Research Article

Jiarui (Alex) Tian*

# A Replication of "The Effect of the Conservation Reserve Program on Rural Economies: Deriving a Statistical Verdict from a Null Finding" (*American Journal of Agricultural Economics*, 2019)

https://doi.org/10.1515/econ-2022-0036
received November 30, 2021; accepted January 28, 2023

**Abstract:** This study replicates the paper "Brown, J. P., Lambert, D. M., & Wojan, T. R. (2019). The effect of the conservation reserve program on rural economies: deriving a statistical verdict from a null finding. *American Journal of Agricultural Economics*, *101*(2), 528–540" and their procedure for calculating the so-called ex post power of statistical tests of significance for regression coefficients. There appears no generally accepted method for calculating ex post power, and Brown, Lambert, and Wojan (BLW) provided a bootstrapping method that can be applied after the parameter of interest is estimated. They recommend researchers to use this procedure to investigate whether a statistically insignificant finding is likely to be due to a low power property of the significance test. This study makes two main contributions. First, it verifies whether the data and code that BLW provided are reliable to reproduce their results. Second, it constructs Monte Carlo experiments to assess the performance of BLW's method. The results indicate that their method produces ex post power estimates that are relatively close to the true power values. Mean power estimates are generally unbiased, and 95% of the estimates lie within +/− 5% points of the true power. In conclusion, my replication provides further evidence of the reliability of BLW's method.

**Keywords:** ex post power, statistical insignificance, Monte Carlo experiments, bootstrapping, replication

**JEL Codes:** C12, C15, C18

## 1 Introduction

This study replicates the report by Brown, Lambert, and Wojan (2019), henceforth BLW. I choose to replicate BLW because they propose a method for calculating ex post power, also referred to as post hoc power, or retrospective power. As the name suggests, ex post power is calculated after empirical analyses are completed. Reliable estimates of ex post power are a potentially valuable addition to the applied econometrician's toolkit. They can help the researcher determine whether a statistically insignificant estimate is due to a negligible effect size or insufficient power. They can also be useful for interpreting statistically significant estimates. Significant estimates in the presence of low power can raise a red flag alerting the researcher to the possibility of Type M error[1] (Gelman & Carlin, 2014). Up to now there has been no generally accepted method for calculating ex post power. The purpose of this replication is to assess the reproducibility and reliability of BLW's method.

A commonly used method, often referred to as "observed power," uses both the estimated effect size and its associated standard error to calculate power. However, this approach is now widely recognized as flawed (Hoenig & Heisey, 2001; Yuan & Maxwell, 2005). Other methods have been used, but they have shortcomings. Skiba and Tobacman (2019) calculate "ex post" power, but they use the same methods employed for ex ante power analyses, which are based on selected summary statistics and predetermined distributional assumptions. Ioannidis et al. (2017) calculated ex post power, but their method is designed to work with meta-analysis and cannot be applied to single studies.

* **Corresponding author: Jiarui (Alex) Tian,** Department of Economics and Finance, University of Canterbury, Christchurch, New Zealand, e-mail: Alex.Tian@pg.canterbury.ac.nz, tel: +64-021-111-0599

1 Type M error (magnitude) is associated with the fact that statistically significant estimates are, on average, larger than the true effect. One source of Type M error is publication bias, where large-magnitude findings are more likely to be reported.

In this space, BLW propose two simulation methods that can be applied to a single dataset: (i) a bootstrap resampling procedure, and (ii) a Bayesian approach that posits a prior distribution for the distribution of effect sizes. My replication focuses on the first of these two methods.

BLW propose their procedure in the context of interpreting the results from a published, economic impact study of the Conservation Reserve Program (CRP; Sullivan et al., 2004). That analysis found that the CRP had a statistically insignificant impact on employment growth. BLW's motivation was to investigate if the statistical insignificance was due to insufficient power. Accordingly, they applied their procedure to the underlying data and concluded that the study was "sufficiently powered" (i.e., statistical power equal to 80%, Ioannnidis et al., 2017) for effect sizes of policy interest. In other words, the authors of the original study interpreted their insignificant results as being unable to address the issue of employment losses, BLW found that the study was sufficiently powered to detect employment losses of a magnitude that would cause the CRP to fail a benefit-cost analysis. This allowed them to re-interpret the original findings as supporting the CRP. Given the potential benefit of being able to produce reliable estimates of ex post power, and the fact that BLW was published in one of the top five journals in agricultural economics leading to significant attention (Bellamare, 2021), its reliability has to be carefully examined. This study pursues such a rule of reliability check through a replication and Monte Carlos experiments.

