ORIGINAL RESEARCH PAPER

Using equivalence testing to get the most out of linear regression -*or*- Equivalence testing for standardized effect sizes in linear regression

Harlan Campbell[a]

[a]University of British Columbia Department of Statistics Vancouver, BC, Canada, V6T 1Z2

ARTICLE HISTORY

Compiled January 9, 2020

ABSTRACT

Determining a lack of association between an outcome variable and a number of different explanatory variables is frequently necessary in order to disregard a proposed model (i.e., to confirm the lack of a meaningful association between an outcome and predictors). Despite this, the literature rarely offers information about, or technical recommendations concerning, the appropriate statistical methodology to be used to accomplish this task. This paper suggests that when using linear regression, researchers use equivalence testing for standardized effect sizes. A simulation study is conducted to examine the type I error rates and statistical power of the tests, and a comparison is made 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

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 an argument 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 power, and while a null-hypothesis significance test can provide evidence to *reject* the null hypothesis, it cannot provide evidence *in favour* of the null. To properly conclude that an association between X and Y is absent (i.e., to confirm the *lack* of an association), the recommended frequentist tool, the equivalence test, is well-suited (Wellek, 2010).

Let θ be the parameter of interest. The equivalence test reverses the question that is asked in a 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 be considered negligible. (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 define the equivalence margin based on what would be considered "negligible" prior to observing any data; see Campbell and Gustafson (2018b) for details. This can often be challenging. 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 effects considered 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 typically report standardized effect sizes to aid with interpretation. For example, in the social sciences, 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 the predictors before fitting the regression. There are other reasons besides arbitrary scales for reporting standardized effects. For example, Nieminen et al. (2013) argues that standardized effect sizes might be helpful for the synthesis of epidemiological studies. And standardization can also help with interpretation: 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 straightforward. In this paper we introduce equivalence testing procedures for standardized effects sizes in a linear regression. We show how to define valid hypotheses and calculate p-values for these tests for two different cases: (1) with *fixed* regressors, and (2) with *random* regressors. In Section 3, we conduct a small simulation study to better understand the test’s operating characteristics and to consider how a frequentist testing scheme compares to a Bayesian testing approach based on Bayes Factors.

2. Equivalence testing for standardized β coefficient parameter

Let us begin by defining some notation. All technical details are presented in the Appendix. 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_{ki}, be the observed value of covariate X_k, for the ith subject, for k in $1, ..., K$;
- X, be the N by $K+1$ 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);
- $R^2_{Y \cdot X}$ 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 kth covariate.

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

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

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: $\widehat{\beta}_k$, $\widehat{y}_i$, $\hat{\epsilon}_i$, $\hat{\sigma}$, for k in 1,..., K, and for i in 1,...,N (see Appendix for details).

A null hypothesis significance test for a specific variable, X_k, ($H_0 : \beta_k = 0$, vs. $H_1 : \beta_k \neq 0$) is typically done with one of two different (yet mathematically identical) tests. Most commonly a t-test is done to calculate a p-value as follows:

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

where we use $p_t(\cdot \quad; df, ncp)$ to denote the 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. In R, we can obtain the p-values 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]
2*pt(abs(beta_hat/SE_beta_hat), N-K-1, 0, lower.tail=FALSE)
```

Alternatively, we can conduct an F-test and we will obtain the very same p-values with:

$$p - \text{value}_k = p_F\left((N-K-1)\frac{\text{diff}R^2_k}{1-R^2_{Y\cdot X}}, 1, N-K-1, 0\right), \text{for } k \text{ in } 1,...,K, \qquad (3)$$

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: $\text{diff}R^2_k = R^2_{Y\cdot X} - R^2_{Y\cdot X_{-k}}$. In R, we can obtain these p-values as follows:

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

Regardless of whether the t-test or the F-test is employed, if p-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, Δ?* Formally, the null and alternative hypotheses for the equivalence test are:

$H_0 : |\beta_k| \geq \Delta,$

$H_1 : |\beta_k| < \Delta.$

Typically, the equivalence test involves two one-sided t-tests (TOST) with two p-values calculated as follows:

$$p\text{-value}_k^{[1]} = p_t\left(\frac{\widehat{\beta_k} - (-\Delta)}{\widehat{SE(\beta_k)}}, N-K-1, 0\right); \text{and} \quad p\text{-value}_k^{[2]} = p_t\left(\frac{-(\widehat{\beta_k} - \Delta)}{\widehat{SE(\beta_k)}}, N-K-1, 0\right). \tag{4}$$

In order to reject this equivalence test null hypothesis, both p-values must be less than α. Alternatively, we can calculate a single p-value as follows:

$$p - \text{value}_k = 1 - p_t\left(\frac{|\widehat{\beta_k}| - \Delta}{\widehat{SE(\beta_k)}}, N-K-1, 0\right); \tag{5}$$

See Counsell and Cribbie (2015) who review equivalence testing procedures for linear regression coefficients. In R, we can obtain the p-values as follows:

