DEPARTMENT OF ECONOMICS AND FINANCE

UC BUSINESS SCHOOL

UNIVERSITY OF CANTERBURY

CHRISTCHURCH, NEW ZEALAND

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)

Jiarui Tian

WORKING PAPER

No. 12/2021

Department of Economics and Finance
UC Business School
University of Canterbury
Private Bag 4800, Christchurch
New Zealand

WORKING PAPER No. 12/2021

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)

Jiarui Tian[1†]

November 2021

Abstract: This study replicates Brown, Lambert, and Wojan (2019) (BLW) and their bootstrapping procedure for calculating ex post power. At the current time there is no generally accepted way of calculating ex post power. BLW provide a novel method for doing this though they provide little justification for their method or evidence of its reliability. My replication makes three contributions. First, it confirms that the data and code provided with their paper is sufficient to reproduce their results. Second, it performs two robustness checks to determine if slight alterations to their procedure affect their results. I determine that including a constant term in their procedure does not affect the results. On the other hand, using a different bootstrapping procedure produces somewhat different results. However, without any ground truth to use as a benchmark, one cannot say which bootstrapping procedure is better. My third contribution is that I use Monte Carlo experiments to assess the performance of BLW's method. My experimental results indicate that their method is unbiased and produces a relatively narrow range of estimates. This suggests that BLW's method may provide a reliable method for researchers to calculate ex post power, though further investigation needs to be done.

Keywords: Ex post power, Statistical insignificance, Monte Carlo experiments, Bootstrapping, Replication

JEL Classifications: C12, C15, C18

Data and Code: All the data and code to reproduce the results of this study are publicly available at OSF: https://osf.io/6ahp7/

Acknowledgements: I acknowledge helpful feedback from participants at 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 previous versions of this paper.

[1] Department of Economics and Finance, University of Canterbury, NEW ZEALAND

† Corresponding author: Jiarui Tian. Email: alex.tian@pg.canterbury.ac.nz

1. Introduction

Null hypothesis significance testing (NHST) is the most frequently used approach to statistical inference in quantitative research. In NHST, a researcher selects an arbitrary alpha value, usually 0.05, then estimates a parameter and calculates the corresponding p-value. If the p-value is less than or equal to the chosen alpha, the null hypothesis is rejected. If the p-value is greater than alpha, the researcher fails to reject the null, and the result is considered statistically inconclusive. As Bausell and Li (2002) note, statistical insignificance may have more to do with the statistical design of a study than the effect itself. Insignificant estimates are expected when the effect being estimated does not exist. Alternatively, they can also arise when a study has too little statistical power to detect that effect.[1] The desire to distinguish between these two possibilities motivates the interest in developing ex post measures of statistical power.

One approach for calculating ex post power that has been widely used, but also severely criticized, relies on the estimated parameter and standard error. This approach is commonly known as "observed power" or "post hoc power". As Hoenig and Heisey (2001) demonstrate, post hoc power has a one-to-one, mathematical relationship with the p-value. As a result, it adds no new information beyond the already known p-value. Yuan and Maxwell (2005) demonstrate that as a measure of true power, post hoc power is generally biased and very imprecise.

To address this deficiency, Brown, Lambert & Wojan (2019), henceforth BLW, propose a method for calculating ex post power that can be used to distinguish insignificance caused by lack of statistical power and insignificance caused by a null effect. BLW apply their method to an analysis of the economic impact of the Conservation Reserve Program (CRP) by Sullivan et al. (2004). Sullivan et al. (2004) tested whether agricultural programs designed to remove environmentally sensitive land from production led to decreased employment in farm-

[1] Statistical power is the probability of detecting an effect, if there is a true effect present to detect.

dependent counties. They found a statistically insignificant relationship between land reduction by the CRP and changes in employment. BLW wanted to determine whether this meant the CRP had no effect, or whether the Sullivan et al. (2004) study did not have sufficient power to detect a possible negative effect. To answer this question, BLW used a bootstrapping procedure. Before explaining this procedure, I first provide context by describing Sullivan et al.'s (2004) analysis of the CRP.

2. Sullivan et al.'s (2004) Analysis of The CRP

Endogenous selection poses a challenge in estimating the effect of CRP participation on employment growth. To address this challenge, Sullivan et al. (2004) used a quasi-experimental, matched pair protocol to compare individual high-CRP counties with similar low-CRP counties. High-CRP counties were those counties that, on average, enrolled a higher percentage of their eligible land in CRP than other types of farms. Conversely, low-CRP counties enrolled a lower percentage of their eligible land in CRP than other types of farms.