My replication proceeds as follows. Section 2 (i) provides the theoretical context for BLW's bootstrapping method, (ii) describes BLW's bootstrap resampling method, and (iii) briefly summarizes the study to which BLW applied their method. Section 3 reports my reproduction of BLW's analysis. Section 4 presents a variant of BLW's bootstrapping method based on residual resampling and explains why it might be expected to be superior. Section 5 describes the Monte Carlo experiments I designed to assess the performance of the two methods and reports the results. Section 6 concludes the study.

# 2 Theoretical Context and Description of BLW's Method

## 2.1 Using Monte Carlo Simulation Methods to Calculate Statistical Power

The use of Monte Carlo simulation methods for calculating statistical power is not new. Programs to implement Monte Carlo simulation methods for power analysis can be found in many statistical software packages such as SAS and Stata (StataCorp, 2021, Wicklin, 2013). Bootstrapping is a computer-based technique similar to Monte Carlo simulation except that it draws repeated samples from the sample itself, as opposed to sampling from a population (Chong & Choo, 2011; Efron & Tibshirani, 1994). Bootstrapping to calculate the statistical power is relatively recent (Kleinman & Huang, 2017). As far as I am aware, BLW is the first study to apply bootstrapping procedures directly to a completed empirical analysis – as opposed to baseline or pilot data – to estimate power.

## 2.2 BLW's Method

Let $Y$ be an outcome variable, $T$ a treatment variable, and $\boldsymbol{C}$ a column vector of $k$ control variables. Let the associated data generating process for $Y$ be given by

$$Y_i = \beta_0 + \beta_T T_i + \boldsymbol{\beta_C'} \boldsymbol{C_i} + \varepsilon_i, \quad i = 1, 2, \ldots N, \tag{1}$$

where $\varepsilon_i$ is an independently and identically distributed error term that is independent of regression variables. A researcher estimates the treatment effect by regressing $Y$ on $T$ controlling for $\boldsymbol{C}$ and obtains estimates $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{\boldsymbol{0}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{T}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{C}}')'$. The associated residuals are defined by $e_i = Y_i - \hat{Y}_i$, where $\hat{Y}_i$ is the predicted value of $Y$ conditional on $T_i$ and $\boldsymbol{C_i}$. $\varepsilon_i$ is independently and identically distributed error term that is independent of regression variables.

Define $\boldsymbol{X_i} = (1, T_i, \boldsymbol{C_i'})'$ and let Beta be the treatment effect size for which the researcher wants to calculate power of a significance test to examine the hypothesis of $\beta_T = 0$. BLW's method samples the $N \times (k + 3)$ matrix $(\boldsymbol{Xe})$ with replacement. Let individual, resampled values of $\boldsymbol{X}$ and $\boldsymbol{e}$ be denoted by $\boldsymbol{X_j^\star}$ and $e_j^\star$, where $j = 1,2,\ldots N^\star$. Note that $N^\star$ need not be the same as $N$.

BLW then created simulated $Y^\star$ values such that $Y_j^\star = \boldsymbol{X_j^{\star\prime}} \hat{\boldsymbol{\beta}}^\star + e_j^\star$, where $\hat{\boldsymbol{\beta}}^\star = (\hat{\beta}_0, \text{Beta}, \hat{\boldsymbol{\beta}}_{\boldsymbol{C}}')'$. They then regressed $\boldsymbol{Y^\star}$ on $\boldsymbol{X^\star}$. This produced an estimate for $\beta_T$ and they noted whether it is statistically significant. This process is repeated 999 times. Ex post power is calculated as the percent of times that the Monte Carlo estimates of $\beta_T$ are statistically significant. BLW applied their method to an economic impact study of the Conservation Reserve Program reported by Sullivan et al. (2004).