```
p1 <- pt(((beta_hat - (-DELTA))/SE_beta_hat), N-K-1, 0, lower.tail=FALSE)
p2 <- pt((-(beta_hat - DELTA)/SE_beta_hat), N-K-1, 0, lower.tail=FALSE)
pval <- max(c(p1,p2))
# or alternatively:
pt((abs(beta_hat) - DELTA)/SE_beta_hat, N-K-1, 0, lower.tail=TRUE)
```

2.1. An equivalence test for standardized regression coefficients

Unfortunately, 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) Δ can be rather challenging. How can one determine the "smallest effect size of interest" in units that have no particular meaning? In these scenarios, it may be preferable to work with standardized regression coefficients.

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

$$\mathcal{B}_k = \beta_k \frac{\sigma_{X_k}}{\sigma_Y}, \tag{6}$$

and can be estimated by:

$$\widehat{\mathcal{B}_k} = \widehat{\beta_k} \frac{\widehat{\sigma_{X_k}}}{\widehat{\sigma_Y}}, \tag{7}$$

where $\widehat{\sigma_{X_k}}$ and $\widehat{\sigma_Y}$ are the observed 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:

$H_0 : |\mathcal{B}_k| \geq \Delta$,

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

The p-value for this equivalence test is obtained by inverting the confidence interval for $\mathcal{B}_k$ (see Appendix for details), and can be calculated as follows, for k in 1,...,K:

$$p - \text{value}_k = p_t \left(\frac{|\widehat{\mathcal{B}_k}|}{\widehat{SE(\mathcal{B}_k)}_{FIX}} ; df = N - K - 1, ncp = \frac{\sqrt{N\left(1 - R^2_{X_k \cdot X_{-k}}\right)}}{\sqrt{\left(1 - R^2_{Y \cdot X}\right)}} \Delta \right) ; \tag{8}$$

where:

$$\widehat{SE(\mathcal{B}_k)}_{FIX} = \sqrt{\frac{(1 - R^2_{Y \cdot X})}{(1 - R^2_{X_k \cdot X_{-k}})(N - K - 1)}}. \tag{9}$$

```
std_beta_hat <- (beta_hat*(apply(X,2,sd)/sd(y)))[-1]
R2XkXk <- unlist(lapply(c(2:(K+1)), function(k) {summary(lm(X[,k]˜X[,-k]))$r.squared}))
SE_std_beta_FIX <- sqrt((1-R2)/((1-R2XkXk)*(N-K-1)))
pt(abs(std_beta_hat)/SE_std_beta_FIX, df=N-K-1, ncp=DELTA*(sqrt(N*(1-R2XkXk)))/sqrt(1-R2), lower.tail=TRUE)
```

This calculation assumes that the covariates, X, are not stochastic, i.e., the covariates are fixed in advance by the researcher. When X is random (i.e., randomly sampled from a larger population of interest) the sampling distribution of $\mathcal{B}_k$ can be substantially different. In the social sciences, the assumption of fixed regressors is often violated and therefore it is important to consider this possibility (Bentler and Lee, 1983).

Yuan and Chan (2011) derive an estimator for the standard error of $\mathcal{B}_k$ which takes into account the additional variance in $\mathcal{B}_k$ that exists as a result of the regressors being random (see Yuan and Chan (2011) eq. 23). This estimator, $\widehat{SE(\mathcal{B}_k)}_{RDM}$, is based on central limit theorem and the delta-method (see Appendix for details and derivation). Jones and Waller (2013) suggest (based on a simulation study) using $\widehat{SE(\mathcal{B}_k)}_{RDM}$ to construct confidence intervals for $\mathcal{B}_k$. Following the same logic, we can make use of $\widehat{SE(\mathcal{B}_k)}_{RDM}$ to calculate a p-value for our equivalence test ($H_0 : |\mathcal{B}_k| \geq \Delta$) when regressors are random:

$$p - value_k = p_t \left(\frac{|\widehat{\mathcal{B}_k}| - \Delta}{\widehat{SE(\mathcal{B}_k)}_{RDM}}, df = N - K - 1, ncp = 0 \right). \tag{10}$$