Sullivan et al. (2004) reported CRP estimates for several models, but complete results were only reported for the long-run local employment growth model. BLW selected this model to replicate and use as a benchmark for their power analysis. In the Sullivan et al. (2004) study, the difference in employment growth over the period 1985 and 2000 between high-CRP (HCRP) and low-CRP (LCRP) counties was estimated using ordinary least squares (OLS) with the following regression specification:

$$y_i \equiv \ln\left(\frac{emp_{i,2000}^{HCRP}}{emp_{i,1985}^{HCRP}}\right) - \ln\left(\frac{emp_{i,2000}^{LCRP}}{emp_{i,1985}^{LCRP}}\right) \tag{1}$$

$$= (Beta \times CRP\ payments\ to\ income\ ratio_i) + \boldsymbol{X_i'\beta} + \varepsilon_i$$

where i indexes a matched pair; *CRP payments to income ratio* is the treatment variable; $Beta$ measures the true (unobserved) treatment effect; $\boldsymbol{X_i'}$ is a $1 \times k$ matrix of matched pair

differences on k control variables consisting of local socioeconomic and agriculture characteristics, and $\boldsymbol{\beta}$ is a $k \times 1$ vector of coefficients. ε is assumed to be an independently and identically distributed random error term with mean zero and constant variance. Sullivan et al. (2005) reported an estimate of $Beta$ equal to 0.007, with a standard error of 0.003, t-statistic of 1.945, and p-value equal to 0.054 (and thus insignificant at the 0.05 level). The statistical insignificance of the estimated treatment effect did not allow the researchers to reach a conclusion regarding a possible adverse employment effect of the CRP program. This is the context in which BLW applied their ex post bootstrapping procedure.

3. BLW's Ex Post Bootstrapping Method

Statistical power measures the probability of obtaining a statistically significant estimate conditional on a specific, true effect size. Statistical power will be greater for larger effect sizes than smaller ones. As their starting point, BLW identified the largest employment loss that would still provide a benefit-cost justification for the Conservation Reserve Program (CRP). They determined that any negative employment effects less than 0.27 in absolute value was sufficient to justify the CRP. Accordingly, they wanted to measure the power of the Sullivan et al. (2004) study to obtain a statistically significant effect given a value for $Beta = -0.027$. In addition, they also investigated statistical power associated with other effect sizes; namely, -0.015, -0.010, -0.005, and -0.001. The latter value was chosen as corresponding to the legislation's requirement to identify any negative employment effect associated with the program.

BLW's procedure involved sequentially inserting $Beta$ values of -0.027, -0.015, -0.010, -0.005, and -0.001 in the estimation framework of Sullivan et al. (2004) to determine their ex post power. BLW's method uses Monte Carlo simulations and consists of several steps. It starts by specifying the data-generating process (DGP) for the simulated datasets. Using Sullivan et

al.'s (2004) sample data and regression specification of Equation (1), BLW fix the coefficient vector for the control variables at the values estimated by Sullivan et al. (2004). The respective treatment effect for the Monte Carlo experiments, $Beta_{MC}$, is then used to construct new values of the dependent variable. These new values are combined with the original, resampled explanatory variables to produce a new estimate of the treatment effect. This process is repeated multiple times. The percentage of estimated treatment effects that are statistically significant provides an estimate of the power of the original estimating equation. BLW calculate the ex post power associated with different possible treatment effects, $Beta_{MC} = (-0.027, -0.015, -0.010, -0.005, -0.001)$.

In their experiments, BLW not only vary the size of the treatment effect, $Beta_{MC}$, they also construct simulated datasets with different numbers of observations. So in addition to replicating the size of the original dataset of 190 observations, they construct simulated samples of 100, 150, 200, 250, and 350 observations. This leads to a total of 30 experiments, one for each combination of effect size $Beta_{MC}$ $(-0.027, -0.015, -0.010, -0.005, -0.001)$ and sample size $(100, 150, 190, 200, 250, 350)$.

FIGURE 1 provides more detail about BLW's method. The first step consists of deciding the $(Beta_{MC}, sample\ size)$ combination to be used for the given experiment. For example, $Beta_{MC}$ might be set equal to -0.015 and sample size at 100 observations. BLW then construct individual observations of a new, simulated dataset by randomly selecting rows of data from the original dataset to generate new values of the dependent variable.

For example, to construct the first observation of the new dataset, the following items are selected from row 32 (a randomly selected row) of Sullivan et al.'s (2004) original data: (i) CRP_{32}; (ii) the 32nd row of the matrix of control variables, $\boldsymbol{X}$; and (iii) the residual, e_{32}. The first observation of the dependent variable is then generated by combining these as follows: $\widehat{Y_1} = -0.015 \cdot CRP_{32} + \hat{\beta}_1 \cdot X_{1,32} + \hat{\beta}_2 \cdot X_{2,32} + \cdots + e_{32}$, where the $\hat{\beta}_i$ come from the

Sullivan et al.'s (2004) original regression. The first observation of the new, simulated dataset is then ($\widehat{Y_1}$, CRP_{32}, $X_{1,32}$, $X_{2,32}$, …, $X_{30,32}$). Suppose the second row randomly selected by BLW's procedure is row 66. Then the corresponding second simulated value of the dependent variable is given by $\widehat{Y_2} = -0.015 \cdot CRP_{66} + \widehat{\beta_1} \cdot X_{1,66} + \widehat{\beta_2} \cdot X_{2,66} + \cdots + e_{66}$, and the corresponding second observation of the new, simulated dataset is then ($\widehat{Y_2}$, CRP_{66}, $X_{1,66}$, $X_{2,66}$, …, $X_{30,66}$). And so on.