## 2.3 The Study by Sullivan et al. (2004)

Sullivan et al. (2004) used a quasi-experimental, matched pair protocol to estimate the effect of the Conservation

Reserve Program on county-level, employment growth data in the US. High-CRP counties were matched with similar low-CRP counties. High-CRP (low-CRP) counties were those counties that, on average, enrolled a higher (lower) percentage of their eligible land in CRP than other types of farms.

Sullivan et al. (2004) reported estimated treatment effects for several models, but complete results were only reported for the long-run local employment growth model. The estimated treatment effect for this model was statistically insignificant. As a result, Sullivan et al. (2004) were unable to reach a conclusion whether the CRP had an adverse impact on the employment growth.

BLW applied their method to Sullivan et al's. (2004) data. They calculated ex post power for the following effect sizes: Beta = (−0.027, −0.015, −0.010, −0.005, and −0.001), where negative values indicated adverse employment effects. Because their method is not restricted by the actual sample size (i.e., $N^*$ need not equal $N$), they not only calculated the power calculations for the original sample size of 190 observations but also for sample sizes of 100, 150, 200, 250, and 350 observations. This generated a total of 30 experiments, one for each combination of effect size Beta = (−0.027, −0.015, −0.010, −0.005, and −0.001), and sample size $N$ = (100, 150, 190, 200, 250, 350).

## 3 Reproduction

My first contribution is to reproduce BLW's results. Table 1 reports three sets of results. The first set of results (column 3/"BLW") is taken directly from Table 4 in BLW. It copies the ex post power results that BLW report in their paper. The second set of results (column 4/"Reproduction-R") are the results that I produced when I used the data and code that BLW provided in their paper. The "R" indicates that their code was written in R.

The final set of results (column 5/"Reproduction-Stata") are the results that I produced when I rewrote their procedure using Stata code. The Center for Open Science (COS) calls this "Author Data Reproduction (ADR)." This type of reproduction uses the original data but re-estimates the models using new analytic code generated by the replicator. It was one of the types of reproductions used in COS's massive SCORE project (Center for Open Science, 2022). It is also a component of good replication practices recommended by the International Initiative for Impact Evaluation (3ie) as part of their Replication Program (International Initiative for Impact Evaluation, 2022).

As noted above, BLW conducted statistical power calculations for 30 pairs of effect and sample size values:

**Table 1:** Replication of BLW's ex post power results

| Beta [1] | *N* [2] | BLW [3](%) | Reproduction-R [4](%) | Reproduction-Stata [5](%) |
|---|---|---|---|---|
| −0.027 | 100 | 99 | 99 | 99 |
| −0.027 | 150 | 100 | 100 | 100 |
| **−0.027** | **190** | **100** | **100** | **100** |
| −0.027 | 200 | 100 | 100 | 100 |
| −0.027 | 250 | 100 | 100 | 100 |
| −0.027 | 300 | 100 | 100 | 100 |
| −0.027 | 350 | 100 | 100 | 100 |
| −0.015 | 100 | 84 | 84 | 85 |
| −0.015 | 150 | 96 | 96 | 96 |
| **−0.015** | **190** | **99** | **99** | **99** |
| −0.015 | 200 | 99 | 99 | 99 |
| −0.015 | 250 | 100 | 100 | 100 |
| −0.015 | 300 | 100 | 100 | 100 |
| −0.015 | 350 | 100 | 100 | 100 |
| −0.010 | 100 | 59 | 59 | 59 |
| −0.010 | 150 | 79 | 79 | 79 |
| **−0.010** | **190** | **88** | **88** | **88** |
| −0.010 | 200 | 90 | 90 | 89 |
| −0.010 | 250 | 96 | 96 | 95 |
| −0.010 | 300 | 98 | 98 | 98 |
| −0.010 | 350 | 99 | 99 | 99 |
| −0.005 | 100 | 24 | 24 | 23 |
| −0.005 | 150 | 33 | 33 | 33 |
| **−0.005** | **190** | **42** | **42** | **42** |
| −0.005 | 200 | 43 | 43 | 42 |
| −0.005 | 250 | 51 | 51 | 53 |
| −0.005 | 300 | 60 | 60 | 60 |
| −0.005 | 350 | 67 | 67 | 67 |
| −0.001 | 100 | 6 | 6 | 6 |
| −0.001 | 150 | 6 | 6 | 6 |
| **−0.001** | **190** | **6** | **6** | **7** |
| −0.001 | 200 | 6 | 6 | 6 |
| −0.001 | 250 | 7 | 7 | 7 |
| −0.001 | 300 | 7 | 7 | 7 |
| −0.001 | 350 | 7 | 8 | 8 |