```
SE_std_beta_RDM <- DEL(X=X[,-1], y=y)$SEs
pt((std_beta_hat-DELTA)/SE_std_beta_RDM, N-K-1, 0, lower.tail=TRUE)
```

2.2. An equivalence test for the increase in the squared multiple correlation coefficient

The increase in the squared multiple correlation coefficient associated with adding a variable in a linear regression model, diffR^2_k, is a commonly used measure for establishing the importance the added variable. In a linear regression model, the R^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^2 statistic's ubiquitous use, its corresponding population parameter, which we will denote as P^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); see Cramer (1987) for a technical discussion. Campbell and Lakens (2019) introduce a non-inferiority test (a one-sided equivalence test) for $R^2_{Y \cdot X}$ in order to test the null hypotheses $H_0 : P^2_{Y \cdot X} \geq \Delta$ vs. $H_1 : P^2_{Y \cdot X} < \Delta$. Things are slightly different for testing diffP^2_k.

Note that the diffR^2_k measure is simply a re-calibration of $\widehat{\mathcal{B}_k}$, such that:

$$\text{diff}R^2_k = \widehat{\mathcal{B}_k}^2(1 - R^2_{X_k \cdot X_{-k}}). \tag{11}$$

Similarly, we have that, for the corresponding population parameter: diff$P^2_k = \mathcal{B}_k{}^2(1 - P^2_{X_k \cdot X_{-k}})$. It may be preferable to consider the magnitude of Δ (what is to be considered a "negligible difference") in terms of diffP^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:

$H_0 : \text{diff}P^2_k \geq \Delta,$

$H_1 : 0 \leq \text{diff}P^2_k < \Delta.$

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

$$p - \text{value}_k = p_t\left(\frac{\sqrt{(N-K-1)\text{diff}R_k^2}}{\sqrt{(1-R^2_{Y\cdot X})}}; df = N-K-1, ncp = \frac{\sqrt{N\Delta}}{\sqrt{\left(1-R^2_{Y\cdot X}\right)}}\right). \tag{12}$$

One can calculate the above p-value in R with the following code:

```
DELTA <- (DELTA_std_beta^2)*(1-R2XkXk)

diffR2 <- (std_beta_hat^2)*(1-R2XkXk)

pt(sqrt((N-K-1)*diffR2)/sqrt(1-R2), N-K-1, sqrt(N*DELTA)/sqrt(1-R2), lower.tail=TRUE)
```

For random regressors we have, for k in 1, ..., K:

$$p - \text{value}_k = p_t\left(\frac{\sqrt{\text{diff}R_k^2} - \sqrt{\Delta}}{\widehat{SE(\mathcal{B}_k)}_{RDM}\sqrt{(1-R^2_{X_k\cdot X_{-k}})}}, df = N-K-1, ncp = 0\right). \tag{13}$$

The above p-value is calculated in R with the following code:

```
pt((sqrt(diffR2) - sqrt(DELTA_R2diff))/ (SE_std_beta_RDM*sqrt(1-R2XkXk)), N-K-1, lower.tail=TRUE)
```

2.3. 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 test in the simulation study considers the following hypothesis test:

$H_0 : |\mathcal{B}_1| \geq \Delta,$
$H_1 : |\mathcal{B}_1| < \Delta.$

We simulated data for each of 48 scenarios ($2\times4\times2\times3$), one for each combination of the following parameters:

- either fixed or random regressors;

- one of four sample sizes: $N = 180$, $N = 540$, $N = 1,000$, or $N = 3,500$;
- one of two designs with $K = 2$, or $K = 4$ binary covariates; with $\beta = (-0.2, 0.1, 0.2)$ or $\beta = (0.2, 0.1, 0.14, -0.1, -0.1)$; and
- one of three variances: $\sigma^2 = 0.05$, $\sigma^2 = 0.15$, or $\sigma^2 = 0.5$.

Note that for the scenarios with "fixed regressors," the covariates are set such that we have an orthogonal, balanced design. With "random regressors," each possible covariate value is chosen at random with equal likelihood. For scenarios with "fixed regressors," we calculated the equivalence test p-value using the formula for fixed regressors (equation 8). For scenarios with "random regressors," we calculated the equivalence test p-value using the formula for random regressors (equation 10).

Depending on the particular values of σ^2, the true population standardized coefficient, $\mathcal{B}_1$, for these data is either: 0.07, 0.12, or 0.20. We also simulated data from an additional 16 scenarios where the regression coefficients where fixed to be $\beta = (-0.2, 0.0, 0.2)$ (for $K = 2$) and $\beta = (0.2, 0.0, 0.14, -0.1, -0.1)$ for $(K = 4)$ so that we could examine situations with $\mathcal{B}_1 = 0$. 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 the psychology literature; see Kühberger et al. (2014), Fraley and 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 0.10.

For each of the total 64 configurations, we simulated 50,000 unique datasets and calculated an equivalence 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 50,000 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 $0.001 \approx \sqrt{0.05(1 - 0.05)/50,000}$); see Morris et al. (2019).

There is no substantial difference between simulation study results with fixed and

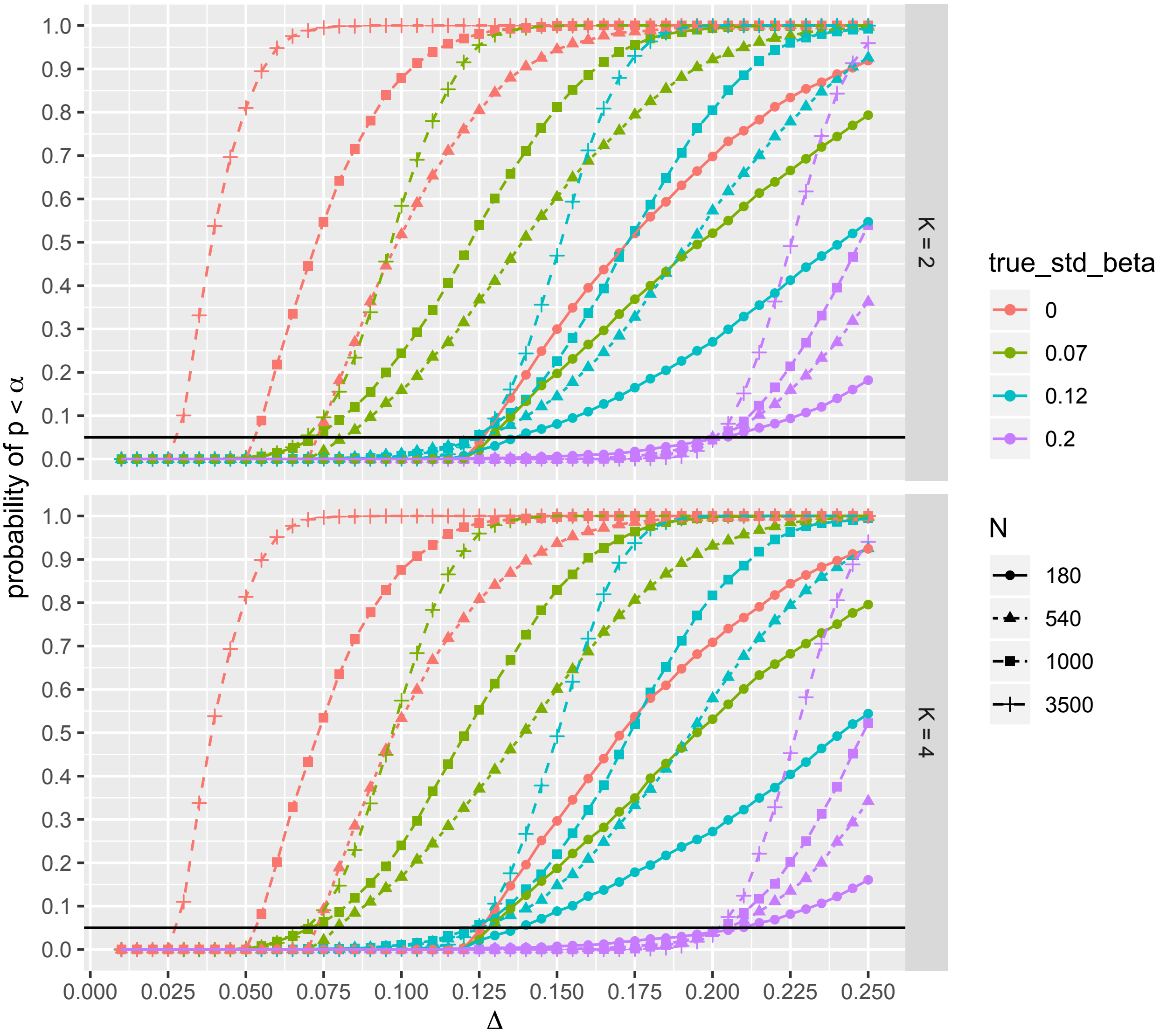

Figure 1. Caption

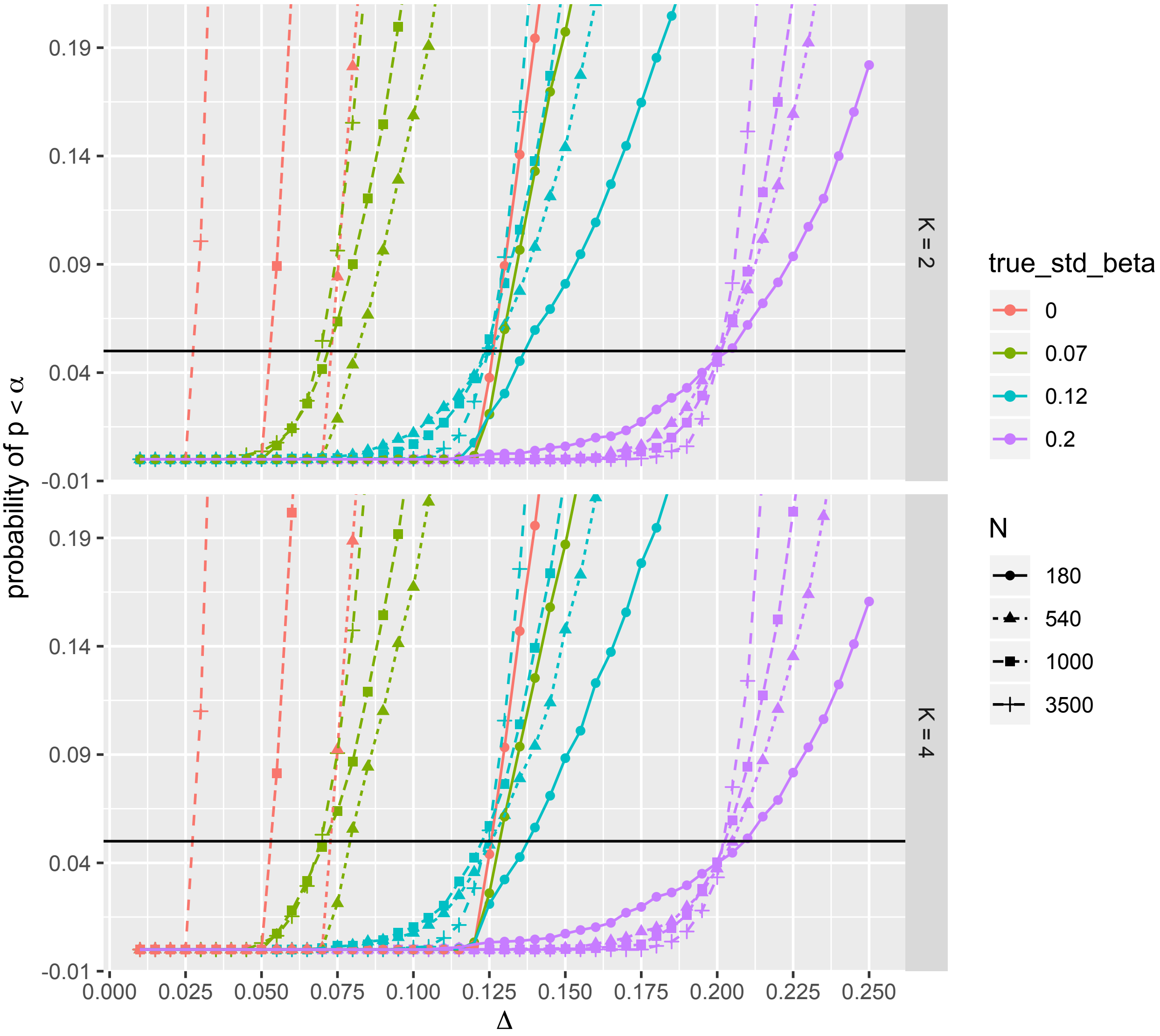

Figure 2. Caption

random regressors. Figure 1 and Figure 2 show results for the simulations with fixed regressors; in the Appendix Figure 3 and Figure 4 show results for the simulations with fixed regressors.

