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 8, 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:

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

where β is a parameter vector of regression coefficients, and σ^2 is the population variance. Least squares estimates for the linear regression model are denoted with: $\widehat{\beta}_k$, $\widehat{y}_i$, $\hat{\epsilon}_i$, $\hat{\sigma}$, 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:

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

where $p_t(\cdot \quad ; df, ncp)$ is 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))
N <- length(y); K <- dim(X)[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)[-1]
```

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

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

where $p_f(\cdot \quad ; df_1, df_2, ncp)$ is the cdf of the non-central F-distribution with df_1 and df_2 degrees of freedom, and non-centrality parameter, ncp (note that $ncp = 0$ corresponds to the *central* F-distribution); and where: $\text{diff}R^2_k = R^2_{Y\cdot X} - R^2_{Y\cdot X_{-k}}$. In R, we can obtain the p-values as follows:

```
lmmod <- summary(lm(y~X))
R2 <- lmmod$r.squared
diffR2k <- unlist(lapply(c(1,2), function(k) {R2-summary(lm(y~X[,-k]))$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 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. 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(abs(beta_hat - (-DELTA)/SE_beta_hat), N-K-1, 0, lower.tail=FALSE)
p2 <- pt(abs((beta_hat - DELTA)/SE_beta_hat), N-K-1, 0, lower.tail=FALSE)
```

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:

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

and can be estimated by:

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

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:

$$p - \text{value}_k = p_t \left(\frac{\widehat{\mathcal{B}_k}}{\widehat{SE(\mathcal{B}_k)}_{FIX}}; df = N - K - 1, ncp = \frac{\sqrt{N\left(1 - R^2_{X_k \cdot X_{-k}}\right)}}{\sqrt{\left(1 - R^2_{Y \cdot X}\right)}} \Delta \right); \tag{7}$$

where:

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

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

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:

$$p - value_k = p_t\left(\frac{\widehat{\mathcal{B}_k} - \Delta}{\widehat{SE(\mathcal{B}_k)}_{RDM}}, df = N - K - 1, ncp = 0\right). \tag{9}$$

```
SE_std_beta_RDM <- DEL(X=X, y=y)$SEs
pt((std_beta_hat-DELTA)/SE_std_beta_RDM, N-K-1, 0)
```

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 $\text{diff}P^2_k$.

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

$$\text{diff}R^2_k = \widehat{\mathcal{B}_k}^2(1 - R^2_{X_k\cdot X_{-k}}). \tag{10}$$

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

$$p - \text{value}_k = p_t\left(\frac{\sqrt{(N-K-1)\text{diff}R^2_k}}{\sqrt{(1-R^2_{Y\cdot X})}}; df = N-K-1, ncp = \frac{\sqrt{N\Delta}}{\sqrt{(1-R^2_{Y\cdot X})}}\right); \tag{11}$$

```
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))
```

and for random regressors as:

$$p - \text{value}_k = p_t\left(\frac{\sqrt{\text{diff}R^2_k} - \sqrt{\Delta}}{\widehat{SE(\mathcal{B}_k)}_{RDM}\sqrt{(1 - R^2_{X_k \cdot X_{-k}})}}, df = N - K - 1, ncp = 0\right). \quad (12)$$

```
pt((sqrt(diffR2) - sqrt(DELTA)) / (SE_std_beta_RDM*sqrt(1-R2XkXk)), N-K-1, 0)
```

2.3. An equivalence test for interaction effects

2.4. 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 (7) or (9) 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: P^2 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., P^2 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). From the outset of their study, researchers can either decide to (1) perform only an equivalence test, or (2) decide to perform both an equivalence test and a NHST (acknowledging the possibility there is a non-zero, but trivial, effect), or (3) plan to only perform an equivalence test if the NHST is not significant (CET). As long as these procedures are chosen and performed transparently