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

harlan.campbell@stat.ubc.ca

**ARTICLE HISTORY**

Compiled August 25, 2021

**Abstract**

We introduce equivalence testing procedures for standardized effect sizes in a linear regression analysis. Such tests are useful for confirming the lack of a meaningful association between a continuous outcome and predictors and may be particularly valuable if the predictors of interest are measured on different and completely arbitrary scales. We consider how to define valid hypotheses and how to calculate $p$-values for these equivalence tests. Via simulation study, we examine type I error rates and statistical power and compare the proposed frequentist equivalence testing with an alternative Bayesian testing approach.

**KEYWORDS**

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

## 1. Introduction

All too often, researchers will conclude that the effect of an explanatory variable, $X$, on an outcome variable, $Y$, is absent when a null-hypothesis significance test (NHST) yields a non-significant $p$-value (e.g., when the $p$-value $> 0.05$). Unfortunately, such a procedure is logically flawed. As the saying goes, "absence of evidence is not evidence of absence" (Hartung et al., 1983; Altman and Bland, 1995). Indeed, a non-significant result can simply be due to insufficient statistical power, and while a NHST can provide evidence to *reject* the null hypothesis, it cannot provide evidence to *accept* the null.

To properly conclude that an association between $X$ and $Y$ is absent or at most negligible (i.e., to confirm the *lack* of an association), the recommended frequentist tool, the equivalence test (also known as the "non-inferiority test" for one-sided testing), is well-suited (Wellek, 2010). Let $\theta$ be the parameter of interest representing the association between $X$ and $Y$. An equivalence test reverses the question that is asked in a NHST. Instead of asking whether we can reject the null hypothesis of no effect, (i.e., reject $H_0 : \theta = 0$), an equivalence test examines whether the magnitude of $\theta$ is at all meaningful by asking: Can we reject the possibility that $\theta$ is as large or larger than our smallest effect size of interest, $\Delta$? The null hypothesis for an equivalence test can therefore be defined as $H_0 : \theta \notin (-\Delta, \Delta)$. In other words, *equivalence* implies that $\theta$ is small enough that any non-zero effect would be at most equal to $\Delta$. To be clear, the interval $(-\Delta, \Delta)$, known as the "equivalence margin," represents the range of values for which $\theta$ can be considered negligible.

In psychology research and in the social sciences, where the practice of equivalence testing is relatively new –but now "rapidly expanding" (Koh and Cribbie, 2013)– there are many questions about how to best conduct and interpret equivalence tests. For instance, defining and justifying the equivalence margin is, for many researchers, one of the "most difficult issues" (Hung et al., 2005). If the margin is too large, any claim of equivalence will be considered meaningless. On the other hand, if the margin is somehow too small, the probability of declaring equivalence will be substantially reduced (Campbell and Gustafson, 2018b; 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).

A researcher should ideally use validated and well-scaled measures where the units of measurement are well understood. However, in many scenarios (and very often in the social sciences), the parameters of interest are measured on different and completely arbitrary scales. This makes the task of defining the equivalence margin more challenging. Without interpretable units of measurement, defining and justifying an appropriate equivalence margin can be all but impossible. How can one determine the "smallest effect size of interest" (Lakens et al., 2018) in units that have no particular meaning?

When working with parameters measured on arbitrary scales, researchers will often prefer to work with standardized effect sizes to aid with interpretation (Baguley, 2009). It therefore stands to reason that, for equivalence testing in such a situation, it would also be preferable to define the equivalence margin in terms of a standardized effect size. Unfortunately, equivalence testing with standardized effects is not always straightforward. Contrary to certain recommendations, one cannot merely define the equivalence margin in terms of a standardized effect size and proceed as normal. For example, as we will demonstrate later, Lakens (2017)'s suggestion that, for a two-sample test for the equivalence in means, one may simply define the equivalence margin in terms of the observed standard deviation is incorrect. The equivalence margin cannot be defined as a function of the observed data as this will invalidate the test. Instead, one must define the parameter of interest to be the standardized parameter, such that the randomness associated with the standardization is properly accounted for.

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. However, there are no established equivalence tests for standardized regression coefficients. Therefore, our objective in this paper is to establish equivalence testing procedures for standardized effect sizes in a linear regression. Note that all of the tests presented in this paper are derived from inverting the non-centrality parameter confidence in-

tervals described by Kelley et al. (2007). In Section 2 we consider how to define valid hypotheses and calculate $p$-values for these tests, and in Section 3 we conduct a small simulation study to better understand the test's operating characteristics.

Several Bayesian methods (e.g., the "default" Bayes Factor (Rouder and Morey, 2012a), the Bayes factor interval null procedure (Morey and Rouder, 2011), and the Bayesian highest density interval region of practical equivalence procedure ("HDI-ROPE") (Kruschke, 2011)) have also been proposed for establishing evidence for the absence of an effect. The pros and cons of frequentist versus Bayesian testing methods are a topic of great debate (see recently, for example: Linde et al. (2021) and Campbell and Gustafson (2021)). While the focus of this paper is frequentist equivalence testing, in Section 4, we briefly review one of the proposed Bayesian alternatives for establishing equivalence of linear regression parameters. Finally, in Section 5, we demonstrate how the various testing methods can be applied in practice with the analysis of two example datasets. We conclude in Section 6 with general recommendations.

## 2. Equivalence testing for regression coefficients

### *2.1. An equivalence test for unstandardized regression coefficients*

Consider a multivariable linear regression where $Y$ is the outcome variable and $X$ is the $N \times (K+1)$ fixed covariate matrix (with a column of 1s for the intercept) (Azen and Budescu, 2009). Going forward, we use the notation $X_{i\times}$ to refer to all $K+1$ values corresponding to the $i$-th observation; and $X_k$ to refer to the $k$-th covariate. Note that the regression may include both categorical and continuous covariates. For example, suppose a researcher is looking to investigate possible predictors of anxiety among high-school students. In this hypothetical study, $Y$ might be a student's score on an anxiety assessment questionnaire; $X_1$ might be a binary variable indicating whether or not the student received counselling services (0 = "did not receive counselling; 1 = "did receive counselling"); $X_2$ might be a continuous covariate corresponding to the student's age in years; and $X_3$ might be a continuous covariate corresponding to the

student's household income in dollars.

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

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

where $\beta = (\beta_0, \beta_1, \ldots, \beta_K)^T$ is a parameter vector of $K$ regression coefficients, and $\sigma^2$ is the population variance parameter. Least squares estimates for the linear regression model are denoted by $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_K})^T$, and $\hat{\sigma}^2$; see equations (17) and (18) in the Appendix for full details.

Recall that the interpretation of the $\beta_k$ coefficient is the rate of change in $Y$ per unit change in $X_k$ (or slope) when all other variables are held constant. For example, in our hypothetical study about anxiety, the $\beta_1$ coefficient would be interpreted as the number of additional points on the anxiety assessment score associated with a student receiving counselling services, the $\beta_2$ coefficient would be interpreted as the number of additional points on the anxiety assessment score associated with every additional year, and the $\beta_3$ coefficient would be interpreted as the number of additional points associated with every additional dollar.

An equivalence test for an unstandardized regression coefficient asks the following question: Can we reject the possibility that $\beta_k$ is as large or larger than our smallest effect size of interest? Formally, the null and alternative hypotheses for the equivalence test are stated as:

$$\begin{aligned} H_0 &: \beta_k \leq \Delta_{k,lower} \quad \text{or} \quad \beta_k \geq \Delta_{k,upper}, \quad \text{vs.} \\ H_1 &: \beta_k > \Delta_{k,lower} \quad \text{and} \quad \beta_k < \Delta_{k,upper}, \end{aligned} \tag{2}$$

where the equivalence margin, $[\Delta_{k,lower}, \Delta_{k,upper}]$, defines the range of values considered negligible, for $k$ in $1, \ldots, K$. Often, the equivalence margin will be symmetrical such that $\Delta_k = \Delta_{k,upper} = -\Delta_{k,lower}$, but this is not necessarily so.

Returning to the hypothetical example, suppose that in order for the impact of

counselling services to be considered at all meaningful, the services would have to be associated with a minimum two point difference on the anxiety assessment questionnaire. In this case, the researcher would simply define $\Delta_1 = 2$ and the equivalence margin for $k = 1$ would be [-2, 2]. For the other covariates, $k = 2$ and $k = 3$, it may not be as simple to define an equivalence margin since $\beta_2$ and $\beta_3$ are measured in terms of "points per year" and "points per dollar". To define an appropriate margin, the researcher would have to ask: What are the minimum meaningful per year and per dollar numbers of points to consider?