As the resampling is done with replacement, it is possible for the same row to be selected more than once. This is also illustrated in FIGURE 1 where I have shown row 32 from the original data is again selected in constructing the third observation of the new, simulated dataset. This process continues until the predetermined size of the simulated dataset is achieved.

Once the new dataset is constructed, the procedure continues by regressing the vector of simulated *Y* values on the reconstituted vector of the treatment variable, *CRP,* along with the respective control variables. It then tests H_0: $Beta = 0$ at $\alpha = 0.05$ and records whether H_0 is rejected. This bootstrapping process is repeated *M* times. The percent of times one obtains a significant estimate for $Beta$ is the ex post power from BLW's method. For example, if *M* = 10,000, and the estimated value of $Beta$ is significant in 3,000 of the simulations, then ex post power is calculated to be 30%. BLW repeat this process for all 30 combinations of $Beta_{MC}$ and sample size values. My first contribution is to see if I can reproduce their results with the data and code they supply with their paper.

4. Replication

Column I of TABLE 1 reproduces the power statistics from Table 4 of BLW's original paper. I replicate their results two ways. First I use the R code and data they supplied with their paper to see if I get the same estimates. The results from this exercise are reported in Column II. I

then take their R code, rewrite it using Stata, and conduct the Monte Carlo experiments with the new code. These results are reported in Column III.

BLW conduct statistical power calculations for a wide range of effect and sample sizes. -0.027 is the effect size they care about the most, because any job loss less than that is "acceptable" from a benefit-cost perspective. Thus they want to be sure the Sullivan et al. (2004) study has sufficient power to obtain a significant estimate if the job loss is that large. For each effect size, they calculate ex post power for sample sizes of 100, 150, 190, 200, 250, 300, and 350. BLW are particularly interested in the results for sample sizes equal to 190, as that is the sample size used in the Sullivan et al. (2004) study. I highlight the corresponding rows in the table. The other sample sizes are included to give a sense of how power changes with sample size for a given effect size.

As expected, ex post power is largest 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 (see the top portion of TABLE 1), BLW calculate that the Sullivan et al. (2004) study has 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 Sullivan et al.'s (2004) study 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 II and III in TABLE 1 report my efforts to replicate BLW's results, first using their R code and then rewriting their program in Stata. Using their R code, I am able to exactly reproduce their results. Using my Stata version of their program, I can reproduce their results with some miniscule differences. For example, when the effect size is -0.015 and the sample size is 100, BLW report an ex post power value of 0.84, but my Stata replication for this case

produces a power value of 0.85. I attribute these differences to rounding and the fact that the random number generators underlying the simulations use different seeds.

Using their code, I obtain 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. This demonstrates that I am correctly implementing their procedure and provides confidence in the analyses that follow in the subsequent sections.

5. Robustness Check

To further test the reliability of the BLW method, I perform some robustness tests. TABLE 2 reports the results of two robustness checks where I (i) add a constant term, and (ii) use a different resampling procedure.

As noted above, Sullivan et al. (2004) used Equation (1) to estimate the effect of the CRP. Their data matrix $\boldsymbol{X_i}$ did not include a constant term. Perhaps this was due to the fact that their dependent variable was in differences and differencing would eliminate the constant term from the untransformed equation. However, it is generally considered bad practice to estimate a regression equation without a constant term. Omitting the constant term forces the regression line to go through the origin. This could bias estimates of the coefficients of the model's variables. To test the robustness of BLW's results, I therefore repeated their analysis, this time adding a constant term to the regressions estimated in the simulations.

In addition, I evaluated whether the specific bootstrap method used by BLW affects their power estimates. There are two main ways to bootstrap by resampling: (1) treat the regressors as random and resample both variables and residuals, or (2) treat the regressors as fixed and resample solely from the residuals of the fitted regression model.

In BLW's method, the X's and residuals are paired with each other and randomly sampled as a set. This method was illustrated in FIGURE 1. An alternative approach fixes the

original X's in each simulation, and randomly samples and matches the residuals. I call this alternative approach BLW[a] and illustrate it in FIGURE 2. The difference is that BLW[a] only "shuffles" the residuals. For example, if there are 190 observations in the original dataset, there will still be 190 observations in the new dataset, but the arrangement of the residuals will be altered, with some residuals potentially appearing more than once.

In FIGURE 2, the first observation of the new, simulated dataset using BLW[a] is $\widehat{Y_1} = Beta_{MC} \cdot CRP_1 + \hat{\beta}_1 \cdot X_{1,1} + \hat{\beta}_2 \cdot X_{2,1} + \cdots + e_{32}$. Contrast this with $\widehat{Y_1} = Beta_{MC} \cdot CRP_{32} + \hat{\beta}_1 \cdot X_{1,32} + \hat{\beta}_2 \cdot X_{2,32} + \cdots + e_{32}$ under BLW. Likewise, the second observation using BLW[a] is $\widehat{Y_2} = Beta_{MC} \cdot CRP_2 + \hat{\beta}_1 \cdot X_{1,2} + \hat{\beta}_2 \cdot X_{2,2} + \cdots + e_{66}$ instead of $\widehat{Y_2} = Beta_{MC} \cdot CRP_{66} + \hat{\beta}_1 \cdot X_{1,66} + \hat{\beta}_2 \cdot X_{2,66} + \cdots + e_{66}$.

