the ordering or the signs of the regression coefficients into account using equality and inequality constraints. In other words, $H_0 : \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$ can be tested against the ‘informative’ hypothesis $H_a : \beta_1 \geq 0, \beta_2 \geq 0, \beta_3 \geq 0, \beta_4 \geq 0, \beta_5 \geq 0$, where at least one of the inequality constraints must be strictly true, whereas the remaining ones may be equalities. Thus, researchers can formulate their hypotheses of interest directly, instead of making a detour via another hypothesis. This implies that researchers can avoid to increase the risk for inflated type I error rates. Furthermore, informative hypothesis testing provides the researcher with greater power compared to regular null hypothesis testing (see, e.g., Vanbrabant et al., 2015).

The method of informative hypothesis testing is especially useful since many research questions in the social and behavioral sciences implicitly include an expectation of the researcher about the sign or the ordering of regression coefficients. Further examples of such research questions are the following: An organizational psychologist might want to assess the effect of 'job satisfaction' and 'workload' on the number of absence days at work. The expectation could be that the former reduces, whereas the latter increases the number of absence days. Or a clinical psychologist might want to evaluate the relevance of 'therapy motivation' or 'functioning of interpersonal support systems' on successful hospital discharge after a stationary psychotherapy program. Here, it could be expected that both increase the rates of successful hospital discharge.

In the context of the generalized linear model, informative hypothesis testing can be conducted by means of modified versions of the Wald, the Score and the LRT (Silvapulle & Sen, 2005). To calculate the *p*-value of these statistics, different approaches have been proposed (Silvapulle & Sen, 2005). Unfortunately, informative hypothesis testing is rarely used despite the extensive literature resources. This may be because software is lacking or because the constrained statistical inference literature is mainly focused on theory. Therefore, we lack knowledge about the practical performance of informative test statistics under different circumstances. This concerns, for example, the number of constraints, even though larger numbers are quite common, especially when multiple regression coefficients are assumed to be in a certain order. Assume, for instance, that a researcher expects the following ordering of five regression coefficients: $\beta_1 > \beta_2, \beta_2 > \beta_3, \beta_3 > \beta_4, \beta_4 > \beta_5$, which includes four constraints. In that case, the researcher might want to know how including more or less constraints in the hypotheses will affect type I error rates.

Furthermore, we do not know the impact of small sample sizes on the performance of informative test statistics, as the literature primarily describes their asymptotic behavior. This is unfortunate, since applied researchers are typically interested to know whether their available sample size suffices to obtain reasonable results. For the standard linear regression model, there exists some literature that focuses on the impact of sample size on the practical performance of informative hypothesis testing (Keck et al., 2021, 2022; Vanbrabant et al., 2015). However, to this point, similar work is missing for the generalized linear model.

In this paper, we aim at closing this gap. We want to assess the performance of various informative test statistics in the context of the generalized linear model by means of simulation studies. We consider the distance statistic (*D*-statistic) as well as the informative test versions of the Wald, the Score and the LRT. Furthermore, we regard different conditions regarding the sample size and the number of constraints. Note that the test statistics that are used in this paper work equally well for all members of the family of generalized linear models. However, we choose to limit our study to logistic and Poisson regression, as these are very widely used.

This paper is structured as follows. First, we briefly review the generalized linear model and discuss ways of parameter estimation. Subsequently, we present 'regular' as well as informative test statistics. Then, we introduce the design of our simulation studies and give an overview of the obtained results. In the subsequent sections, we present an empirical data example and explain informative hypothesis testing with non-linear constraints. We finish with a short discussion. All R code that was used is available on the OSF project site for this paper.[1]

2. Generalized linear regression model

The generalized linear model has been described by, for example, Agresti (2003), Agresti (2018), McCullagh and Nelder (1989), Nelder and Wedderburn (1972). It differs from the linear regression model in two aspects. First, it can handle non-normally distributed outcomes and second, it models non-linear functions of the mean response variable Y (Agresti, 2003). Three features constitute the generalized linear model, namely a random component, a systematic component and a link function.

The random component refers to the response variable Y and its probability distribution from the exponential family. The systematic component specifies a linear function of the explanatory variables and is called the linear predictor:

$$\beta_0 x_{i0} + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \sum_{j=0}^{p} \beta_j x_{ij}. \tag{3}$$

Usually $x_{i0} = 1$, which makes β_0 the intercept of the model.

The link function $g()$ connects the systematic and the random component:

$$g[E(Y_i)] = \sum_{j=0}^{p} \beta_j x_{ij}, \tag{4}$$

where $g()$ is a possibly non-linear monotone differentiable function. Since we focus on logistic and Poisson regression in this paper, we present only these models in more depth.

The logistic regression model is a generalized linear model with a Bernoulli or binomial random component. If Y_i has a Bernoulli distribution, the distribution is specified by the parameter π_i, which is the probability of success $P(Y_i = 1)$, while $1 - \pi_i$ represents the probability of failure. The canonical link function is a logit link function:

$$g(\pi_i) = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right), \quad (0 < \pi_i < 1), \tag{5}$$

where the value of π_i changes with the values of the explanatory variables:

$$\pi_i = \frac{\exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)}. \tag{6}$$

The Poisson regression model is a generalized linear model with a count random component. If Y_i has a Poisson distribution, the distribution is specified by the parameter μ_i, which represents the expected count. The canonical link function is a log link function:

$$g(\mu_i) = \log(\mu_i), \qquad (\mu_i > 0), \tag{7}$$

where the value of μ_i changes with the values of the explanatory variables:

$$\mu_i = \exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right). \tag{8}$$

Given a random sample of n observations $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, the maximum likelihood (ML) estimates of the regression coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)'$ are usually obtained using iteratively reweighted least squares or (quasi-)Newton methods. This can be done using standard software, for example the `glm()` function in R (R Core Team, 2020). To compute standard errors, we also need the information matrix:

$$\boldsymbol{I}_1 = \frac{1}{n}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}, \tag{9}$$

where $\boldsymbol{X}$ is the design matrix and $\boldsymbol{W}$ is a diagonal matrix of size $n \times n$ whose elements depend on the model (Agresti, 2003, p. 135).