Recall that there is a one-to-one correspondence between an equivalence test and a confidence interval (CI); see Dixon et al. (2018) for details. As such, an equivalence test can be constructed by simply inverting a confidence interval. For example, we will reject the above null hypothesis ($H_0 : \beta_k \leq \Delta_{k,lower} \quad \text{or} \quad \beta_k \geq \Delta_{k,upper}$), at a $\alpha = 0.05$ significance level, whenever a (1 - $2\alpha$) = 90% CI for $\beta_k$ fits entirely within $[\Delta_{k,lower}, \Delta_{k,upper}]$. Inverting the CI for $\beta_k$ leads to two one-sided $t$-tests (TOST) with the following $p$-values:

$$p_k^{[1]} = F_t\left(\frac{\widehat{\beta_k} - \Delta_{k,lower}}{SE(\widehat{\beta_k})}, N - K - 1\right), \quad \text{and} \quad p_k^{[2]} = F_t\left(\frac{\Delta_{k,upper} - \widehat{\beta_k}}{SE(\widehat{\beta_k})}, N - K - 1\right), \tag{3}$$

for $k$ in 0,...,$K$; where $F_t(\ \cdot\ ; df)$ denotes the cumulative distribution function (cdf) of the $t$-distribution with $df$ degrees of freedom, and where $SE(\widehat{\beta_k}) = \hat{\sigma}\sqrt{[(X^TX)^{-1}]_{kk}}$. In order to reject the equivalence test null hypothesis ($H_0 : \beta_k \leq \Delta_{k,lower} \quad \text{or} \quad \beta_k \geq \Delta_{k,upper}$), both $p$-values, $p_k^{[1]}$ and $p_k^{[2]}$, must be less than $\alpha$. As such, for the $k$-th regression coefficient, $\beta_k$, a single overall $p$-value for the equivalence test can be calculated as: $p\text{-value}_k = \max(p_k^{[1]}, p_k^{[2]})$.

### *2.2. An equivalence test for standardized regression coefficients*

It has been previously suggested that the bounds of an equivalence margin can be defined in terms of a sample estimates; see (Lakens, 2017) and Lakens et al. (2020). This is incorrect. To illustrate why, suppose that, for a two-sample test for the difference

in means, $\mu_d$, one were to define a symmetric equivalence margin, $[-\Delta, \Delta]$, in terms of the observed standard deviation, $\hat{\sigma}$, such that $\Delta = 0.5 \times \hat{\sigma}$.

Recall that, in order for a hypothesis test to be valid, the hypotheses must be statements about the unobserved parameters and not about the observed data. Therefore, since the hypotheses for the test in our example, $H_0 : |\mu_d| \geq 0.5 \times \hat{\sigma}$, vs. $H_1 : |\mu_d| < 0.5 \times \hat{\sigma}$, are defined as functions of the observed data (i.e., in terms of $\hat{\sigma}$), the test is invalid. Instead, the correct procedure is to define the parameter of interest, $\theta$, to be the standardized effect size, e.g. define $\theta = \mu_d/\sigma$. Then, one can define the margin on the standardized scale without invalidating the hypotheses. To be clear, $H_0 : |\theta| \geq 0.5$ vs. $H_1 : |\theta| < 0.5$ is a completely valid equivalence test, while $H_0 : |\mu_d| \geq 0.5 \times \hat{\sigma}$, vs. $H_1 : |\mu_d| < 0.5 \times \hat{\sigma}$ is invalid. While in practice, the difference between these may be small, it should nevertheless be acknowledged since one should always (ideally) take into account for the uncertainty in the sample standard deviation estimate.

Returning now to linear regression, in order to define the equivalence margin in terms of a standardized effect size, we will define the parameter of interest to be $\mathcal{B}_k$, the standardized regression coefficient. Standardizing a regression coefficient is done by multiplying the unstandardized regression coefficient, $\beta_k$, 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:

$$\mathcal{B}_k = \beta_k \frac{s_k}{\sigma_Y}, \tag{4}$$

where $\sigma_Y = (\beta^T \text{Cov}(X)\beta + \sigma^2)^{\frac{1}{2}}$ and $s_k$ is the standard deviation of $X_k$. This can be estimated by:

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

where and $\widehat{\sigma_Y}$ is the standard deviation of $y$. Note that in the psychometric literature, standardized regression coefficients are often known as Beta-coefficients, while the

conventional unstandardized regression coefficients are often called B-coefficients.

An equivalence test for $\mathcal{B}_k$ can be defined by the following null and alternative hypotheses:

$$H_0 : \mathcal{B}_k \leq \Delta_{k,lower} \quad \text{or:} \quad \mathcal{B}_k \geq \Delta_{k,upper}, \quad \text{vs.}$$
$$H_1 : \mathcal{B}_k > \Delta_{k,lower} \quad \text{and:} \quad \mathcal{B}_k < \Delta_{k,upper},$$

where the equivalence margin is $[\Delta_{k,lower}, \Delta_{k,upper}]$. We make use of noncentrality interval estimation (NCIE) (Smithson, 2001) to construct the equivalence test. By inverting a confidence interval for $\mathcal{B}_k$ (see Kelley et al. (2007) for details), we obtain the following, for $k$ in 1,...,$K$:

$$\mathbb{p}_k^{[1]} = F_t\left(\frac{\widehat{\mathcal{B}_k}}{SE(\widehat{\mathcal{B}_k})}; df = N-K-1, ncp = \Delta_{k,lower}\frac{\sqrt{N\left(1-R^2_{X_kX_{-k}}\right)}}{\sqrt{1-((1-R^2_{X_kX_{-k}})\Delta^2_{k,lower}+R^2_{YX_{-k}})}}\right), \tag{6}$$

and:

$$\mathbb{p}_k^{[2]} = F_t\left(\frac{-\widehat{\mathcal{B}_k}}{SE(\widehat{\mathcal{B}_k})}; df = N-K-1, ncp = -\Delta_{k,upper}\frac{\sqrt{N\left(1-R^2_{X_kX_{-k}}\right)}}{\sqrt{1-((1-R^2_{X_kX_{-k}})\Delta^2_{k,upper}+R^2_{YX_{-k}})}}\right),$$

with:

$$SE(\widehat{\mathcal{B}_k}) = \sqrt{\frac{(1-R^2_{YX})}{(1-R^2_{X_kX_{-k}})(N-K-1)}}, \tag{7}$$

where $R^2_{YX}$ is the coefficient of determination from the linear regression of $Y$ predicted from $X$; $R^2_{X_kX_{-k}}$ is the coefficient of determination from the linear regression of $X_k$ predicted from the remaining $K-1$ regressors; $R^2_{YX_{-k}}$ is the coefficient of determination from the linear regression of $Y$ predicted from all but the $k$-th covariate; and $F_t(\ \cdot\ ; df, ncp)$ denotes the cdf of the non-central $t$-distribution with $df$ degrees of freedom and non-centrality parameter $ncp$. (Note that when $ncp = 0$, the non-central $t$-distribution is equivalent to the central $t$-distribution.) For the $k$-th covariate, the null hypothesis, $H_0 : \mathcal{B}_k \leq \Delta_{k,lower}$ or: $\mathcal{B}_k \geq \Delta_{k,upper}$, is rejected if and only if the

$p$-value, $p\text{-value}_k = max(\mathbb{p}_k^{[1]}, \mathbb{p}_k^{[2]})$, is less than $\alpha$.

### *2.3. 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 of the added variable (Dudgeon, 2017). We define $\text{diff}R_k^2 = R_{YX}^2 - R_{YX_{-k}}^2$. In a linear regression model, the $R_{YX}^2$ 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 $R_{YX}^2$ statistic's ubiquitous use, its corresponding population parameter, which we will denote as $P_{YX}^2$, 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).

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

$$\text{diff}R_k^2 = \widehat{\mathcal{B}_k}^2(1 - R_{X_k X_{-k}}^2). \tag{8}$$

Similarly, we have that for the corresponding population parameter: $\text{diff}P_k^2 = \mathcal{B}_k{}^2(1 - P_{X_k X_{-k}}^2)$.

It may be preferable to consider an effect size (and what can be considered a "negligible difference") in terms of $\text{diff}P_k^2$ instead of in terms of $\mathcal{B}_k$. If this is the case, one can conduct a non-inferiority test (a one-sided equivalence test), for $k$ in 1,...,$K$, with the following hypotheses:

$H_0 : 1 > \text{diff}P_k^2 \geq \Delta_k, \quad \text{vs.}$
$H_1 : 0 \leq \text{diff}P_k^2 < \Delta_k.$

The $p$-value for this non-inferiority test is obtained by replacing $\widehat{\mathcal{B}}_k$ with

$\sqrt{\text{diff}R_k^2/(1-R^2_{X_kX_{-k}})}$ and can be calculated, for fixed regressors as follows:

$$p\text{−value}_k = 1-F_t\left(\frac{\sqrt{(N-K-1)\text{diff}R_k^2}}{\sqrt{(1-R^2_{YX})}}; df = N-K-1, ncp = \frac{\sqrt{N\Delta}}{\sqrt{\left(1-\Delta+R^2_{X_kX_{-k}}\right)}}\right). \tag{9}$$

In related work, Campbell and Lakens (2021) introduce a non-inferiority test (a one-sided equivalence test) to test the null hypotheses:

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

Note that when $K = 1$, we have that: $\text{diff}P_1^2 = P^2_{YX} = {\mathcal{B}_1}^2$, and as such, the three tests will be entirely equivalent, assuming that the desired equivalence bounds used for the equivalence test of $\mathcal{B}_1$ are symmetric (i.e., assuming that $\Delta_{1,upper} = -\Delta_{1,lower}$).