When $\mathcal{B}_1 = 0.20$, we see that when the equivalence bound Δ equals the true effect size (i.e., 0.07, 0.12, or 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 becomes possible to correctly conclude equivalence. For smaller values of $\mathcal{B}_1$ (i.e., for $\mathcal{B}_1 = 0.07$ and $\mathcal{B}_1 = 0.12$), when the equivalence bound Δ equals the true effect size, the test is overly conservative for small N. This is due to the fact that with small N, the sampling variance of the estimate $\mathcal{B}_1$ will be too large to reject $H_0 : |\mathcal{B}_1| \geq \Delta$. (i.e., a 90% CI will be too large to fit within the equivalence margin).

As expected, the power of the test increases with larger values of Δ, larger values of N, and smaller values of K. Also, in order for the test to have substantial power, $\mathcal{B}$ must be substantially smaller than Δ.

3. Comparison to a Bayesian alternative

3.1. Conditional equivalence testing

Ideally, a researcher uses the non-inferiority 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 conduct the equivalence test (i.e., calculate a second p-value, p_2, using equation (8) or (10) if the NHST fails to reject the null. Such a two-stage sequential testing scheme has recently been put forward by Campbell and Gustafson (2018a) 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 greater 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 non-inferiority, i.e., $\mathcal{B}_k$ is at most negligible. If both p-values are large, the result is inconclusive: there is 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).

3.2. Bayes Factor testing for linear regression

For linear regression models, based on the work of Liang et al. (2008), Rouder and Morey (2012) 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)). Here we refer to the null model ("Model 0") and alternative (full) model ("Model 1") as:

$$\text{Model } 0: Y_i \sim \quad Normal(X_{i,-k}^T\beta_{-k}, \sigma^2), \qquad \forall i = 1, ..., N; \tag{14}$$

$$\text{Model } 1: Y_i \sim \quad Normal(X_{i,\cdot}^T\beta, \sigma^2), \qquad \forall i = 1, ..., N; \tag{15}$$

where β_{-k} ($X_{i,-k}$) is the vector (matrix) of regression coefficients (covariates), with the kth 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:

$$BF_{10} = \frac{Pr(Data \,|\, Model\, 1)}{Pr(Data \,|\, Model\, 0)}, \tag{16}$$

with the "10" subscript indicating that the full 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 full model.

