ORIGINAL RESEARCH PAPER

Equivalence testing for standardized effect sizes in linear regression

Harlan Campbell

University of British Columbia, Department of Statistics

Vancouver, British Columbia, Canada

ARTICLE HISTORY

Compiled April 1, 2020

ABSTRACT

In this paper, we introduce equivalence testing procedures for standardized effect sizes in a linear regression. We show how to define valid hypotheses and calculate p-values for these tests. Such tests are necessary to confirm the lack of a meaningful association between an outcome and predictors. A simulation study is conducted to examine type I error rates and statistical power. We also compare using equivalence testing as part of a frequentist testing scheme with an alternative Bayesian testing approach. The results indicate that the proposed equivalence test is a potentially useful tool for "testing the null."

KEYWORDS

equivalence testing, non-inferiority testing, linear regression, standardized effect sizes

CONTACT Harlan Campbell. Email: harlan.campbell@stat.ubc.ca

1. Introduction

Let θ be the parameter of interest. An equivalence test reverses the question that is asked in a null hypothesis significance test (NHST). Instead of asking whether we can reject the null hypothesis of no effect, e.g., $H_0 : \theta = 0$, an equivalence test examines whether the magnitude of θ is at all meaningful: *Can we reject the possibility that θ is as large or larger than our smallest effect size of interest, Δ?*

The null hypothesis for an equivalence test is defined as $H_0 : \theta \notin [-\Delta, \Delta]$. In other words, *equivalence* implies that θ is small enough that any non-zero effect would be at most equal to Δ. The interval $[-\Delta, \Delta]$ is known as the equivalence margin and represents a range of values for which θ is considered negligible. The value of Δ is sometimes known as the "smallest effect size of interest" (Lakens 2017). Note that the equivalence margin need not necessarily be symmetric, i.e., we could have $H_0 : \theta \notin [\Delta_1, \Delta_2]$, where $\Delta_1 \neq -\Delta_2$.

In order for one to conduct an equivalence test, one must ideally define the equivalence margin prior to observing any data. This can often be challenging; see Campbell & Gustafson (2018*b*). Indeed, for many researchers, defining and justifying the equivalence margin is one of the "most difficult issues" (Hung et al. 2005). If the margin is too large, then any claim of equivalence will be considered meaningless. If the margin is somehow too small, then the probability of declaring equivalence will be substantially reduced; see Wiens (2002). While the margin is ideally based on some objective criteria, these can be difficult to justify, and there is generally no clear consensus among stakeholders (Keefe et al. 2013).

To make matters worse, in many scenarios (and very often in the social sciences), the parameters of interest are measured on different and completely arbitrary scales. Without interpretable units of measurement, the task of defining and justifying an appropriate equivalence margin is even more challenging. How can one determine the "smallest effect size of interest" in units that have no particular meaning?

Researchers working with variables measured on arbitrary scales will often report standardized effect sizes to aid with interpretation. For example, for linear regression

analyses, reporting standardized regression coefficients is quite common (West et al. 2007, Bring 1994) and can be achieved by normalizing the outcome variable and all predictor variables before fitting the regression. In the psychometric literature, standardized regression coefficients are known as Beta-coefficients while the conventional unstandardized regression coefficients are called B-coefficients.

There are many reasons besides the need to overcome arbitrary scales for reporting standardized effects. For example, Nieminen et al. (2013) argue that standardized effect sizes might be helpful for the synthesis of epidemiological studies. Standardization can also help with interpretation of a regression analysis: subtracting the mean can improve the interpretation of main effects in the presence of interactions, and dividing by the standard deviation will ensure that all predictors are on a common scale.

Unfortunately, equivalence testing of standardized effects is not so straightforward. In this paper, we introduce equivalence testing procedures for standardized effect sizes in a linear regression. We show how to define valid hypotheses and calculate p-values for these tests. To the best of our knowledge, such tests have not been detailed elsewhere. In Section 3, we conduct a small simulation study to better understand the test's operating characteristics and in Section 4 we consider how a frequentist testing scheme compares to a Bayesian testing approach based on Bayes Factors. We demonstrate how the testing methods can be applied in practice with the analysis of an example dataset in Section 5, and conclude in Section 6.

2. Equivalence testing for standardized $\boldsymbol{\beta}$ coefficient parameter

Let us begin by defining some required notation. Let:

- N, be the number of observations in the observed data;
- K, be the number of explanatory variables in the linear regression model;
- y_i, be the observed value of random variable Y for the ith subject;
- X, be the $N \times K+1$ fixed covariate matrix (with a column of 1s for the intercept;

we use the notation $X_{i,\cdot}$ to refer to all $K+1$ values corresponding to the ith subject; and X_k to refer to the k-th covariate);

- $R^2_{Y\cdot X}$ is the coefficient of determination from the linear regression where Y is the dependent variable predicted from X;
- $R^2_{X_k \cdot X_{-k}}$ is the coefficient of determination from the linear regression model where X_k is the dependent variable predicted from the remaining $K-1$ regressors; and
- $R^2_{Y\cdot X_{-k}}$ is the coefficient of determination from the linear regression where Y is the dependent variable predicted from all but the k-th covariate.

We operate under the standard linear regression assumption that observations in the data are independent and normally distributed with:

$$Y_i \sim \ Normal(X_{i,\cdot}^T \beta, \sigma^2), \qquad \forall\, i = 1, ..., N; \tag{1}$$

where β is a parameter vector of regression coefficients, and σ^2 is the population variance. Least squares estimates for the linear regression model are denoted with $\hat{\sigma}$, and $\widehat{\beta}_k$, for k in 0,...,K; see equations (21) and (22) in the Appendix for details. Let us first consider a standard NHST for the k-th covariate, X_k:

$H_0 : \beta_k = 0$, vs.
$H_1 : \beta_k \neq 0$.

Typically one conducts one of two different (yet mathematically identical) tests. Most commonly a t-test is done to calculate a p-value as follows:

$$p\text{-value}_k = 2 \cdot p_t \left(\frac{|\widehat{\beta_k}|}{SE(\widehat{\beta_k})}, N - K - 1, 0 \right), \text{for } k \text{ in } 0,...,K, \tag{2}$$

where we use $p_t(\quad \cdot \quad ; df, ncp)$ to denote the cumulative distribution function (cdf) of the non-central t-distribution with df degrees of freedom and non-centrality parameter

ncp; and where: $\widehat{SE(\beta_k)} = \hat{\sigma}\sqrt{[(X^TX)^{-1}]_{kk}}$. Note that when $ncp = 0$, the non-central t-distribution is equivalent to the central t-distribution.

Alternatively, we can conduct an F-test and, for k in 1,...,K, we will obtain the very same p-value with:

$$p\text{-value}_k = p_F\left((N-K-1)\frac{\text{diff}R_k^2}{1-R_{Y\cdot X}^2}, 1, N-K-1, 0\right), \tag{3}$$

where $p_f(\cdot \quad ; df_1, df_2, ncp)$ is the cdf of the non-central F-distribution with df_1 and df_2 degrees of freedom, and non-centrality parameter, ncp (note that $ncp = 0$ corresponds to the *central* F-distribution); and where: $\text{diff}R_k^2 = R_{Y\cdot X}^2 - R_{Y\cdot X_{-k}}^2$. Regardless of whether the t-test or the F-test is employed, if $p\text{-value}_k < \alpha$, we reject the null hypothesis of $H_0 : \beta_k = 0$ against the alternative $H_0 : \beta_k \neq 0$.

An equivalence test asks a different question: *Can we reject the possibility that β_k is as large or larger than our smallest effect size of interest?* Formally, the null and alternative hypotheses for the equivalence test are:

$$H_0 : \beta_k \leq \Delta_1 \quad \text{or:} \quad \beta_k \geq \Delta_2,$$
$$H_1 : \beta_k > \Delta_1 \quad \text{and:} \quad \beta_k < \Delta_2,$$

where the equivalence margin is $[\Delta_1, \Delta_2]$ and defines the range of values considered negligible. Often, one has a symmetric margin with $\Delta_1 = -\Delta_2$, but this is not necessarily the case.

Recall that there is a one-to-one correspondence between an equivalence test and a confidence interval (CI); see Wellek (2017). For example, we will reject the above H_0 at a $\alpha = 0.05$ significance level whenever the 90% (= 1 - 2α) CI for β_k fits entirely within $[\Delta_1, \Delta_2]$. As such, an equivalence test can be constructed by simply inverting a confidence interval. To obtain a p-value for the equivalence test above, one conducts two one-sided t-tests (TOST) and calculates the two following p-values:

$$p_k^{[1]} = p_t\left(\frac{\widehat{\beta_k} - \Delta_1}{SE(\widehat{\beta_k})}, N-K-1, 0\right); \quad \text{and} \quad p_k^{[2]} = p_t\left(\frac{\Delta_2 - \widehat{\beta_k}}{SE(\widehat{\beta_k})}, N-K-1, 0\right), \tag{4}$$

for k in 0,...,K. In order to reject this equivalence test null hypothesis, both p-values, $p_k^{[1]}$ and $p_k^{[2]}$, must be less than α. As such, a single overall p-value for the equivalence test is calculated as: $p\text{-value}_k = max(p_k^{[1]}, p_k^{[2]})$.

2.1. An equivalence test for standardized regression coefficients

In many scenarios (and very often in the social sciences), the variables considered are measured on different and completely arbitrary scales. Without interpretable units of measurement, defining (and justifying) the equivalence margin can be rather challenging. How can one determine the "smallest effect size of interest" in units that have no particular meaning? In these scenarios, rather than conduct a standard equivalence test (with equation (4)), it may be preferable to work with standardized regression coefficients.