### *2.4. 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 null hypothesis significance test (NHST) (i.e., calculate a $p$-value, $p_{NHST}$, using equation (19) or (20)) and only proceed to the equivalence test (i.e., calculate a second $p$-value, $p_{EQUIV}$, using equation (6)) if the NHST fails to reject the null. Such a two-stage sequential testing scheme has been discussed by Seaman and Serlin (1998) and more recently by Campbell and Gustafson (2018a) under the name of "conditional equivalence testing" (CET).

Under the two-step CET scheme, if the first $p$-value, $p_{NHST}$, is less than the type 1 error $\alpha$-threshold (e.g., if $p_{NHST} < 0.05$), one concludes with a "positive" finding: $\mathcal{B}_k$ is significantly different than 0. On the other hand, if the first $p$-value, $p_{NHST}$, is greater than $\alpha$ and the second $p$-value, $p_{EQUIV}$, is smaller than $\alpha$ (e.g., if $p_{NHST} \geq 0.05$ and $p_{EQUIV} < 0.05$), one concludes with a "negative" finding: there is evidence of statistically significant equivalence, i.e., $\mathcal{B}_k$ is at most negligible. If both

$p$-values are larger than $\alpha$, the result is inconclusive: there are insufficient data to support either finding. In this paper, we are not advocating for (or against) the CET approach, but simply use it to facilitate a comparison with Bayes Factor testing (see Section 4) which also categorizes outcomes as either positive, negative or inconclusive.

## 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, as described in Section 2.2. The equivalence test in the simulation study involves a symmetric equivalence margin, $[-\Delta, \Delta]$, such that: $H_0 : |\mathcal{B}_1| \geq \Delta$, vs. $H_1 : |\mathcal{B}_1| < \Delta$. In the simulation study, $p$-values for this test are calculated from equation (6).

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 $\beta = (-0.20, 0.10, 0.20)^T$ or $\beta = (0.20, 0.10, 0.14, -0.10, -0.10)^T$; and
- one of three variances: $\sigma^2 = 0.05$ , $\sigma^2 = 0.15$, or $\sigma^2 = 0.50$.

The outcome variable was simulated from a Normal distribution such that $Y_i \sim \text{Normal}(X_{i\times}^T \beta, \sigma^2), \forall\ i = 1, ..., N$. Depending on the specific value of $\sigma^2$, 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 $\beta = (-0.20, 0.00, 0.20)^T$, for $K = 2$, and $\beta = (0.20, 0.00, 0.14, -0.10, -0.10)^T$, for $K = 4$. For all of these additional scenarios, $\sigma^2$ 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 and Vazire, 2014; Marszalek et al., 2011)) and so that we would obtain three unique values for

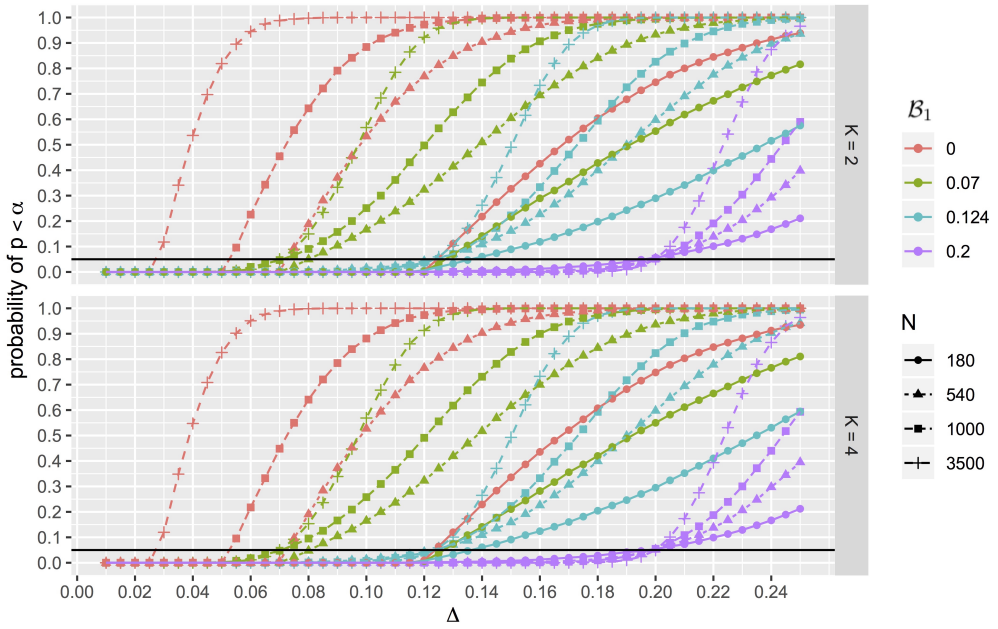


**Figure 1.** Simulation Study 1 - Upper panel shows results for $K = 2$; Lower panel shows results for $K = 4$. The solid horizontal black line indicates the desired type 1 error of $\alpha = 0.05$.

$\mathcal{B}_1$ 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 $\Delta$ (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)).

The simulation study was done in R statistical software using the default simulation routines (R Core Team, 2020) and the code is available in the Supplementary Material. Figure 1 and Figure 2 show results from the simulation study. Note that Figure 2 is an "inset" (i.e., a "magnified portion") of Figure 1 to allow a focus on the type 1 error rate.

First, let us consider results obtained when $\mathcal{B}_1 = 0.200$. We see that when the equivalence bound, $\Delta$, also equals 0.200 (i.e., when $\mathcal{B}_1 = \Delta = 0.200$), the type 1 error rate is exactly 0.05, as it should be, for all $N$. This situation represents the boundary

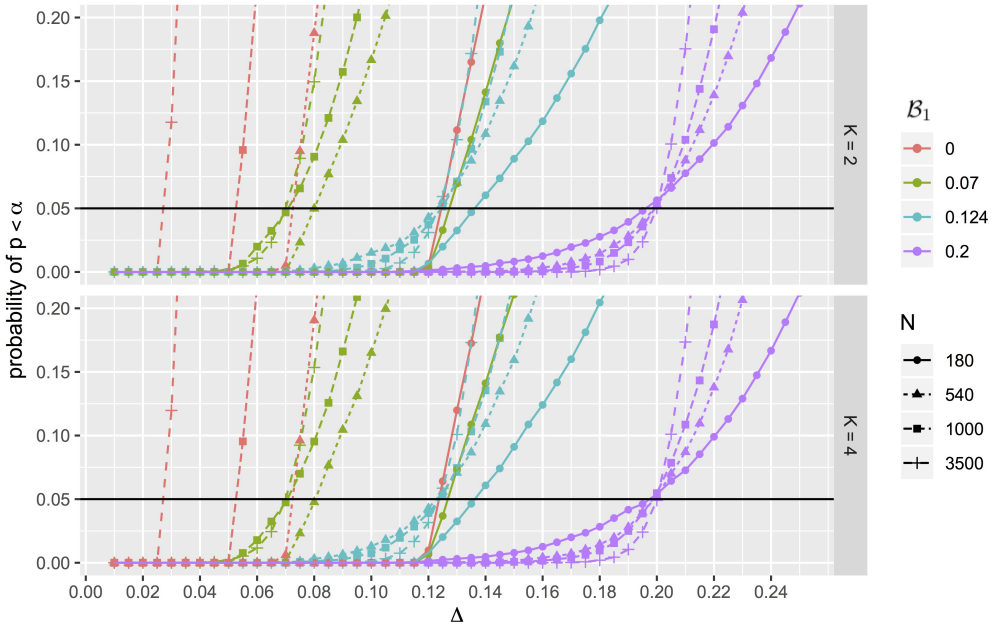


**Figure 2.** Simulation Study 1 - Note that this Figure is an "inset" (i.e., a "magnified portion") of Figure 1. Upper panel shows results for $K = 2$; lower panel shows results for $K = 4$. Both plots are presented with a restricted vertical-axis to better show the type 1 error rates. The solid horizontal black line indicates the desired type 1 error of $\alpha = 0.05$.

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 $\sigma^2$ is relatively large, the 90% CI for $\mathcal{B}_1$ may be far too wide to fit entirely within the equivalence margin.

In order for the test to have any substantial power, $\mathcal{B}_1$ must be substantially smaller than $\Delta$. It is important to note that limited sample sizes may prevent the proposed equivalence test from having sufficient power to rule out any truly "negligible" effect. As expected, the power of the test increases with larger values of $N$, smaller

values of $K$, and larger values of $\Delta$. In some cases, the power is strictly zero. For example, when $\mathcal{B}_1 = 0.00$, $N = 180$ and $K = 2$, the $\Delta$ must be greater or equal to 0.10 for there to any possibility of rejecting $H_0$. Otherwise, for $\Delta < 0.10$, the power is zero; see Figure 2.

In the Appendix, we show results from an alternate version of the simulation study where the covariates are unbalanced and are correlated. The results obtained from this alternate version are very similar.

## 4. A Bayesian alternative for establishing equivalence in a linear regression

As discussed in the Introduction, there are a multitude of different Bayesian methods available for establishing equivalence. Rouder and Morey (2012a)'s proposed "default" Bayes factors (based on the work of Liang et al. (2008)) is one approach that has proven to be particularly popular in psychology research for linear regression models (Etz, 2015; Morey et al., 2015). We briefly review the default Bayes factors approach for linear regression in order to consider how it might compare to the frequentist equivalence tests proposed.