There is a reason why I am interested in investigating the performance of this alternative variant of BLW. BLW does not consider how changes in the composition of the data matrix affect ex post power. Suppose X is a binary variable taking values 0 and 1. Suppose that in a dataset of 10,000 observations, X = 0 for half the sample and X = 1 for the other half. Consider the DGP: $Y = 1000 + 3.92 \cdot X + error,$ where $error \sim N(0, \sigma^2)$, and $\sigma = 100$. As shown below, the corresponding standard error of the slope coefficient from OLS estimation of this equation is 2. Accordingly, approximately half of the sample *t*-values will be greater than 1.96 ($= \frac{3.96}{2}$), and half will be less than 1.96, so that the corresponding statistical power will be 50%.

Now consider the case where individual *X* values are resampled with replacement. Because the same X values can be resampled more than once, it is possible that a simulated dataset could consist of an unequal number of 0's and 1's, say, 7,500 observations with X = 1 and 2,500 observations with X = 0. Using the same DGP as above, it is straightforward to show that BLW's method would produce an ex post power value of 43% in this case. In other words, BLW's ex post power method depends on the particular configuration of *X* values drawn from the random selection procedure. If, however, we are interested in calculating ex post power for

the specific configuration of *X* values in the original data, these should be fixed at their original values. This is the motivation behind the alternative bootstrapping procedure, BLW[a].

The results of my robustness checks are reported in TABLE 2. The first column reproduces the Stata results from TABLE 1 to facilitate comparison. The second column ("Robustness 1") reports the results when I repeat BLW's procedure, this time adding a constant term when estimating the regressions in the simulations. The third column ("Robustness2") employs the variant BLW[a] procedure described above. Note that the BLW[a] procedure always has 190 observations because it fixes the dataset at its original values, without resampling. Only the residuals are resampled.

As one can see from TABLE 2, including a constant term makes no appreciable difference. I conclude that BLW, at least in this instance, is robust to the exclusion of a constant term. In contrast, the alternative resampling method BLW[a] does make a small difference. For example, when $Beta_{MC} = -0.01$ 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 $Beta_{MC} = -0.005$ and $Beta_{MC} = -0.001$.

The differences between BLW and BLW[a] raise a number of questions. Without some ground truth to compare to, it is difficult to say which method is "better". However, a full performance assessment lies beyond the purview of this replication. Such an assessment would go beyond the example of Sullivan et al. (2004) and consider other error structures, such as clustering. Once clustering is brought into the analysis, additional bootstrapping methods, such as wild-cluster bootstrapping (Cameron et al., 2008; Roodman, 2019), should also be considered. A full exploration of this issue would lead me away from my focus of replicating BLW. Nevertheless, a focused performance assessment of BLW's method could provide insight into the value of their method. I pursue this in the next section.

6. A Performance Analysis of BLW

Somewhat surprisingly, BLW do not provide any discussion to justify their method, only citing Cameron & Trivedi's (2006) classic textbook, *Microeconometric Analysis*. To assess the performance of BLW's method, I conduct Monte Carlo experiments where I create a data generating process (DGP) with known *Power*, then see how well BLW's method is able to estimate it. I start with a simple DGP:

$$y_i = 100 + ES \cdot x_i + error_i, \quad (2)$$

where ES is the effect size and error ~ N (0, 100^2). I set a sample size of 10,000 observations, with half receiving the treatment (x=1) and half not (x=0).

Starting from a significance level $\alpha = 0.05$ and $N = 10{,}000$, it is easy to complete the conversion from the variance of the error term to the standard error of the estimated effect, $s.e.(\widehat{ES})$:

$$s.e.(\widehat{ES}) = \sqrt{\frac{\sigma^2}{\sum(x_i - \bar{x})^2} \times \frac{N-1}{N-2}}, \quad (3)$$

where σ^2 is the variance of the error term, $\sum(x_i - \bar{x})^2$ is the sum of squared deviations of *x*, and *N* is the sample size. Given that half of the observations receive the treatment and half do not, *x*={0,1}, $\bar{x} = 0.5$, $(x - \bar{x}) = \pm 0.5$, and $(x - \bar{x})^2 = 0.25$. Thus, $\sum(x_i - \bar{x})^2 = 10{,}000 \times 0.25 = 2500$. It follows that $s.e.(\widehat{ES}) = \sqrt{\frac{10000}{2500} \times \frac{9999}{9998}} \approx 2$.