Note that it has been previously suggested that the "bounds [of an equivalence margin] can be defined in raw scores or in a standardized difference" (Lakens 2017). For example, in a two-sample test for the difference in means, μ_d, one could supposedly define equivalence to be a difference within half standard deviation, i.e., define $\Delta = 0.5 \times \hat{\sigma}$, where $\hat{\sigma}$ is the pooled standard deviation estimated from the data. This is problematic.

Recall that hypotheses are statements about parameters and not about the observed data. Hypotheses must be nonrandom statements. As such, Δ being defined as a function of the data invalidates the proposed hypotheses; e.g., $H_0 : |\mu_d| > 0.5 \times \hat{\sigma}$, is not a valid hypothesis. The correct method is to define the parameter of interest to be the standardized effect size (e.g., μ_d/σ is the parameter of interest) and then define the margin on the standardized scale; e.g., $H_0 : |\mu_d| > 0.5 \times \sigma$. While in practice, the difference between these two H_0 statements may seem minuscule, it should neverthe-

less be acknowledged. Going forward, we define the parameter of interest to be $\mathcal{B}_k$, the standardized regression coefficient.

The process of standardizing a regression coefficient can proceed by multiplying the unstandardized regression coefficient, β_k, by the ratio of the standard deviation of X_k to the standard deviation of Y. The population standardized regression coefficient parameter, $\mathcal{B}_k$, for k in 1,...,K, is defined as:

$$\mathcal{B}_k = \beta_k \frac{s_k}{\sigma_Y}, \tag{5}$$

and can be estimated by:

$$\widehat{\mathcal{B}_k} = \widehat{\beta_k} \frac{s_k}{\widehat{\sigma_Y}}, \tag{6}$$

where s_k and $\widehat{\sigma_Y}$ are the standard deviations of X_k and y, respectively. An equivalence test for $\mathcal{B}_k$ can be defined by the following null and alternative hypotheses:

$$H_0 : \mathcal{B}_k \leq \Delta_1 \quad \text{or:} \quad \mathcal{B}_k \geq \Delta_2,$$
$$H_1 : \mathcal{B}_k > \Delta_1 \quad \text{and:} \quad \mathcal{B}_k < \Delta_2.$$

We make use of noncentrality interval estimation (NCIE); see Smithson (2001). By inverting a confidence interval for $\mathcal{B}_k$ (see Kelley et al. (2007) for details), we can obtain the following, for k in 1,...,K:

$$p_k^{[1]} = p_t \left(\frac{\widehat{\mathcal{B}_k}}{SE(\widehat{\mathcal{B}_k})}; df = N - K - 1, ncp = \Delta_1 \frac{\sqrt{N\left(1 - R^2_{X_k \cdot X_{-k}}\right)}}{\sqrt{1 - ((1 - R^2_{X_k \cdot X_{-k}})\Delta_1^2 + R^2_{Y \cdot X_{-k}})}} \right), \text{ and:} \tag{7}$$

$$p_k^{[2]} = p_t \left(\frac{-\widehat{\mathcal{B}_k}}{SE(\widehat{\mathcal{B}_k})}; df = N - K - 1, ncp = -\Delta_2 \frac{\sqrt{N\left(1 - R^2_{X_k \cdot X_{-k}}\right)}}{\sqrt{1 - ((1 - R^2_{X_k \cdot X_{-k}})\Delta_2^2 + R^2_{Y \cdot X_{-k}})}} \right),$$

where:

$$SE(\widehat{\mathcal{B}_k}) = \sqrt{\frac{(1 - R^2_{Y \cdot X})}{(1 - R^2_{X_k \cdot X_{-k}})(N - K - 1)}}. \tag{8}$$

In order to reject this equivalence test null hypothesis, $p\text{-value}_k = max(p_k^{[1]}, p_k^{[2]})$ must be less than α.

2.2. An equivalence test for the increase in R^2

The increase in the squared multiple correlation coefficient associated with adding a variable in a linear regression model, $\text{diff}R_k^2$, is a commonly used measure for establishing the importance the added variable (Dudgeon 2017). As stated earlier, $\text{diff}R_k^2 = R^2_{Y \cdot X} - R^2_{Y \cdot X_{-k}}$. In a linear regression model, the $R^2_{Y \cdot X}$ is equal to the square of the Pearson correlation coefficient between the observed and predicted outcomes (Nagelkerke et al. 1991, Zou et al. 2003). Despite the $R^2_{Y \cdot X}$ statistic's ubiquitous use, its corresponding population parameter, which we will denote as $P^2_{Y \cdot X}$, as in Cramer (1987), is rarely discussed. When considered, it is sometimes is known as the "parent multiple correlation coefficient" (Barten 1962) or the "population proportion of variance accounted for" (Kelley et al. 2007).

Campbell & Lakens (2020) introduce a non-inferiority test (a one-sided equivalence test) to test the null hypotheses:

$H_0 : 1 > P^2_{Y \cdot X} \geq \Delta$, vs.
$H_1 : 0 \leq P^2_{Y \cdot X} < \Delta$.

For the increase in variance explained by the k-th covariate, $\text{diff}R^2_k$, when $K = 1$ (i.e., for simple linear regression), we have that: $\text{diff}R^2_k = R^2_{Y \cdot X} = \mathcal{B}^2_k$. When $K > 1$, things are not so simple. In general, we have that the $\text{diff}R^2_k$ measure is a re-calibration of $\widehat{\mathcal{B}_k}$ (see Dudgeon (2017)), such that:

$$\text{diff}R^2_k = \widehat{\mathcal{B}_k}^2(1 - R^2_{X_k \cdot X_{-k}}). \tag{9}$$

Similarly, we have that for the corresponding population parameter: $\text{diff}P^2_k = {\mathcal{B}_k}^2(1 - P^2_{X_k \cdot X_{-k}})$. It may be preferable to consider an effect size (and what can be considered a "negligible difference") in terms of $\text{diff}P^2_k$ instead of in terms of $\mathcal{B}_k$. If this is the case, one can conduct a non-inferiority test, for k in 1,...,K, with the following hypotheses:

$H_0 : 1 > \text{diff}P^2_k \geq \Delta$, vs.
$H_1 : 0 \leq \text{diff}P^2_k < \Delta$.

The p-value for this non-inferiority test is obtained by replacing $\widehat{\mathcal{B}}_k$ with $\sqrt{\text{diff}R^2_k/(1 - R^2_{X_k \cdot X_{-k}})}$ and can be calculated, for fixed regressors as follows:

$$p\text{−value}_k = 1 - p_t\left(\frac{\sqrt{(N - K - 1)\text{diff}R^2_k}}{\sqrt{(1 - R^2_{Y \cdot X})}}; df = N - K - 1, ncp = \frac{\sqrt{N\Delta}}{\sqrt{\left(1 - \Delta + R^2_{X_k \cdot X_{-k}}\right)}}\right). \tag{10}$$

3. Simulation Study 1

We conducted a simple simulation study in order to better understand the operating characteristics of the proposed equivalence test for standardized regression coefficients and to confirm that the test has correct type 1 error rates. The equivalence test in the simulation study involves a symmetric equivalence margin, $[-\Delta, \Delta]$, and the following hypotheses:

$H_0 : |\mathcal{B}_1| \geq \Delta$, vs.

$H_1 : |\mathcal{B}_1| < \Delta$.

The design of our simulation study is similar to the simulation study of Campbell & Lakens (2020). We simulated data for each of 24 scenarios ($4 \times 2 \times 3$), one for each combination of the following parameters:

- one of four sample sizes: $N = 180$, $N = 540$, $N = 1,000$, or $N = 3,500$;
- one of two orthogonal, balanced designs with $K = 2$, or $K = 4$ binary covariates; with $\beta = (-0.20, 0.10, 0.20)$ or $\beta = (0.20, 0.10, 0.14, -0.10, -0.10)$; and
- one of three variances: $\sigma^2 = 0.05$, $\sigma^2 = 0.15$, or $\sigma^2 = 0.50$.

Depending on the specific value of σ^2, the true population standardized coefficient, $\mathcal{B}_1$, for these data is either: 0.070, 0.124, or 0.200. In order to examine situations with $\mathcal{B}_1 = 0$, we also simulated data from an additional 8 scenarios where the regression coefficients were fixed to be $\beta = (-0.20, 0.00, 0.20)$, for $K = 2$, and $\beta = (0.20, 0.00, 0.14, -0.10, -0.10)$, for $K = 4$. For all of these additional scenarios, σ^2 was set equal to 0.5.

Parameters for the simulation study were chosen so that we would consider a wide range of values for the sample size (representative of those sample sizes commonly used in large psychology studies; see Kühberger et al. (2014), Fraley & Vazire (2014), and Marszalek et al. (2011)). We also wished to obtain three unique values for $\mathcal{B}_1$ approximately evenly spaced between 0 and 0.20.

For each of the total 32 configurations, we simulated 10,000 unique datasets and calculated an equivalence test p-value with each of 49 different values of Δ (ranging from 0.01 to 0.25). We then calculated the proportion of these p-values less than $\alpha = 0.05$. We specifically chose to conduct 10,000 simulation runs so as to keep computing time within a reasonable limit while also reducing the amount of Monte Carlo standard error to a negligible amount (for looking at type 1 error with $\alpha = 0.05$, Monte Carlo SE will be approximately $0.002 \approx \sqrt{0.05(1 - 0.05)/10,000}$); see Morris et al. (2019).