The Bayes Factor, $BF_{10}$, is defined as the probability of the data under the alternative model relative to the probability of the data under the null model:

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

with the "10" subscript indicating that the alternative model (i.e., "Model 1") is being compared to the null model (i.e., "Model 0"). Interpretation of the Bayes factor is straightforward. 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 (i.e., define $\Pr(\ Model\ 0)$ and $\Pr(\ Model\ 1)$) for all model parameters (Consonni et al., 2008). Rouder and Morey (2012a) suggest using a Jeffreys-Zellner-Siow (JZS) "objective prior" and provide and overview of its

various advantages; see also Heck (2019). A scaled version of this prior, whereby the $r$ scale parameter is set equal to a specific value allows one to specify prior beliefs about the expected effect size. For instance, For instance, the BayesFactor package uses the scaled-JZS prior with a default $r = \sqrt{2}/4$, so as to correspond to a prior belief in a medium effect size.

To test the $k$-th regression coefficient, $\beta_k$, in a multivariable linear regression model, one computes a Bayes factor for a model that includes the $k$-th covariate against a model that does not, such that:

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

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

where $\beta_{-k}$ ($X_{i,-k}$) is the vector (matrix) of regression coefficients (covariates), with the $k$-th coefficient (covariate) omitted. If this Bayes factor were to be above a certain threshold (e.g. if $BF_{10} > 6$), one would conclude that $\beta_k$ is different than 0 (i.e., evidence in support of Model 1). On the other hand, if this Bayes factor were to be bellow a certain threshold (e.g. if $BF_{10} < 1/6$), one would conclude that there is evidence that $\beta_k = 0$ (i.e., evidence in support of Model 0). A threshold of 3 (or 1/3) can be considered "moderate evidence," a threshold of 6 (or 1/6) can be considered "strong evidence," and a threshold of 10 (or 1/10) can be considered "very strong evidence" (Jeffreys, 1961).

Note that, just like frequentist testing of standardized regression coefficients, Bayes factor testing of regression coefficients is scale invariant such that the Bayes factor is entirely independent of the units of the regression coefficients. To be clear, the the Bayes factor will not change if the model variables ($Y$, $X_k$, for $k = 1, \ldots, K$) are measured in different units. Also note that some recent work (Tendeiro and Kiers, 2019) has argued that the posterior odds, rather than the Bayes factor, should be used for Bayesian null hypothesis testing. Rouder and Morey (2012a) discuss this in a section of their paper entitled "Adding value through prior odds."

In the Appendix we conduct a second simulation study (see Simulation Study

2) to compare a CET frequentist testing scheme (based on NHST and equivalence testing of standardized regression coefficients) to the Bayesian testing approach based on Rouder and Morey (2012a)'s default Bayes factors. The results of the simulation study suggest that, given the same data, both approaches will often arrive at the same overall conclusion (i.e., both approaches will obtain either a positive, negative or inconclusive result). The level of agreement however is highly sensitive to the choice of equivalence margin and the choice of the Bayes factor evidence threshold. While we do not consider the impact of selecting different priors with the Bayes factors, it is reasonable to assume that the level of agreement between Bayes factor testing and frequentist testing will also be rather sensitive to the chosen priors, particularly when $N$ is small; see Berger (2013).

## 5. Practical Examples

### *5.1. 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 and 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, 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. For illustration purposes, we consider both a simple linear regression ($K = 1$) and a multivariable linear regression ($K = 6$).

### *5.1.1. A simple linear regression*

Consider a simple linear regression (i.e., $Y \sim \text{Normal}(\beta_0 + \beta_1 X_1, \sigma^2)$) for the association between salary ($Y$, measured in \$) and sex ($X_1$, where "0" corresponds to "female," and "1" corresponds to "male."). Standard least squares estimation results in the following parameter estimates: $\widehat{\beta}_0 = 101002$, $\text{SE}(\widehat{\beta}_0) = 4809$, and $\widehat{\beta}_1 = 14088$, $\text{SE}(\widehat{\beta}_1) = 5065$ (see equation (17)); $\widehat{\sigma} = 30034.61$ (see equation (18)); $\widehat{\mathcal{B}}_1 = 0.14$ (see equation (5)), $\text{SE}(\widehat{\mathcal{B}}_1) = 0.05$ (see equation (7)); and $R^2_{YX} = \text{diff}R^2_1 = 0.019$.

We can 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. Suppose that any difference of less than $\Delta = \$5,000$ is considered negligible. Then a $p$-value for the equivalence test, $H_0 : |\beta_1| \geq 5000$ vs. $H_1 : |\beta_1| < 5000$, can be calculated using equation (3). We obtain a $p$-value = 0.963 and therefore fail to reject the equivalence test null hypothesis.

If it were not possible to determine a specific number of dollars to be considered negligible, we could conduct an equivalence test for the standardized regression coefficient, $\mathcal{B}_1$. Suppose we consider very small effect sizes to be negligible and define an equivalence margin as [-0.10, 0.10]. Then we can calculate a $p$-value for $H_0 : |\mathcal{B}_1| \geq 0.10$ vs. $H_1 : |\mathcal{B}_1| < 0.10$ as per equation (6). We obtain

$p$-value $= max(\mathbb{p}_1^{[1]}, \mathbb{p}_1^{[2]}) = 0.780$, where:

$$\begin{aligned}
\mathbb{p}_1^{[1]} &= F_t\left(\frac{\widehat{\mathcal{B}_k}}{SE(\widehat{\mathcal{B}_k})}; df = N-K-1, ncp = \Delta_{1,lower}\frac{\sqrt{N\left(1-R^2_{X_kX_{-k}}\right)}}{\sqrt{1-((1-R^2_{X_kX_{-k}})\Delta^2_{1,lower}+R^2_{YX_{-k}})}}\right) \\
&= F_t\left(\frac{0.14}{0.05}; df = 397-1-1, ncp = -0.10\frac{\sqrt{397\,(1-0)}}{\sqrt{1-((1-0)\times 0.01+0)}}\right) \\
&= F_t\left(2.78, df = 395, ncp = -2.00\right) < 0.001 \\
&< 0.001 \qquad (13)
\end{aligned}$$

$$\begin{aligned}
\mathbb{p}_1^{[2]} &= F_t\left(\frac{-\widehat{\mathcal{B}_k}}{SE(\widehat{\mathcal{B}_k})}; df = N-K-1, ncp = -\Delta_{1,upper}\frac{\sqrt{N\left(1-R^2_{X_kX_{-k}}\right)}}{\sqrt{1-((1-R^2_{X_kX_{-k}})\Delta^2_{1,upper}+R^2_{YX_{-k}})}}\right) \\
&= F_t\left(-2.78, df = 395, ncp = -2.00\right) = 0.780. \\
&= 0.780. \qquad (14)
\end{aligned}$$

Alternatively, we could 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 ($H_0 : \text{diff}P_1^2 \geq \Delta$ vs. $H_1 : \text{diff}P_1^2 < \Delta$). Setting $\Delta = 0.10^2 = 0.01$, we obtain a $p$-value (as per equation (9)) identical to what we obtained with the previous test:

$$\begin{aligned}
p-\text{value} &= 1 - F_t\left(\frac{\sqrt{(N-K-1)\text{diff}R_k^2}}{\sqrt{(1-R^2_{YX})}}; df = N-K-1, ncp = \frac{\sqrt{N\Delta}}{\sqrt{\left(1-\Delta+R^2_{X_kX_{-k}}\right)}}\right) \\
&= 1 - F_t\left(\frac{\sqrt{(397-1-1)0.019}}{\sqrt{(1-0.019)}}; df = 397-1-1, ncp = \frac{\sqrt{397\times 0.01}}{\sqrt{(1-0.01+0)}}\right) \\
&= \quad 0.780. \qquad (15)
\end{aligned}$$

Bayes factors are easy to compute as well. With the BayesFactor package and the "regressionBF" function (with the default prior-scale $r = \sqrt{2}/4$), we obtain a $BF_{10} =$

4.5 which suggests that the alternative model (= the model with "sex" 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.[1]

### *5.1.2. A multivariable linear regression*

Now consider a multivariable linear regression model, with $K = 6$:

$$Y \sim \text{Normal}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6, \sigma^2), \qquad (16)$$

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 1 lists parameter estimates obtained by standard least squares estimation and Table 2 lists the $p$-values for each of the hypothesis tests we consider as well as Bayes factors. The Bayes factors are calculated using the "regressionBF" function from the BayesFactor package (with the default prior-scale $r = \sqrt{2}/4$). We obtain a Bayes factor for $k = 1$ of $B_{10} = 1/3.9$, indicating only moderate evidence in favour of the null model. This corresponds to an "inconclusive" result with a Bayes factor threshold of 6, or 10 (or any threshold higher than 3.9). The result for $k = 1$ 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

[1]The apparent contradiction can be explained by the fact that the two "default BF" functions are using different "default priors." The "regressionBF" function (as we are using it, see Appendix) 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 Section 13 of Rouder et al. (2012) for details.

