frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 14 October 2022
DOI 10.3389/fpsyg.2022.899165

OPEN ACCESS

EDITED BY
Begoa Espejo,
University of Valencia, Spain

REVIEWED BY
Joshua Ray Tanzer,
Lifespan, United States
Katerina M. Marcoulides,
University of Minnesota Twin Cities,
United States

*CORRESPONDENCE
Caroline Keck
caroline.keck@UGent.be

SPECIALTY SECTION
This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

RECEIVED 18 March 2022
ACCEPTED 31 August 2022
PUBLISHED 14 October 2022

CITATION
Keck C, Mayer A and Rosseel Y (2022)
Overview and evaluation of various
frequentist test statistics using
constrained statistical inference in the
context of linear regression.
Front. Psychol. 13:899165.
doi: 10.3389/fpsyg.2022.899165

Overview and evaluation of various frequentist test statistics using constrained statistical inference in the context of linear regression

Caroline Keck[1]*, Axel Mayer[2] and Yves Rosseel[1]

[1]Department of Data Analysis, Ghent University, Ghent, Belgium, [2]Psychological Methods and Evaluation, Bielefeld University, Bielefeld, Germany

Within the framework of constrained statistical inference, we can test informative hypotheses, in which, for example, regression coefficients are constrained to have a certain direction or be in a specific order. A large amount of frequentist informative test statistics exist that each come with different versions, strengths and weaknesses. This paper gives an overview about these statistics, including the Wald, the LRT, the Score, the $\bar{F}$- and the D-statistic. Simulation studies are presented that clarify their performance in terms of type I and type II error rates under different conditions. Based on the results, it is recommended to use the Wald and $\bar{F}$-test rather than the LRT and Score test as the former need less computing time. Furthermore, it is favorable to use the degrees of freedom corrected rather than the naive mean squared error when calculating the test statistics as well as using the $\bar{F}$- rather than the $\bar{\chi}^2$-distribution when calculating the p-values.

KEYWORDS

informative hypothesis testing, constrained statistical inference, informative test statistics, type I error rates, naive mean squared error, corrected mean squared error, $\bar{F}$-distribution, $\bar{\chi}^2$-distribution

Introduction

Imagine a researcher wants to examine a novel psychotherapy program. A randomized experiment is set up with three treatment groups. One is a control group ($X = 0$), one participates in an established, standard psychotherapy program ($X = 1$) and one participates in the novel psychotherapy program ($X = 2$). No covariates are considered. The researcher is interested in the group means of the dependent variable Y, which denotes the score on a mental health questionnaire. Studies like this are usually conducted to show the superiority of the novel treatment over the standard treatment, as well as the superiority of the standard treatment over the control group. Thus, the researcher assumes that $\mu_2 > \mu_1 > \mu_0$. However, following classical null hypothesis testing procedures, we usually first test a hypothesis like $H_0 : \mu_2 = \mu_1 = \mu_0$ against H_a : not H_0, that is, not all three means are equal.

If we can reject H_0 in favor of H_a, a second step often follows, in which we execute pairwise comparisons to determine which means are equal and which means are not equal. This implies multiple testing, which brings along the risk of an inflated type I error rate. The framework of constrained statistical inference (Silvapulle and Sen, 2005; Hoijtink, 2012) allows us to test so-called informative hypotheses, meaning that we can test the null hypothesis $H_0 : \mu_2 = \mu_1 = \mu_0$ against the ordered hypothesis $H_a : \mu_2 > \mu_1 > \mu_0$ in a single step. Thus, in contrast to classical null hypothesis testing, researchers have the advantages that they can formulate their hypotheses of interest directly, instead of making a detour via another hypothesis, while additionally avoiding to increase the risk for inflated type I error rates.

Informative hypothesis testing can be conducted by means of the Bayesian (see, e.g., Hoijtink et al., 2008; Hoijtink, 2012) as well as the frequentist (see, e.g., Barlow et al., 1972; Robertson et al., 1988; Silvapulle and Sen, 2005) approach, where the latter is the focus of this paper. The Bayesian approach is implemented in the R (R Core Team, 2020) package bain (Gu et al., 2020). The frequentist approach is implemented in SAS/STAT® by means of the PLM procedure (for instructions, see Chapter 87 of SAS Institute Inc., 2015) as well as in several R packages including restriktor (Vanbrabant, 2020) and ic.infer (Grömping, 2010). Recent work of Keck et al. (2021) also demonstrated how to integrate informative hypothesis testing into the EffectLiteR (Mayer and Dietzfelbinger, 2019) package.

Restriktor and ic.infer use a broad range of test statistics, which are presented in Silvapulle and Sen (2005). However, research in the field of constrained statistical inference often uses the famous $\bar{F}$-statistic (see, e.g., Kuiper and Hoijtink, 2010; Vanbrabant et al., 2015) and neglects the distance statistic (D-statistic). Furthermore, each test statistic comes in various versions, for example depending on which estimate is used for the mean squared error or the variance-covariance matrix, and oftentimes, it is not obvious which software program uses which test statistic. There are also different options regarding the distributions that can be used to compute the p-values ($\bar{\chi}^2, \bar{F}$). At the same time, small sample properties of informative test statistics are mostly unknown. Finally, simulation studies that examine the performance of informative test statistics are lacking in the constrained statistical inference literature.

The aim of this paper is twofold. First, we want to give an overview of a broad range of different informative test statistics, including the Wald test, the likelihood-ratio test (LRT), the Score test, the $\bar{F}$- and the D-statistic as well as their different versions. Second, we want to clarify how those test statistics perform when sample and effect sizes, hypotheses and the distribution used for calculating the p-values vary. Note that we only consider the regression setting, where all variables are observed. The paper is structured as follows: We start by presenting the univariate linear regression model to explain all necessary terminology that is used in the following section, where we define the test statistics. These test statistics include "regular" as well as informative test statistics to illustrate the link between them. We also discuss different versions of these test statistics. Subsequently, we report about simulation studies that we conducted. We introduce the design of the studies, that included a broad range of sample sizes as well as effect sizes, and we outline type I and type II error rates. We conclude with a short discussion. Supplementary materials are provided and will be referenced throughout the paper.

Univariate linear regression model

The univariate linear regression model for an observation i can be defined as:

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \quad (1)$$

where y_i is the value of the response variable for observation $i = 1, 2, ..., n$, x_{i0} is 1 and $x_{i1}, ..., x_{ip}$ are the values of the p regressors for observation i, which are assumed to be fixed (in terms of repeated sampling). $\beta_0, ..., \beta_p$ are the regression coefficients and ε_i is a residual for observation i. In matrix notation, the model can be written as $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ is called the design matrix.

This regression model relies on several assumptions. First, we assume that the expected value of ε_i is zero. That is, $E(\varepsilon_i) = 0$ for all i. In matrix notation, this is expressed as $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$, which implies that $E(\boldsymbol{y}) = \boldsymbol{X\beta}$, meaning that there is a linear relationship between $E(\boldsymbol{y})$ and the columns of $\boldsymbol{X}$. Second, we assume that x_i is non-stochastic and $\boldsymbol{X}$ is of full column rank. Third, we assume that the error term has a constant variance: $Var(\varepsilon_i) = \sigma^2_\varepsilon$ for all i. This implies that $Var(y_i) = \sigma^2_\varepsilon$ for all i. Fourth, we assume that the covariance of any two error terms is zero, that is $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all (i, j), where $i \neq j$.

The model can be estimated by means of different approaches such as ordinary least squares (OLS) or maximum likelihood (ML). It can be shown that under the presented assumptions, the OLS estimates of $\boldsymbol{\beta}$ are BLUE (best linear unbiased estimators, see, e.g., Seber and Lee, 2012). Using an example including four predictors, the following model is fitted:

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i, \quad (2)$$

and $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$, and $\hat{\beta}_4$ are obtained via OLS estimation. We may be interested in hypotheses concerning a single parameter like $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ or we might be interested in hypotheses about nested model comparisons like $H_0 : \beta_1 = 0 \wedge \beta_2 = 0$ vs. $H_a : \beta_1 \neq 0 \vee \beta_2 \neq 0 \vee \beta_3 \neq 0 \vee \beta_4 \neq 0$. We can compute various important quantities that are used in hypothesis testing and that are characterized by a hat on top of it. Note that the hat indicates that estimation of the model parameters takes place in an unrestricted way, which will change once we test informative hypotheses. First, an unbiased estimator for the mean squared error is:

$$\hat{\sigma}_{\varepsilon}^{2} = \hat{S}^{2}_{corrected} = \frac{\widehat{RSS}}{n-k}, \tag{3}$$

where k is the column rank of $\boldsymbol{X}$ and $\widehat{RSS}$ is the estimated residual sum of squares $\sum_{i=1}^{n} \hat{e}_i^2$, where $\hat{e}_i = y_i - \hat{y}_i$ and $\hat{y}_i$ are the model predicted values of the response variable. Note that by considering k, we yield a small-sample correction for the mean squared error, as opposed to simply using:

$$\hat{S}^{2}_{naive} = \frac{\widehat{RSS}}{n}, \tag{4}$$

which corresponds to the maximum likelihood estimator of σ_{ε}^{2}.