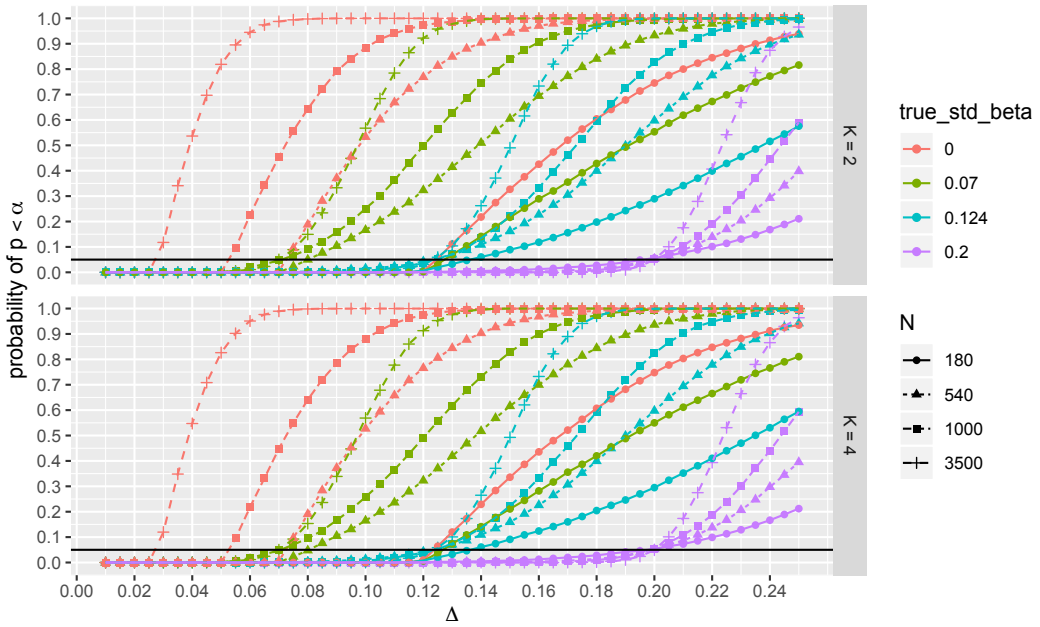

Figure 1. Simulation Study 1 - Upper panel shows results for $K = 2$; Lower panel shows results for $K = 4$. The solid horizontal black line indicates the desired type 1 error of $\alpha = 0.05$.

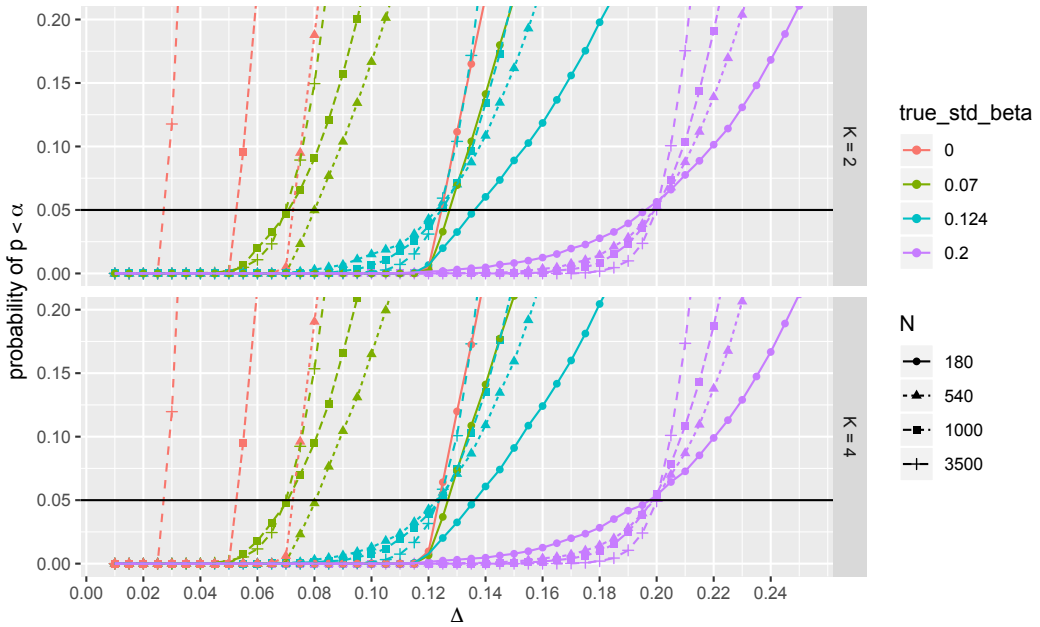

Figure 2. Simulation Study 1 - Upper panel shows results for $K = 2$; lower panel shows results for $K = 4$. Both plots are presented with a restricted vertical-axis to better show the type 1 error rates. The solid horizontal black line indicates the desired type 1 error of $\alpha = 0.05$.

3.0.1. Simulation Study 1 Results-

Figure 1 and Figure 2 show results from the simulation study. In the Appendix, we show results from an alternate version of the simulation study where the covariates are unbalanced and are correlated.

When $\mathcal{B}_1 = 0.200$, we see that when the equivalence bound Δ equals the true effect size (i.e., when $\mathcal{B}_1 = \Delta = 0.20$), the type 1 error rate is exactly 0.05, as it should be, for all N. This situation represents the boundary of the null hypothesis. As the equivalence bound increases beyond the true effect size (i.e., $\Delta > \mathcal{B}_1$), the alternative hypothesis is then true and it is more and more likely we will correctly conclude equivalence.

For smaller values of $\mathcal{B}_1$ (i.e., for $\mathcal{B}_1 = 0.070$ and $\mathcal{B}_1 = 0.124$), when the equivalence bound equals the true effect size (i.e., when $\mathcal{B}_1 = \Delta$), the test is conservative, particularly for small N. Even when $\Delta > \mathcal{B}_1$, the equivalence test may reject the null hypothesis for less than 5% of cases. For example, when $N = 180$, $\mathcal{B}_1 = 0.124$ and $K = 2$, the rejection rate is only 0.020 when $\Delta = 0.125$. This is due to the fact that with a small N, the sampling variance of $\widehat{\mathcal{B}}_1$ may be far too large to reject $H_0 : |\mathcal{B}_1| \geq \Delta$. Consider that, when N is small and σ^2 is relatively large, the 90% CI for $\mathcal{B}_1$ may be far too wide to fit entirely within the equivalence margin.

4. Comparison to a Bayesian alternative

4.1. Conditional equivalence testing

Ideally, a researcher uses an equivalence test to examine a preregistered hypothesis concerning the absence of a meaningful effect. However, in practice, one might first conduct a NHST (i.e., calculate a p-value, p_1, using equation (2) or (3)) and only proceed to the equivalence test (i.e., calculate a second p-value, p_2, using equation (7)) if the NHST fails to reject the null. Such a two-stage sequential testing scheme has recently been put forward by Campbell & Gustafson (2018*a*) under the name of "conditional equivalence testing" (CET).

Under the proposed CET scheme, if the first p-value, p_1, is less than the type 1 error α-threshold (e.g., if $p_1 < 0.05$), one concludes with a "positive" finding: $\mathcal{B}_k$ is significantly different than 0. On the other hand, if the first p-value, p_1, is greater than α and the second p-value, p_2, is smaller than α (e.g., if $p_1 \geq 0.05$ and $p_2 < 0.05$), one concludes with a "negative" finding: there is evidence of a statistically significant equivalence, i.e., $\mathcal{B}_k$ is at most negligible. If both p-values are larger than α, the result is inconclusive: there are insufficient data to support either finding. In this paper, we are not advocating for (or against) CET, but simply use it to facilitate a comparison with Bayes Factor testing (which also categorizes outcomes as either positive, negative or inconclusive).

4.2. Bayes Factor testing for linear regression

For linear regression models, based on the work of Liang et al. (2008), Rouder & Morey (2012*a*) propose using Bayes Factors (BFs) to determine whether the data support the inclusion of a particular variable in the model. This is a common approach used in psychology studies (e.g., see the tutorial of Etz (2015)). In other related Bayesian work, Lu & Westfall (2019) consider how to calculate Bayesian credible intervals for standardized linear regression coefficients.

Here we will consider using BFs and refer to the null model ("Model 0") and alternative model ("Model 1") as:

$$\text{Model } 0 : Y_i \sim \quad Normal(X_{i,-k}^T \beta_{-k}, \sigma^2), \qquad \forall i = 1, ..., N; \tag{11}$$

$$\text{Model } 1 : Y_i \sim \quad Normal(X_{i,\cdot}^T \beta, \sigma^2), \qquad \forall i = 1, ..., N; \tag{12}$$

where β_{-k} ($X_{i,-k}$) is the vector (matrix) of regression coefficients (covariates), with the k-th coefficient (covariate) omitted.

We define the Bayes Factor, BF_{10}, as the probability of the data under the

alternative model relative to the probability of the data under the null model:

$$BF_{10} = \frac{Pr(Data \,|\, Model\, 1)}{Pr(Data \,|\, Model\, 0)}, \tag{13}$$

with the "10" subscript indicating that the alternative model (i.e., "Model 1") is being compared to the null model (i.e., "Model 0"). The BF can be easily interpreted. For example, a BF_{10} equal to 0.20 indicates that the null model is five times more likely than the alternative model.

Bayesian methods require one to define appropriate prior distributions for all model parameters. Rouder & Morey (2012*a*) suggest using "objective priors" for linear regression models and explain in detail how one may implement this approach. Defining prior distributions can often be controversial, as their choice can substantially influence the posterior when few data are available; see Lambert et al. (2005), Berger (2013). We will not discuss the thorny issue of prior specification in detail, and instead point interested readers to Consonni et al. (2008) who provide an in-depth overview of how to specify prior distributions for linear models.

Using the BayesFactor package in `R` (Morey et al. 2015) with the function `regressionBF()`, one can easily obtain BFs corresponding to a given linear regression dataset. (See also the baymedr package in `R` (Linde & van Ravenzwaaij 2019)). Since we can also calculate frequentist p-values (see equations (2), (3), (7)) corresponding to any given linear regression dataset, the frequentist and Bayesian approaches can be compared in a relatively straightforward way. We will explore this in Simulation Study 2.