conclude that, when controlling for observed confounders, there are insufficient data to support either an association, or the lack of an association, between sex and salary.

| $k$ | covariate | $\beta_k$ | $SE(\hat{\beta}_k)$ | $\mathcal{B}_k$ | $SE(\widehat{\mathcal{B}_k})$ |
|---|---|---|---|---|---|
| 0 | intercept | 65955.23 | 4588.60 | - | - |
| 1 | sex (male) | 4783.49 | 3858.67 | 0.047 | 0.038 |
| 2 | years since Ph.D. | 535.06 | 240.99 | 0.228 | 0.103 |
| 3 | years of service | -489.52 | 211.94 | -0.210 | 0.091 |
| 4 | discipline (applied) | 14417.63 | 2342.88 | 0.237 | 0.039 |
| 5 | rank (Asst. Prof.) | 12907.59 | 4145.28 | 0.157 | 0.050 |
| 6 | rank (Prof.) | 45066.00 | 4237.52 | 0.700 | 0.066 |
| | | | $\hat{\sigma} = 22538.65$ | | $R^2_{Y,X} = 0.455$ |

**Table 1.** Parameter estimates obtained by standard least squares estimation for the full multivariable linear regression model.

| $k$ | $\mathcal{B}_k$ | $p_{NHST}$ | $p_{EQUIV}$ $\Delta = 0.10$ | $BF_{10}$ $r = \sqrt{2}/4$ | CET conclusion $\alpha = 0.05$ | Bayesian conclusion BF threshold = 6 |
|---|---|---|---|---|---|---|
| 1 | 0.05 | 0.216 | 0.076 | 1/3.9 | Inconclusive | Inconclusive |
| 2 | 0.23 | 0.027 | 0.892 | 1.4 | Positive | Inconclusive |
| 3 | -0.21 | 0.021 | 0.885 | 1.7 | Positive | Inconclusive |
| 4 | 0.24 | $< 0.001$ | 1.000 | $6.5 \times 10^6$ | Positive | Positive |
| 5 | 0.16 | 0.002 | 0.868 | 13.6 | Positive | Positive |
| 6 | 0.70 | $< 0.001$ | 1.000 | $1.8 \times 10^{20}$ | Positive | Positive |

**Table 2.** Calculated values and conclusions for both frequentist and Bayesian testing for the salaries multivariable linear regression model.

### *5.2. Factors in hominid brain evolution*

As a second example, we follow Rouder and Morey (2012a) (and Heck (2019)) in reanalyzing a dataset first presented by Bailey and Geary (2009). The dataset consists of a information from a sample of $N = 175$ hominid crania aged between 10 thousand and 1.9 million years. The sample includes a total of $N = 175$ observations with cranium capacity ranging from $475cm^3$ to $1880cm^3$. The outcome of interest is the cranium capacity, a continuous variable, measured in $cm^3$, which ranges from $475cm^3$ to $1880cm^3$. We also consider 4 covariates of interest:

(1) the local climate variation ($X_1$ : "local climate", in degrees Celsius ranging between 8 and 47), which is measured as the difference between the highest mean monthly high and the lowest mean monthly low temperature, in degrees Celsius, for a location near to where the fossil cranium was discovered;

(2) the global average temperature ($X_2$ : "global climate", in standard deviations, ranging between 0.21 and 0.43), as indicated by the standard deviation of the $\delta^{18}O$ value (a metric calculated based on the record of oxygen isotopes) for the 200,000 years before the fossil was dated;

(3) the parasite load, ($X_3$ : "parasites", integer between 0 and 7), which is measured as the total number of types of parasites considered to be potentially harmful to humans in the region where the fossil cranium was discovered; and

(4) the population density, ($X_4$ : "pop. density", integer between 13 and 141) of the group the hominid lived within, as measured by " a population density proxy using the number of individuals living in surrounding areas during the hominids ancestral history".

We fit the data with a multivariable linear regression: $Y \sim \text{Normal}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4, \sigma^2)$. Table 3 lists parameter estimates obtained by standard least squares estimation and Table 4 lists the $p$-values for each of the hypothesis tests we consider as well as Bayes factors. The Bayes factors are calculated using the "regressionBF" function from the BayesFactor package (with the default prior-scale $r = \sqrt{2}/4$). Note that the Bayes factors we obtain are somewhat different than those obtained by Rouder and Morey (2012a). This is due to the fact that Rouder and Morey (2012a) use a prior-scale of $r = 1$, whereas we use $r = \sqrt{2}/4$. With a prior-scale of $r = 1$, one obtains Bayes factors of 1/12.9, $9.4 \times 10^7$, 1/4.4, and $6.3 \times 10^{13}$, for $k = 1, \ldots, 4$, respectively.

With $\alpha = 0.05$ and a Bayes factor threshold of 6, each of the frequentist CET conclusions match each of the Bayesian conclusions. Regardless of what approach is used, one can conclude that the regression analysis indicates evidence for the effect of population density ($p_{NHST} < 0.001$, $BF_{01} = 5.1 \times 10^{13}$ (with $r = \sqrt{2}/4$)) and of global climate ($p_{NHST} < 0.001$, $BF_{01} = 8.9 \times 10^7$ (with $r = \sqrt{2}/4$)), and evidence for a lack of effect of local climate ($p_{EQUIV}$=0.014 (with $\Delta$=0.10), $BF_{01} = 1/11$ (with $r = \sqrt{2}/4$)). Results are inconclusive with regards to the effect (or lack thereof) of parasites ($p_{NHST}$= 0.144, $p_{EQUIV}$=0.276 (with $\Delta$=0.10), $BF_{01} = 1/3.8$ (with $r = \sqrt{2}/4$)).

| $k$ | covariate | $\beta_k$ | $SE(\hat{\beta}_k)$ | $\mathcal{B}_k$ | $SE(\widehat{\mathcal{B}_k})$ |
|---|---|---|---|---|---|
| 0 | intercept | 261.70 | 97.72 | - | - |
| 1 | local climate | 0.15 | 1.62 | 0.004 | 0.045 |
| 2 | global climate | 1871.75 | 271.82 | 0.360 | 0.052 |
| 3 | parasites | -9.26 | 6.30 | -0.073 | 0.049 |
| 4 | pop. density | 4.43 | 0.48 | 0.531 | 0.058 |
| | | | $\hat{\sigma} = 168.7$ | | $R^2_{Y,X} = 0.711$ |

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

| $k$ | $\mathcal{B}_k$ | $p_{NHST}$ | $p_{EQUIV}$ $\Delta = 0.10$ | $BF_{10}$ $r = \sqrt{2}/4$ | CET conclusion $\alpha = 0.05$ | Bayesian conclusion BF threshold = 6 |
|---|---|---|---|---|---|---|
| 1 | 0.004 | 0.927 | 0.014 | 1/11 | Negative | Negative |
| 2 | 0.360 | $< 0.001$ | 1.000 | $8.9 \times 10^7$ | Positive | Positive |
| 3 | -0.073 | 0.144 | 0.276 | 1/3.8 | Inconclusive | Inconclusive |
| 4 | 0.531 | $< 0.001$ | 1.000 | $5.1 \times 10^{13}$ | Positive | Positive |

**Table 4.** Calculated values and conclusions for both frequentist and Bayesian testing for the Bailey multivariable linear regression model.

## 6. Conclusion

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

Equivalence tests may improve current research practices by allowing researchers to falsify their predictions concerning the presence of an effect. In this sense, equivalence testing provides a more formal approach to the "good-enough principle" (Serlin et al., 1993). To be clear, effect sizes need not be dimensionless (or standardized) in order to be meaningful (Kelley and Preacher, 2012). However, 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.

Calculating statistical power (i.e., sample size calculations) for any of the pro-

posed equivalence tests will unfortunately require simulation since the power of these tests depend largely on the distribution of the observed variables. Simulation studies are also required to calculate the statistical power of Bayes factor testing procedures; see Schönbrodt and Wagenmakers (2018).

As Rouder and Morey (2012b) 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 and Gustafson, 2018b).

Note that our non-inferiority test for the increase in the squared multiple correlation coefficient (diff$P_k^2$) 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.

We wish to emphasize that the use of equivalence/non-inferiority tests should not rule out the complementary use of confidence intervals. Indeed, confidence intervals can be extremely useful for highlighting the stability (or lack of stability) of a given estimator, whether that be the $\beta_k$, $\mathcal{B}_k$, $diffP_k^2$ or any other statistic (Fidler et al., 2004). Perhaps one advantage of equivalence/non-inferiority testing over confidence intervals may be that testing can improve the interpretation of null results (Parkhurst, 2001; Hauck and Anderson, 1986). By clearly distinguishing between what is a "neg-

ative" versus an "inconclusive" result, equivalence testing serves to simplify the long "series of searching questions" necessary to evaluate a "failed outcome" (Pocock and Stone, 2016). The best interpretation of data might be obtained when using both tools together.