To obtain $\tilde{\beta}$, the vector of inequality constrained regression coefficients, different approaches can be followed. We created custom functions using constrained optimization algorithms (Nocedal & Wright, 1999). More specifically, we employed a quasi-Newton method with box constraints, as implemented in the R function `nlminb()`. Note that this approach can only handle informative hypotheses specifying the sign of regression coefficients (for instance $H_a: \beta_1 > 0$). There are, however, other types of informative hypotheses (Hoijtink, 2012), for example assuming a certain ordering of the regression coefficients (such as $H_a: \beta_1 \geq \beta_2 \geq \beta_3$).

A more flexible approach is implemented in the R package restriktor (Vanbrabant, 2020). It includes ML estimation comparable to iteratively reweighted least squares (IRLS), where the least squares solver is replaced by a quadratic program. This approach can handle all kinds of informative hypotheses, as long as they can be expressed as a linear function of the model parameters.

3. Hypothesis testing

In this section, we present regular test statistics used in classical null hypothesis testing, as well as informative test statistics used in informative hypothesis testing. Note that the test statistics from classical null hypothesis testing are denoted by means of a 'reg' subscript, whereas the test statistics used in informative hypothesis testing are specified by means of an 'info' subscript. Furthermore, $\boldsymbol{R}$ is the constraint matrix specifying the linear combination of regression coefficients expressing the hypothesis of interest.

For example, assume the researcher wants to test whether the effect of 'physical activity' (x_4) is more than twice as large as the effect of 'access to health care' (x_1) and the effect of 'mental health' (x_3) is more than twice as large as the effect of 'use of preventive services' (x_2) after standardizing all variables. In that case, the researcher aims to test $H_0: \beta_4 = 2\beta_1, \beta_3 = 2\beta_2$ against $H_a: \beta_4 \neq 2\beta_1, \beta_3 \neq 2\beta_2$, in the context of classical null hypothesis testing, or $H_a: \beta_4 \geq 2\beta_1, \beta_3 \geq 2\beta_2$ in the context of informative hypothesis testing. Then, in both the regular and the informative case, the rows of $\boldsymbol{R}$ are specified as

$$\boldsymbol{r}_1' = \begin{pmatrix} 0 & -2 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{10}$$

$$\boldsymbol{r}_2' = \begin{pmatrix} 0 & 0 & -2 & 1 & 0 & 0 \end{pmatrix}, \tag{11}$$

leading to the full constraint matrix:

$$\boldsymbol{R} = \begin{pmatrix} 0 & -2 & 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 1 & 0 & 0 \end{pmatrix} \tag{12}$$

with row rank $h = 2$.

3.1. Classical null hypothesis testing

The test statistics from classical null hypothesis testing that will be presented include the Wald, the Score and the LRT. These large sample test statistics are explained in Buse (1982) and are defined as follows:

$$Wald_{reg} = n(\boldsymbol{R}\hat{\boldsymbol{\beta}})'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}), \tag{13}$$

$$Score_{reg} = \frac{1}{n}\boldsymbol{S}(\bar{\boldsymbol{\beta}})'\bar{\boldsymbol{I}}_1^{-1}\boldsymbol{S}(\bar{\boldsymbol{\beta}}), \tag{14}$$

$$LRT_{reg} = -2[\ell(\bar{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}})], \tag{15}$$

where $\boldsymbol{S}(\bar{\boldsymbol{\beta}}) = \frac{\partial}{\partial\boldsymbol{\beta}}\ell(\bar{\boldsymbol{\beta}})$ is the score function evaluated at $\bar{\boldsymbol{\beta}}$, the vector of equality constrained estimates, $\ell(\bar{\boldsymbol{\beta}})$ is the log-likelihood evaluated at $\bar{\boldsymbol{\beta}}$ and $\ell(\hat{\boldsymbol{\beta}})$ is the log-likelihood evaluated at $\hat{\boldsymbol{\beta}}$. All three test statistics follow asymptotically a χ^2-distribution under the null hypothesis with $df = h$, if the model is correct.

The Wald, the Score and the LRT are asymptotically equivalent. But it has been shown that the values of the Wald test are always slightly larger than the values of the LRT, which in turn are always slightly larger than the values of the Score test (Buse, 1982, p. 157). Thus, using the same critical χ^2 value, the tests may have different power properties. This can be one aspect guiding the choice between them. Another aspect may be the computational resources it needs to compute the three tests. For the Wald test, we need to fit the unconstrained model, whereas for the Score test, we need to fit the equality constrained model and for the LRT, we need to fit both the unconstrained and the equality constrained model. Oftentimes, fitting the unconstrained model takes the least amount of time, which is why the Wald test is chosen frequently. However, in some cases, for example if the equality constrained model has a lot less parameters than the unconstrained model, it may need less computational resources to compute the equality constrained model compared to the unconstrained model. Furthermore, the LRT is, in contrast to the Wald and Score test, scale-invariant (see, e.g., Lehmann, 1986).