The variance-covariance matrix of the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ is usually computed as:

$$VCOV(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\hat{\boldsymbol{I}}_1^{-1}, \tag{5}$$

where $\hat{\boldsymbol{I}}_1$ is the unit information matrix:

$$\hat{\boldsymbol{I}}_1 = \frac{1}{n\,\hat{S}^{2}_{corrected}}\boldsymbol{X}'\boldsymbol{X}. \tag{6}$$

Note that if certain model assumptions are violated, for example if the error term does not have a constant variance, robust versions of the standard errors (Huber, 1967; White, 1980) and the variance-covariance matrix (Zeileis, 2006) can be used.

We can also test hypotheses about linear or non-linear combinations of regression parameters, like $H_0 : \beta_1 + \beta_2 = 0 \land \beta_3 + \beta_4 = 0$ vs. $H_a : \beta_1 + \beta_2 \neq 0 \lor \beta_3 + \beta_4 \neq 0$. Note that in this paper, we will focus only on hypotheses containing linear combinations of regression coefficients. These combinations are specified by means of the $\boldsymbol{R}$-matrix and each part of the hypothesis can be expressed as a row in $\boldsymbol{R}$:

$$\boldsymbol{r}_1' = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \end{pmatrix}, \tag{7}$$

$$\boldsymbol{r}_2' = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 \end{pmatrix}, \tag{8}$$

leading to the full constraint matrix:

$$\boldsymbol{R} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}. \tag{9}$$

Then the hypothesis of interest can be expressed as $H_0 : \boldsymbol{R\beta} = \boldsymbol{0}$ vs. $H_a : \boldsymbol{R\beta} \neq \boldsymbol{0}$. Note that all kinds of hypotheses, including the single parameter case and comparisons of nested models, as discussed before, can be expressed by means of the $\boldsymbol{R}$-matrix.

In case our hypothesis of interest contains inequality constraints, like $H_a : \beta_1 + \beta_2 > 0 \lor \beta_3 + \beta_4 > 0$, $\boldsymbol{R}$ still looks the same, but we need to fit a model where we enforce the inequality constraints on the regression coefficients. This can be done by means of quadratic programming, for example using the subroutine `solve.QP()` of the R package quadprog (Turlach and Weingessel, 2019). It implements the dual method of Goldfarb and Idnani (1982, 1983). If we apply this method in the linear regression context, it has the following form (see "Data Sheet 1" in the Supplementary materials for further explanations):

$$min(-\boldsymbol{y}'\boldsymbol{X\beta} + \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta}) \qquad \text{with the constraints } \boldsymbol{R\beta} \geq \boldsymbol{\beta}_0. \tag{10}$$

Note that all quantities based on an inequality constrained model are denoted by a tilde on top of them. Assume that the unconstrained estimates $\hat{\boldsymbol{\beta}}'$ are $(0.100 \quad -0.130 \quad 0.100 \quad -0.240 \quad 0.250)$, but the inequality constrained estimates $\tilde{\boldsymbol{\beta}}'$ may be $(0.110 \quad -0.110 \quad 0.120 \quad -0.230 \quad 0.240)$, where the estimates of β_0, β_3 and β_4 may also change slightly, even though they already satisfied the constraints in the unrestricted estimation. The restricted estimation will also lead to different residuals than the unrestricted estimation.

If our hypothesis of interest contains equality constraints, for example $H_a : \beta_1 + \beta_2 = 0 \lor \beta_3 + \beta_4 = 0$, the equality constrained estimates $\bar{\boldsymbol{\beta}}$ can also be found via quadratic programming. Note that here, H_a from informative hypothesis testing equals H_0 from classical null hypothesis testing. Similarly, all estimated quantities with a bar on top are both the quantities from the equality constrained fit in informative hypothesis testing and the quantities obtained based on H_0 in classical null hypothesis testing, which are in fact equality constrained estimates as well. The corresponding mean squared error terms for the inequality and equality constrained case are defined as follows:

$$\tilde{S}^{2}_{corrected} = \frac{\widetilde{RSS}}{n-k}, \tag{11}$$

$$\tilde{S}^{2}_{naive} = \frac{\widetilde{RSS}}{n}, \tag{12}$$

$$\bar{S}^{2}_{corrected} = \frac{\overline{RSS}}{n-(k-h)}, \tag{13}$$

$$\bar{S}^{2}_{naive} = \frac{\overline{RSS}}{n}, \tag{14}$$

where $\widetilde{RSS}$ is the residual sum of squares of the inequality constrained fit $\sum_{i=1}^{n} \tilde{e}_i^2$, where $\tilde{e}_i = y_i - \tilde{y}_i$ and $\tilde{y}_i$ are the model predicted values of the response variable. Furthermore, $\overline{RSS}$ is the residual sum of squares under the equality constrained fit $\sum_{i=1}^{n} \bar{e}_i^2$, where $\bar{e}_i = y_i - \bar{y}_i$ and $\bar{y}_i$ are the model predicted values of the response variable. Finally, h is the row rank of $\boldsymbol{R}$.

Similarly, we can define the unit information matrices of the inequality and equality constrained fits:

$$\tilde{\boldsymbol{I}}_1 = \frac{1}{n\,\tilde{S}^2_{corrected}}\boldsymbol{X}'\boldsymbol{X}, \tag{15}$$

$$\bar{\boldsymbol{I}}_1 = \frac{1}{n\,\bar{S}^2_{corrected}}\boldsymbol{X}'\boldsymbol{X}. \tag{16}$$

Note that $\boldsymbol{X}$ from the inequality constrained fit equals $\boldsymbol{X}$ from the unconstrained fit. The estimates $\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\beta}}$ as well as the corresponding mean squared error terms and unit information matrices are used in the test statistics that are presented in the subsequent section.

Hypothesis testing

In order to give a broad overview about different test statistics, we present regular test statistics used in classical null hypothesis testing, as well as informative test statistics used in informative hypothesis testing. Note that an overview table containing all test statistics is provided at the end of each section. All test statistics can be applied in the setting of linear regression. "Data Sheet 2" in the Supplementary materials shows how these test statistics are implemented in R code.

Classical null hypothesis testing

The test statistics from classical null hypothesis testing that we will explain include the Wald test, the LRT, the Score test, the F-test as well as the t-test. The large sample test statistics, that is the Wald test, the LRT and the Score test, can be defined as follows Buse (1982):

$$Wald = n(\boldsymbol{R}\hat{\boldsymbol{\beta}})'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}), \tag{17}$$

$$LRT = -2 \cdot [\ell(\bar{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}})], \tag{18}$$

$$Score = \frac{1}{n}\boldsymbol{S}(\bar{\boldsymbol{\beta}})'\bar{\boldsymbol{I}}_1^{-1}\boldsymbol{S}(\bar{\boldsymbol{\beta}}), \tag{19}$$

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

Note that all three test statistics implicitly depend on S^2 in the information matrices (see Equation 6) and in the log-likelihoods. In the regression setting, since we always know what the residual degrees of freedom are, we can use $\hat{S}^2_{corrected}$ instead of $\hat{S}^2_{naive}$ to obtain the corrected instead of naive test statistic versions. That way, we can use the F-distribution with $df_1 = h, df_2 = n - p$ to obtain the p-values, which is more precise in small samples compared to the χ^2-distribution.

Note that the LRT, the Wald and the Score test are asymptotically equivalent. However, 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, which can be one aspect guiding the choice between them. Another aspect may be the time it takes 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 equality constrained model. In many cases, fitting the unconstrained model takes the least amount of time, which is why the Wald test is chosen often. However, in some cases, for example if the equality constrained model has a lot less parameters than the unconstrained model, it may be faster to fit the equality constrained model compared to the unconstrained model.

The F-test can be calculated as Seber and Lee (2012, p. 100):

$$F_{corrected} = \frac{\frac{1}{h}[\overline{RSS} - \widehat{RSS}]}{\hat{S}^2_{corrected}}. \tag{20}$$

Another test statistic version results from using $\hat{S}^2_{naive}$ instead of $\hat{S}^2_{corrected}$, which we denote as F_{naive}. Seber and Lee (2012, p. 100) show that $F_{corrected}$ can be re-written to contain the unit information matrix:

$$F^{info}_{corrected} = \frac{n}{h}(\boldsymbol{R}\hat{\boldsymbol{\beta}})'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}), \tag{21}$$

where the superscript "info" refers to the information matrix. When $\hat{S}^2_{naive}$ instead of $\hat{S}^2_{corrected}$ is used in constructing the unit information matrix, we call this test statistic F^{info}_{naive}. If the model is specified correctly, $F_{corrected}$ follows an F-distribution with $df_1 = h, df_2 = n - k$ under the null hypothesis.