Finally, note that we only considered equivalence tests based on inverting NCIE-based confidence intervals. It would certainly be worthwhile in future research to consider equivalence tests based on alternative approximations for the sampling variability of standardized regression coefficients; see Jones and Waller (2013) and Yuan and Chan (2011). We also see research potential to expand equivalence testing for standardized regression coefficients in logistic regression models and time-to-event models. Such work will help to "extend the arsenal of confirmatory methods rooted in the frequentist paradigm of inference" (Wellek, 2017).

## References


Altman, D. G. and 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., and McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, (567):305–307; https://doi.og/10.1038/d41586–019–00857–9.

Azen, R. and Budescu, D. (2009). Applications of multiple regression in psychological research. *The SAGE handbook of quantitative methods in psychology*, pages 285–310.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, 100(3):603–617; https://doi.org/10.1348/000712608X377117.

Bailey, D. H. and Geary, D. C. (2009). Hominid brain evolution. *Human Nature*, 20(1):67–79.

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., and Trafimow, D. (2019). The replacement for hypothesis testing. In *International Conference of the Thailand Econometrics Society*, pages 3–17. Springer.

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

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

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

Campbell, H. and Gustafson, P. (2021). re: Linde et al.(2021)–the bayes factor, hdi-rope and frequentist equivalence testing are actually all equivalent. *arXiv preprint arXiv:2104.07834*.

Campbell, H. and Lakens, D. (2021). Can we disregard the whole model? omnibus non-inferiority testing for r2 in multi-variable linear regression and in anova. *British Journal of Mathematical and Statistical Psychology*, 74(1):64–89.

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.

Dixon, P. M., Saint-Maurice, P. F., Kim, Y., Hibbing, P., Bai, Y., and Welk, G. J. (2018). A primer on the use of equivalence testing for evaluating measurement agreement. *Medicine and science in sports and exercise*, 50(4):837.

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*.

Fidler, F., Thomason, N., Cumming, G., Finch, S., and Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, 15(2):119–126.

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. and 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.

Ghashim, E. and Boily, P. (2018). A ggplot2 Primer. *Data Action Lab - Data Science Report Series*.

Hartung, J., Cottrell, J. E., and Giffin, J. P. (1983). Absence of evidence is not evidence of absence. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 58(3):298–299.

Hauck, W. W. and Anderson, S. (1986). A proposal for interpreting and reporting negative studies. *Statistics in Medicine*, 5(3):203–209.

Heck, D. W. (2019). A caveat on the savage–dickey density ratio: the case of computing bayes factors for regression parameters. *British Journal of Mathematical and Statistical Psychology*, 72(2):316–333.

Hung, H., Wang, S.-J., and 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. and 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., and 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.

Kelley, K. and Preacher, K. J. (2012). On effect size. *Psychological Methods*, 17(2):137.

Koh, A. and Cribbie, R. (2013). Robust tests of equivalence for k independent groups. *British Journal of Mathematical and Statistical Psychology*, 66(3):426–434; https://doi.org/10.1111/j.2044–8317.2012.02056.x.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, 6(3):299–312.

Kühberger, A., Fritz, A., and 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.

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., and Dienes, Z. (2020). Improving inferences about null effects with bayes factors and equivalence tests. *The Journals of Gerontology: Series B*, 75(1):45–57.

Lakens, D., Scheel, A. M., and Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269.

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

Linde, M., Tendeiro, J., Selker, R., Wagenmakers, E.-J., and van Ravenzwaaij, D. (2021). Decisions about equivalence: a comparison of TOST, HDI-ROPE, and the Bayes Factor. *accepted for publication in Psychological Methods*.

Marszalek, J. M., Barber, C., Kohlhart, J., and 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. and Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16(4):406–419.

Morey, R. D., Rouder, J. N., Jamil, T., and Morey, M. R. D. (2015). Package bayesfactor. *URLh http://cran/r-projectorg/web/packages/BayesFactor/BayesFactor pdf i (accessed 1006 15)*.

Morris, T. P., White, I. R., and 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.

Parkhurst, D. F. (2001). Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation equivalence tests improve the logic of significance testing when demonstrating similarity is important, and reverse tests can help show that failure to reject a null hypothesis does not support that hypothesis. *Bioscience*, 51(12):1051–1057.

Pocock, S. J. and Stone, G. W. (2016). The primary outcome fails -what next? *New England Journal of Medicine*, 375(9):861–870.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rouder, J. N. and Morey, R. D. (2012a). Default Bayes factors for model se-

lection in regression. *Multivariate Behavioral Research*, 47(6):877–903; DOI: 10.1080/00273171.2012.734737.

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

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

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

Schönbrodt, F. D. and Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic bulletin & review*, 25(1):128–142.

Seaman, M. A. and Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological methods*, 3(4):403.

Serlin, R. C., Lapsley, D. K., Keren, G., and Lewis, C. (1993). *Rational appraisal of psychological research and the good-enough principle.* A handbook for data analysis in the behavioral sciences: Methodological issues. Hillsdale, NJ.

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.*

Tendeiro, J. N. and Kiers, H. A. (2019). A review of issues about null hypothesis bayesian testing. *Psychological Methods*, 24(6):774.

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority.* Chapman and Hall/CRC.

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., and 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. and Chan, W. (2011). Biases and standard errors of standardized regression coefficients. *Psychometrika*, 76(4):670–690.

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

## 7. Appendix

### *7.1. Least squares estimation*

For completeness, we provide details and notation for least squares estimation in a standard linear regression model. We define:

$$\widehat{\beta}_k = ((X^T X)^{-1} X^T y)_k, \text{for } k \text{ in } 1,..., K; \text{ and} \tag{17}$$

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{N} (\hat{\epsilon}_i^2)/(N - K - 1)}, \tag{18}$$

where $\hat{\epsilon}_i = \widehat{y}_i - y_i$, and $\widehat{y}_i = X_{i\times}^T \widehat{\beta}$, for $i$ in 1,..., $N$. A standard NHST for the $k$-th covariate, $X_k$, is stated as:

$H_0 : \beta_k = 0$, 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:

$$p\text{-value}_k = 2 \times F_t\left(\frac{|\widehat{\beta_k}|}{SE(\widehat{\beta_k})}, N-K-1\right), \text{for } k \text{ in } 0,...,K, \tag{19}$$

where we use $F_t(\quad \cdot \quad ; df)$ to denote the cdf of the $t$-distribution with $df$ degrees of freedom, and where: $\widehat{SE(\beta_k)} = \hat{\sigma}\sqrt{[(X^TX)^{-1}]_{kk}}$. Alternatively, we can conduct an $F$-test and, for $k$ in 1,...,$K$, we will obtain the very same $p$-value with:

$$p\text{-value}_k = p_F\left((N-K-1)\frac{\text{diff}R^2_k}{1-R^2_{YX}}, 1, N-K-1\right), \tag{20}$$

where $p_f(\cdot \quad ; df_1, df_2)$ is the cdf of the $F$-distribution with $df_1$ and $df_2$ degrees of freedom, and where: $\text{diff}R^2_k = R^2_{YX} - R^2_{YX_{-k}}$. 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 $H_0 : \beta_k \neq 0$.

### *7.2. 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 $\beta = (-0.20, 0.10, 0.19)^T$ 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 $\beta = (-0.20, 0.00, 0.19)^T$ 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:

$$cor(X_1, X_2, X_3, X_4) = \begin{pmatrix} 1 & 0.4 & 0.3 & 0 \\ 0.4 & 1 & 0.4 & 0.3 \\ 0.3 & 0.4 & 1 & 0.4 \\ 0 & 0.3 & 0.4 & 1 \end{pmatrix}, \tag{21}$$

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 $\beta = (0.20, 0.10, 0.14, -0.12, -0.14)^T$ 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 $\beta = (0.20, 0.00, 0.14, -0.12, -0.14)^T$ to correspond to $\mathcal{B}_1 = 0.000$, with $\sigma^2 = 0.50$.

Figures 3 and 4 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.

### *7.3. Simulation Study 2*

By means of a simple simulation study, we compared a CET frequentist testing scheme (based on NHST and equivalence testing of standardized regression coefficients) to the Bayesian approach based on Rouder and Morey (2012a)'s default BFs. In the simulation study, frequentist conclusions are based on the CET procedure by setting $\Delta$ equal to either 0.05, or 0.10, or 0.25; and with $\alpha$=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). All priors required for calculating the BF are set by simply selecting the default settings of the `regressionBF()` function (with $r = \sqrt{2}/4$); see Morey et al. (2015).