NOTE: The values in the table report the ex post statistical power associated with the effect size given in column 1. Column 2 reports the size of the individual datasets used in the Monte Carlo simulations. Bold values indicate that the sample size is the same as the original dataset. Note that the BLW's bootstrapping procedure allows the simulated datasets to be smaller/larger than the original. Column 3 copies the values for ex post statistical power reported in BLW (Table 4). Column 4 reports the values I produced when I used BLW's data and code, originally written in R. Column 5 reports the values I obtained when I rewrote their procedure using Stata code.

Beta = (−0.027, −0.015, −0.010, −0.005, and −0.001) and $N$ = (100, 150, 190, 200, 250, and 350). −0.027, which is the maximum loss in employment growth that BLW determined, would be acceptable in a benefit-cost analysis of the CRP. They included $N$ = 190 as this was the size of Sullivan et al's. (2004) original study. Accordingly, I have

highlighted in bold these experiments in the table. But they also include other sample sizes to better understand the role of sample size on power in this setting. Likewise, they considered effect sizes smaller than −0.027 to observe how well the sample design detects smaller negative employment effects of the CRP.

As expected, ex post power is greatest for the larger (in absolute value) effect sizes and larger sample sizes. For an effect size of −0.027 and a sample size of 190, BLW calculated that the Sullivan et al. (2004) study had statistical power approximately equal to 100%. In other words, if the job loss associated with the CRP was large enough to reject the CRP on benefit-cost grounds, there is virtually a 100% likelihood that the study of Sullivan et al. (2004) would have produced a statistically significant estimate of this effect. The fact that they did not obtain a statistically significant estimate leads BLW to conclude that the job loss was smaller than this.

Columns 4 and 5 of Table 1 report my efforts to reproduce BLW's results, first using their R code and then rewriting their program in Stata. Using their R code, I was able to exactly reproduce their results. Using my Stata version of their program, I reproduced their results with only miniscule differences. For example, when the effect size was −0.015 and the sample size was 100, BLW reported an ex post power value of 84%, but my Stata replication produced a power value of 85%. I attribute these differences to rounding and the fact that the random number generators underlying the simulations use different seeds.

In conclusion, using BLW's data and code, I obtained results that are identical, or approximately identical, to the results published in their paper. The same holds when I rewrite their program and use STATA rather than R.

## 4 Extension

There is more than one approach to bootstrapping (Brown et al., 2019; Efron, 1982; Kleinman & Huang, 2017). In fact, there are at least two potential problems with BLW's approach. First, when restricting oneself to the same size as the original sample, resampling with replacement throws away information. When observations are sampled more than once, other observations are left out of the reconstituted dataset. This represents a loss of information.

Second, changing the dataset changes the power of the sample design. For example, consider a binary treatment variable and suppose half of the original dataset received treatment and half did not. Now consider an extreme case where resampling resulted in a reconstituted dataset where only one-fourth of the original dataset received treatment. The sample design of the reconstituted dataset would have lower power than that of the original sample. In fact, every reconstituted sample where the percent of treated observations was other than 50% would have lower power. Thus, by changing the nature of the dataset, BLW's method can introduce bias in estimates of ex post power.

An alternative procedure that leaves the original dataset unchanged is residual resampling (Wicklin, 2018). Residual resampling works the same way as BLW's method, except it only resamples the residuals with replacement. It then pairs the resampled residuals with the original observations. This addresses the two shortcomings of BLW's method, though it should be noted that it can only generate datasets with the same number of observations as the original. Unlike BLW's method, it cannot generate datasets with more or less observations than the original. However, this is not so much a disadvantage when the researcher's aim is to determine the power of a given estimate in a specific regression. I call this alternative bootstrapping procedure BLW$^a$.