Once $s.e.(\widehat{ES})$ is known, then the *ES* corresponding to any given *Power* value can be calculated using the following formula (Djimeu & Houndolo, 2016):

$$ES = (t_{Power,v} + t_{1-\frac{\alpha}{2},v}) \times s.e.(\widehat{ES}) \quad (4)$$

where (i) $t_{1-\frac{\alpha}{2},v}$ is the value of the t-distribution with *v* degrees of freedom such that Prob($t \leq t_{1-\frac{\alpha}{2},v}$) $= \left(1 - \frac{\alpha}{2}\right) \times 100\%$; and (ii) $t_{Power,v}$ is the corresponding value such that Prob($t \leq$

$t_{Power,v}) = Power$. For example, given *Power* equal to 40%, α=0.05 and *N*=10,000, it follows that $t_{1-\frac{\alpha}{2},v} = 1.96$ and $t_{0.4}, v = -0.25$. Thus, $ES = (1.96 + (-0.25)) \times 2 = 3.42$. In this manner I calculate effect sizes, *ES*, for Equation (3) that correspond to *Power* values of 20%, 30%, …, 80%, 90%.

I need to make a small modification to calculate *ES* when *Power* is 10%. For small values of *Power,* Equation (4) produces values of *ES* that are somewhat too large. This happens because Equation (4) only looks at one tail of the *t*-distribution. Equation (5) gives the correct formula to calculate *ES* when *Power* is small:

$$1 - \phi\left(t_{1-\frac{\alpha}{2},v} - \frac{|ES|}{s.e.(\widehat{ES})}\right) + \phi\left(t_{\frac{\alpha}{2},v} - \frac{|ES|}{s.e.(\widehat{ES})}\right) = Power, \quad (5)$$

where ϕ is the cumulative distribution function of the t-distribution with v degrees of freedom. If α=0.05, $s.e.(\widehat{ES}) = 2$, $Power = 10\%$, and v is very large, then Equation (5) yields *ES* = 1.31. Using the one-tailed formula in Equation (4) produces a value for *ES* = 1.36, which is slightly larger. For larger values of *Power*, say 20%, the difference between the two equations becomes negligible and I can ignore the other tail.

For a given *Power* and corresponding *ES* value, I run a Monte Carlo experiment where I generate 10,000 observations using the DGP in Equation (2). I then use these simulated data to estimate the regression equation, $y_i = \beta_0 + \beta_1 \cdot x_i + error_i$, and save the residuals, $\widehat{error_i}$. The next step consists of resampling the 10,000 observations (x_i, $\widehat{error_i}$) with replacement, and creating 10,000 corresponding $\widehat{y_i}$ values as per BLW's procedure. I then take the simulated $\widehat{y_i}$ and resampled x_i values and run a regression of $\widehat{y_i}$ on x_i. This produces a single estimate of β_1, and I record whether it is statistically significant. I repeat this process 999 times until I have 1,000 estimates of β_1. The percentage of β_1 estimates that are statistically significant provides one estimate of *Power* for that experiment.

I then repeat the experiment above, generating 10,000 new observations from the DGP in Equation (2), using the same *ES* value as before. This produces a new set of residuals, $\widehat{error}_i$, which I then resample with replacement to create new simulated $\hat{y}_i$ values. I regress these on the resampled x_i values to get a fresh estimate for β_1. I repeat this process until I have a completely new set of 1,000 estimates of β_1. The associated percentage of significant estimates produces a second estimate of *Power.*

The whole process above is repeated again and again until I obtain 1,000 *Power* estimates. By comparing the mean and 95% sample interval of these 1000 estimates with the true *Power* value, I can assess how well BLW's method performs. I do this for each *Power* value, *Power = 10%, 20%, 30%, ..., 80%,* and *90%.*

The results are reported in TABLE 3. Each row summarizes the experimental results for a given true *Power* and corresponding *ES* value consisting of 1,000 *Power* estimates, which in turn are each based on 1,000 estimates of β_1. Column (I) reports the true *Power* value. Column (II) reports the average *Power* value for the sample of 1,000 *Power* estimates. It provides a measure of bias. Column (III) reports a 95% "confidence interval", where the lower and upper bounds are set equal to the 0.025 and 0.975 percentile values of the distribution of 1,000 estimated *Power* values. It provides a measure of precision.

My simulation results provide evidence that BLW's method performs exceptionally well on the dimension of bias. As shown in Column II, average *Power* values are very close, if not exactly equal, to the true *Power* values. Column III assesses how close the individual *Power* estimates are to the true values. For example, when true *Power* is 10%, 95% of the *Power* estimates produced by BLW's method lie between 8.3% and 11.9%. Performance is similar for other *Power* values. When *Power* is 50%, the 95% sample interval ranges from 46.9% to 53.0%. For 80% *Power,* the corresponding interval is bounded by 77.1% and 82.6%. I view these measures as highly favourable to BLW's procedure. While the results in TABLE 3 derive

from a very simple data environment with a spherical error structure, they provide some evidence that BLW's procedure could provide reliable estimates of ex post power.