The one-sample t-test is defined as Allen(1997, p. 67):

$$t = \frac{\hat{\beta} - \bar{\beta}}{SE_{\hat{\beta}}}, \tag{22}$$

where $\bar{\beta}$ is the value of β under the null hypothesis and $SE_{\hat{\beta}}$ is the standard error of $\hat{\beta}$. Under the null hypothesis, t is t-distributed with $df = n - k$, if the model is correct. Note that if $h = 1$ the t- and F-statistic are related in a certain way, which is $t^2 = F$.

It is widely known that the one-sample t-test can be used for testing both two-sided hypotheses like $H_0 : \beta = 0$ against $H_a : \beta \neq 0$ as well as one-sided hypotheses like $H_0 : \beta = 0$ against $H_a : \beta > 0$ or $H_a : \beta < 0$. The test statistic stays the same in both cases, but the p-value is computed differently. That is, when testing a two-sided hypothesis, half of the significance level is allocated to each side of the t-distribution, whereas when testing a one-sided hypothesis, all of it is allocated to one side of the t-distribution. That means that the cut-off levels, denoting from which point on the t-statistic can be considered to be significant, change. The two-sided p-value, which is the default output of most statistical software, simply adds up the probabilities of the negative and positive version of the observed t-value (t_{obs}), independently of whether it was in fact positive or negative:

$$\begin{aligned} p_{two-sided} &= 2 \cdot P(t > |t_{obs}|) \\ &= P(t > t_{obs}) + P(t < -t_{obs}). \end{aligned} \tag{23}$$

Since the t-distribution is symmetric, $P(t > t_{obs})$ is the same as $P(t < -t_{obs})$. When we are interested in the one-sided p-value and $H_a : \beta > 0$, the p-value is obtained as:

$$p_{one-sided} = P(t > t_{obs}), \tag{24}$$

whereas if $H_a : \beta < 0$, the p-value is obtained as:

$$p_{one-sided} = P(t < t_{obs}). \tag{25}$$

Note that in case the obtained t-value is a positive number and we are interested in $H_a : \beta > 0$ or in case t is a negative number and we are interested in $H_a : \beta < 0$, the one-sided p-value can be obtained by dividing the two-sided p-value by 2.

In summary, the t-statistic is a special case, since this statistic from the classical null hypothesis testing framework can be used for testing an informative hypothesis, as long as the hypothesis only contains one parameter. If we are interested in more than one parameter, we can no longer use the t-statistic, but have to use an informative test statistic. Table 1 shows an overview about all presented regular test statistics.

TABLE 1 Overview of all presented regular test statistics.

Regular test statistics	Formula
$LRT_{naive/corrected}$	$-2 \cdot [\ell(\bar{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}})]$
$Wald_{naive/corrected}$	$n(\boldsymbol{R}\hat{\boldsymbol{\beta}})'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}})$
$Score_{naive/corrected}$	$\frac{1}{n}\boldsymbol{S}(\bar{\boldsymbol{\beta}})'\bar{\boldsymbol{I}}_1^{-1}\boldsymbol{S}(\bar{\boldsymbol{\beta}})$
$F_{naive/corrected}$	$\frac{\frac{1}{h}[\overline{RSS}-\widehat{RSS}]}{\hat{S}^2_{naive/corrected}}$
t	$\frac{\hat{\beta}-\bar{\beta}}{SE_{\hat{\beta}}}$

Informative hypothesis testing

Informative test statistics are often a modified version of the regular test statistics. In case the model is correct, the large sample informative test statistics, including the LRT, the Wald test, the Score test and the D-statistic, asymptotically follow a $\bar{\chi}^2$-distribution under the null hypothesis, which is a mixture of χ^2-distributions. The small sample informative test statistic, that is the $\bar{F}$-statistic, follows an $\bar{F}$-distribution under the null hypothesis, if the model is correctly specified. The $\bar{F}$-distribution is a mixture of F-distributions. Note that similar to classical null hypothesis testing, we can use the corrected instead of naive mean squared error to obtain the large sample test statistics. In that way, we can calculate the p-values by means of the $\bar{F}$-distribution instead of the $\bar{\chi}^2$-distribution to obtain more precise results in small sample sizes.

The $LRT_{corrected}$ test statistic can be calculated as follows Silvapulle and Sen (2005, p. 157):

$$LRT_{corrected} = -2 \cdot [\ell(\bar{\boldsymbol{\beta}}) - \ell(\tilde{\boldsymbol{\beta}})], \tag{26}$$

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 $\tilde{\boldsymbol{\beta}}$. $\ell(\bar{\boldsymbol{\beta}})$ has been calculated using $\bar{S}^2_{corrected}$ and $\ell(\tilde{\boldsymbol{\beta}})$ has been calculated using $\tilde{S}^2_{corrected}$. If $\bar{S}^2_{naive}$ and $\tilde{S}^2_{naive}$ were used instead, we would obtain LRT_{naive}.

The Wald statistic can be found in Silvapulle and Sen (2005, p. 154):

$$Wald^{info}_{corrected} = \frac{n}{\hat{S}^2_{corrected}}(\boldsymbol{R}\tilde{\boldsymbol{\beta}})'(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\beta}}), \tag{27}$$

where $\boldsymbol{W} = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$. The Wald version where we use $\hat{S}^2_{naive}$ instead of $\hat{S}^2_{corrected}$ is called $Wald^{info}_{naive}$. Both versions implicitly contain $\hat{\boldsymbol{I}}_1$ (see Equation 6), which can also be replaced by $\tilde{\boldsymbol{I}}_1$. Note that $Wald^{info}_{naive}$ will give different results, especially in small sample sizes, due to the missing correction. Assuming $VCOV(\hat{\boldsymbol{\beta}})$ is defined as in Equation 5, we can re-write the Wald statistic as:

$$Wald^{VCOV} = [\boldsymbol{R}\tilde{\boldsymbol{\beta}}]'[\boldsymbol{R}\ VCOV(\hat{\boldsymbol{\beta}})\ \boldsymbol{R}']^{-1}[\boldsymbol{R}\tilde{\boldsymbol{\beta}}], \tag{28}$$

which is identical to $Wald^{info}_{corrected}$. Note that we can also replace $VCOV(\hat{\boldsymbol{\beta}})$ by a more robust sandwich-estimator, which is not commonly done in the applied literature.

The D-statistic is calculated as follows (Silvapulle and Sen, 2005, p. 159):

$$D_{corrected} = \frac{2 \cdot n}{\hat{S}^2_{corrected}}[d(\bar{\boldsymbol{\beta}}) - d(\tilde{\boldsymbol{\beta}})], \tag{29}$$

where $d(\bar{\boldsymbol{\beta}})$ and $d(\tilde{\boldsymbol{\beta}})$ are the values of the following two functions at their solutions (see "Data Sheet 3" in the Supplementary materials for further information):

$$f(\boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \qquad \text{under the constraint } \boldsymbol{R\beta} = \boldsymbol{0}, \tag{30}$$

$$f(\boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \qquad \text{under the constraint } \boldsymbol{R\beta} \geq \boldsymbol{0}. \tag{31}$$

When minimizing these functions, we treat $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{W}$ as known constants. Note that in the regression case, $D_{corrected}$ is identical to $Wald^{info}_{corrected}$ and $Wald^{VCOV}$, as long as $\hat{S}^2_{corrected}$ is used. In contrast, if we switch to using $\hat{S}^2_{naive}$, we obtain D_{naive}, in which case $D_{naive} = Wald^{info}_{naive}$.

The $\bar{F}$-statistic can be found in (Silvapulle and Sen, 2005, p. 29):

$$\bar{F}_{corrected} = \frac{\overline{RSS} - \widetilde{RSS}}{\hat{S}^2_{corrected}}. \tag{32}$$

According to Silvapulle and Sen (2005, p. 29), including the constant $\frac{1}{h}$ from the regular F-statistic in the $\bar{F}$-statistic is not necessary, as it does not affect the results. Again, when using $\hat{S}^2_{naive}$ instead of $\hat{S}^2_{corrected}$, we obtain $\bar{F}_{naive}$. We can re-write the $\bar{F}$-statistic similarly to how we re-wrote the F-statistic. Assuming that we use $\hat{S}^2_{corrected}$ to compute the unit information matrix, we obtain:

$$\bar{F}^{info}_{corrected} = n(\boldsymbol{R}\tilde{\boldsymbol{\beta}})'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\beta}}). \tag{33}$$

Again, $\hat{\boldsymbol{I}}_1$ can be replaced by $\tilde{\boldsymbol{I}}_1$.