3.2. Informative hypothesis testing

Often, the informative test statistics are a modified version of the regular test statistics. For example, $Wald_{info}$ can be found in Silvapulle and Sen (2005, p. 154):

$$Wald_{info} = n(\boldsymbol{R}\tilde{\boldsymbol{\beta}})'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\beta}}). \tag{16}$$

It uses the same constraint matrix $\boldsymbol{R}$ but a different vector of regression coefficients compared to the regular Wald test. While $Wald_{reg}$ uses $\hat{\boldsymbol{\beta}}$, the vector of unconstrained regression coefficients, $Wald_{info}$ uses $\tilde{\boldsymbol{\beta}}$, the vector of inequality constrained regression coefficients. $Score_{info}$ can be computed as follows (Silvapulle & Sen, 2005, p. 159):

$$Score_{info} = \frac{1}{n}[\boldsymbol{S}(\tilde{\beta}) - \boldsymbol{S}(\bar{\beta})]'\hat{\boldsymbol{I}}_1^{-1}[\boldsymbol{S}(\tilde{\beta}) - \boldsymbol{S}(\bar{\beta})]. \tag{17}$$

Again, $Score_{reg}$ and $Score_{info}$ use the same $\boldsymbol{R}$ matrix, but $Score_{reg}$ uses $\hat{\beta}$ and $Score_{info}$ uses $\tilde{\beta}$. LRT_{info} is defined as (Silvapulle & Sen, 2005, p. 157):

$$LRT_{info} = -2[\ell(\bar{\boldsymbol{\beta}}) - \ell(\tilde{\boldsymbol{\beta}})], \tag{18}$$

where $\ell(\bar{\boldsymbol{\beta}})$ is the log-likelihood evaluated at $\bar{\boldsymbol{\beta}}$ and $\ell(\tilde{\boldsymbol{\beta}})$ is the log-likelihood evaluated at $\tilde{\boldsymbol{\beta}}$. The difference between LRT_{reg} and LRT_{info} is that the former uses $\ell(\hat{\boldsymbol{\beta}})$, whereas the latter uses $\ell(\tilde{\boldsymbol{\beta}})$.

Finally, lesser known is the D-statistic. It is calculated as (Silvapulle & Sen, 2005, p. 159):

$$D_{info} = 2n[d(\bar{\boldsymbol{\beta}}) - d(\tilde{\boldsymbol{\beta}})], \tag{19}$$

where $d(\bar{\boldsymbol{\beta}})$ and $d(\tilde{\boldsymbol{\beta}})$ are the values of the following two functions at their solutions:

$$f(\boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\hat{\boldsymbol{I}}_1(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad \text{under the constraint } \boldsymbol{R\beta} = \boldsymbol{0}, \tag{20}$$

$$f(\boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\hat{\boldsymbol{I}}_1(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad \text{under the constraint } \boldsymbol{R\beta} \geq \boldsymbol{0}. \tag{21}$$

When minimizing these functions, we treat $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{I}}_1$ as known constants. Note that to compute the D-statistic, we have to use quadratic programming.

In case the model is correct, the informative Wald, Score and LRT as well as the D-statistic asymptotically follow a $\bar{\chi}^2$-distribution under the null hypothesis. This is a mixture of χ^2-distributions.

3.2.1. *P-values*

Silvapulle and Sen (2005) present two approaches for calculating the p-value of informative test statistics. In the first part of this paper, where the informative hypothesis of interest can be expressed as a linear function of the regression coefficients, we use the approach where we first calculate the weights $w_0, \ldots, w_q$ of the $\bar{\chi}^2$ mixture distribution (Silvapulle & Sen, 2005, p. 79). The sum of the weights from 0 to q is one, where q is the rank of $\boldsymbol{X}$ under the null hypothesis. Once we have computed the weights, the p-value of the observed $\bar{\chi}^2$-value ($\bar{\chi}^2_{obs}$) is obtained as follows (Silvapulle & Sen, 2005, p. 86):

$$\Pr(\bar{\chi}^2 \geq \bar{\chi}^2_{obs}) = \sum_{i=0}^{q} w_i(H_0, H_a)\Pr[(h - q + i)\chi^2_{h-q+i} \geq \bar{\chi}^2_{obs}]. \tag{22}$$

The second approach to calculate the p-value of informative test statistics is described and demonstrated in Keck et al. (2021). We use this approach in the second part of this paper, where the informative hypothesis of interest is expressed as a non-linear function of the regression coefficients. Both approaches are explained in more detail in the document ‘A-pValues.pdf’ on the OSF project site.

Note that if the hypothesis of interest only refers to the sign of one regression coefficient or to the sign of one quantity of interest, which is defined as a function of regression coefficients, we have a special case. Then the informative p-value equals the regular (non-informative) p-value divided by 2.

4. Simulation studies

We conducted several simulation studies using logistic and Poisson regression. The goal was to compare the presented informative test statistics in terms of their type I error rates under different conditions. One of the design factors in our simulation studies was sample size. This was of interest, as it is only known how the test statistics behave asymptotically, but not in finite sample sizes. The other design factor was the number of constraints that was included in the informative hypothesis. This coincides with h, the row rank of $\boldsymbol{R}$. As a benchmark, the presented regular test statistics were also included in the simulation studies. By means of these simulation studies, we would like to give applied users a sense of what they might expect when using informative hypothesis testing in the context of the generalized linear model.