4.3. Simulation Study 2

We wish to compare a frequentist testing scheme based on NHST and equivalence testing to the Bayesian approach based on BFs by means of a simple simulation study. Our main interest is in determining how often will the frequentist and Bayesian approaches

arrive at the same conclusion.

Frequentist conclusions are based on the CET procedure and by setting Δ equal to either 0.05, or 0.10, or 0.25; and with α=0.05. Bayesian conclusions are based on an evidence threshold of either 3, 6, or 10. A threshold of 3 can be considered "moderate evidence," a threshold of 6 can be considered "strong evidence," and a threshold of 10 can be considered "very strong evidence" (Jeffreys 1961). Note that for the simulation study here we examine only the "fixed-n design" for BF testing; see Schönbrodt & Wagenmakers (2016) for details. Also note that all priors required for calculating the BF were set by simply selecting the default settings of the `regressionBF()` function (with rscaleCont = "medium"); see Morey et al. (2015).

We simulated datasets for 36 unique scenarios. We varied over the following:

- one of twelve sample sizes: $N = 20$, $N = 33$, $N = 55$, $N = 90$, $N = 149$, $N = 246$, $N = 406$, $N = 671$, $N = 1,109$, $N = 1,832$, $N = 3,027$, or $N = 5,000$;
- one of two designs with $K = 4$ binary covariates (with an orthogonal, balanced design), with either $\beta = (0.20, 0.10, 0.14, -0.10, -0.10)$ or $\beta = (0.20, 0.00, 0.14, -0.10, -0.10)$;
- one of two variances: $\sigma^2 = 0.50$, or $\sigma^2 = 1.00$.

Note that for the $\beta = (0.20, 0.00, 0.14, -0.10, -0.10)$ design, we only consider one value for $\sigma^2 = 1.00$. Depending on the particular design and σ^2, the true standardized regression coefficient, $\mathcal{B}_1$, for these data is either $\mathcal{B}_1 = 0.00$, $\mathcal{B}_1 = 0.05$, or $\mathcal{B}_1 = 0.07$.

4.3.1. Simulation Study 2 Results-

For each simulated dataset, we obtained frequentist p-values, BFs, and declared the result to be positive, negative or inconclusive, accordingly. Results are presented in Figures 3, 4 and 5 and are based on 150 distinct simulated datasets per scenario.

We are particularly interested in how often the two approaches will reach the same overall conclusion (positive, negative or inconclusive). Table 4.3.1 displays the the average rate of agreement between the Bayesian and frequentist methods. *Aver-*

aging over all 36 scenarios, how often on average will the Bayesian and frequentist approaches reach the same conclusion given the same data?

	BF threshold=3	BF threshold=6	BF threshold=10
$\Delta = 0.25$	0.67	0.47	0.39
$\Delta = 0.10$	0.85	0.76	0.69
$\Delta = 0.05$	0.77	0.85	0.84

Table 1. Averaging over all 36 scenarios and over the 150 Monte Carlo simulations per scenario, how often on average did the Bayesian and frequentist approaches reach the same conclusion? Numbers in the above table represent the average proportion of simulated datasets (averaged over $36 \times 150 = 5,400$ unique datasets) for which the Bayesian and frequentist methods arrive at the same conclusion.

Two observations merit comment:

- The Bayesian testing scheme is always less likely to deliver a positive conclusion (see how the dashed blue curve is always higher than the solid blue curve). In the scenarios like the ones we considered, the BF may require larger sample sizes for reaching a positive conclusion and thus may be considered "less powerful" in a traditional frequentist sense.

- With $\Delta = 0.10$ or $\Delta = 0.25$, and a BF threshold of 6 or 10, the BF testing scheme requires substantially more data to reach a negative conclusion than the frequentist scheme; see dashed orange lines in Figures 4, and 5 - panels 2, 3, 5, 6, 8, and 9. Note that, the probability of reaching a negative result with CET will never exceed 0.95 since the NHST is performed first (before the equivalence test).

Based on our comparison of BFs and frequentist tests, we can confirm that, given the same data, both approaches will often provide one with the same overall conclusion. The level of agreement however is highly sensitive to the choice of Δ and the choice of the BF evidence threshold, see Table 4.3.1. While we did not consider the impact of selecting different priors with the BFs, it is reasonable to assume that the level of agreement between BFs and frequentist tests will also be rather sensitive to the chosen priors, particularly when N is small; see Berger (2013).

We observed the highest level of agreement, recorded at 0.85, when $\Delta = 0.10$ with the BF evidence threshold equal to 3, and when $\Delta = 0.05$ with the BF evidence threshold equal to 6. In contrast, when $\Delta = 0.25$ and the BF evidence threshold is 10,

the two approaches will deliver the same conclusion less than 40% of the time. Table 4.3.1 shows that the two approaches never arrived at entirely contradictory conclusions for the same dataset. In not a single case (amongst the 5,400 datasets) did we observe one approach arrive at a positive conclusion while the other approach arrived at a negative conclusion, when faced with the same exact data.

The results of the simulation study are reassuring since they suggest that the conclusions obtained from frequentist and Bayesian testing will very rarely lead to substantial disagreements.

	BF threshold=3	BF threshold=6	BF threshold=10
$\Delta = 0.25$	0.00	0.00	0.00
$\Delta = 0.10$	0.00	0.00	0.00
$\Delta = 0.05$	0.00	0.00	0.00

Table 2. Averaging over all 36 scenarios and over all 150 Monte Carlo simulations per scenario, how often on average did the Bayesian and frequentist approaches strongly disagree in their conclusion? Numbers in the above table represent the average proportion of simulated datasets (averaged over $36 \times 150 = 5,400$ unique datasets) for which the Bayesian and frequentist methods arrived at completely opposite (one positive and one negative) conclusions.

5. Practical Example: Evidence for gender bias or the lack thereof in academic salaries

In order to illustrate the various testing methods, we turn to the "Salaries" dataset (from R CRAN package car; see Fox et al. (2012)) to use as an empirical example. This dataset has been used as an example in other work: as an example for "anti-NHST" statistical inference in Briggs et al. (2019); and as an example for data visualization methods in Moon (2017) and Ghashim & Boily (2018).

The data consist of a sample of salaries of university professors collected during the 2008-2009 academic year. In addition to the posted salaries (a continuous variable,

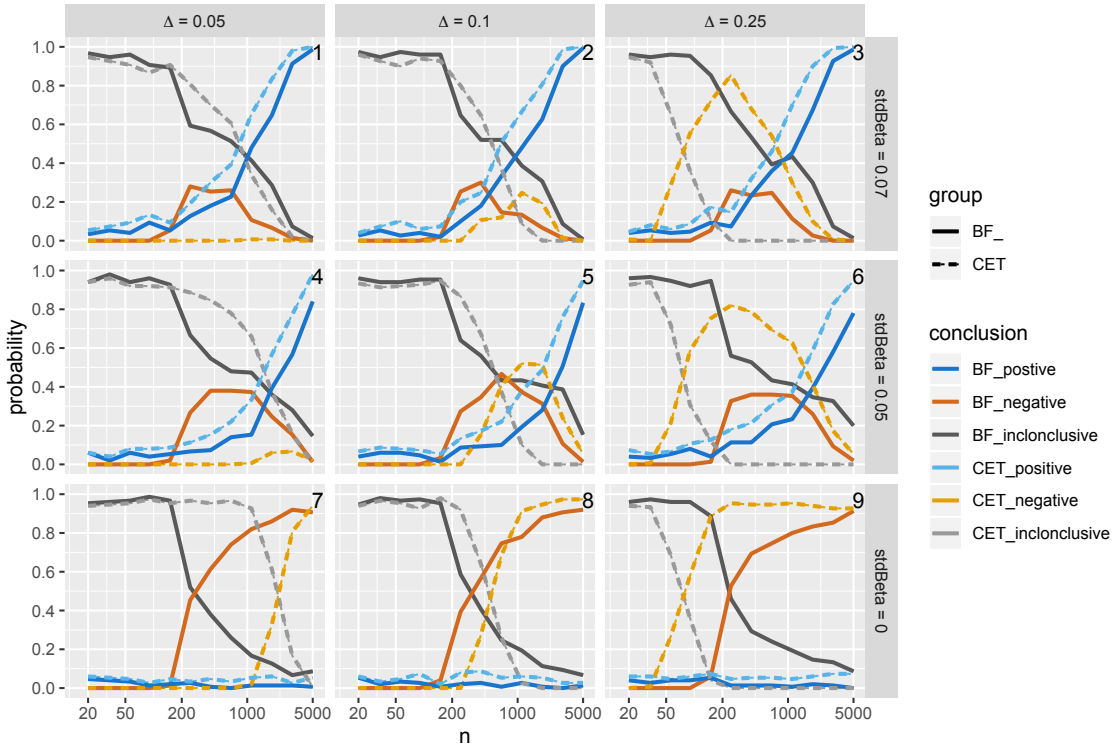

Figure 3. Simulation study 2, complete results for BF threshold of 3. The probability of obtaining each conclusion by Bayesian testing scheme (JZS-BF with fixed sample size design, BF threshold of 3:1) and CET ($\alpha = 0.05$). Each panel displays the results of simulations with for different values of Δ and $\mathcal{B}_1$. Note that all solid lines and the dashed blue line do not change for different values of Δ.