There are various versions of the Score statistic. $Score^U_{corrected}$ can be found in Silvapulle and Sen (2005, p. 159):

$$Score^U_{corrected} = \frac{1}{n \cdot \hat{S}^2_{corrected}} \boldsymbol{U}'(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}')^{-1}\boldsymbol{U}, \tag{34}$$

where $\boldsymbol{U} = \boldsymbol{R}\boldsymbol{W}^{-1}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]$. When using $\hat{S}^2_{naive}$ as compared to $\hat{S}^2_{corrected}$, we obtain $Score^U_{naive}$. Another version of the Score statistic, $Score^{null-info}_{corrected}$, is defined as follows Silvapulle and Silvapulle (1995, p. 342):

$$Score^{null-info}_{corrected} = \frac{1}{n}[\boldsymbol{S}(\bar{\boldsymbol{\beta}}) - \boldsymbol{S}(\tilde{\boldsymbol{\beta}})]'\bar{\boldsymbol{I}}_1^{-1}[\boldsymbol{S}(\bar{\boldsymbol{\beta}}) - \boldsymbol{S}(\tilde{\boldsymbol{\beta}})], \tag{35}$$

where $\bar{\boldsymbol{I}}_1$ has been calculated by means of $\bar{S}^2_{corrected}$ (see Equation 13). In contrast, if we use $\bar{S}^2_{naive}$, we obtain $Score^{null-info}_{naive}$.

Furthermore, $Score^{info}_{corrected}$ can be calculated as Silvapulle and Sen (2005, p. 166):

$$Score^{info}_{corrected} = \frac{1}{n}\boldsymbol{P}'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}\boldsymbol{P}, \tag{36}$$

where $\boldsymbol{P} = \boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]$ and $\hat{\boldsymbol{I}}_1$ is calculated using $\hat{S}^2_{corrected}$ and can be replaced by either $\tilde{\boldsymbol{I}}_1$ or $\bar{\boldsymbol{I}}_1$. If we use $\hat{S}^2_{naive}$ to calculate $\hat{\boldsymbol{I}}_1$, we obtain $Score^{info}_{naive}$. Silvapulle and Sen (2005, p. 166) mention another way to express $Score^{info}_{corrected}$:

$$Score^{info,Robertson}_{corrected} = \frac{1}{n}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]'\hat{\boldsymbol{I}}_1^{-1}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})], \tag{37}$$

where the superscript "Robertson" indicates that this is the version defined by Robertson et al. (1988), $\hat{\boldsymbol{I}}_1$ is calculated using $\hat{S}^2_{corrected}$ and can be replaced by either $\tilde{\boldsymbol{I}}_1$ or $\bar{\boldsymbol{I}}_1$. Assuming that $VCOV(\hat{\boldsymbol{\beta}})$ is defined as in Equation 5, $Score^{info}_{corrected}$ can be re-written as:

$$Score^{VCOV} = \boldsymbol{V}'[\boldsymbol{R}\ VCOV(\hat{\boldsymbol{\beta}})\ \boldsymbol{R}']^{-1}\boldsymbol{V}, \tag{38}$$

TABLE 2 Overview of all presented informative test statistics.

Informative test statistics	Formulas
$LRT_{naive/corrected}$	$-2 \cdot [\ell(\bar{\boldsymbol{\beta}}) - \ell(\tilde{\boldsymbol{\beta}})]$
$Wald^{info}_{naive}$	$\frac{n}{\hat{S}^2_{naive}}(\boldsymbol{R}\tilde{\boldsymbol{\beta}})'(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\beta}})$
$Wald^{info}_{corrected} = Wald^{VCOV}$	$\frac{n}{\hat{S}^2_{corrected}}(\boldsymbol{R}\tilde{\boldsymbol{\beta}})'(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\beta}})$
	$= [\boldsymbol{R}\tilde{\boldsymbol{\beta}}]'[\boldsymbol{R}\ VCOV(\hat{\boldsymbol{\beta}})\ \boldsymbol{R}']^{-1}[\boldsymbol{R}\tilde{\boldsymbol{\beta}}]$
$D_{naive/corrected}$	$\frac{2 \cdot n}{\hat{S}^2_{naive/corrected}}[d(\bar{\boldsymbol{\beta}}) - d(\tilde{\boldsymbol{\beta}})]$
$\bar{F}_{naive}$	$\frac{\overline{RSS} - \widetilde{RSS}}{\hat{S}^2_{naive}}$
$\bar{F}_{corrected} = \bar{F}^{info}_{corrected}$	$\frac{\overline{RSS} - \widetilde{RSS}}{\hat{S}^2_{corrected}}$
	$= n(\boldsymbol{R}\tilde{\boldsymbol{\beta}})'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\beta}})$
$Score^U_{naive/corrected}$	$\frac{1}{n \cdot \hat{S}^2_{naive/corrected}}\boldsymbol{U}'(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}')^{-1}\boldsymbol{U}$
$Score^{null-info}_{naive/corrected}$	$\frac{1}{n}[\boldsymbol{S}(\bar{\boldsymbol{\beta}}) - \boldsymbol{S}(\tilde{\boldsymbol{\beta}})]'\bar{\boldsymbol{I}}_1^{-1}[\boldsymbol{S}(\bar{\boldsymbol{\beta}}) - \boldsymbol{S}(\tilde{\boldsymbol{\beta}})]$
$Score^{info}_{naive} = Score^{info,Robertson}_{naive}$	$\frac{1}{n}\boldsymbol{P}'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}\boldsymbol{P}$
	$= \frac{1}{n}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]'\hat{\boldsymbol{I}}_1^{-1}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]$
$Score^{info}_{corrected} = Score^{info,Robertson}_{corrected}$	$\frac{1}{n}\boldsymbol{P}'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}\boldsymbol{P}$
$= Score^{VCOV}$	$= \frac{1}{n}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]'\hat{\boldsymbol{I}}_1^{-1}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]$
	$= \boldsymbol{V}'[\boldsymbol{R}\ VCOV(\hat{\boldsymbol{\beta}})\ \boldsymbol{R}']^{-1}\boldsymbol{V}$

where $\boldsymbol{V} = \boldsymbol{R}\ VCOV(\hat{\boldsymbol{\beta}})\ [\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]$, again allowing for a more robust sandwich-estimator of $VCOV(\hat{\boldsymbol{\beta}})$ to be inserted. Table 2 gives an overview about all the informative test statistics that were presented.

P-values

There are two approaches for calculating the p-value of informative test statistics (Silvapulle and Sen, 2005). In this paper, we use the approach where we first calculate the weights of the respective mixture distribution ($\bar{\chi}^2, \bar{F}$). Note that the sum of the weights from 0 to q is one, where q is the rank of $\boldsymbol{X}$ under the null hypothesis.

If the residuals of our data are normally distributed, we can use the multivariate normal probability function as well as the `ic.weight()` function of the R package ic.infer (Grömping, 2010) to compute the weights. These calculations are also implemented in the R package restriktor (Vanbrabant, 2020). Once we have computed the weights, the p-values of the observed $\bar{\chi}^2$-value ($\bar{\chi}^2_{obs}$) and of the observed $\bar{F}$-value ($\bar{F}_{obs}$) are obtained as follows Silvapulle and Sen (2005, pp. 86 and 99):

$$\Pr(\bar{\chi}^2 \geq \bar{\chi}^2_{obs}) = \sum_{i=0}^{q} w_i(H_0, H_a)\Pr[(h-q+i)\chi^2_{h-q+i} \geq \bar{\chi}^2_{obs}], \tag{39}$$

$$\Pr(\bar{F} \geq \bar{F}_{obs}) = \sum_{i=0}^{q} w_i(H_0, H_a)\Pr[(h-q+i)F_{h-q+i,n-p} \geq \bar{F}_{obs}]. \tag{40}$$

TABLE 3 Type I error rates when using R_1 and applying the test statistics as outlined in the referenced books.