Column 4 of Table 2 reports the results of applying the BLW$^a$ method to the data in BLW. Column 3 copies the Stata results from Table 1 to facilitate comparison. Using the alternative BLW$^a$ method does make a small difference. For example, when Beta = −0.01 and sample size = 190, BLW's method produces an ex post power estimate of 89% (see "Reproduction" column). BLW$^a$ produces an ex post power estimate of 94%. Similar differences are observed for Beta = −0.005and Beta = −0.001. This raises the question, which estimate is correct?

Without some ground truth to compare to, one cannot say which method is "better." While a full performance assessment lies beyond the scope of this replication, I performed a limited performance analysis that stays within the research design of BLW. Specifically, I constructed a series of Monte Carlo experiments where the data and specification are the same as those used by BLW. I set the model parameters such that I know the true power of the ordinary least squares (OLS) estimates of the treatment effect. I then compared BLW and BLW$^a$ on the basis of bias, sample range, and mean squared error (MSE).

## 5 Assessing the Performance of BLW and BLW$^a$

I conducted Monte Carlo experiments where I created a data generating process with known power to see how well the two methods are able to estimate it. My

**Table 2:** Extension: BLW[a]

| Beta [1] | *N* [2] | Reproduction-Stata [3](%) | BLW[a] [4](%) |
|---|---|---|---|
| −0.027 | 100 | 99 | — |
| −0.027 | 150 | 100 | — |
| **−0.027** | **190** | **100** | **100** |
| −0.027 | 200 | 100 | — |
| −0.027 | 250 | 100 | — |
| −0.027 | 300 | 100 | — |
| −0.027 | 350 | 100 | — |
| −0.015 | 100 | 85 | — |
| −0.015 | 150 | 96 | — |
| **−0.015** | **190** | **99** | **100** |
| −0.015 | 200 | 99 | — |
| −0.015 | 250 | 100 | — |
| −0.015 | 300 | 100 | — |
| −0.015 | 350 | 100 | — |
| −0.010 | 100 | 59 | — |
| −0.010 | 150 | 79 | — |
| **−0.010** | **190** | **88** | **94** |
| −0.010 | 200 | 89 | — |
| −0.010 | 250 | 95 | — |
| −0.010 | 300 | 98 | — |
| −0.010 | 350 | 99 | — |
| −0.005 | 100 | 23 | — |
| −0.005 | 150 | 33 | — |
| **−0.005** | **190** | **42** | **49** |
| −0.005 | 200 | 42 | — |
| −0.005 | 250 | 53 | — |
| −0.005 | 300 | 60 | — |
| −0.005 | 350 | 67 | — |
| −0.001 | 100 | 6 | — |
| −0.001 | 150 | 6 | — |
| −0.001 | 190 | 7 | 9 |
| −0.001 | 200 | 6 | — |
| −0.001 | 250 | 7 | — |
| −0.001 | 300 | 7 | — |
| −0.001 | 350 | 8 | — |

NOTE: The values in the table report the ex post statistical power associated with the effect size given in column 1. Column 2 reports the size of the individual datasets used in the Monte Carlo simulations. Bold values indicate that the sample size is the same as the original dataset. While BLW's bootstrapping procedure allows the simulated datasets to be smaller/larger than the original, the alternative bootstrapping method, BLW[a], restricts the simulated datasets to have the same number of observations as the original. Column 4 reports ex post power estimates for the BLW[a] bootstrapping procedure, which is coded in Stata. Column 3 reproduces the Stata-coded BLW estimates of power from Table 1 (cf. column 5) to facilitate a comparison of the two procedures.

experiments are tailored to the BLW/Sullivan et al. (2004) dataset. Their data consist of 190 observations. The dependent variable, $Y_i$, is a measure of county-level employment growth; and the treatment variable, $T_i$, is "CRP payments to income ratio." There are a total of 30 control variables.

The OLS regression equation for which BLW calculated ex post power for the estimated treatment effect is reported in Table 3 of their paper. While the estimated variance of the error term is not reported in the table, I was able to exactly reproduce their results and determine that $\hat{\sigma}_\varepsilon^2 = 0.022870$. It follows from this that $\text{var}(\hat{\beta}_T)$ can be identified by the corresponding term along the main diagonal of $\hat{\sigma}_\varepsilon^2(\boldsymbol{X'X})^{-1}$(=0.000011256).