Bayesian methods require one to define appropriate prior distributions for all model parameters. Rouder and Morey (2012) suggest using "objective priors" for linear regression models and explain in detail how one may implement this approach. We will not discuss the 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 a BF corresponding to Y and X. (See also the baymedr package in `R` (Linde and van Ravenzwaaij, 2019)). Since we can also calculate frequentist p-values corresponding to values for Y and X, the frequentist and Bayesian approaches can be compared in a relatively straightforward way. We will explore this in Simulation Study 2.

3.3. Simulation Study 2

We wish to compare the frequentist testing we are proposing to the Bayesian approach based on BFs by means of a simple simulation study. How often will the frequentist and Bayesian approaches arrive at the same conclusion?

Frequentist conclusions were based on setting Δ equal to either 0.05, or 0.10, or 0.25; and with α=0.05. BF conclusions were based on an evidence threshold of either 3, 6, or 10. A threshold of 3 can be considered "substantial evidence," a threshold of 6 can be considered "strong evidence," and a threshold of 10 can be considered "very strong evidence" (Wagenmakers et al., 2011). Note that for the simulation study here we examine only the "fixed-n design" for BF testing; see Schönbrodt and Wagenmakers (2016) for details. Also note that, as in Section 3, 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 48 unique scenarios. We considered the following parameters:

- 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 $\beta = (0.2, 0.1, 0.14, -0.1, -0.1)$ or $\beta = (0.2, 0.0, 0.14, -0.1, -0.1)$;
- one of two variances: $\sigma^2 = 1.0$, or $\sigma^2 = 0.5$.

Note that for the $\beta = (0.2, 0.0, 0.14, -0.1, -0.1)$ design, we only consider one value for $\sigma^2 = 1.0$. Depending on the particular design and σ^2, the true coefficient of determination for these data is either $\mathcal{B}_1 = 0.00$, $\mathcal{B}_1 = 0.05$, or $\mathcal{B}_1 = 0.07$. For this second simulation study, we only considered the case of fixed regressors.

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 XX, XX and XX and are based on 5,000 distinct simulated datasets per scenario. We are also interested in how often the two approaches will reach the same overall conclusion: *averaging over all 48 scenarios, how often on average will the Bayesian and frequentist approaches reach the same conclusion given the same data?* Table 3.3 displays the the average rate of agreement between the Bayesian and frequentist methods.