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

- one of twelve sample sizes: $N = 20$, $N = 33$, $N = 55$, $N = 90$, $N = 149$, $N = 246$,

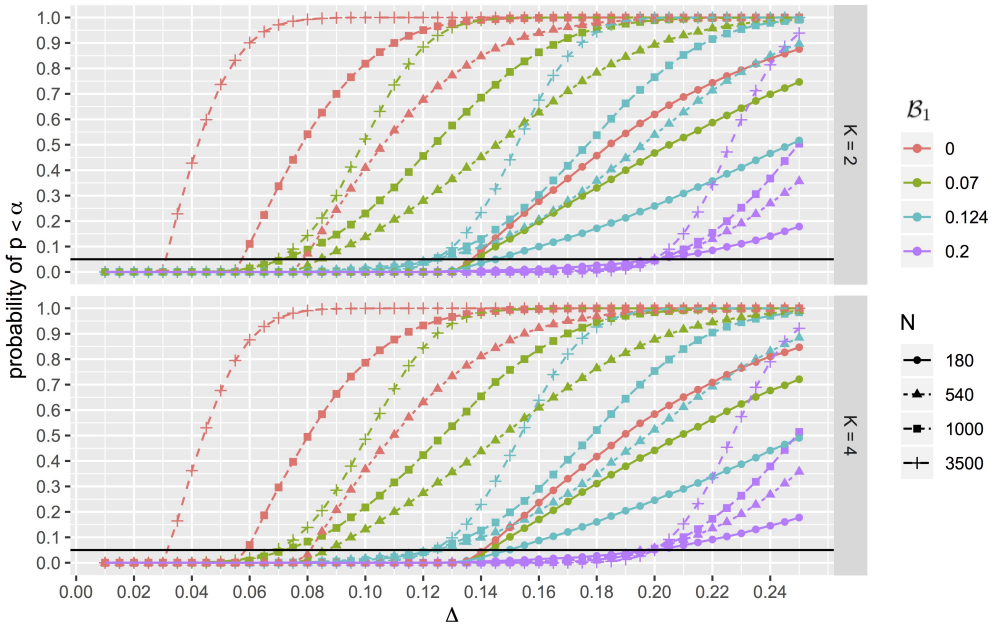


**Figure 3.** Simulation Study 1 (alternative settings) - Upper panel shows results for $K = 2$; Lower panel shows results for $K = 4$. The solid horizontal black line indicates the desired type 1 error of $\alpha = 0.05$.

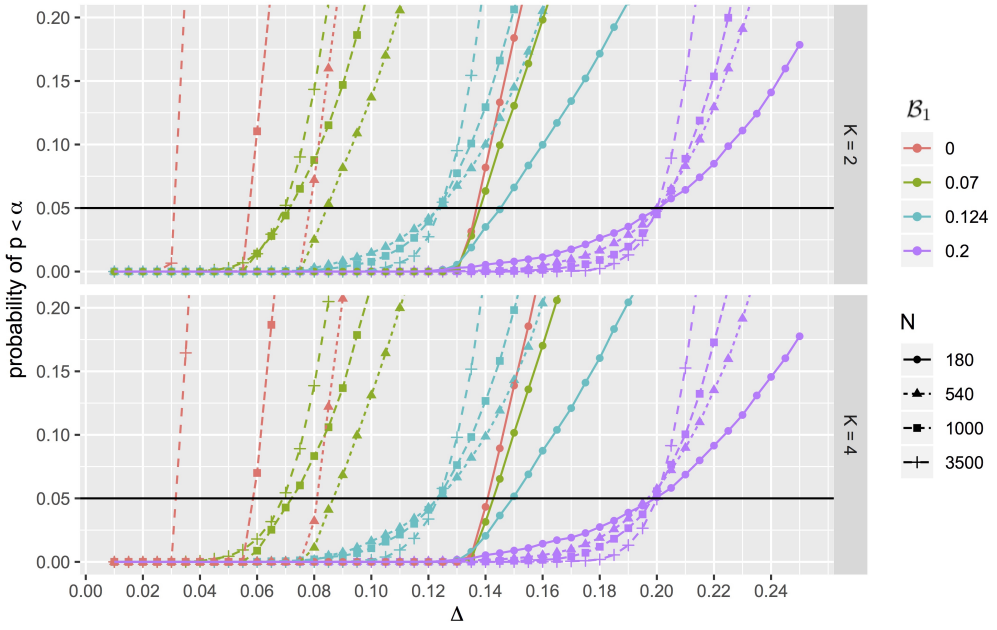


**Figure 4.** Simulation Study 1 (alternative settings) - Upper panel shows results for $K = 2$; lower panel shows results for $K = 4$. Both plots are presented with a restricted vertical-axis to better show the type 1 error rates. The solid horizontal black line indicates the desired type 1 error of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

$N = 406$, $N = 671$, $N = 1,109$, $N = 1,832$, $N = 3,027$, or $N = 5,000$;

- one of two designs with $K = 4$ binary covariates (with an orthogonal, balanced design), with either $\beta = (0.20, 0.10, 0.14, -0.10, -0.10)^T$ or $\beta = (0.20, 0.00, 0.14, -0.10, -0.10)^T$;
- one of two variances: $\sigma^2 = 0.50$, or $\sigma^2 = 1.00$.

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

We compared the CET frequentist testing scheme (based on NHST and equivalence testing) to the Bayesian approach based on BFs by means of a simple simulation study. These are two philosophically different approaches. Our main interest was in determining whether or not the methods yield similar overall trends.

Frequentist conclusions are based on the CET procedure by setting $\Delta$ equal to either 0.05, or 0.10, or 0.25; and with $\alpha$=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 Bayes factor testing; see Schönbrodt and Wagenmakers (2016) for details. Also note that all priors required for calculating the BF are set by simply selecting the default settings of the `regressionBF()` function (with $r = \sqrt{2}/4$); 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 $\beta = (0.20, 0.10, 0.14, -0.10, -0.10)^T$ or $\beta = (0.20, 0.00, 0.14, -0.10, -0.10)^T$;
- one of two variances: $\sigma^2 = 0.50$, or $\sigma^2 = 1.00$.

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

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 5, 6 and 7 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 7.3 displays the the average rate of agreement between the Bayesian and frequentist methods. *Averaging 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 5.** 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 6, and 7 - 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 $\Delta$ and the choice of the BF evidence threshold, see Table 7.3. 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. 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. This is reassuring since it suggests that conclusions obtained from frequentist and Bayesian testing will very rarely lead to substantial disagreements.

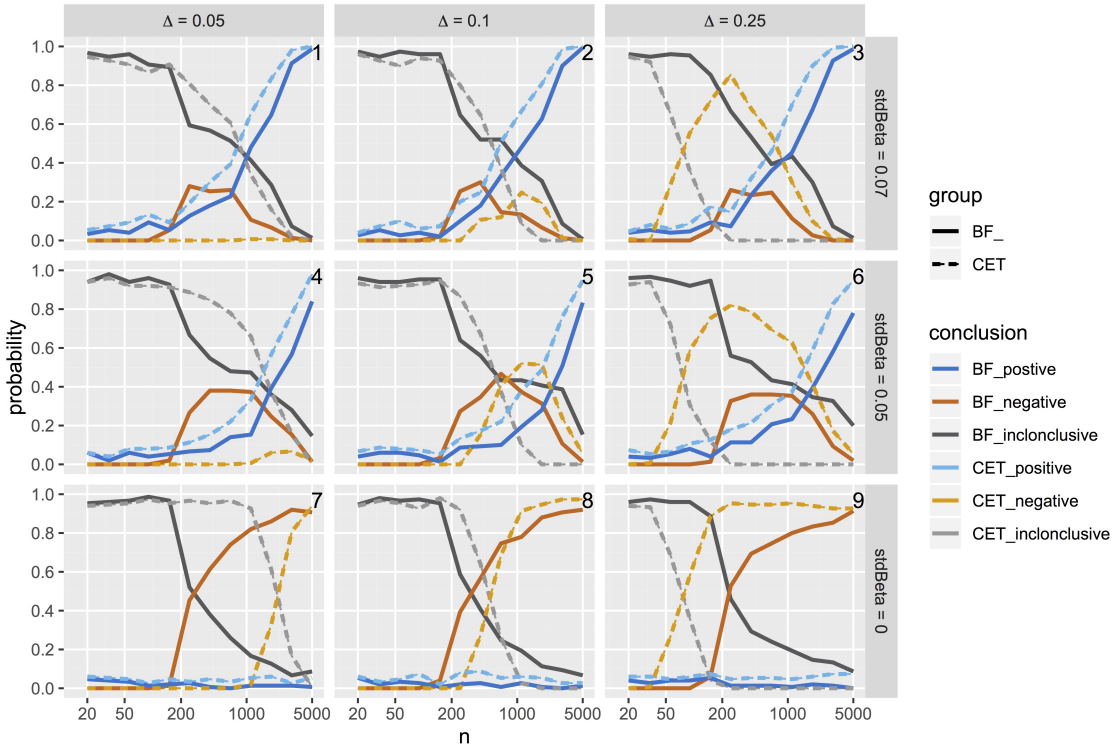


**Figure 5. Simulation study 2, complete results for BF threshold of 3.** The probability of obtaining each conclusion by Bayesian testing scheme (JZS-BF with fixed sample size design, BF threshold of 3:1) and CET ($\alpha = 0.05$). Each panel displays the results of simulations with for different values of $\Delta$ and $\mathcal{B}_1$. Note that all solid lines and the dashed blue line do not change for different values of $\Delta$.