4.1. Design

The model we used had a single outcome Y and five predictors $x_1, \ldots, x_5$. We assumed that all predictors were continuous and normally distributed. Since we are interested in type I error, we generated data under the null hypothesis and thus set all regression coefficients $\beta_0, \ldots, \beta_5$ to 0. For the logistic regression, Y was sampled from the Binomial distribution with probability .50. For the Poisson regression, Y was sampled from the Poisson distribution with $\mu = 1$. Sample sizes of 10, 25, 50, 100, 200, 300, 400, 500, 1000, 2000, 10000 were examined. This way, we regarded small sample sizes that are typical in the social sciences (see, e.g., Van de Schoot & Miočević, 2020) as well as medium and large sample sizes. Even though the large sample sizes are unrealistic, they provide insights into the pattern emerging when increasing the sample size. Furthermore, we considered three constraint matrices, namely

$$\boldsymbol{R}_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{23}$$

$$\boldsymbol{R}_2 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{24}$$

and

$$\boldsymbol{R}_3 = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}. \tag{25}$$

The first constraint matrix represents the hypothesis that only β_1 is greater than zero: $H_a\colon \beta_1 > 0$. The second constraint matrix states that at least one regression coefficient, except the intercept, is greater than zero: $H_a\colon \beta_1 \geq 0, \beta_2 \geq 0, \beta_3 \geq 0, \beta_4 \geq 0, \beta_5 \geq 0$.

And the third constraint matrix specifies a hypothesis, where regression parameters are assumed to be in a certain order, namely H_a: $\beta_1 \geq \beta_2, \beta_2 \geq \beta_3, \beta_3 \geq \beta_4, \beta_4 \geq \beta_5$. These constraint matrices were chosen to consider a typical range of the number of regression parameters (see, e.g., Vanbrabant et al., 2015).

For each condition, we generated data, fitted the model and computed the test statistics as well as the corresponding p-values. This was repeated 10000 times. Subsequently, we defined the type I error rate as the proportion of p-values that were lower than the significance level. In this paper, we used $\alpha = .05$.

In the following sections, the results are presented for logistic and Poisson regression. Note that we only report the results of the simulation studies using the first two constraint matrices. The reason is that using the second and third constraint matrices led to very similar results with absolute differences smaller than 0.01, given that the flexible approach for parameter estimation that is implemented in restriktor was used. This intuitively makes sense, since both matrices include five constraints. Furthermore, the constraints of $\boldsymbol{R}_3$, such as $\beta_1 \geq \beta_2$ can be re-formulated to $\beta_1 - \beta_2 \geq 0$, which makes them very similar to the constraints used in $\boldsymbol{R}_2$, such as $\beta_1 \geq 0$. In the end, all matrices are based on inequality constraints. The interested reader is referred to the R scripts of the simulation studies using $\boldsymbol{R}_3$, which can be found on the OSF project site ('B-orderingExampleLogistic.R' and 'C-orderingExamplePoisson.R'). Note that the files 'D-weights.R' and 'E-lavUvpois.R' are also needed to run the script. The latter belongs to the R package lavaan (Rosseel, 2012), but at the time of writing, the functions are not included (yet) in the current public version of the package.

4.2. Type I results logistic regression

Tables 1 and 2 show the type I error rates resulting from running the simulation studies using logistic regression. When considering $\boldsymbol{R}_1$, we can see that LRT_{reg} and $Score_{reg}$ show too large type I error rates for $N = 50$ and below, whereas $Wald_{reg}$ shows too small type I error rates for $N = 25$ and below. In contrast, when considering $\boldsymbol{R}_2$, $Wald_{reg}$ shows too small type I error rates for $N = 200$ and below and LRT_{reg} shows too large type I error rates already for $N = 100$ and below. Thus, in the context of logistic regression, LRT_{reg} and $Wald_{reg}$ seem to be sensitive to the number of constraints. This does not apply to $Score_{reg}$, which only shows too small type I error rates for $N = 25$ and below when using $\boldsymbol{R}_2$.

Table 1. Type I error rates in logistic regression when using $\boldsymbol{R}_1$. Bold values are above .06 and underlined values are below .04.

N	LRT_{reg}	LRT_{info}	$Wald_{reg}$	$Wald_{info}$	D_{info}	$Score_{reg}$	$Score_{info}$
10000	0.051	0.050	0.051	0.050	0.050	0.051	0.050
2000	0.048	0.046	0.048	0.046	0.046	0.048	0.046
1000	0.054	0.049	0.053	0.048	0.048	0.054	0.050
500	0.053	0.050	0.052	0.049	0.049	0.052	0.051
400	0.054	0.050	0.052	0.049	0.049	0.053	0.051
300	0.052	0.050	0.050	0.049	0.049	0.052	0.053
200	0.055	0.050	0.051	0.049	0.049	0.054	0.053
100	0.059	0.055	0.052	0.052	0.052	0.057	**0.062**
50	**0.071**	**0.066**	0.046	0.053	0.053	**0.063**	**0.084**
25	**0.097**	**0.079**	0.021	0.044	0.044	**0.073**	**0.115**
10	**0.234**	**0.130**	0.008	0.000	0.003	**0.128**	**0.353**

Table 2. Type I error rates in logistic regression when using $\boldsymbol{R}_2$. Bold values are above .06 and underlined values are below .04.