7. 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 investigate Browne, Lambert, and Wojan's (2019) (BLW) procedure for calculating ex post power. While ex ante power calculations are commonly done in many fields, these basically consist of estimating the standard error of the estimated treatment effect in advance. Whether these ex ante estimates are sufficient to produce reliable estimates is power is unknown. It would be useful to compare them with ex post estimates of power. Unfortunately, there is at the current time no generally accepted way of calculating ex post power. BLW provide a novel method for doing this though they provide little justification for their method or evidence of its reliability.

It is in that context that my replication of their work makes three contributions. First, it confirms that the data and code they provide with their paper is sufficient to reproduce their results. Second, it performs two robustness checks to determine if slight alterations to their procedure affect their results. I determine that including a constant term in their procedure does not affect the results. On the other hand, using a different bootstrapping procedure does produce somewhat different results. However, without any ground truth to use as a benchmark for comparison, one cannot say which bootstrapping procedure is better.

My third contribution is that I use Monte Carlo experiments to assess the performance of BLW's method on the basis of biasedness and precision. My experimental results indicate that their method is unbiased and produces a relatively narrow range of estimates. These experimental results suggest that BLW's method may provide a reliable method for researchers

to calculate ex post power. However, my experiments employ a relatively simple data generating process. Future research should investigate whether these promising initial results extend to more complicated, and realistic, data environments.

REFERENCES

Bausell, R. B., & Li, Y. F. (2002). *Power analysis for experimental research: a practical guide for the biological, medical and social sciences*. Cambridge University Press.

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.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3), 414-427.

Cameron, C., and P. Trivedi. 2006. *Microeconometric Analysis*. Cambridge: Cambridge University Press.

Djimeu, E. W., & Houndolo, D. G. (2016). Power calculation for causal inference in social science: sample size and minimum detectable effect determination. *Journal of Development Effectiveness, 8*(4), 508-527.

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.

Levine, M., & Ensom, M. H. (2001). Post hoc power analysis: an idea whose time has passed?. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 21*(4), 405-409.

Roodman, D., Nielsen, M. Ø., MacKinnon, J. G., & Webb, M. D. (2019). Fast and wild: Bootstrap inference in Stata using boottest. *The Stata Journal*, 19(1), 4-60.

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).

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.

TABLE 1
Replication of BLW's Ex Post Power Results

$Beta_{MC}$	*n*	*BLW (I)*	*Reproduction-R (II)*	*Reproduction-Stata (III)*
-0.027	100	0.99	0.99	0.99
-0.027	150	1.00	1.00	1.00
-0.027	190	1.00	1.00	1.00
-0.027	200	1.00	1.00	1.00
-0.027	250	1.00	1.00	1.00
-0.027	300	1.00	1.00	1.00
-0.027	350	1.00	1.00	1.00
-0.015	100	0.84	0.84	0.85
-0.015	150	0.96	0.96	0.96
-0.015	190	0.99	0.99	0.99
-0.015	200	0.99	0.99	0.99
-0.015	250	1.00	1.00	1.00
-0.015	300	1.00	1.00	1.00
-0.015	350	1.00	1.00	1.00
-0.010	100	0.59	0.59	0.59
-0.010	150	0.79	0.79	0.79
-0.010	190	0.88	0.88	0.88
-0.010	200	0.90	0.90	0.89
-0.010	250	0.96	0.96	0.95
-0.010	300	0.98	0.98	0.98
-0.010	350	0.99	0.99	0.99
-0.005	100	0.24	0.24	0.23
-0.005	150	0.33	0.33	0.33
-0.005	190	0.42	0.42	0.42
-0.005	200	0.43	0.43	0.42
-0.005	250	0.51	0.51	0.53
-0.005	300	0.60	0.60	0.60
-0.005	350	0.67	0.67	0.67
-0.001	100	0.06	0.06	0.06

$Beta_{MC}$	*n*	*BLW (I)*	*Reproduction-R (II)*	*Reproduction-Stata (III)*
-0.001	150	0.06	0.06	0.06
-0.001	190	0.06	0.06	0.07
-0.001	200	0.06	0.06	0.06
-0.001	250	0.07	0.07	0.07
-0.001	300	0.07	0.07	0.07
-0.001	350	0.07	0.08	0.08

NOTE: Column I shows the statistical power described in BLW's paper (Page 10, Table 4). Column II shows that statistical powers I produce when using BLW's R code. Column III shows that statistical powers I get after reprogramming BLW's method in Stata. All results in the table are presented to two decimal places.

TABLE 2
Robustness Checks

$Beta_{MC}$	*n*	*Reproduction (Stata)*	*Robustness1 (Constant Term)*	*Robustness2 (BLW^{a} Method)*
-0.027	100	0.99	0.99	----
-0.027	150	1.00	1.00	----
-0.027	190	1.00	1.00	1.00
-0.027	200	1.00	1.00	----
-0.027	250	1.00	1.00	----
-0.027	300	1.00	1.00	----
-0.027	350	1.00	1.00	----
-0.015	100	0.85	0.83	----
-0.015	150	0.96	0.96	----
-0.015	190	0.99	0.99	1.00
-0.015	200	0.99	0.99	----
-0.015	250	1.00	1.00	----
-0.015	300	1.00	1.00	----
-0.015	350	1.00	1.00	----
-0.010	100	0.59	0.58	----
-0.010	150	0.79	0.78	----
-0.010	190	0.89	0.87	0.94
-0.010	200	0.89	0.88	----
-0.010	250	0.95	0.95	----
-0.010	300	0.98	0.98	----
-0.010	350	0.99	0.99	----
-0.005	100	0.23	0.24	----
-0.005	150	0.33	0.32	----
-0.005	190	0.42	0.42	0.49
-0.005	200	0.42	0.42	----
-0.005	250	0.53	0.52	----
-0.005	300	0.60	0.59	----
-0.005	350	0.67	0.66	----
-0.001	100	0.06	0.06	----
-0.001	150	0.06	0.06	----