	BF=3	BF=6	BF=10
$\Delta = 0.25$	0.77	0.54	0.44
$\Delta = 0.10$	0.74	0.82	0.72
$\Delta = 0.05$	0.63	0.79	0.84

Table 1. Agreement

Three observations merit comment:

4. Practical Example: ....

XXXX

XXXX

XXXX

XXXX

This result, p-value =, suggests that we can confidently reject the null hypothesis that XXX. We therefore conclude that the data are most compatible with no important effect. For comparison, the Bayesian testing scheme we considered in Section 3 obtains a Bayes Factor of $B_{10} = XXX$. The `R` code for these calculations is presented in the Appendix.

5. Conclusion

In this paper we presented a statistical method for equivalence testing of standardized effect sizes in linear regression. We also considered how frequentist non-inferiority testing, and equivalence testing more generally, offer an attractive alternative to Bayesian methods for "testing the null." We recommend that all researchers specify an appropriate equivalence margin and plan to use the proposed equivalence tests in the event that a standard NHST fails to reject the null. Or in cases when the sample size is very large, the non-inferiority test can be useful to detect effects that are statistically significant but not meaningful.

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 $\widehat{\mathcal{B}}$, or diffR^2_k or any other statistic. 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 "negative" 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). In our opinion, the best interpretation of data will be when using both tools together and our proposal simply serves to “extend the arsenal of confirmatory methods rooted in the frequentist paradigm of inference” (Wellek, 2017).

There is a great risk of bias in the scientific literature if researchers only rely on statistical tools that can reject null hypotheses, but do not have access to statistical tools that allow them to reject the presence of meaningful effects. Most recently, Amrhein et al. (2019) express great concern with the the practice of statistically non-significant results being “interpreted as indicating ‘no difference’ or ‘no effect’ ” (Amrhein et al., 2019); see also Altman and Bland (1995). Equivalence tests provide one approach to improve current research practices by allowing researchers to falsify their predictions concerning the presence of an effect.

6. Appendix

Least squares estimates for the linear regression model are:

- $\widehat{\beta}_k = ((X^T X)^{-1} X^T y)_k$, for k in 1,..., K;
- $\widehat{y}_i = X_{i,\cdot}^T \widehat{\beta}$, for i in 1,..., N;
- $\hat{\epsilon}_i = \widehat{y}_i - y_i$, for i in 1,..., N; and
- $\hat{\sigma} = \sqrt{\sum_{i=1}^{N} (\hat{\epsilon}_i^2)/(N-K-1)}$.

6.1. Linear Regression: further details and R-code.

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.

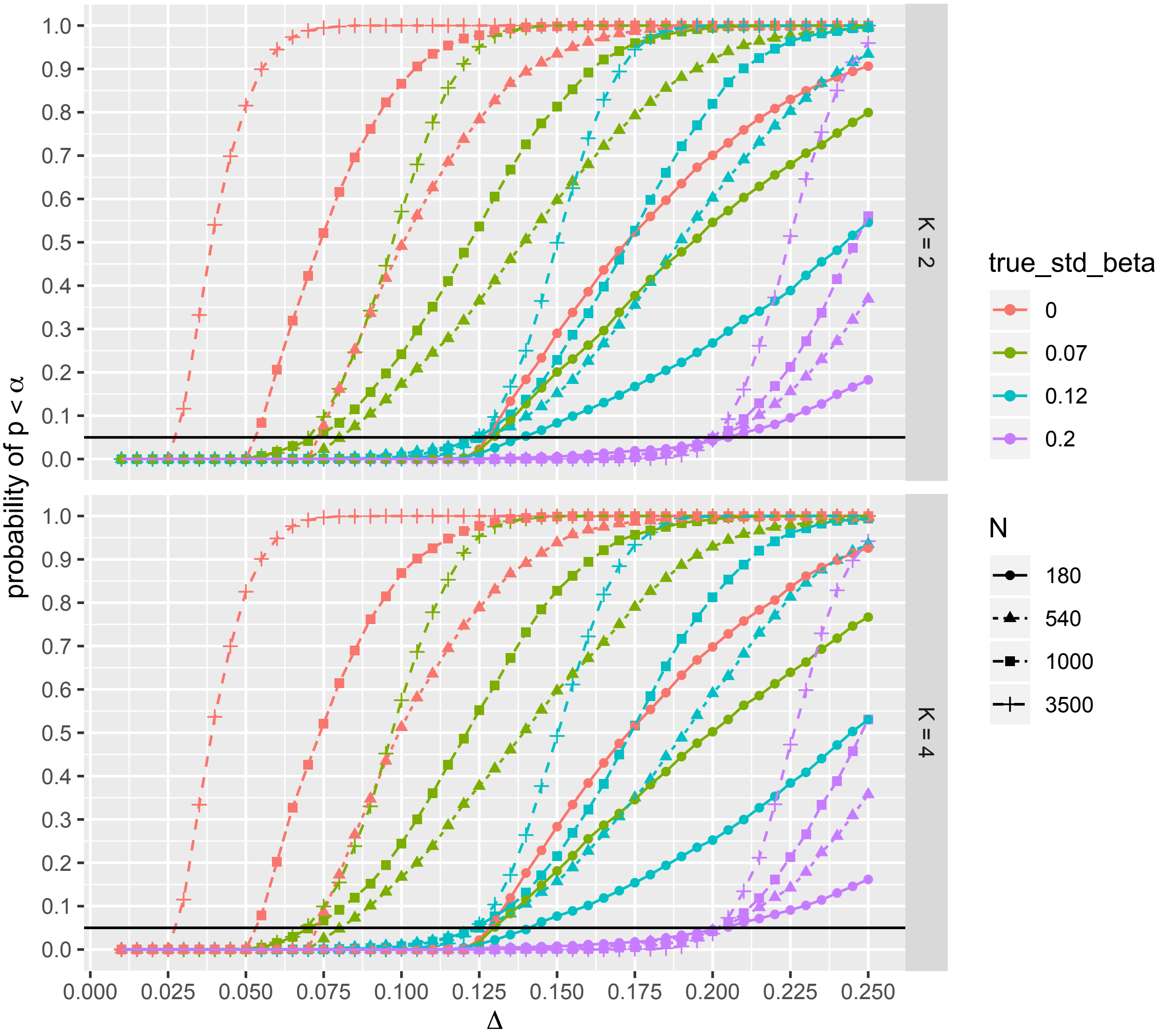

Figure 3. Caption

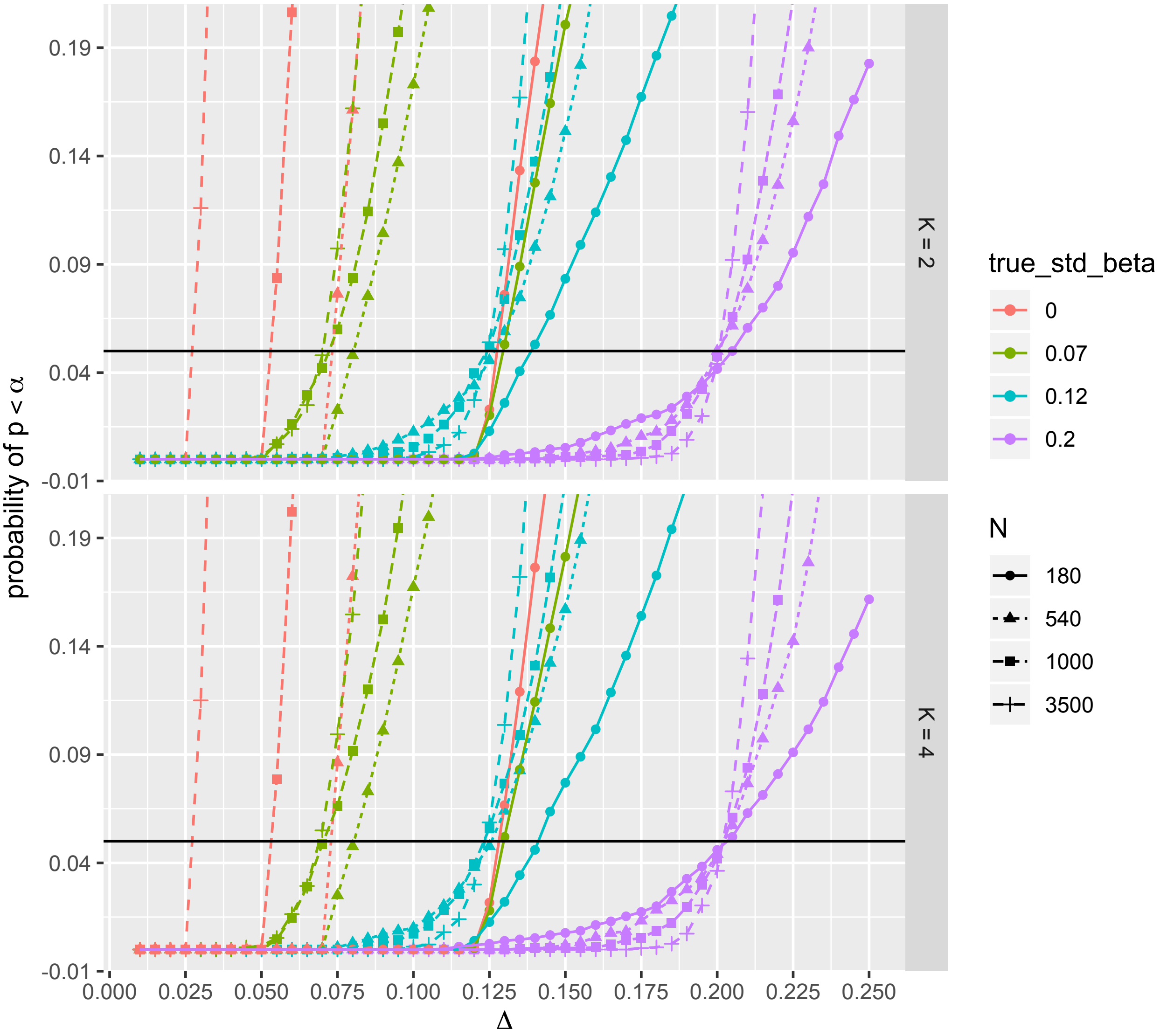

Figure 4. Caption

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.

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

Bentler, P. M. and Lee, S.-Y. (1983). Covariance structures under polynomial constraints: Applications to correlation and alpha-type structural models. *Journal of Educational Statistics*, 8(3):207–222.

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 Lakens, D. (2019). Can we disregard the whole model? *BJMSP*.

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

Counsell, A. and Cribbie, R. A. (2015). Equivalence tests for comparing correlation and regression coefficients. *British Journal of Mathematical and Statistical Psychology*, 68(2):292–309.

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

Etz, A. (2015). Using bayes factors to get the most out of linear regression: A practical guide using r. *The Winnower*.

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.

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.

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.

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.

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.

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. and 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*.

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.

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

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.

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

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.

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

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

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., and Van Der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on bem (2011).

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*.

West, S. G., Aiken, L. S., Wu, W., and Taylor, A. B. (2007). Multiple regression: Applications of the basics and beyond in personality research.

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.