n	$LRT_{corr.}$	$LRT_{restr.}$	$Wald^{info}_{naive}$	$Wald^{info}_{corr.}$ $Wald^{VCOV}$ $D_{corr.}$	$Score^{U}_{corr.}$	$Score^{null-info}_{corr.}$	$Score^{null-info}_{restr.}$	$Score^{info}_{corr.}$ $Score^{VCOV}$	$\bar{F}_{corr.}$ $\bar{F}_{restr.}$	$F_{corr.}$	$t_{one-s.}$	$t_{two-s.}$
10000	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.050	0.047	0.050
2000	0.054	0.054	0.054	0.054	0.057	0.054	0.054	0.054	0.054	0.060	0.054	0.060
1000	0.057	0.057	0.058	0.057	0.058	0.057	0.057	0.057	0.057	0.067	0.057	0.067
500	0.058	0.058	0.058	0.055	0.053	0.055	0.055	0.055	0.055	0.049	0.055	0.049
100	0.054	0.054	0.056	0.051	0.054	0.050	0.048	0.048	0.049	0.043	0.049	0.043
50	0.057	0.058	0.060	0.054	**0.066**	0.051	0.044	0.044	0.049	0.044	0.049	0.044
25	**0.074**	**0.074**	**0.092**	**0.066**	**0.089**	0.057	0.047	0.045	0.057	0.057	0.057	0.057
10	**0.125**	**0.112**	**0.186**	**0.098**	**0.169**	0.054	0.002	0.000	0.054	**0.061**	0.054	**0.061**

The test statistics are abbreviated as follows: $LRT_{corrected}$ as $LRT_{corr.}$, $LRT_{restriktor}$ as $LRT_{restr.}$, $Wald^{info}_{corrected}$ as $Wald^{info}_{corr.}$, $D_{corrected}$ as $D_{corr.}$, $Score^{U}_{corrected}$ as $Score^{U}_{corr.}$, $Score^{null-info}_{corrected}$ as $Score^{null-info}_{corr.}$, $Score^{null-info}_{restriktor}$ as $Score^{null-info}_{restr.}$, $Score^{info}_{corrected}$ as $Score^{info}_{corr.}$, $\bar{F}_{corrected}$ as $\bar{F}_{corr.}$, $F_{corrected}$ as $F_{corr.}$, $t_{one-sided}$ as $t_{one-s.}$ and $t_{two-sided}$ as $t_{two-s.}$. Bold values are above 0.06 and underlined values are below 0.04.

TABLE 4 Type I error rates when using R_2 and applying the test statistics as outlined in the referenced books.

n	$LRT_{corr.}$	$LRT_{restr.}$	$Wald^{info}_{naive}$	$Wald^{info}_{corr.}$ $Wald^{VCOV}$ $D_{corr.}$	$Score^{U}_{corr.}$	$Score^{null-info}_{corr.}$	$Score^{null-info}_{restr.}$	$Score^{info}_{corr.}$ $Score^{VCOV}$	$\bar{F}_{corr.}$ $\bar{F}_{restr.}$	$F_{corr.}$
10000	0.052	0.052	0.052	0.052	0.049	0.052	0.052	0.052	0.052	0.049
2000	0.048	0.048	0.050	0.048	0.052	0.048	0.048	0.048	0.048	0.046
1000	0.051	0.051	0.051	0.051	0.052	0.051	0.047	0.047	0.051	0.058
500	0.059	0.059	**0.062**	0.060	0.059	0.059	0.057	0.057	0.059	0.048
100	0.057	0.056	**0.070**	**0.061**	**0.078**	0.055	0.053	0.051	0.056	0.058
50	0.051	0.046	**0.090**	0.060	**0.099**	0.044	0.039	0.035	0.048	0.055
25	**0.068**	0.055	**0.135**	**0.083**	**0.119**	0.052	0.027	0.010	**0.064**	0.055
10	**0.069**	0.011	**0.416**	**0.163**	**0.334**	0.024	0.001	0.000	0.054	**0.061**

The test statistics are abbreviated as follows: $LRT_{corrected}$ as $LRT_{corr.}$, $LRT_{restriktor}$ as $LRT_{restr.}$, $Wald^{info}_{corrected}$ as $Wald^{info}_{corr.}$, $D_{corrected}$ as $D_{corr.}$, $Score^{U}_{corrected}$ as $Score^{U}_{corr.}$, $Score^{null-info}_{corrected}$ as $Score^{null-info}_{corr.}$, $Score^{null-info}_{restriktor}$ as $Score^{null-info}_{restr.}$, $Score^{info}_{corrected}$ as $Score^{info}_{corr.}$, $\bar{F}_{corrected}$ as $\bar{F}_{corr.}$, $F_{corrected}$ as $F_{corr.}$. Bold values are above 0.06 and underlined values are below 0.04.

It can be expected that the p-values are very similar, irrespective of whether they are calculated based on the $\bar{\chi}^2$- or $\bar{F}$-distribution, as long as sample sizes are large. However, for small sample sizes, the $\bar{F}$-distribution should yield more accurate results.

Simulation studies

We conducted several simulation studies to examine the impact of different conditions on the performance of the presented test statistics in terms of type I and type II error rates. We were interested in the effects of sample and effect sizes, the number of regression parameters considered in H_a as well as the distribution used for calculating the p-values. Our main motivation was to provide a reference framework for applied researchers who wish to test informative hypotheses, helping them to chose the optimal test statistic(s) in the present situation.

Design

We generated a design matrix $\boldsymbol{X}$, including data for five regression coefficients $\boldsymbol{\beta}' = (\beta_1\ \beta_2\ \beta_3\ \beta_4\ \beta_5)$ and considered effect sizes of $f^2 = 0.02$ (small), $f^2 = 0.10$ (medium) and $f^2 = 0.35$ (large) and sample sizes of $10, 25, 50, 100, 500, 1000, 2000,$ and 10000. For examining the type I error rate, we generated a random outcome Y, whereas for examining the type II error rate, we fixed all βs to 0.1 and generated y with a random error term that was specific for the effect size used. Since $f^2 = \frac{R^2}{1-R^2}$, where R^2 is the determination coefficient, we can calculate the error terms of y by plugging in the f^2-specific value of R^2 in

$$S_y^2 = [\boldsymbol{\beta}\ Cor(\boldsymbol{X})\ \boldsymbol{\beta}] \times \frac{1-R^2}{R^2}, \tag{41}$$

where $Cor(\boldsymbol{X})$ is the correlation matrix of the design matrix $\boldsymbol{X}$. The number of replications was 1000.

TABLE 5 Type I error rates when using R_1, $\hat{S}^2_{corrected}$ (or $\tilde{S}^2_{corrected}$, $\bar{S}^2_{corrected}$) and the $\bar{F}$-distribution for calculating the p-value.

n	$LRT_{corr.}$	$Wald^{info}_{corr.}$ $D_{corr.}$	$Score^{U}_{corr.}$
10000	0.047	0.047	0.047
2000	0.054	0.054	0.057
1000	0.057	0.057	0.058
500	0.055	0.055	0.053
100	0.052	0.049	0.052
50	0.055	0.049	**0.065**
25	**0.067**	0.057	**0.080**
10	**0.084**	0.054	**0.126**

The test statistics are abbreviated as follows: $LRT_{corrected}$ as $LRT_{corr.}$, $Wald^{info}_{corrected}$ as $Wald^{info}_{corr.}$, $D_{corrected}$ as $D_{corr.}$ and $Score^{U}_{corrected}$ as $Score^{U}_{corr.}$. Bold values are above 0.06 and underlined values are below 0.04.

TABLE 6 Type I error rates when using R_2, $\hat{S}^2_{corrected}$ (or $\tilde{S}^2_{corrected}$, $\bar{S}^2_{corrected}$) and the $\bar{F}$-distribution for calculating the p-value.

n	$LRT_{corr.}$	$Wald^{info}_{corr.}$ $D_{corr.}$	$Score^{U}_{corr.}$
10000	0.052	0.052	0.049
2000	0.048	0.048	0.051
1000	0.050	0.051	0.051
500	0.059	0.059	0.058
100	0.055	0.056	**0.072**
50	0.043	0.048	**0.085**
25	0.042	**0.064**	**0.097**
10	0.006	0.054	**0.161**

The test statistics are abbreviated as follows: $LRT_{corrected}$ as $LRT_{corr.}$, $Wald^{info}_{corrected}$ as $Wald^{info}_{corr.}$, $D_{corrected}$ as $D_{corr.}$ and $Score^{U}_{corrected}$ as $Score^{U}_{corr.}$. Bold values are above 0.06 and underlined values are below 0.04.

We considered two different kinds of $\boldsymbol{R}$ matrices, where the first one was defined as follows:

$$\boldsymbol{R}_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{42}$$

This represents the hypothesis that only β_1 is greater than zero: $H_a : \beta_1 > 0$. The second $\boldsymbol{R}$ matrix was defined as:

$$\boldsymbol{R}_2 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{43}$$

stating that at least one of the regression coefficients, except the intercept, are greater than zero: $H_a : \beta_1 > 0 \vee \beta_2 > 0 \vee \beta_3 > 0 \vee \beta_4 > 0 \vee \beta_5 > 0$.