$Beta_{MC}$	*n*	*Reproduction (Stata)*	*Robustness1 (Constant Term)*	*Robustness2 (BLW^a Method)*
-0.001	190	0.07	0.07	0.09
-0.001	200	0.06	0.07	----
-0.001	250	0.07	0.07	----
-0.001	300	0.07	0.07	----
-0.001	350	0.08	0.08	----

NOTE: "*Reproduction (Stata)*" identifies statistical powers I get after reprogramming BLW's method in Stata. "*Robustness1*" means the first robustness check where I include a constant term. "*Robustness2*" means I apply the alternative bootstrapping method BLW^a as described in the text. All results in the table are reported to two decimal places.

TABLE 3
Replication of BLW1 Power with DGP Results

ES	*True Power (I)*	*BLW's Power (II)*	*95% Sample Interval (III)*
1.31	0.100	0.100	(0.083, 0.119)
2.24	0.200	0.200	(0.175, 0.225)
2.87	0.300	0.300	(0.270, 0.328)
3.42	0.400	0.399	(0.367, 0.430)
3.92	0.500	0.499	(0.469, 0.530)
4.43	0.600	0.599	(0.567, 0.633)
4.97	0.700	0.699	(0.670, 0.728)
5.60	0.800	0.799	(0.771, 0.826)
6.48	0.900	0.900	(0.879, 0.919)

NOTE: Column I shows the statistical power I set in my data generating process (DGP), which is the true power. Column II shows that statistical powers I get using BLW's method. Column III shows a 95% sample interval of BLW's power.