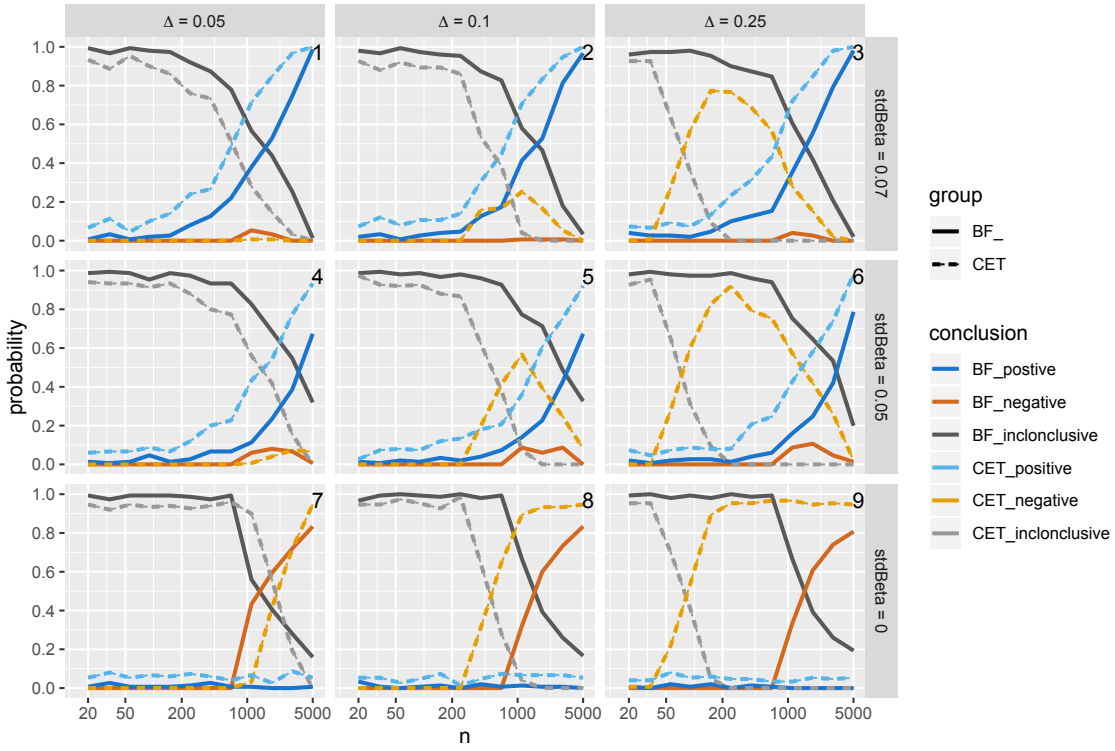

Figure 4. Simulation study 2, complete results for BF threshold of 6. The probability of obtaining each conclusion by Bayesian testing scheme (JZS-BF with fixed sample size design, BF threshold of 6:1) and CET ($\alpha = 0.05$). Each panel displays the results of simulations with for different values of Δ and $\mathcal{B}_1$. Note that all solid lines and the dashed blue line do not change for different values of Δ.

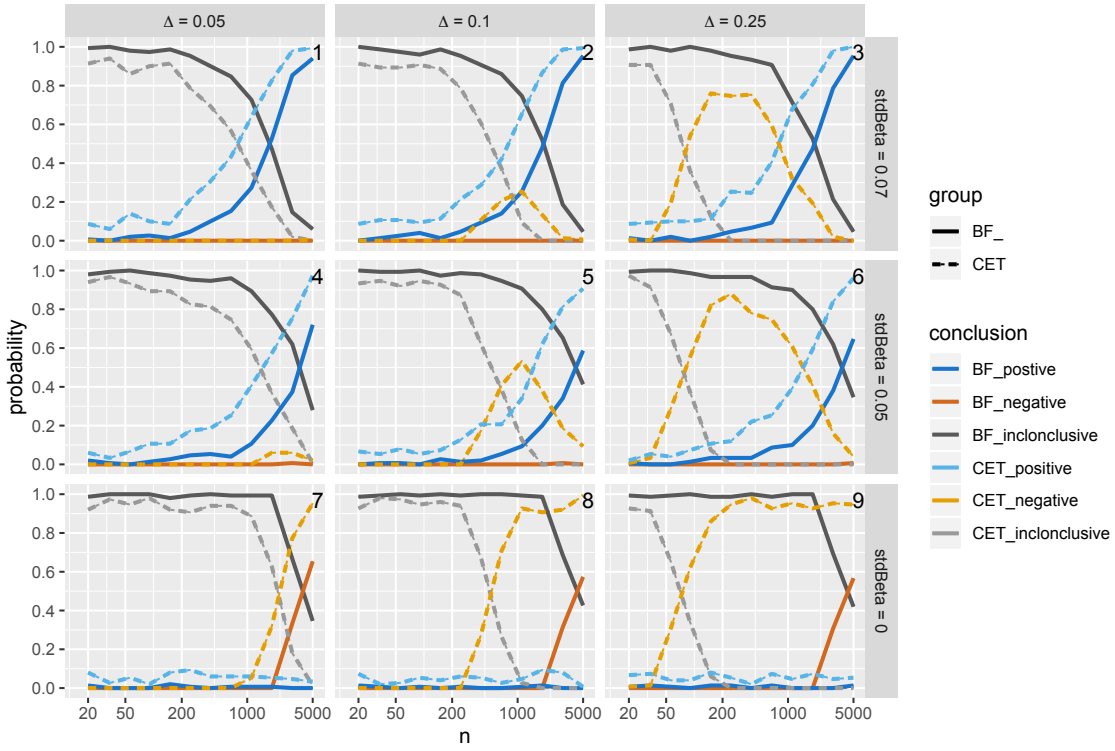

Figure 5. Simulation study 2, complete results for BF threshold of 10. The probability of obtaining each conclusion by Bayesian testing scheme (JZS-BF with fixed sample size design, BF threshold of 10:1) and CET ($\alpha = 0.05$). Each panel displays the results of simulations with for different values of Δ and $\mathcal{B}_1$. Note that all solid lines and the dashed blue line do not change for different values of Δ.

in $US), the data includes 5 additional variables of interest:

(1) sex (2 categories: (1) Female, (2) Male);

(2) years since Ph.D. (continuous, in years);

(3) years of service (continuous, in years);

(4) discipline (2 categories: (1) theoretical, (2) applied).

(5) academic rank (3 categories: (1) Asst. Prof. , (2) Assoc. Prof., (3) Prof.);

The sample includes a total of $N = 397$ observations with 358 observations from male professors and 39 observations from female professors. The minimum measured salary is $57,800, the maximum is $231,545, and the median salary is $107,300.

A primary question of interest is whether there is a difference between the salary of a female professor and a male professor when accounting for possible observed confounders: rank, years since Ph.D., years of service, and discipline. The mean salary for male professors in the sample is $115,090, while the mean salary for female professors in the sample is $101,002.

Let us begin by first conducting a simple linear regression ($K = 1$) for the association between salary (Y, measured in $) and sex ($X_1$, where "0" corresponds to "female," and "1" corresponds to "male."):

$$Y \sim \mathcal{N}(\beta_0 + \beta_1 X_1, \sigma^2). \tag{14}$$

We obtain the following parameter estimates by standard least squares estimation:

- $\widehat{\beta}_0 = 101002$, $\text{SE}(\widehat{\beta}_0) = 4809$, and $\widehat{\beta}_1 = 14088$, $\text{SE}(\widehat{\beta}_1) = 5065$ (see equation (21));
- $\widehat{\sigma} = 30034.61$, (see equation (22));
- $\widehat{\mathcal{B}}_1 = 0.14$, $\text{SE}(\widehat{\mathcal{B}}_1) = 0.05$ (see equations (6) and (8)); and

- $R^2_{Y \cdot X} = \text{diff}R^2_1 = 0.019$.

We can calculate p-values for a number of different hypothesis tests associated with the simple linear regression model.

- Calculating a p-value for the standard NHST ($H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$) is done by either t-test of F-test (see equations (2) and (3)) and we obtain $p = 0.006$. Since $p < 0.05$, we can reject H_0.

- We can also conduct an equivalence test to determine if the difference in salaries between male and female professors is at most no more than some negligible amount. We might claim that any difference of less than $\Delta = \$5,000$ is negligible, but this specific choice is clearly somewhat arbitrary. Indeed, any choice for Δ will be subjective and it will be difficult to justify what amount of money is too small to be considered meaningful. A p-value for the equivalence test, $H_0 : |\beta_1| \geq 5000$ vs. $H_1 : |\beta_1| < 5000$, can be calculated using equation (4). We obtain $p = 0.963$ and can therefore reject the equivalence test null hypothesis. There is insufficient evidence to suggest that any difference is at most negligible (with $\Delta = \$5,000$).

- We can also conduct an equivalence test for the coefficient of determination, (Campbell & Lakens 2020). In this case, we might simply set $\Delta = 0.01$. The choice of $\Delta = 0.01$ represents the belief that any association between sex and salary explaining less than 1% of the variability in the data would be considered negligible. For reference, Cohen (1988) describes a $R^2 = 0.0196$ as "a modest enough amount, just barely escaping triviality"; and more recently, Fritz et al. (2012) consider associations explaining "1% of the variability" as "trivial."