To compute the test statistics, we used $\hat{S}^2_{naive}$ and $\hat{S}^2_{corrected}$ as well as $\tilde{S}^2_{naive}$, $\tilde{S}^2_{corrected}$, $\bar{S}^2_{naive}$ and $\bar{S}^2_{corrected}$ and to compute the p-values, we used the $\bar{\chi}^2$- as well as the $\bar{F}$-distribution. In addition to the manual calculation of the test statistics, we also included the test statistics as reported by the R package restriktor.

Type I results

Test statistics were first applied the way they are presented in the referenced literature. That is, $Wald^{info}_{naive}$ makes use of $\hat{S}^2_{naive}$, whereas all other test statistics make use of $\hat{S}^2_{corrected}$ (or $\tilde{S}^2_{corrected}$, $\bar{S}^2_{corrected}$). For calculating the p-values, the $\bar{\chi}^2$-distribution is used for $LRT_{corrected}$, $Wald^{info}_{naive}$, $Wald^{info}_{corrected}$, $Wald^{VCOV}$, $D_{corrected}$, $Score^{U}_{corrected}$, $Score^{null-info}_{corrected}$, $Score^{info}_{corrected}$ and $Score^{VCOV}$. The $\bar{F}$-distribution is used for calculating the p-values for the $\bar{F}$-statistic, the F-distribution is used for calculating the p-values for the F-statistic and the t-distribution is used for calculating the p-values for the t-statistic. Note that restriktor always uses $\hat{S}^2_{corrected}$ (or $\tilde{S}^2_{corrected}$, $\bar{S}^2_{corrected}$) for all available test statistics and always calculates the p-value based on the $\bar{F}$-distribution. Tables 3, 4 show the results.

We can observe that when using $\boldsymbol{R}_1$ (see Table 3), that is when testing a hypothesis concerning only one regression parameter, type I error rates are identical between F and $t_{two-sided}$ as well as between $\bar{F}$ and $t_{one-sided}$, showing the link between classical null hypothesis testing and informative hypothesis testing. When using $\boldsymbol{R}_2$ (see Table 4), that is when testing a hypothesis concerning multiple regression parameters, problems with type I error rates seem to occur earlier as compared to when using $\boldsymbol{R}_1$. More specifically, problematic type I error rates occur as early as with $n = 500$ or $n = 100$ when using $\boldsymbol{R}_2$, but only start occurring with $n = 50$ or $n = 25$ when using $\boldsymbol{R}_1$. Apart from that, $Score^{U}_{corrected}$ and $Wald^{info}_{naive}$ show the highest type I error rates for both $\boldsymbol{R}$ matrices, whereas $\bar{F}$ and $\bar{F}_{restriktor}$ show the most appropriate type I error rates for both $\boldsymbol{R}$ matrices. This is because the $\bar{F}$-distribution is more precise in small sample sizes as compared to the $\bar{\chi}^2$-distribution.

When using the $\bar{F}$-distribution instead of the $\bar{\chi}^2$-distribution when calculating the p-value for all test statistics, type I error rates are closer to the nominal level when sample sizes get smaller. This can be seen in Tables 5, 6 where a selection of test statistics are shown.

Furthermore, it can be observed that when using $\boldsymbol{R}_1$, type I error rates increase when using $LRT_{corrected}$ and $Score^{U}_{corrected}$ and $n = 10$ in contrast to $n = 25$. The same can only be observed for $Score^{U}_{corrected}$ when using $\boldsymbol{R}_2$, but not for $LRT_{corrected}$, where the type I error rate decreases quite substantially instead.

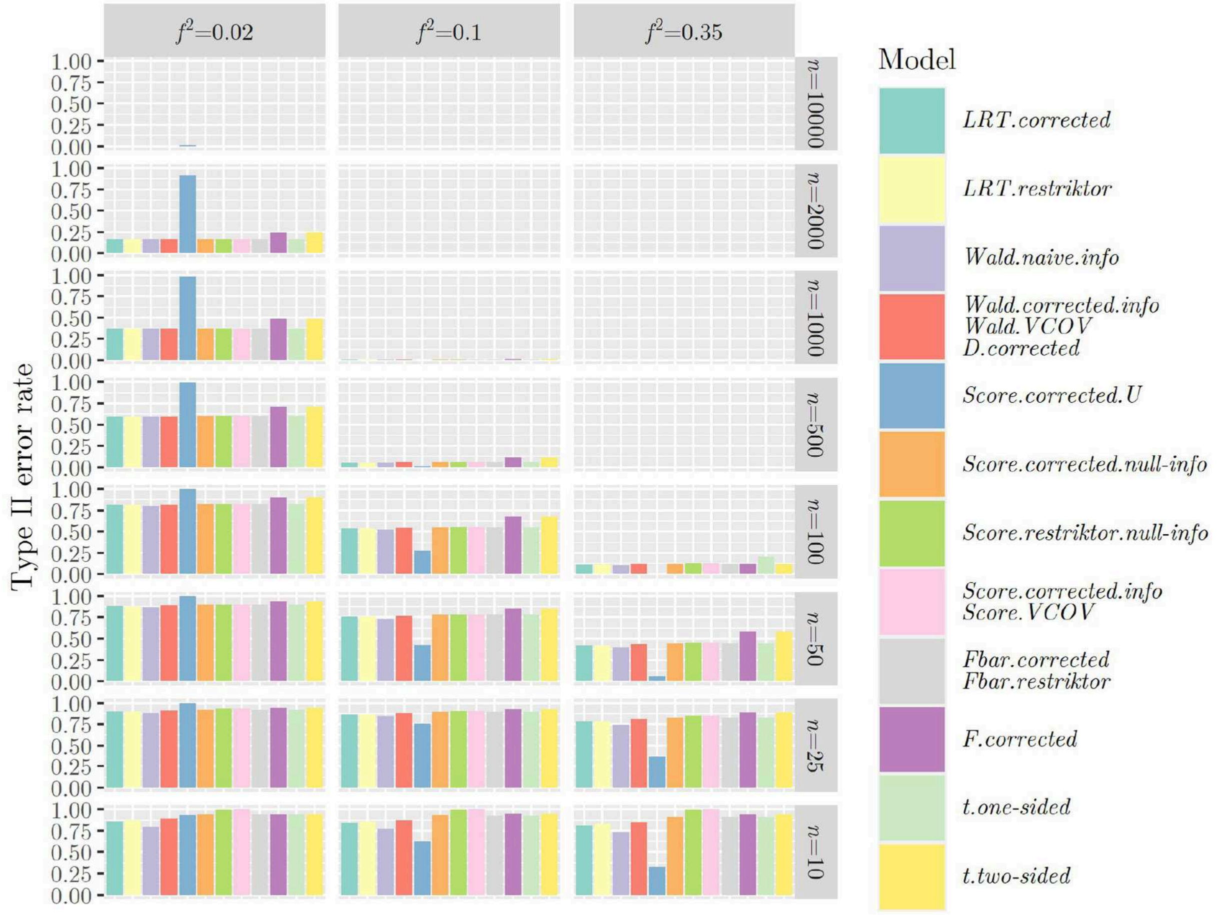

FIGURE 1
Type II error rates when using $\boldsymbol{R}_1$ and applying the test statistics as outlined in the referenced books.

More results can be found in "Data Sheet 4" in the Supplementary materials.

Type II results

Figures 1, 2 show the type II error rates when applying the test statistics as in the referenced books.

Once more, we can observe that when using $\boldsymbol{R}_1$ (see Figure 1), that is when testing a hypothesis concerning only one regression parameter, type II error rates are identical between F and $t_{two-sided}$ as well as between $\bar{F}$ and $t_{one-sided}$, showing the link between classical null hypothesis testing and informative hypothesis testing. When using $\boldsymbol{R}_2$ (see Figure 2), that is when testing a hypothesis concerning multiple regression parameters, problems with type II error rates seem to occur later (in terms of sample size) as compared to when using $\boldsymbol{R}_1$. This was the other way around regarding the type I error rate and it demonstrates the nature of the relationship between type I and type II error rates: If one goes down, the other one goes up and vice versa.

The same mechanism can be observed when using the $\bar{F}$-distribution instead of the $\bar{\chi}^2$-distribution when calculating the p-value for all test statistics (Figures 3, 4): Type II error rates are increased in small sample sizes, since type I error rates had improved, that is, decreased. Again, further results can be found in "Data Sheet 5" in the Supplementary materials.