N	LRT_{reg}	LRT_{info}	$Wald_{reg}$	$Wald_{info}$	D_{info}	$Score_{reg}$	$Score_{info}$
10000	0.051	0.053	0.051	0.053	0.053	0.051	0.053
2000	0.049	0.049	0.048	0.048	0.047	0.049	0.050
1000	0.049	0.048	0.044	0.046	0.045	0.048	0.051
500	0.052	0.049	0.044	0.044	0.044	0.050	0.054
400	0.052	0.049	0.045	0.046	0.045	0.050	0.060
300	0.054	0.053	0.041	0.048	0.046	0.049	**0.066**
200	0.057	0.050	0.038	0.040	0.038	0.049	**0.065**
100	**0.063**	0.057	0.024	0.034	0.030	0.047	**0.096**
50	**0.077**	**0.062**	0.002	0.017	0.012	0.043	**0.150**
25	**0.113**	**0.074**	0.000	0.000	0.000	0.037	**0.241**
10	**0.599**	**0.207**	0.008	0.000	0.008	0.000	**0.847**

The sensitivity for the number of constraints is something that we can also observe for all informative test statistics except LRT_{info}. That is, LRT_{info} shows too large type I error rates for $N = 50$ and below, no matter whether $\boldsymbol{R}_1$ or $\boldsymbol{R}_2$ is used. In contrast, $Wald_{info}$ and D_{info} only show too small type I error rates for $N = 10$ when using $\boldsymbol{R}_1$. When using $\boldsymbol{R}_2$, $Wald_{info}$ shows too small type I error rates already for $N = 100$ and below and D_{info} shows too small type I error rates already for $N = 200$ and below. In conclusion, it seems that most regular as well as informative test statistics, except $Score_{reg}$ and LRT_{info}, are sensitive to the number of constraints and the sample size.

4.3. Type I results Poisson regression

Tables 3 and 4 show the type I error rates resulting from running the simulation studies using Poisson regression. First of all, it can be observed that the results show more stable type I error rates for all test statistics, compared to logistic regression. The regular test statistics do not show a sensitivity for the number of constraints. More specifically, when using $\boldsymbol{R}_1$, LRT_{reg} shows a too large type I error rate and $Wald_{reg}$ shows a too small type I error rate only at a sample size of $N = 10$. $Score_{reg}$ even shows no problematic type I error rate at all. When using $\boldsymbol{R}_2$, LRT_{reg} shows no problematic type I error rate at all, but $Wald_{reg}$ shows too small type I error rates for $N = 25$ and below and $Score_{reg}$ shows a too small type I error rate for $N = 10$.

Table 3. Type I error rates in Poisson regression when using $\boldsymbol{R}_1$. Bold values are above .06 and underlined values are below .04.

N	LRT_{reg}	LRT_{info}	$Wald_{reg}$	$Wald_{info}$	D_{info}	$Score_{reg}$	$Score_{info}$
10000	0.049	0.051	0.049	0.051	0.051	0.049	0.051
2000	0.050	0.051	0.050	0.051	0.051	0.050	0.051
1000	0.051	0.052	0.051	0.052	0.052	0.051	0.052
500	0.050	0.049	0.050	0.049	0.049	0.050	0.049
400	0.050	0.048	0.050	0.048	0.048	0.050	0.048
300	0.051	0.045	0.050	0.045	0.045	0.050	0.045
200	0.051	0.047	0.051	0.048	0.048	0.051	0.045
100	0.050	0.045	0.049	0.046	0.046	0.050	0.040
50	0.049	0.051	0.050	0.057	0.057	0.051	0.039
25	0.052	**0.061**	0.053	**0.065**	**0.065**	0.055	0.051
10	**0.067**	**0.063**	0.024	0.034	0.034	0.054	**0.106**

Table 4. Type I error rates in Poisson regression when using $\boldsymbol{R}_2$. Bold values are above .06 and underlined values are below .04.

N	LRT_{reg}	LRT_{info}	$Wald_{reg}$	$Wald_{info}$	D_{info}	$Score_{reg}$	$Score_{info}$
10000	0.053	0.054	0.053	0.054	0.054	0.053	0.054
2000	0.051	0.051	0.051	0.051	0.051	0.051	0.051
1000	0.052	0.051	0.052	0.051	0.051	0.052	0.051
500	0.050	0.047	0.051	0.047	0.047	0.051	0.047
400	0.050	0.050	0.050	0.050	0.050	0.050	0.051
300	0.050	0.048	0.050	0.047	0.047	0.050	0.050
200	0.049	0.049	0.050	0.047	0.047	0.050	0.050
100	0.050	0.049	0.051	0.050	0.049	0.050	0.047
50	0.051	0.052	0.048	0.054	0.053	0.051	0.051
25	0.054	0.051	0.033	0.050	0.043	0.049	0.055
10	0.050	0.040	0.003	0.014	0.007	0.030	**0.181**

Moreover, the informative test statistics also do not show a sensitivity for the number of constraints. That is, when using $\boldsymbol{R}_1$, LRT_{info}, $Wald_{info}$ and D_{info} show type I error rates that deviate from the nominal level for $N = 25$ and below. In contrast, $Score_{info}$ already shows one slightly too small type I error rate for $N = 50$. When using $\boldsymbol{R}_2$, $Wald_{info}$, D_{info} and $Score_{info}$ only show problematic type I error rates for $N = 10$. LRT_{info} even shows no problematic type I error rates at all.

To conclude, the sensitivity for the number of constraints that was observed for most regular and informative test statistics in the context of logistic regression could not be observed in the context of Poisson regression. Furthermore, type I error rates were also more stable concerning the sample size when using Poisson regression compared to logistic regression.

5. Empirical data example

The 'DoctorVisits' data set (Cameron, 1986; Cameron & Trivedi, 1998; Mullahy, 1997) comes with the AER package (Kleiber & Zeileis, 2022) and contains data about Australian health service use. The sample size is $n = 5190$. For the empirical data example of this paper, we chose the number of doctor visits as the dependent variable. Age, the number of illnesses and the number of days with reduced activity due to illness served as the predictors. The computations were conducted both for Poisson and logistic regression. In the latter case, the dependent variable was dichotomized, meaning that all values greater than zero were re-coded to 1, whereas 0 values stayed the same.