## 5.1 Step One: Calculating an Effect Size for Every Power Value

In order to simulate data for the Monte Carlo experiments, I first determined the effect size that corresponds to a given power value from the equation below:

$$\text{Beta}_{\text{Power}} = \left(t_{\text{Power},\upsilon} + t_{1-\frac{\alpha}{2},\upsilon}\right) \times \sqrt{\text{var}(\hat{\beta}_T)}, \qquad (2)$$

where $t_{\text{Power},\upsilon}$ and $t_{1-\frac{\alpha}{2},\upsilon}$ are the respective $t$ values from the cumulative t distribution with $\upsilon$ degrees of freedom such that Power = Prob($t < t_{\text{Power},\upsilon}$) and $\left(1-\frac{\alpha}{2}\right) = \text{Prob}\left(t < t_{1-\frac{\alpha}{2},\upsilon}\right)$.

For example, if Power = 0.50, $\left(1-\frac{\alpha}{2}\right) = 0.975$, and $\upsilon = 159$, then $t_{\text{Power},\upsilon} = 0$ and $t_{1-\frac{\alpha}{2},\upsilon} = 1.975$. In this case, $\text{Beta}_{0.50} = 1.975 \times 0.003355 = 0.0066261$. If Power =0.10 and $\left(1-\frac{\alpha}{2}\right) = 0.975$, then $t_{\text{Power},\upsilon} = -1.2869$ and $t_{1-\frac{\alpha}{2},\upsilon} = 1.975$. In this case, $\text{Beta}_{0.10} = 0.6881 \times 0.003355 = 0.0023086$. In this way, Beta values can be calculated for Power = (0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, and 0.90).

## 5.2 Step Two: Simulating a Dataset

I began by setting Power = 0.10. Having estimated $\hat{\sigma}_\varepsilon^2$ and calculated $\text{Beta}_{0.10}$, I then simulated an artificial dataset modelled after BLW/Sullivan et al. (2004):

$$Y_i^\star = \boldsymbol{X_i'}\hat{\boldsymbol{\beta}}^\star + \varepsilon_i^\star,\ i = 1, 2, \ldots 190, \qquad (3)$$

where $\boldsymbol{X_i}$ is the data from BLW; $\hat{\boldsymbol{\beta}}^\star = (\hat{\beta}_0', \text{Beta}_{0.10}, \hat{\boldsymbol{\beta}}_{\boldsymbol{C}}')'$, $\hat{\beta}_0$, and $\hat{\boldsymbol{\beta}}_{\boldsymbol{C}}$ are the estimates from Table 3 in BLW, and $\text{Beta}_{0.10}$ comes from STEP ONE; and $\varepsilon_i^\star$ is a draw from a normal distribution with mean value 0 and variance $\hat{\sigma}_\varepsilon^2$.

## 5.3 Step Three: Estimating a Regression

Once I have simulated an artificial dataset consisting of 190 observations of $Y_i^\star$ and $\boldsymbol{X_i}$, I estimated an OLS regression and collected the estimates $\hat{\boldsymbol{\beta}}$ and residuals $\boldsymbol{e}$.

**Table 3:** Performance assessment of BLW and BLW[a]

| Power [1](%) | $Beta_{Power}$ [2] | Mean value [3](%) | Lower bound [4](%) | Upper bound [5](%) | MSE [6] |
|---|---|---|---|---|---|
| **Panel A:** BLW | | | | | |
| 10 | 0.00237 | 10.1 | 8.8 | 11.3 | 0.006 |
| 20 | 0.00389 | 20.2 | 18.2 | 22.1 | 0.010 |
| 30 | 0.00499 | 30.1 | 27.8 | 32.3 | 0.012 |
| 40 | 0.00592 | 40.1 | 37.9 | 42.2 | 0.010 |
| 50 | 0.00679 | 50.3 | 47.6 | 52.6 | 0.012 |
| 60 | 0.00767 | 60.0 | 57.8 | 62.4 | 0.012 |
| 70 | 0.00860 | 69.9 | 67.6 | 72.2 | 0.012 |
| 80 | 0.00970 | 80.1 | 78.7 | 82.5 | 0.012 |
| 90 | 0.01122 | 89.9 | 88.5 | 91.6 | 0.008 |
| **Panel B:** BLW[a] | | | | | |
| 10 | 0.00237 | 10.1 | 7.5 | 12.8 | 0.012 |
| 20 | 0.00389 | 20.1 | 16.8 | 23.5 | 0.017 |
| 30 | 0.00499 | 29.4 | 24.4 | 34.4 | 0.020 |
| 40 | 0.00592 | 40.4 | 36.2 | 45.0 | 0.022 |
| 50 | 0.00679 | 50.8 | 46.1 | 55.4 | 0.022 |
| 60 | 0.00767 | 59.9 | 55.9 | 63.6 | 0.021 |
| 70 | 0.00860 | 70.4 | 66.2 | 74.6 | 0.019 |
| 80 | 0.00970 | 79.9 | 76.5 | 88.5 | 0.016 |
| 90 | 0.01122 | 90.4 | 88.1 | 93.2 | 0.012 |