Discussion

In this paper, we gave an overview of a large number of different informative test statistics, including their different versions. Furthermore, we clarified how those test statistics perform in terms of type I and type II error rates under different conditions by means of simulation studies in the context of linear regression. We considered varying sample and effect sizes as well as two different constraint matrices, where one

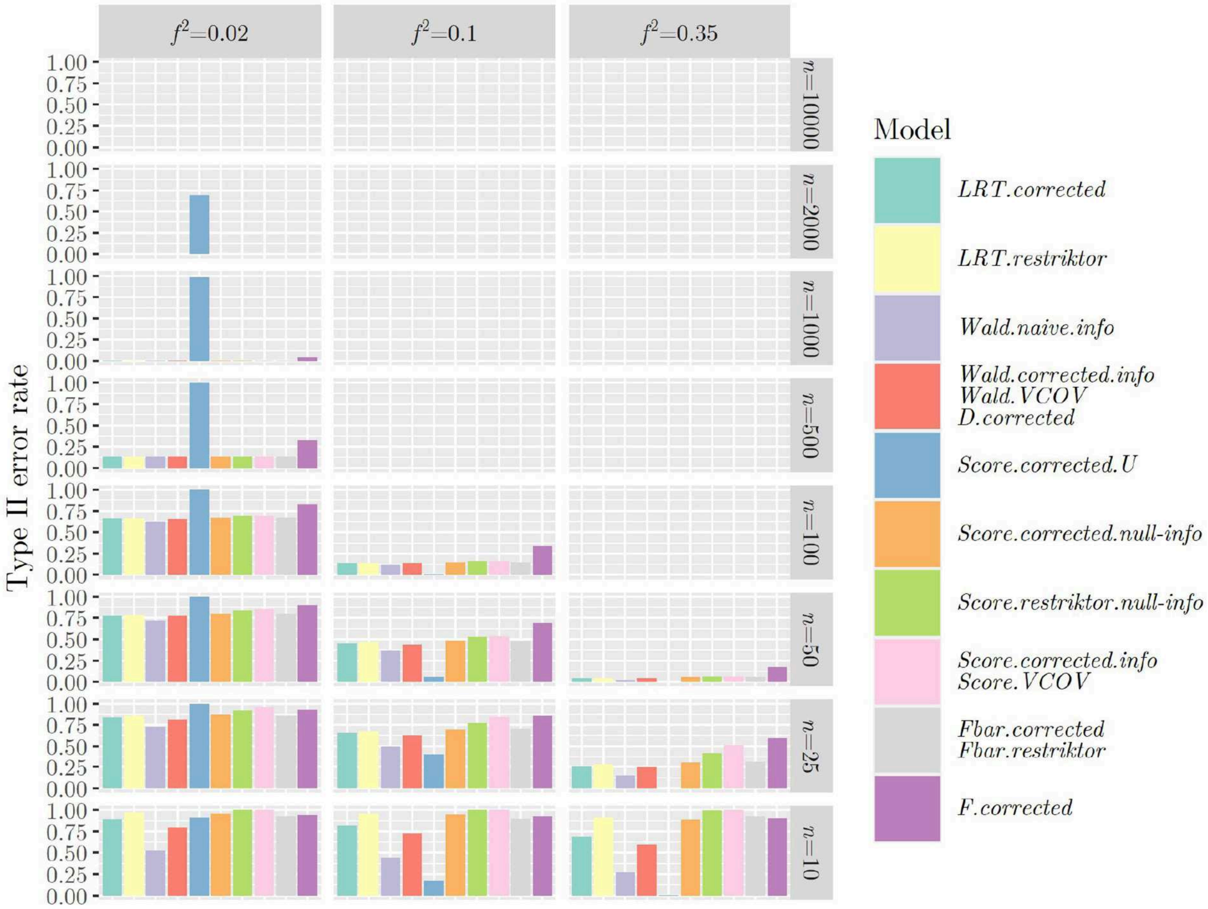

FIGURE 2
Type II error rates when using $\boldsymbol{R}_2$ and applying the test statistics as outlined in the referenced books.

specified a hypothesis about one parameter and the other one specified a hypothesis about multiple parameters. Moreover, we considered the naive and corrected mean squared errors of the unconstrained, inequality and equality constrained models as part of the test statistics as well as the $\bar{\chi}^2$- and $\bar{F}$-distribution to calculate the p-values.

Based on our findings, the following recommendations can be made. Considering the time it takes to compute the informative test statistics, both the Wald and the $\bar{F}$-test versions are favorable, since they only need fitting of the inequality constrained model to obtain $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{I}}_1$. Even if we do not use $\tilde{\boldsymbol{I}}_1$ but use $\hat{\boldsymbol{I}}_1$ instead, the increase in time is small in the context of linear regression. The Score test and the LRT versions are less favorable, since they require fitting both the inequality constrained as well as the equality constrained model to obtain $\tilde{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\beta}}$, as well as the respective unit information matrices or log-likelihoods.

The D-statistic versions only require fitting the unconstrained model to obtain $\hat{\boldsymbol{\beta}}$. However, we then additionally need to compute the two functions $d(\bar{\boldsymbol{\beta}})$ and $d(\tilde{\boldsymbol{\beta}})$, which is as time-consuming as fitting the inequality constrained model. Thus, there is no advantage of using the D-statistic versions over the Wald and the $\bar{F}$-test versions in the context of linear regression. However, if the regression model was non-linear, computing the two functions would be significantly less computationally expensive than fitting the inequality constrained model.

Moreover, we recommend using the corrected mean squared error versions in the test statistics as well as using the $\bar{F}$-distribution for calculating the p-values, if sample sizes are small. This seems to keep type I error rates closer to the nominal level compared to using the naive mean squared error versions and using the $\bar{\chi}^2$-distribution for calculating the p-value. An additional interesting finding was that the relationship between LRT, Wald and Score test values that has been found in the unconstrained context also holds in the constrained context. That is, Wald test values are always slightly larger than LRT

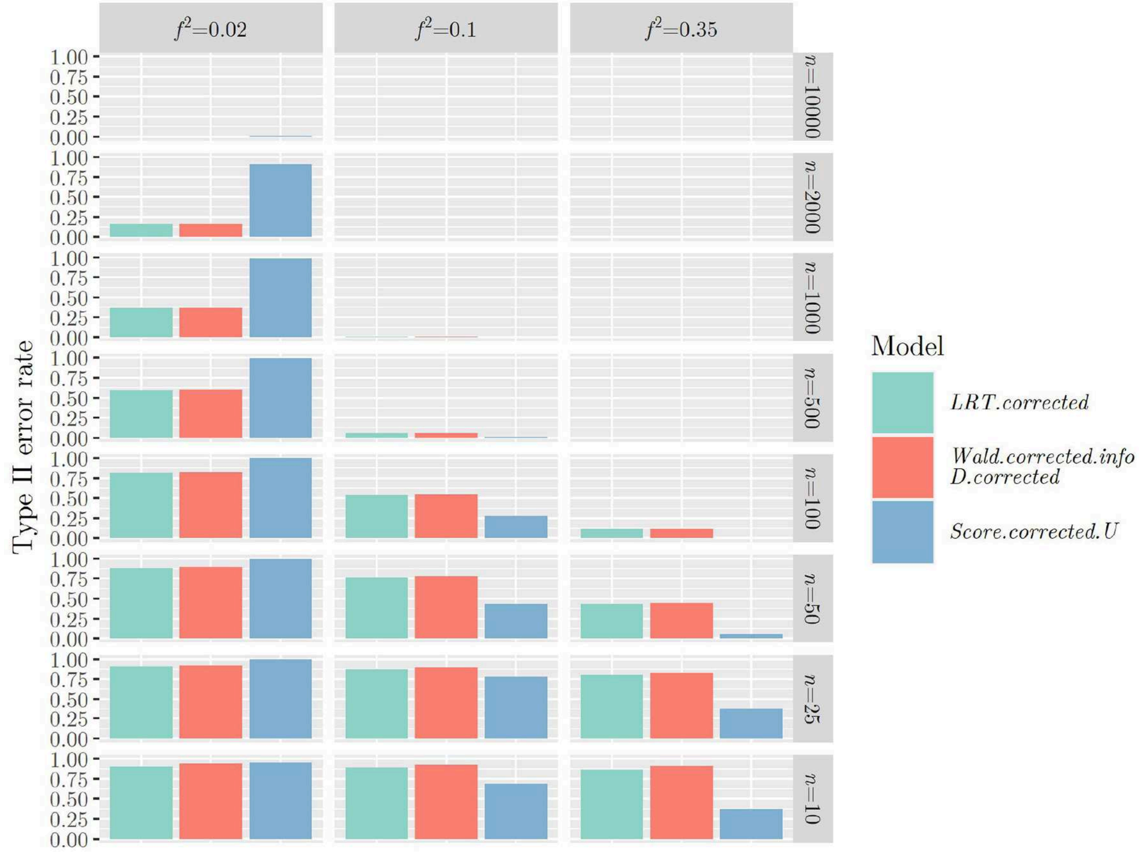

FIGURE 3
Type II error rates when using $\boldsymbol{R}_1$, $\hat{S}^2_{corrected}$ (or $\tilde{S}^2_{corrected}$, $\bar{S}^2_{corrected}$) and the $\bar{F}$-distribution for calculating the p-value.

values, which in turn are always slightly larger than Score test values.