  As per Campbell & Lakens (2020), we calculate the p-value ($H_0 : P^2_{Y \cdot X} \geq 0.01$ vs. $H_1 : P^2_{Y \cdot X} < 0.01$) as follows:

$$
\begin{aligned}
p\text{-value} &= p_f\left(F; K, N-K-1, \frac{N\Delta}{(1-\Delta)}\right) \\
&= p_f\left(7.73; 1, 397-1-1, \frac{397\cdot 0.01}{(1-0.01)}\right) \\
&= 0.780,
\end{aligned} \tag{15}
$$

where:

$$
\begin{aligned}
F &= \frac{R^2_{Y,X}/K}{(1-R^2_{Y,X})/(N-K-1)} \\
&= \frac{0.019/1}{(1-0.019)/(397-1-1)} \\
&= 7.73.
\end{aligned} \tag{16}
$$

- We can also conduct an equivalence test for $\text{diff}P_1^2$, the increase in the coefficient of determination attributable to including the sex variable in the model. We consider $H_0 : \text{diff}P_1^2 \geq \Delta$ vs. $H_1 : \text{diff}P_1^2 < \Delta$. Since $K = 1$, this will be identical to the equivalence test for $P^2_{Y\cdot X}$ above. We set $\Delta = 0.01$ and obtain a p-value as per equation (17):

$$
\begin{aligned}
p-\text{value} &= 1 - p_t\left(\frac{\sqrt{(N-K-1)\text{diff}R_k^2}}{\sqrt{(1-R^2_{Y\cdot X})}}; df = N-K-1, ncp = \frac{\sqrt{N\Delta}}{\sqrt{\left(1-\Delta+R^2_{X_k\cdot X_{-k}}\right)}}\right) \\
&= 1 - p_t\left(\frac{\sqrt{(397-1-1)0.019}}{\sqrt{(1-0.019)}}; df = 397-1-1, ncp = \frac{\sqrt{397\times 0.01}}{\sqrt{(1-0.01+0)}}\right) \\
&= \quad 0.780.
\end{aligned} \tag{17}
$$

- Finally, we can conduct an equivalence test for the standardized regression coefficient, $\mathcal{B}_1$. Since $K = 1$, this will be identical to the equivalence tests for $P^2_{Y\cdot X}$ and $\text{diff}P_1^2$ above, with $\Delta = 0.10$ $(= \sqrt{0.01})$. We calculate the p-value

($H_0 : |\mathcal{B}_1| \geq 0.10$ vs. $H_0 : |\mathcal{B}_1| < 0.10$) as per equation (7):

p-value $= max(p_1^{[1]}, p_1^{[2]}) = 0.780$, where:

$$
\begin{aligned}
p_1^{[1]} &= p_t \left(\frac{\widehat{\mathcal{B}_k}}{SE(\widehat{\mathcal{B}_k})}; df = N - K - 1, ncp = \Delta_1 \frac{\sqrt{N\left(1 - R^2_{X_k \cdot X_{-k}}\right)}}{\sqrt{1 - ((1 - R^2_{X_k \cdot X_{-k}})\Delta_1^2 + R^2_{Y \cdot X_{-k}})}} \right) \\
&= p_t \left(\frac{0.14}{0.05}; df = 397 - 1 - 1, ncp = -0.10 \frac{\sqrt{397\,(1 - 0)}}{\sqrt{1 - ((1 - 0) \times 0.01 + 0)}} \right) \\
&= p_t \left(2.78, df = 395, ncp = -2.00\right) \\
&< 0.001 \qquad (18)
\end{aligned}
$$

$$
\begin{aligned}
p_1^{[2]} &= p_t \left(\frac{-\widehat{\mathcal{B}_k}}{SE(\widehat{\mathcal{B}_k})}; df = N - K - 1, ncp = -\Delta_2 \frac{\sqrt{N\left(1 - R^2_{X_k \cdot X_{-k}}\right)}}{\sqrt{1 - ((1 - R^2_{X_k \cdot X_{-k}})\Delta_2^2 + R^2_{Y \cdot X_{-k}})}} \right) \\
&= p_t \left(-2.78, df = 395, ncp = -2.00\right) \\
&= 0.780. \qquad (19)
\end{aligned}
$$

We can also calculate several BFs of interest. With the BayesFactor package, using default priors and the "regressionBF" function, we obtain a $BF_{10} = 4.52$ which suggests that the alternative model (= the model with "sex" variable included) is about four and a half times more likely than the null model (= the intercept only model). Note that we obtain the identical result using the "linearReg.R2stat" function. However, when using the "lmBF", we obtain a value of $BF_{10} = 6.21$ which suggests that the alternative model is about 6 times more likely than the null model. Both functions are comparing the two very same models so this result is surprising. The apparent contradiction can be explained by the fact that the two "default BF" functions are using

different "default priors." The "regressionBF" function assumes "sex" is a continuous variable, while the "lmBF" function assumes that "sex" is a categorical variable. The "default priors" are defined accordingly, in different ways. This may strike one as rather odd, since both models are numerically identical. However, others see logic in such practice: Rouder et al. (2012) suggest researchers "be mindful of some differences when considering categorical and continuous covariates" and "recommend that researchers choose priors based on whether the covariate is categorical or continuous"; see Rouder et al. (2012), Section 13 for details.

Let us now consider the multivariable linear regression model, with $K = 6$:

$$Y \sim \mathcal{N}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6, \sigma^2), \tag{20}$$

where $X_1 = 0$ corresponds to "female," and $X_1 = 1$ corresponds to "male"; X_2 corresponds to years since Ph.D.; X_3 corresponds to years of service; $X_4 = 0$ corresponds to "theoretical," and $X_4 = 1$ corresponds to "applied"; and where ($X_5 = 0$, $X_6 = 0$) corresponds to "Asst. Prof.", ($X_5 = 1$, $X_6 = 0$) corresponds to "Assoc. Prof.", and ($X_5 = 0$, $X_6 = 1$) corresponds to "Prof.".

Table 3 lists parameter estimates obtained by standard least squares estimation for the full multivariable linear regression model. Table 4 lists the p-values for each of the hypothesis tests we consider. We also calculate a Bayes Factor (using the "regressionBF" function) comparing the full model (with $K = 6$, and X_1, X_2, X_3, X_4, X_5, and X_6) to the null model (with $K = 5$, and X_2, X_3, X_4, X_5, and X_6). We obtain a Bayes Factor of $B_{10} = 0.260 = 1/3.86$, indicating only moderate evidence in favour of the null model. This would correspond to an "inconclusive" result for a BF threshold of 6, or 10 (or any threshold higher than 3.86). The result from CET would also be "inconclusive" (for $\alpha = 0.05$ and $\Delta = 0.10$), since both the NHST p-value ($= 0.216$) and the equivalence test p-value ($= 0.076$) are larger than $\alpha = 0.05$. As such, we conclude that there are insufficient data to support either an association, or the lack of an association, between sex and salary.

k	β_k	$SE(\hat{\beta}_k)$	$\mathcal{B}_k$	$SE(\widehat{\mathcal{B}_k})$
0	65955.23	4588.60	-	-
1	4783.49	3858.67	0.05	0.04
2	535.06	240.99	0.23	0.10
3	-489.52	211.94	-0.21	0.09
4	14417.63	2342.88	0.24	0.04
5	12907.59	4145.28	0.16	0.05
6	45066.00	4237.52	0.70	0.07
		$\hat{\sigma} = 22538.65$		$R^2_{Y,X} = 0.455$

Table 3. Parameter estimates obtained by standard least squares estimation for the full multivariable linear regression model.

test			Δ	p_1
$H_0:$	$\beta_1 = 0$	vs. $H_1 : \beta_1 \neq 0$	-	0.216
$H_0:$	$\lvert\beta_1\rvert \geq \Delta$	vs. $H_1 : \lvert\beta_1\rvert < \Delta$	5000	0.478
$H_0:$	$\text{diff}P_1^2 \geq \Delta$	vs. $H_1 : \text{diff}P_1^2 < \Delta$	0.01	0.232
$H_0:$	$\lvert\mathcal{B}_1\rvert \geq \Delta$	vs. $H_0 : \lvert\mathcal{B}_1\rvert < \Delta$	0.10	0.076

Table 4. Hypothesis tests for association, or lack thereof, between "salary" and "sex," from the full multivariable linear regression model. R-code to obtain these results is presented in the Appendix.

6. Conclusion

Researchers require statistical tools that allow them to reject the presence of meaningful effects; see Altman & Bland (1995) and more recently Amrhein et al. (2019). In this paper we present just such a tool: an equivalence test for standardized effect sizes in linear regression analyses. Equivalence testing is a formal version of what

Equivalence tests may improve current research practices by allowing researchers to falsify their predictions concerning the presence of an effect. Moreover, expanding equivalence testing to standardized effect sizes can help researchers conduct equivalence tests by facilitating what is often a very challenging task: defining an appropriate equivalence margin. While the use of "default equivalence margins" based on standardized effect sizes cannot be whole-heartily recommended for all cases, their use is not unlike the use of "default priors" for Bayesian inference which have indeed proven useful to researchers in many scenarios.

Via simulation study, we considered how frequentist equivalence testing offers an attractive alternative to Bayesian methods for "testing the null" in the linear regression context. Depending on how they are configured, testing based on BFs and based on equivalence testing may operate very similarly for making "trichotomous significance-

testing decisions" (i.e., for determining if the evidence is "positive," "negative," or "inconclusive").

As Rouder & Morey (2012*b*) note when discussing default BFs: "Subjectivity should not be reflexively feared. Many aspects of science are necessarily subjective. [...] Researchers justify their subjective choices as part of routine scientific discourse, and the wisdom of these choices are evaluated as part of routine review." The same sentiment applies to frequentist testing. Researchers using equivalence testing should be prepared to justify their choice for the equivalence margin based on what effect sizes are considered negligible. That being said, equivalence tests for standardized effects may help researchers in situations when what is "negligible" is particularly difficult to determine. They may also help establish generally acceptable levels for standard margins in the literature (Campbell & Gustafson 2018*b*).

Note that our non-inferiority test for the increase in the squared multiple correlation coefficient (diffP_k^2) in a standard multivariable linear regression is limited to comparing two models for which the difference in degrees of freedom is 1. In other words, the test is not suitable for comparing two nested models where the difference is more than a single variable. For example, with the salaries data we considered, we cannot use the proposed test to compare a "smaller model" with only "sex" as a covariate, with a "larger model" that includes "sex," "discipline" and "rank," as covariates. A more general equivalence test for comparing two nested models will be considered in future work; Tan Jr (2012) is an excellent resource for this undertaking.

Also, note that we only considered equivalence tests based on inverting NCIE-based confidence intervals. It would certainly be worthwhile to consider equivalence tests based on alternative approximations for the sampling variability of standardized regression coefficients; see Jones & Waller (2013) and Yuan & Chan (2011). Finally, going forward, we wish to expand equivalence testing for standardized regression coefficients in logistic regression models and time-to-event models, in order to further "extend the arsenal of confirmatory methods rooted in the frequentist paradigm of inference" (Wellek 2017).