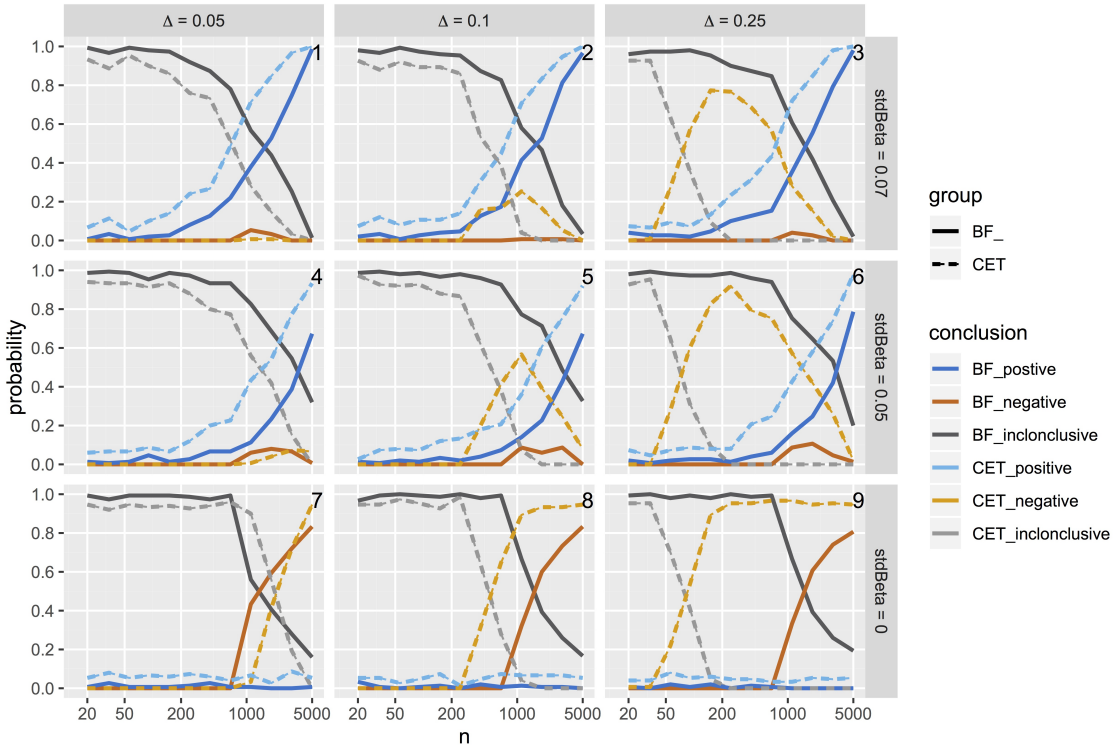


**Figure 6. Simulation study 2, complete results for BF threshold of 6.** The probability of obtaining each conclusion by Bayesian testing scheme (JZS-BF with fixed sample size design, BF threshold of 6:1) and CET ($\alpha = 0.05$). Each panel displays the results of simulations with for different values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

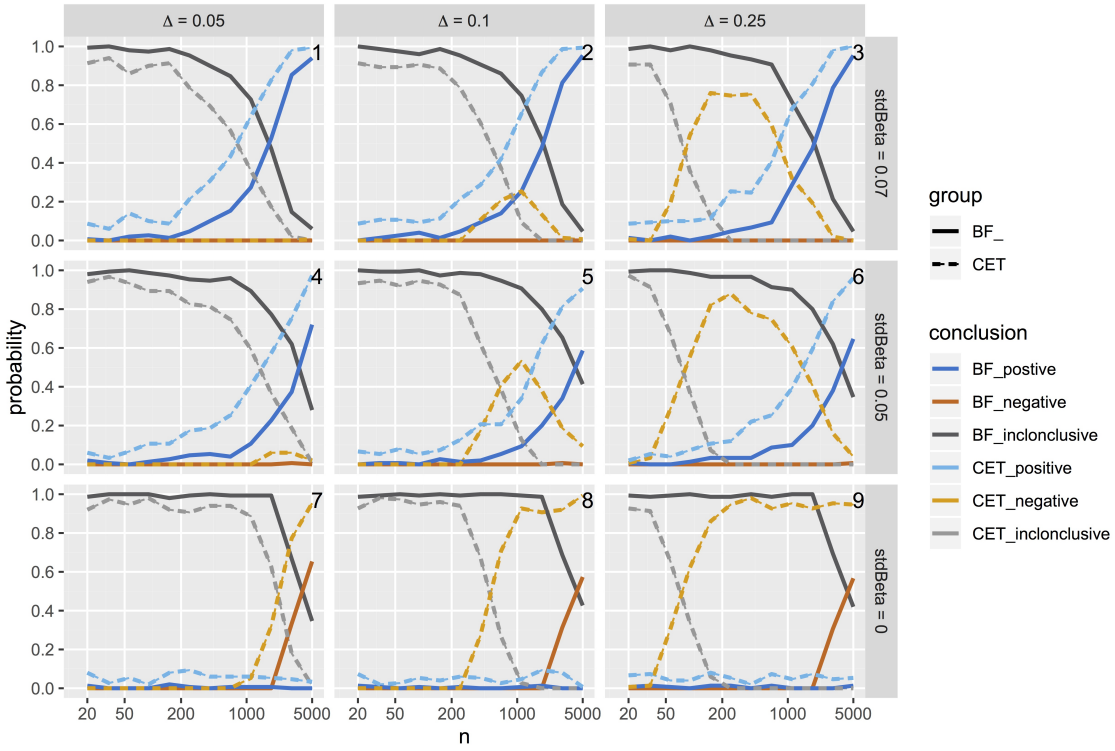


**Figure 7. Simulation study 2, complete results for BF threshold of 10.** The probability of obtaining each conclusion by Bayesian testing scheme (JZS-BF with fixed sample size design, BF threshold of 10:1) and CET ($\alpha = 0.05$). Each panel displays the results of simulations with for different values of $\Delta$ and $\mathcal{B}_1$. Note that all solid lines and the dashed blue line do not change for different values of $\Delta$.

### *7.4. R-code for calculating p-values*

Consider any random data:

```
y <- rnorm(50);
X <- cbind(1,rnorm(50));
```

In R, we can obtain the $p$-value from equation (19) as follows:

```
lmmod <- summary(lm(y~X[,-1]));
N <- length(y);
K <- dim(cbind(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 (20) 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 (3) as follows:

```
DELTA<-t(matrix(rep(c(-0.1,0.1),K+1),,K+1))
p1<-p2<-pval<-vector()
for(k in 1:(K+1)){
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 (5)) in R as follows:

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

and obtain the $p$-values from equation (6) in R with the following code:

```
SE_beta_FIX<-R2YdotX<-R2YdotXmink<-R2XkdotXminK<-p1<-p2<-pval<-vector()
for(k in 1:K){
```

```
if(K>1){ Xmink <- cbind(cbind(X[,-1])[,-k])}
if(K==1){ Xmink <- rep(1,N)}

R2XkdotXminK[k] <- (summary(lm(cbind(X[,-1])[,k]~ Xmink)))$r.squared;
R2YdotXmink[k]  <- summary(lm(y~ Xmink))$r.squared;
R2YdotX[k]      <- (summary(lm(y~ cbind(X[,-1]))))$r.squared
SE_beta_FIX[k]  <- sqrt( (1-R2YdotX[k])/( (1-R2XkdotXminK[k])*(N-K-1)));

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 (9) in R with the following code:

```
DELTA <- rep(0.01, K);
for(k in 1:K){
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);
}
```

### *7.5. R-code for Salaries example*

Results for the Salaries analysis example can be obtained with the following R-code:

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

```
### simple linear regression:
y <- Salaries$salary
X <- model.matrix(lm(salary~ sex, data=Salaries))

# equivalence test for regression coef (Delta=5000), p-val = 0.9632451
equivBeta(Y = y, Xmatrix = X[,-1], DELTA = 5000)$pval[2]

# equivalence test for standardized regression coef (Delta=0.10), p-val = 0.7804196
equivstandardBeta(Y = y, Xmatrix = X[,-1], DELTA = 0.10)$pval

# equivalence test for diffP2 (Delta=0.01), p-val = 0.7804196
equivdiffP2(Y = y, Xmatrix = X[,-1], DELTA = 0.01)$pval

library(BayesFactor);
sdata<-data.frame(salary = Salaries$salary, sex = as.numeric(Salaries$sex)-1);

regressionBF(salary ~ sex, data = sdata)
# 4.525

linearReg.R2stat(N = 397, p = 1, R2 = summary(lm(salary ~ sex, data=Salaries))$r.squared, simple = TRUE)
# 4.525

lmBF(salary ~ sex, data = Salaries)
#  6.177

lmBF(salary ~ sex, data = sdata)
# 4.525

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

summary((lm(salary ~  sex + yrs.since.phd  + yrs.service+ discipline + rank, data=Salaries)))
equivstandardBeta(Y = y, Xmatrix = X[,-1], DELTA = 0.10)$pval

BFs <- (BFstandardBeta(Y= y, Xmatrix=X[,-1])$BF)
BFs
```

Note that the meaning of $\sigma_Y$ for random regressors is different than for fixed

regressors. For random regressors, we have that: $\sigma_Y = (\beta^{'}\Sigma_{XX}\beta + \sigma^2)^{\frac{1}{2}}$. For fixed regressors, we have $\sigma_Y = (\beta^{'}S_{XX}\beta + \sigma^2)^{\frac{1}{2}}$; see Yuan and Chan (2011).