NOTE: Column 1 reports the true power. Column 2 reports the effect size that corresponds to that power (cf. equation (2)). Column 3 reports the mean estimated power over the 1,000 Monte Carlo experiments. Columns 4 and 5 report the 2.5 and 97.5% quantile values of the 1,000 estimates of power. Column 6 reports the MSE, defined as the average squared difference between the estimated power and the true power.

### 5.4 Step Four: Use the BLW and BLW[a] Methods to Estimate *Power*

I constructed the matrix ($\boldsymbol{Xe}$) and used the BLW and BLW[a] methods as described above to generate a *Power* estimate for the estimated treatment effect, $\hat{\beta}_T$, in STEP THREE.

### 5.5 Step Five: Repeat Step Four to Obtain 999 More *Power* Estimates for the Case when True *Power* = 0.10

Having generated one *Power* estimate each using BLW and BLW[a], I then repeated Step Four 999 more times until I had generated a total of 1,000 *Power* estimates for the case when true *Power* = 0.10. Note that a million regressions are estimated to obtain 1,000 *Power* estimates for a single true *Power* value.

### 5.6 Step Six: Repeat Steps Two through Five for *Power* Values 0.20, 0.30,...,0.90

Having obtained a sample of *Power* estimates for true *Power* = 0.10, I then repeated the whole process to get samples of estimates for true *Power* values = 0.20 through 0.90. With each sample of 1,000 estimates, I calculated the mean estimated *Power* value, the 90% sample interval which ranges from the 5th to the 95th percentile values, and the MSE of the estimates. This allowed me to both determine the absolute and relative performance of the two ex post estimators of *Power*, BLW and BLW[a].

Table 3 reports the results. The top panel reports the experimental results using BLW's method, and the bottom panel does the same for the alternative, BLW[a] method. For BLW's method, when true power is 10%, we calculate $Beta_{0.10}$ = 0. 00237. Following Steps Two through Five produces 1,000 estimates of ex post power. The mean of those estimates is 10.1%, with a 95% sample interval ranging between 8.8 and 11.3%. The associated MSE is 0.006. When I used the BLW[a] method, I again obtained a mean ex post power estimate of 10.1%. However, the 95% sample interval is wider, ranging from 7.5 to 12.8%, with an MSE of 0.012.

The results are similar when true power = 20%. BLW's method produces a mean ex post power estimate of 20.2%, with a 95% sample interval of (18.2, 22.1%). While BLW[a] has a smaller sample bias with a mean power estimate of 20.1%, it has a wider 95% sample interval of (16.8, 23.5%). This results in BLW having a lower MSE (0.010 vs 0.017). In fact, BLW has a narrower 95% sample interval and a smaller MSE for every true power value. Thus, for at least the set of Monte Carlo experiments in the table, the BLW method outperforms the BLW[a] method.

## 6 Conclusion

Replication plays, or should play, a fundamental role in any empirical science. To be able to independently confirm previously published results is critical for establishing a solid foundation for future research to build on. In this replication, I investigated BLW's procedure for calculating ex post power. While ex ante power calculations are commonly done in many fields, there is no generally accepted method for calculating ex post power. Into this space, BLW proposed a novel bootstrapping method. As an illustration, they applied their method to a study that produced a statistically insignificant estimate and showed how ex post power analysis can be

used to ascertain whether the insignificance was due to a negligible effect size or insufficient power. BLW's method provides a potentially very valuable tool for researchers.