Acknowledgements Thank you to Prof. Paul Gustafson for the helpful advice with preliminary drafts and thank you also to Prof. Daniël Lakens and Prof. Ken Kelley for insightful discussions and encouragement.

7. Appendix

Least squares estimates for the linear regression model are:

$$\widehat{\beta}_k = ((X^T X)^{-1} X^T y)_k, \text{for } k \text{ in } 1,..., K; \tag{21}$$

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{N} (\hat{\epsilon}_i^2)/(N - K - 1)}, \tag{22}$$

where $\hat{\epsilon}_i = \widehat{y}_i - y_i$, and $\widehat{y}_i = X_{i,\cdot}^T \widehat{\beta}$, for i in 1,..., N.

7.1. Simulation Study 1 - alternative settings

We conducted a second version of Simulation Study 1 with correlated non-balanced covariates. Specifically, for $K = 2$, we sampled correlated binary variables in such a way so that $cor(X_1, X_2) = 0.40$, and so that half of the X_1 values are equal to 1 and only a quarter of the X_2 values are equal to 1. We set $\beta = (-0.20, 0.10, 0.19)$ to correspond to $\mathcal{B}_1 = 0.070$, 0.124, and 0.200, with $\sigma^2 = 0.50, 0.15$ and 0.05, respectively. We set $\beta = (-0.20, 0.00, 0.19)$ to correspond to $\mathcal{B}_1 = 0.000$, with $\sigma^2 = 0.50$.

With $K = 4$, we sampled correlated binary variables in such a way that the correlation between the four variables was:

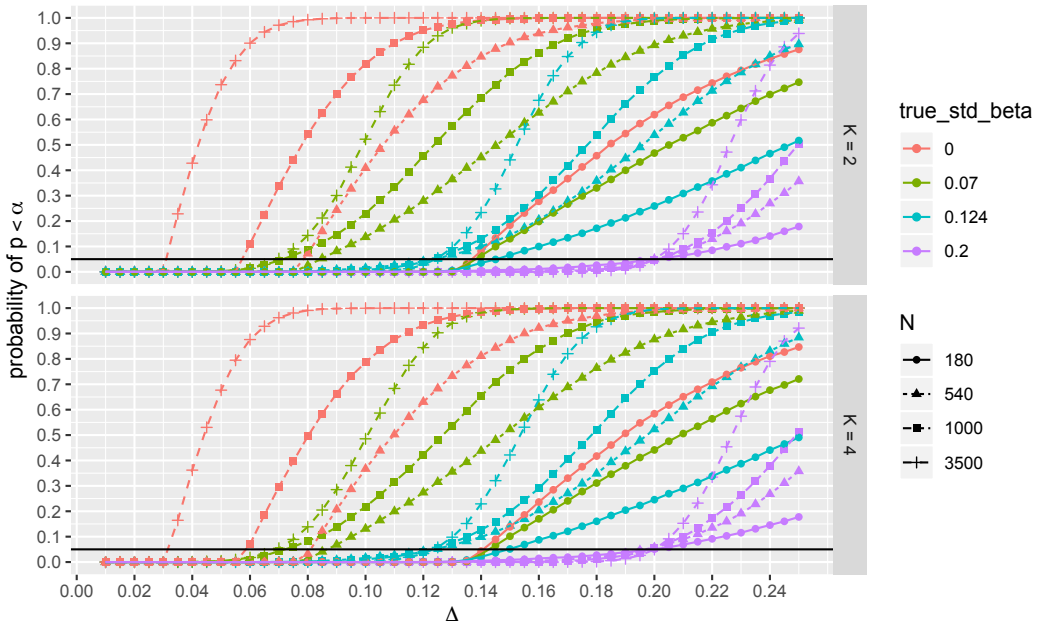

Figure 6. Simulation Study 1 (alternative settings) - Upper panel shows results for $K = 2$; Lower panel shows results for $K = 4$. The solid horizontal black line indicates the desired type 1 error of $\alpha = 0.05$.

$$cor(X_1, X_2, X_3, X_4) = \begin{pmatrix} 1 & 0.4 & 0.3 & 0 \\ 0.4 & 1 & 0.4 & 0.3 \\ 0.3 & 0.4 & 1 & 0.4 \\ 0 & 0.3 & 0.4 & 1 \end{pmatrix}, \tag{23}$$

and so that half of the X_1 , and the X_4 values are equal to 1 and only a quarter of the X_2 and the X_3 values are equal to 1. We set $\beta = (0.20, 0.10, 0.14, -0.12, -0.14)$ to correspond to $\mathcal{B}_1 = 0.070$, 0.124, and 0.200, with $\sigma^2 = 0.50, 0.15$ and 0.05, respectively. We set $\beta = (0.20, 0.00, 0.14, -0.12, -0.14)$ to correspond to $\mathcal{B}_1 = 0.000$, with $\sigma^2 = 0.50$.

Figures 6 and 7 plot the results. Results are similar to those obtained with orthogonal, balanced designs. The only difference to note is that, as one might expect, power is much lower with correlated non-balanced covariates.

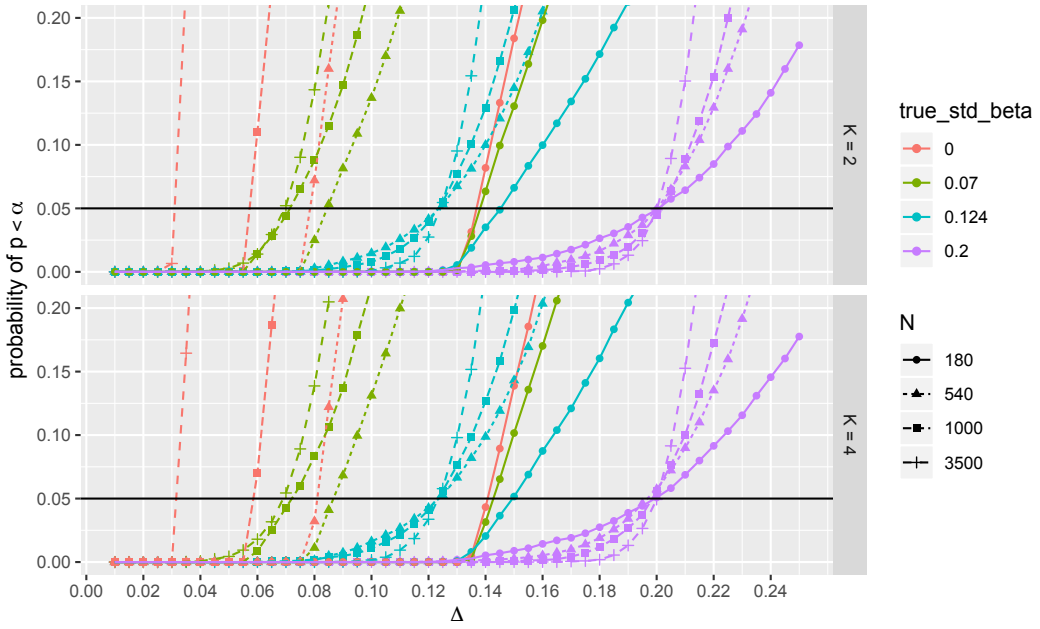

Figure 7. Simulation Study 1 (alternative settings) - Upper panel shows results for $K = 2$; lower panel shows results for $K = 4$. Both plots are presented with a restricted vertical-axis to better show the type 1 error rates. The solid horizontal black line indicates the desired type 1 error of $\alpha = 0.05$.

7.2. R-code for calculating p-values

In R, we can obtain the p-value from equation (2) as follows:

```
lmmod <- summary(lm(y~X[,-1]))
N <- length(y); K <- dim(X[,-1])[2]
beta_hat <- lmmod$coef[,1]; SE_beta_hat <- lmmod$coef[,2]
pval <- 2*pt(abs(beta_hat/SE_beta_hat), N-K-1, 0, lower.tail=FALSE)
```

and the p-value from equation (3) as follows:

```
R2 <- lmmod$r.squared
diffR2k <- unlist(lapply(c(2:(K+1)), function(k) {R2-summary(lm(y~X[,-k]))$r.squared}))
pval <- pf((N-K-1)*(diffR2k/(1-R2)), 1, N-K-1, 0, lower.tail=FALSE)
```

We can obtain the p-values from equation (4) as follows:

```
p1[k] <- pt((beta_hat[k] - DELTA[k,1])/SE_beta_hat[k], N-K-1, 0, lower.tail=FALSE)
p2[k] <- pt((-beta_hat[k] + DELTA[k,2])/SE_beta_hat[k], N-K-1, 0, lower.tail=FALSE)
pval[k] <- max(c(p1[k], p2[k]))
```

We can obtain the estimated standardized regression coefficients (equation (6)) in R

as follows:

```
b_vec <- (beta_hat*(apply(X,2,sd)/sd(y)))[-1]
```

and obtain the p-values from equation (7) in R with the following code:

```
P2_1 = (1-R2XkdotXminK[k])*DELTA[k,1]^{2} + R2YdotXmink[k]
ncp_1 = sqrt(N*(1-R2XkdotXminK[k])) * (DELTA[k,1]/sqrt( 1 - P2_1 ))

P2_2 = (1-R2XkdotXminK[k])*DELTA[k,2]^{2} + R2YdotXmink[k]
ncp_2 = sqrt(N*(1-R2XkdotXminK[k])) * (-DELTA[k,2]/sqrt( 1 - P2_1 ))

p1[k] <- pt(b_vec[k]/SE_beta_FIX[k], N-K-1, ncp=ncp_1, lower.tail=FALSE)
p2[k] <- pt(-b_vec[k]/SE_beta_FIX[k], N-K-1, ncp=ncp_2, lower.tail=FALSE)
pval[k] <- max(c(p1[k], p2[k]))
```