Following the design of the simulation studies, the following constraint matrix was used:

$$\boldsymbol{R} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{26}$$

stating that at least one regression coefficient, except the intercept, is greater than zero. Using Poisson regression, we obtained $D_{info} = 1640.05, p < .001$ and using logistic regression, we obtained $D_{info} = 504.66, p < .001$. The R code of the empirical data example can be found on the OSF project site ('F-doctorVisits.R').

6. Informative hypothesis testing with non-linear constraints

Informative hypotheses can not only be specified for linear combinations of the regression coefficients, that is $\boldsymbol{R\beta}$, but also for non-linear combinations of the regression coefficients, namely $\boldsymbol{c}(\boldsymbol{\beta})$. In the latter case, $\boldsymbol{c}(\boldsymbol{\beta})$ can either be a scalar- or a vector-based non-linear function of the model parameters that computes a quantity (or a vector of quantities) of interest. Important examples for such non-linear functions include 'effects of interest' such as risk ratios, odds ratios, conditional effects, and average or marginal effects. Keck et al. (2021) already demonstrated how to test informative hypotheses about various effects of interest in the context of the linear regression model. In the context of the generalized linear model, we need to use generalizations of the informative test statistics for this purpose. The generalized Wald test is defined as follows (Silvapulle & Sen, 2005, p. 166):

$$Wald_{info,gen} = n\boldsymbol{c}(\tilde{\boldsymbol{\beta}})'[\boldsymbol{C}(\tilde{\boldsymbol{\beta}})\tilde{I}_1^{-1}\boldsymbol{C}(\tilde{\boldsymbol{\beta}})']^{-1}\boldsymbol{c}(\tilde{\boldsymbol{\beta}}), \tag{27}$$

where $\boldsymbol{c}$ is a non-linear function of $\tilde{\boldsymbol{\beta}}$, $\boldsymbol{C}$ is the Jacobian matrix of $\boldsymbol{c}$ and $\tilde{I}_1$ the unit information matrix. Note that if $\boldsymbol{c}(\boldsymbol{\beta})$ was linear, it would be equal to $\boldsymbol{R\beta}$ in $Wald_{info}$ (see Equation 16).

The generalized D-statistic (Silvapulle & Sen, 2005, p. 164) can be computed as

$$D_{info,gen} = 2n[d(\bar{\boldsymbol{\beta}}) - d(\tilde{\boldsymbol{\beta}})], \tag{28}$$

where $d(\bar{\boldsymbol{\beta}})$ and $d(\tilde{\boldsymbol{\beta}})$ are the values of the following two functions at their solutions:

$$f(\boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\hat{\boldsymbol{I}}_1(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad \text{under the constraint } \boldsymbol{c}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{0}, \tag{29}$$

$$f(\boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\hat{\boldsymbol{I}}_1(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad \text{under the constraint } \boldsymbol{c}(\tilde{\boldsymbol{\beta}}) \geq \boldsymbol{0}. \tag{30}$$

The constraints are different compared to the ones that are used for D_{info}. In this case, we have to use non-linear optimization methods to compute the D-statistic.

The other generalized informative test statistics still look the same as presented before. That is, the generalized Score test can be computed as (Silvapulle & Sen, 2005, p. 166):

$$Score_{info,gen} = \frac{1}{n}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]'\hat{\boldsymbol{I}}_1^{-1}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})], \tag{31}$$

and the generalized LRT is calculated as follows (Silvapulle & Sen, (2005, p. 164):

$$LRT_{info,gen} = -2[\ell(\bar{\boldsymbol{\beta}}) - \ell(\tilde{\boldsymbol{\beta}})]. \tag{32}$$

To compute $\tilde{\boldsymbol{\beta}}$, we can employ non-linear optimization algorithms (see, e.g., Nocedal & Wright, 1999). An example for this can be found on the OSF project site in 'G-estimationExampleNonlin.R', where data of next section's empirical data example is used. The additional R files 'H-elrEffects.R' and 'E-lavUvpois.R' are needed. The former belongs to the R package EffectLiteR (Mayer & Dietzfelbinger, (2019); Mayer et al. (2016)), but at the time of writing, the functions are not included (yet) in the current public version of this package.

6.1. Empirical data example

To demonstrate informative hypothesis testing concerning effects of interest, we present a second empirical data example, where we use data from the ACTIVE study[2] (Ball et al., 2002; Jobe et al., 2001; Tennstedt et al., 2005). The ACTIVE study is a large randomized controlled trial designed to examine the effectiveness of cognitive interventions among older adults. It also served as an empirical data example in Kiefer and Mayer (2021a, 2021b), where effects on count outcomes with non-normal as well as latent covariates are discussed.