I made three contributions with this replication. First, I verified that the data and code that BLW provided in their paper are reliable to reproduce their results. Second, I identified two shortcomings in their method that could impact the performance of their method. As a result, I proposed an alternative bootstrapping procedure (BLW$^a$). My third contribution used Monte Carlo experiments to assess the performance of BLW's original method and compare it to BLW$^a$. Despite its shortcomings, BLW outperformed BLW$^a$.

In terms of absolute performance, BLW performed well. For true power values ranging from 10 to 90%, BLW's method produced mean ex post power estimates that were very close to the true values. Further, the estimated power values were relatively closely clustered around their true values. For example, the 95% sample intervals always lay within 5% points of the true power values on either side. So, for example, when true power was 50%, 95% of the estimated ex post power values lay between 47.6 and 52.6%.

In conclusion, my replication provides further evidence of the reliability of BLW's method. A limitation of my replication is that it stayed closely within the confines of BLW's empirical application; specifically, it assumed a data generating process characterized by a linear specification with independently and homoscedastically distributed error terms. Future research should investigate whether these initial results extend to more complicated and realistic data environments.

**Acknowledgements:** I acknowledge the helpful feedback from participants of the New Zealand Association of Economists 2019 conference. Special thanks go to Tom Coupé and W. Robert Reed, the supervisors of my thesis, for their input on my research. I especially thank W. Robert Reed for providing generous editorial support for this article. Finally, I thank the careful comments of the reviewers whose constructive criticisms resulted in a much-improved manuscript.

**Funding information:** The author declares that he has no relevant or material financial interests that relate to the research described in this article.

**Conflict of interest:** Author states no conflict of interest.

**Article note:** As part of the open assessment, reviews and the original submission are available as supplementary files on our website.

**Data availability statement:** All the data and code necessary to reproduce the results in this study are available at: https://osf.io/6ahp7/.

## References

Bellamare, M. (2021, June 30). Top 5 agricultural economics journals–2021 Edition (Updated). *Marc F. Bellemare*. http://marcfbellemare.com/wordpress/13856.

Brown, J. P., Lambert, D. M., & Wojan, T. R. (2019). The effect of the conservation reserve program on rural economies: Deriving a statistical verdict from a null finding. *American Journal of Agricultural Economics*, *101*(2), 528–540.

Center for Open Science. (2022). *Non-HSR project definitions*. https://osf.io/upywe?view_only=495a1c72f0df4ccd9492962ae38d65e4.

Chong, S. F., & Choo, R. (2011). Introduction to bootstrap. *Proceedings of Singapore Healthcare*, *20*(3), 236–240.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Society of Industrial and Applied Mathematics CBMS-NSF Monographs, *38*. ISBN 0-89871-179-7.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55*(1), 19–24.

International Initiative for Impact Evaluation. (2022). *Replication studies*. https://www.3ieimpact.org/evidence-hub/replication-studies-status.

Ioannidis, J. P., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal*, *127*(October), F236–F265. doi: 10.1111/ecoj.12461.

Kleinman, K., & Huang, S. S. (2017). Calculating power by bootstrap, with an application to cluster-randomized trials. *EGEMs,* (*Generating Evidence & Methods to improve patient outcomes*). *4*(1), 1–18. doi: 10.13063/2327-9214.1202.

Skiba, P. M., & Tobacman, J. (2019). Do payday loans cause bankruptcy?. *The Journal of Law and Economics*, *62*(3), 485–519.

StataCorp. (2021). *Stata 17. Power, precision, and sample-size reference manual*. College Station, TX: Stata Press.

Sullivan, P., Hellerstein, D., Hansen, L., Johansson, R., Koenig, S., Lubowski, R. N., & Bucholz, S. (2004). The conservation reserve program: Economic implications for rural America. *USDA-ERS Agricultural Economic Report*, 834.

Wicklin, R. (2013, May 30). *Using simulation to estimate the power of a statistical test. SAS Blogs*. https://blogs.sas.com/content/iml/2013/05/30/simulation-power.html.

Wicklin, R. (2018, October 29). *Bootstrap regression estimates: Residual resampling. SAS Blogs*. https://blogs.sas.com/content/iml/2018/10/29/bootstrap-regression-residual-resampling.html.

Yuan, K. H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, *30*(2), 141–167.