Finally, one can calculate the p-value from equation (17) in R with the following code:

```
R2XkdotXminK[k] <- (summary(lm(cbind(X[,-1])[,k]~ Xmink)))$r.squared
ncp_1<-sqrt(N*DELTA[k])/sqrt(1-DELTA[k]+  R2XkdotXminK[k])
pval[k] <- pt(sqrt((N-K-1)*diffR2k[k])/sqrt(1-R2), N-K-1, ncp=ncp_1, lower.tail=TRUE)
```

7.3. R-code for Salaries example

Results shown in Table 4 can be obtained with the following R-code:

```
library(carData)
library(RCurl)
script <- getURL("https://raw.githubusercontent.com/harlanhappydog/EquivTestStandardReg/master/EquivTestStandardReg.R"
eval(parse(text = script))

y <- Salaries$salary
X <- model.matrix(lm(salary~ discipline + rank + yrs.since.phd + yrs.service + sex, data=Salaries))

# NSHT, p-val = 0.216
summary((lm(salary ~ rank + discipline + yrs.since.phd + yrs.service + sex, data=Salaries)))

# equivalence test for regression coef (Delta=5000), p-val = 0.478
equivBeta(Y = y, Xmatrix = X[,-1], DELTA = 5000)
```

```
# equivalence test for diffP2 (Delta=0.01), p-val = 0.232
equivdiffP2(Y = y, Xmatrix = X[,-1], DELTA = 0.01)

# equivalence test for standardized regression coef (Delta=0.10), p-val = 0.076
equivstandardBeta(Y = y, Xmatrix = X[,-1], DELTA = 0.10)
```

References

Altman, D. G. & Bland, J. M. (1995), 'Statistics notes: Absence of evidence is not evidence of absence', *The BMJ* **311**(7003), 485; https://doi.org/10.1136/bmj.311.7003.485.

Amrhein, V., Greenland, S. & McShane, B. (2019), 'Scientists rise up against statistical significance', *Nature* (567), 305–307; https://doi.og/10.1038/d41586–019–00857–9.

Barten, A. (1962), 'Note on unbiased estimation of the squared multiple correlation coefficient', *Statistica Neerlandica* **16**(2), 151–164.

Berger, J. O. (2013), *Statistical decision theory and Bayesian analysis*, Springer Science & Business Media.

Briggs, W. M., Nguyen, H. T. & Trafimow, D. (2019), The replacement for hypothesis testing, *in* 'International Conference of the Thailand Econometrics Society', Springer, pp. 3–17.

Bring, J. (1994), 'How to standardize regression coefficients', *The American Statistician* **48**(3), 209–213.

Campbell, H. & Gustafson, P. (2018*a*), 'Conditional equivalence testing: An alternative remedy for publication bias', *PLoS ONE* **13**(4), e0195145; https://doi.org/10.1371/journal.pone.0195145.

Campbell, H. & Gustafson, P. (2018*b*), 'What to make of non-inferiority and equivalence testing with a post-specified margin?', *arXiv preprint arXiv:1807.03413* .

Campbell, H. & Lakens, D. (2020), 'Can we disregard the whole model?', *in press - British Journal of Mathematical and Statistical Psychology* .

Consonni, G., Veronese, P. et al. (2008), 'Compatibility of prior specifications across linear models', *Statistical Science* **23**(3), 332–353.

Cramer, J. S. (1987), 'Mean and variance of R2 in small and moderate samples', *Journal of Econometrics* **35**(2-3), 253–266.

Dudgeon, P. (2017), 'Some improvements in confidence intervals for standardized regression coefficients', *Psychometrika* **82**(4), 928–951; https://doi.org/10.1007/s11336–017–9563–z.

Etz, A. (2015), 'Using bayes factors to get the most out of linear regression: A practical guide using R', *The Winnower* .
URL: *https://thewinnower.com/papers/using-bayes-factors-to-get-the-most-out-of-linear-regression-a-practical-guide-using-r*

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S. et al. (2012), 'Package car', *Vienna: R Foundation for Statistical Computing* .

Fraley, R. C. & Vazire, S. (2014), 'The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power', *PLoS ONE* **9**(10), e109019.

Fritz, C. O., Morris, P. E. & Richler, J. J. (2012), 'Effect size estimates: current use, calculations, and interpretation.', *Journal of Experimental Psychology: General* **141**(1), 2; https://doi.org/10.1037/a0024338.

Ghashim, E. & Boily, P. (2018), 'A ggplot2 Primer', *Data Action Lab - Data Science Report Series* .
URL: *https://www.data-action-lab.com/wp-content/uploads/2018/11/DSRS$_G$GP2.pdf*

Hung, H., Wang, S.-J. & O'Neill, R. (2005), 'A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials', *Biometrical Journal* **47**(1), 28–36.

Jeffreys, H. (1961), *The theory of Probability*, OUP Oxford.

Jones, J. A. & Waller, N. G. (2013), 'Computing confidence intervals for standardized regression coefficients.', *Psychological Methods* **18**(4), 435.

Keefe, R. S., Kraemer, H. C., Epstein, R. S., Frank, E., Haynes, G., Laughren, T. P., Mcnulty, J., Reed, S. D., Sanchez, J. & Leon, A. C. (2013), 'Defining a clinically meaningful effect for the design and interpretation of randomized controlled trials', *Innovations in Clinical Neuroscience* **10**(5-6 Suppl A), 4S.

Kelley, K. et al. (2007), 'Confidence intervals for standardized effect sizes: Theory, application, and implementation', *Journal of Statistical Software* **20**(8), 1–24.

Kühberger, A., Fritz, A. & Scherndl, T. (2014), 'Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size', *PLoS ONE* **9**(9), e105825.

Lakens, D. (2017), 'Equivalence tests: a practical primer for t-tests, correlations, and meta-analyses', *Social Psychological and Personality Science* **8**(4), 355–362.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R. & Jones, D. R. (2005), 'How vague is vague? a simulation study of the impact of the use of vague prior distributions in mcmc using winbugs', *Statistics in medicine* **24**(15), 2401–2428.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. & Berger, J. O. (2008), 'Mixtures of g-priors for Bayesian variable selection', *Journal of the American Statistical Association* **103**(481), 410–423.

Linde, M. & van Ravenzwaaij, D. (2019), 'baymedr: An R Package for the Calculation of Bayes Factors for Equivalence, Non-Inferiority, and Superiority Designs', *arXiv preprint arXiv:1910.11616* .

Lu, Y. & Westfall, P. (2019), 'Simple and flexible Bayesian inferences for standardized regression coefficients', *Journal of Applied Statistics* pp. 1–35.

Marszalek, J. M., Barber, C., Kohlhart, J. & Cooper, B. H. (2011), 'Sample size in psychological research over the past 30 years', *Perceptual and Motor Skills* **112**(2), 331–348.

Moon, K.-W. (2017), *Learn "ggplot2" Using Shiny App*, Springer International Publishing.

Morey, R. D., Rouder, J. N., Jamil, T. & Morey, M. R. D. (2015), 'Package 'BayesFactor", *URL http://cran/r-projectorg/web/packages/BayesFactor/BayesFactor* .

Morris, T. P., White, I. R. & Crowther, M. J. (2019), 'Using simulation studies to evaluate statistical methods', *Statistics in Medicine* **38**(11), 2074–2102.

Nagelkerke, N. J. et al. (1991), 'A note on a general definition of the coefficient of determination', *Biometrika* **78**(3), 691–692.

Nieminen, P., Lehtiniemi, H., Vähäkangas, K., Huusko, A. & Rautio, A. (2013), 'Standardised regression coefficient as an effect size index in summarising findings in epidemiological studies', *Epidemiology, Biostatistics and Public Health* **10**(4).

Rouder, J. N. & Morey, R. D. (2012*a*), 'Default Bayes factors for model selection in regression', *Multivariate Behavioral Research* **47**(6), 877–903; DOI: 10.1080/00273171.2012.734737.

Rouder, J. N. & Morey, R. D. (2012*b*), 'Default Bayes factors for model selection in regression', *Multivariate Behavioral Research* **47**(6), 877–903.

Rouder, J. N., Morey, R. D., Speckman, P. L. & Province, J. M. (2012), 'Default bayes factors for anova designs', *Journal of Mathematical Psychology* **56**(5), 356–374.

Schönbrodt, F. D. & Wagenmakers, E.-J. (2016), 'Bayes factor design analysis: Planning for compelling evidence', *Psychonomic Bulletin & Review* pp. 1–15.

Smithson, M. (2001), 'Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals', *Educational and Psychological Measurement* **61**(4), 605–632.

Tan Jr, L. (2012), 'Confidence intervals for comparison of the squared multiple correlation coefficients of non-nested models', *Thesis submitted in partial fulfillment of the requirements for the degree in Master of Science; The University of Western Ontario* .

Wellek, S. (2017), 'A critical evaluation of the current p-value controversy', *Biometrical Journal* **59**(5), 854–872.

West, S. G., Aiken, L. S., Wu, W. & Taylor, A. B. (2007), 'Multiple regression: Applications of the basics and beyond in personality research.', *Handbook of research methods in personality psychology (p. 573 – 601). The Guilford Press* .

Wiens, B. L. (2002), 'Choosing an equivalence limit for noninferiority or equivalence studies', *Controlled Clinical Trials* **23**(1), 2–14.

Yuan, K.-H. & Chan, W. (2011), 'Biases and standard errors of standardized regression coefficients', *Psychometrika* **76**(4), 670–690.

Zou, K. H., Tuncali, K. & Silverman, S. G. (2003), 'Correlation and simple linear regression', *Radiology* **227**(3), 617–628.