For our example, we consider two levels of the treatment variable X, where $X = 0$ is the control group and $X = 1$ denotes the group that receives memory training. This leads to a sample size of $n = 1401$. Additionally, we consider the continuous predictor variable Z, which is a depression score. The outcome variable Y is a count variable, which describes the performance on an inductive reasoning assessment. We are interested to test $H_0\colon AE_{10} = 0$ against $H_a\colon AE_{10} > 0$. In other words, we would like to test if the average effect of memory training is greater than zero. An average effect can be defined as the unconditional expectation of the difference between expected outcomes under treatment and under control, that is $E[E(Y|X = 1, Z) - E(Y|X = 0, Z)]$. In our example, we use a Poisson regression, where the regressors are related to the logarithm of the conditional expectation of the count outcome (see also Equation 8):

$$\log [E(Y|X, Z)] = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ, \tag{33}$$

$$E(Y|X, Z) = \exp (\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ). \tag{34}$$

Consequently, the average effect can be defined as

$$AE = E[\exp (\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ)] - E[\exp (\beta_0 + \beta_2 Z)], \tag{35}$$

which can be estimated as follows:

$$\widehat{AE}(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n} \exp (\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z_i + \hat{\beta}_3 x_i z_i) - \exp (\hat{\beta}_0 + \hat{\beta}_2 z_i). \tag{36}$$

As discussed before, this is a non-linear function of the regression coefficients $\boldsymbol{\beta}$. When testing $H_0\colon AE_{10} = 0$ against $H_a\colon AE_{10} > 0$, we obtain $Wald_{info,gen} = 1.80$, $p = .09$. The computations can be found in the R script 'I-averageEffect.R' on the OSF project site. Since the hypothesis of interest only concerns one effect of interest, the p-value equals half of the value of the regular (non-informative) p-value (0.18). For more complicated informative hypotheses, we demonstrate how to compute the p-value in the R script 'J-pValueSimulation.R', which can also be found on the OSF project site.

7. Discussion

In this paper, we reported about simulation studies examining the impact of the sample size and the number of constraints when performing informative hypothesis testing in the context of the generalized linear model. We focused on logistic and Poisson regression, as these are widely used. As a benchmark, we also included regular test statistics in the simulation studies. We considered various sample sizes as well as three different constraint matrices. One included a single constraint, whereas the other two included five constraints.

Our findings are different for logistic and Poisson regression. That is, using logistic regression, most regular and informative test statistics were sensitive to the number of constraints and the sample size. This could be concluded since type I error rates were closer to the nominal level in small sample sizes when considering only one constraint in the hypothesis, in contrast to five constraints. Using Poisson regression, the sensitivity to the number of constraints and the sample size was much less pronounced for both regular and informative test statistics.

The following general recommendations can be made for applied researchers who want to use informative hypothesis testing. First, the more computationally intensive test statistics seem to perform better in terms of type I error rates compared to the less computationally intensive test statistics. That is, the LRT, which requires fitting of both the inequality and the equality constrained model, performs better than the Score test, which requires fitting of the inequality constrained model, and which, in turn, performs better than the Wald test, which only requires fitting of the unconstrained model.

Specifically, in the context of logistic regression, applied researchers should use sample sizes that are larger than $n = 50$ when dealing with a single constraint and larger than $n = 100$ when dealing with multiple constraints. In the context of Poisson regression, applied researchers should use sample sizes that are larger than $n = 25$ when dealing with a single or with multiple constraints.

By means of using $\boldsymbol{R}_1$, $\boldsymbol{R}_2$ and $\boldsymbol{R}_3$ in our simulation studies, we only investigated hypotheses including inequality and equality constraints concerning regression coefficients. However, there are further options to formulate informative hypotheses (Hoijtink, 2012). For example, effect sizes can be included as follows:

$$H_a\colon \beta_1 > \beta_2 + d \cdot \sigma, \tag{37}$$

where d is an effect size according to Cohen (1988) and σ is the standard deviation. Furthermore, ‘about equality’ constraints can be used to test informative hypotheses such as

$$H_a\colon |\beta_1 - \beta_2| < d \cdot \sigma, \tag{38}$$

which corresponds to

$$H_a\colon \beta_1 - \beta_2 \geq -d \cdot \sigma,\ \ \beta_1 - \beta_2 \leq d \cdot \sigma. \tag{39}$$

Range constraints are a generalization of ‘about equality’ constraints. They can be used to test informative hypotheses such as

$$H_a\colon \beta_1 - \beta_2 \geq \eta_1, \beta_1 - \beta_2 \leq \eta_2, \tag{40}$$

where the difference between β_1 and β_2 is supposed to lie in an interval with lower bound η_1 and upper bound η_2. Of course, combinations of different types of informative hypotheses are also possible.

Even though these hypotheses are formulated in a different way compared to the ones assessed in our paper, we expect that they should lead to similar results concerning type I error rates in future simulation studies. More specifically, we expect that the number of constraints as well as the sample size will remain the decisive factors. This is because in the end, all these different options to formulate informative hypotheses are based on equality and inequality constraints in their null and alternative hypotheses. Still, this needs to be confirmed in future simulation studies.

When designing these future simulation studies, it should be noted that constraint matrices in informative hypothesis testing must be of full row rank, that is, they must have less or an equal number of rows compared to columns. This way, it is made sure that the product of $\boldsymbol{R}\tilde{\boldsymbol{I}}_1^{-1}\boldsymbol{R}$ in the Wald statistic can be inverted (see Equation 16). In contrast, a rank-deficient $\boldsymbol{R}$ would lead to a rank-deficient product, which could not be inverted.

Furthermore, future research should repeat the simulation studies of this paper for other members of the family of generalized linear models. In this paper, we refrain from drawing conclusions about other generalized linear models, such as gamma regression, that were not considered in our simulation studies.

Another interesting future research endeavour would be to assess the impact of different numbers of predictors on type I error rates. It might be that a higher number of constraints is less problematic for a smaller number of predictors compared to a larger number of predictors. This could be, for example, three constraints on three parameters compared to three constraints on thirty parameters.

As mentioned before, informative hypothesis testing provides the researcher with greater power compared to regular null hypothesis testing (see, e.g., Vanbrabant et al., 2015). The question how much power is gained exactly by using informative hypothesis testing in the generalized linear model compared to using regular null hypothesis testing has to be examined in future simulation studies. For these studies, our paper has provided some groundwork to refer to.