The limitations of our simulation studies include the following aspects. We treated all variables as manifest, even though variables of interest in the social and behavioral sciences are often latent in nature. Furthermore, we solely generated normal data despite the fact that violations against the normality assumption occur regularly. Moreover, we used orthogonal predictors without interactions albeit this is rarely the case in the social and behavioral sciences. And lastly, we only included the regular versions of the standard errors and the variance-covariance matrix. Future research should thus repeat the simulation studies in the context of Structural Equation Modeling (SEM) to take into account latent variables. Furthermore, the impact of non-normal data as well as correlated predictors with interactions and using the robust versions of the standard errors and the variance-covariance matrix should be examined. It may be that under these conditions, type I and type II error rates deviate from the results presented in this paper. Moreover, the properties of informative test statistics, especially concerning the D-statistic, should also be investigated in the context of non-linear models.

Finally, research in the social and behavioral sciences is often not only interested in inference concerning regression coefficients, but also regarding effects of interest. These effects may be average or conditional treatment effects, which are defined as a linear or non-linear combination of regression coefficients. The EffectLiteR approach (Mayer et al., 2016) provides a framework and R package for the estimation of average and conditional effects of a discrete treatment variable on a continuous outcome variable, conditioning on categorical and continuous covariates. Keck et al. (2021) already demonstrated how to integrate informative hypothesis testing into the EffectLiteR framework in the context of linear regression. The present paper provides interested readers who want to apply informative hypothesis testing concerning regression coefficients or effects of interest with practical information regarding test statistics as well as type I and type II error rates.

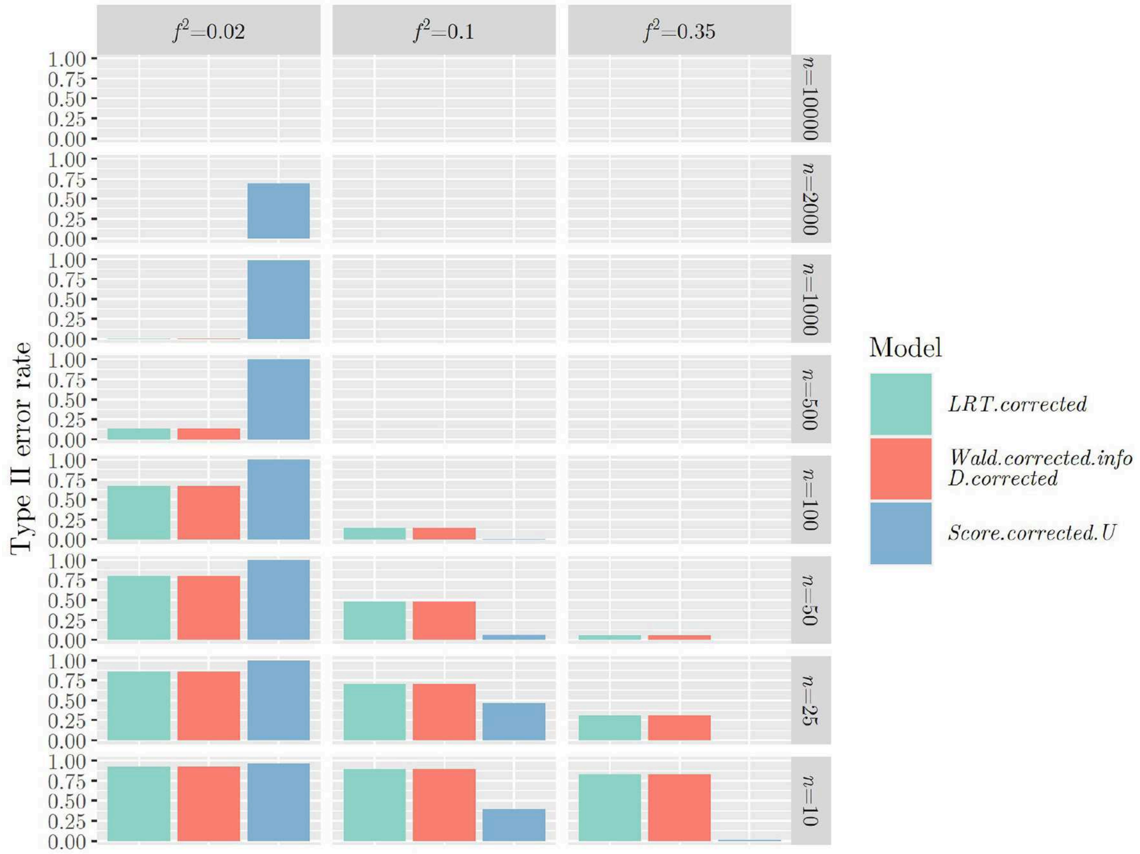

FIGURE 4
Type II error rates when using $\boldsymbol{R}_2$, $\hat{S}^2_{corrected}$ (or $\tilde{S}^2_{corrected}$, $\bar{S}^2_{corrected}$) and the $\bar{F}$-distribution for calculating the p-value.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

CK and YR contributed to conception and design of the study. CK performed the statistical analysis and wrote the first draft of the manuscript. CK, YR, and AM wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work has been supported by the Research Foundation Flanders (FWO, Grant No. G002819N to YR and AM).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.899165/full#supplementary-material

References

Allen, M. P. (1997). *Understanding Regression Analysis*. Boston, MA: Springer.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions*. New York, NY: Wiley.

Buse, A. (1982). The likelihood ratio, Wald and Lagrange multiplier tests: An expository note. *Am. Stat*. 36, 153–157. doi: 10.1080/00031305.1982.10482817

Goldfarb, D., and Idnani, A. (1982). *Dual and Primal-Dual Methods for Solving Strictly Convex Quadratic Programs*, Berlin: Springer.

Goldfarb, D., and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Math. Program*. 27, 1–33. doi: 10.1007/BF02591962

Grömping, U. (2010). Inference with linear equality and inequality constraints using R: The package ic.infer. *J. Stat. Softw*. 33, 1–31. doi: 10.18637/jss.v033.i10

Gu, X., Hoijtink, H., Mulder, J., Van Lissa, C. J., Van Zundert, C., Jones, J., et al. (2020). *Bain: Bayes factors for informative hypotheses*. R package version 0.2.4.

Hoijtink, H. (2012). *Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists*. Boca Raton, FL: Chapman & Hall/CRC.

Hoijtink, H., Klugkist, I., and Boelen, P. A. (2008). *Bayesian Evaluation of Informative Hypotheses*. New York, NY: Springer.

Huber, P. (1967). "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, Statistics* (Berkeley, CA: University of California Press), 221–233.

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

Kuiper, R. M., and Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory approaches. *Psychol. Methods* 15, 69–86. doi: 10.1037/a0018720

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

Mayer, A., Dietzfelbinger, L., Rosseel, Y., and Steyer, R. (2016). The EffectLiteR approach for analyzing average and conditional effects. *Multivariate Behav. Res*. 5, 374–391. doi: 10.1080/00273171.2016.1151334

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R foundation for statistical computing.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. New York, NY: Wiley.

SAS Institute Inc. (2015). *Sas/stat ® 14.1 User'S Guide*. Cary, NC: SAS Institute Inc..

Seber, G. A. F., and Lee, A. J. (2012). *Linear Regression Analysis*. Hoboken, NJ: Wiley.

Silvapulle, M. J., and Sen, P. K. (2005). *Constrained Statistical Inference: Order, Inequality, and Shape Restrictions*. Hoboken, NJ: Wiley.

Silvapulle, M. J., and Silvapulle, P. (1995). A Score test against one-sided alternatives. *J. Am. Stat. Assoc*. 90, 342–349. doi: 10.1080/01621459.1995.10476518

Turlach, B. A., and Weingessel, A. (2019). *Quadprog: Functions to solve quadratic programming problems*. R package version 1.5–8.

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

Vanbrabant, L., Van de Schoot, R., and Rosseel, Y. (2015). Constrained statistical inference: Sample-size tables for ANOVA and regression. *Front. Psychol*. 5, 1565. doi: 10.3389/fpsyg.2014.01565

White, H. (1980). A heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity. *Econometrica* 48, 817–838. doi: 10.2307/1912934

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *J. Stat. Softw*. 16, 1–16. doi: 10.18637/jss.v016.i09