The simulation studies of this paper are complemented by an empirical data example using the 'DoctorVists' data set. Furthermore, we discussed informative hypothesis testing with non-linear constraints. In this case, we have to use generalizations of the Wald and the *D*-statistic. We demonstrated this by means of a second empirical data example using data from the ACTIVE study. Here, we showed how informative hypotheses about effects of interest can be computed.

We have provided extensive R code on this paper's OSF project site, which can be adapted by researchers. We hope that the paper, together with the R code, makes it easier for interested readers to use this technology in the future.

Notes

1. https://osf.io/6svrm/?view_only=e3ee3ecd9cb442b1bf0eb175f319a815
2. The data can be downloaded from https://doi.org/10.3886/E128941.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work has been supported by the Research Foundation Flanders (FWO, grant G002819N to Yves Rosseel and Axel Mayer).

Data availability statement

The data set 'DoctorVisits' that has been used in the first empirical data example can be obtained via the R package AER. The data set of the ACTIVE study that has been used in the second empirical data example can be downloaded from https://doi.org/10.3886/E128941. The data sets that were used in the simulation studies are available upon request from the corresponding author. Furthermore, the scripts 'A-orderingExampleLogistic.R' and 'B-orderingExamplePoisson.R' on the OSF project site (https://osf.io/6svrm/?view_only=e3ee3ecd9cb442b1bf0eb175f319a815) show the data generating mechanism.

References

Agresti, A. (2003). *Categorical data analysis*. John Wiley & Sons.

Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.

Ball, K., Berch, D. B., Helmers, K. F., Jobe, J. B., Leveck, M. D., Marsiske, M., …S. L. Willis (2002). Effects of cognitive training interventions with older adults. *Journal of the American Medical Association*, *288*(18), 2271. doi:10.1001/jama.288.18.2271

Buse, A. (1982). The likelihood ratio, Wald and Lagrange multiplier tests: an expository note. *The American Statistician*, *36*(3), 153–157. doi:10.2307/2683166

Cameron, A. C. (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, *1*(1), 29–53. doi:10.1002/jae.3950010104

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge University Press.

CDC. (2020). Leading health indicators.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Chapman & Hall/CRC.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Jobe, J. B., Smith, D. M., Ball, K., Tennstedt, S. L., Marsiske, M., Willis, S. L., …Kleinman, K. (2001). ACTIVE. *Controlled Clinical Trials*, *22*(4), 453–479. doi:10.1016/S0197-2456(01)00139-8

Keck, C., Mayer, A., & Rosseel, Y. (2021). Integrating informative hypotheses into the EffectLiteR framework. *Methodology*, *17*(4), 307–325. doi:10.5964/meth.7379

Keck, C., Mayer, A., & Rosseel, Y. (2022). Overview and evaluation of various frequentist test statistics using constrained statistical inference in the context of linear regression. *Frontiers in Psychology*. doi:10.3389/fpsyg.2022.899165. 16648714

Kiefer, C., & Mayer, A. (2021a). Accounting for latent covariates in average effects from count regressions. *Multivariate Behavioral Research*, *56*(4), 579–594. doi:10.1080/00273171.2020.1751027

Kiefer, C., & Mayer, A. (2021b). Treatment effects on count outcomes with non-normal covariates. *British Journal of Mathematical and Statistical Psychology*, *74*(3), 513–540. doi:10.1111/bmsp.12237

Kleiber, C., & Zeileis, A. (2022). *AER: Applied econometrics with R*. R package version 1.2-10.

Lehmann, E. (1986). *Testing statistical hypotheses*. Springer.

Mayer, A., & Dietzfelbinger, L. (2019). *EffectLiteR: Average and conditional effects*. R package version 0.4-4.

Mayer, A., Dietzfelbinger, L., Rosseel, Y., & Steyer, R. (2016). The EffectLiteR approach for analyzing average and conditional effects. *Multivariate Behavioral Research*, *51*(2–3), 374–391. doi:10.1080/00273171.2016.1151334

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. Chapman & Hall.

Mullahy, J. (1997). Heterogeneity, excess zeros and the structure of count data models. *Journal of Applied Econometrics*, *12*, 337–350. doi:10.1002/(SICI)1099-1255(199705)12:33.0.CO;2-G

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, *135*(3), 370–384. doi:10.2307/2344614

Nocedal, J., & Wright, S. J. (1999). *Numerical optimization*. Springer.

R Core Team (2020). *R: A language and environment for statistical computing*. R foundation for statistical computing.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi:10.18637/jss.v048.i02

Silvapulle, M. J., & Sen, P. K. (2005). *Constrained statistical inference: Order, inequality, and shape restrictions*. Wiley.

Tennstedt, S., Morris, J., Unverzagt, F., Rebok, G., Willis, S., Ball, K., & Marsiske, M. (2005). ACTIVE (Advanced Cognitive Training for Independent and Vital Elderly), United States, 1999–2001. *Interuniversity Consortium for Political and Social Research.*

Vanbrabant, L. (2020). *Restriktor: Constrained statistical inference*. R package version 0.2-800.

Vanbrabant, L., Van de Schoot, R., & Rosseel, Y. (2015). Constrained statistical inference: Sample-size tables for ANOVA and regression. *Frontiers in Psychology*, *5*, 1–8. doi:10.3389/fpsyg.2014.01565

Van de Schoot, R., & Miočević, M., (Ed.). (2020). *Small sample size solutions: A guide for applied researchers and practitioners*. Routledge.