FIGURE 1. Schematic Diagram of Original BLW Bootstrapping Method

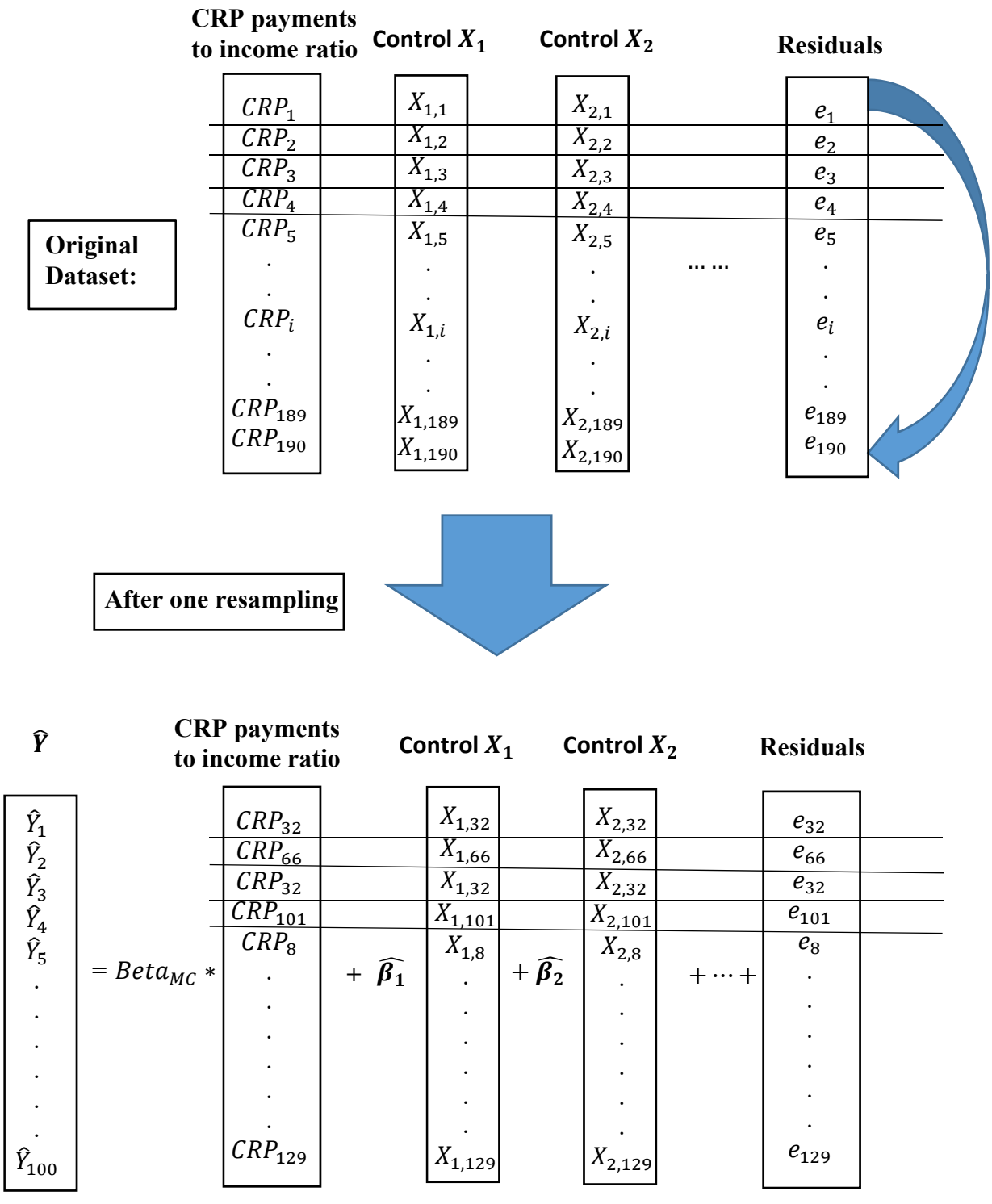
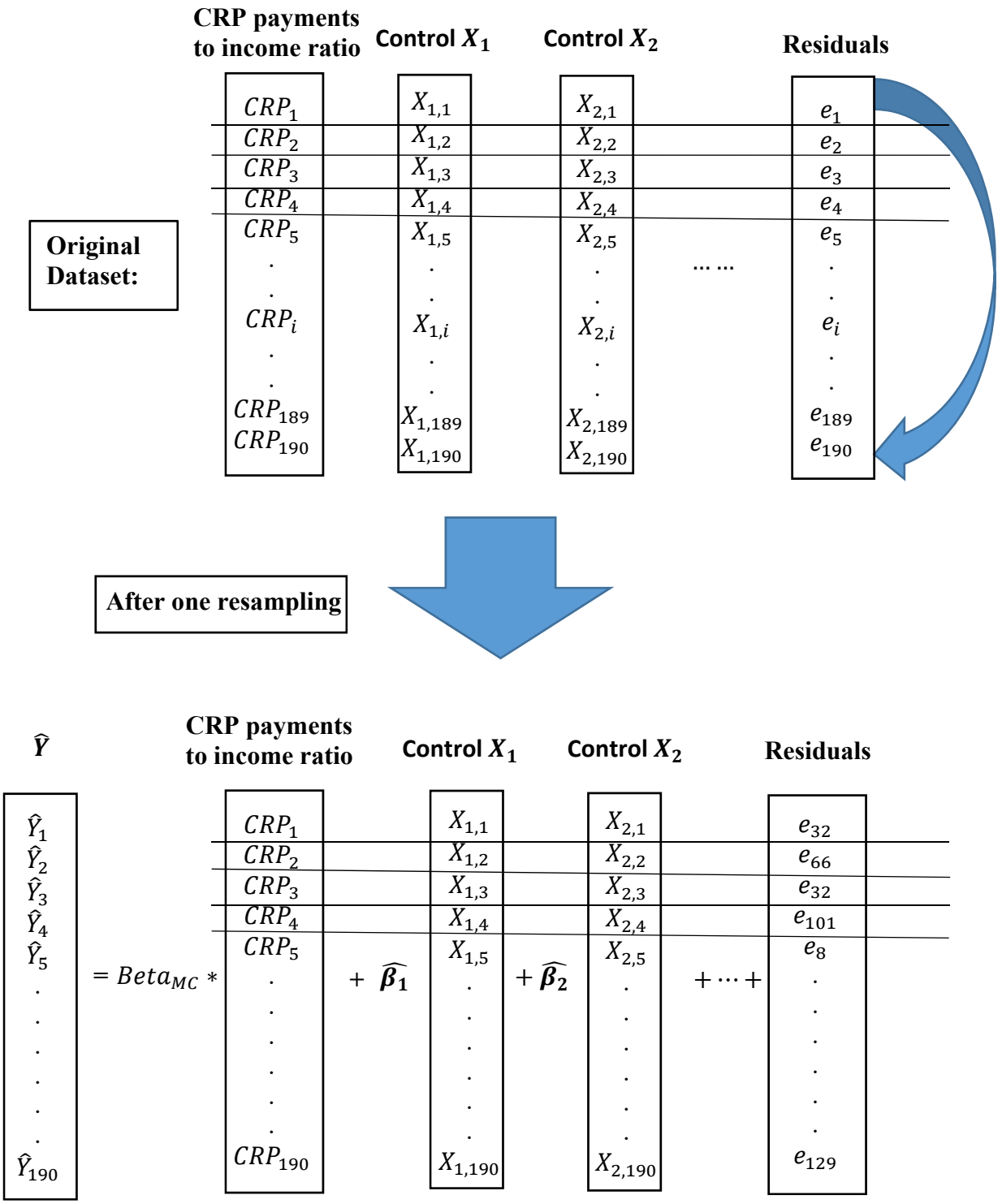
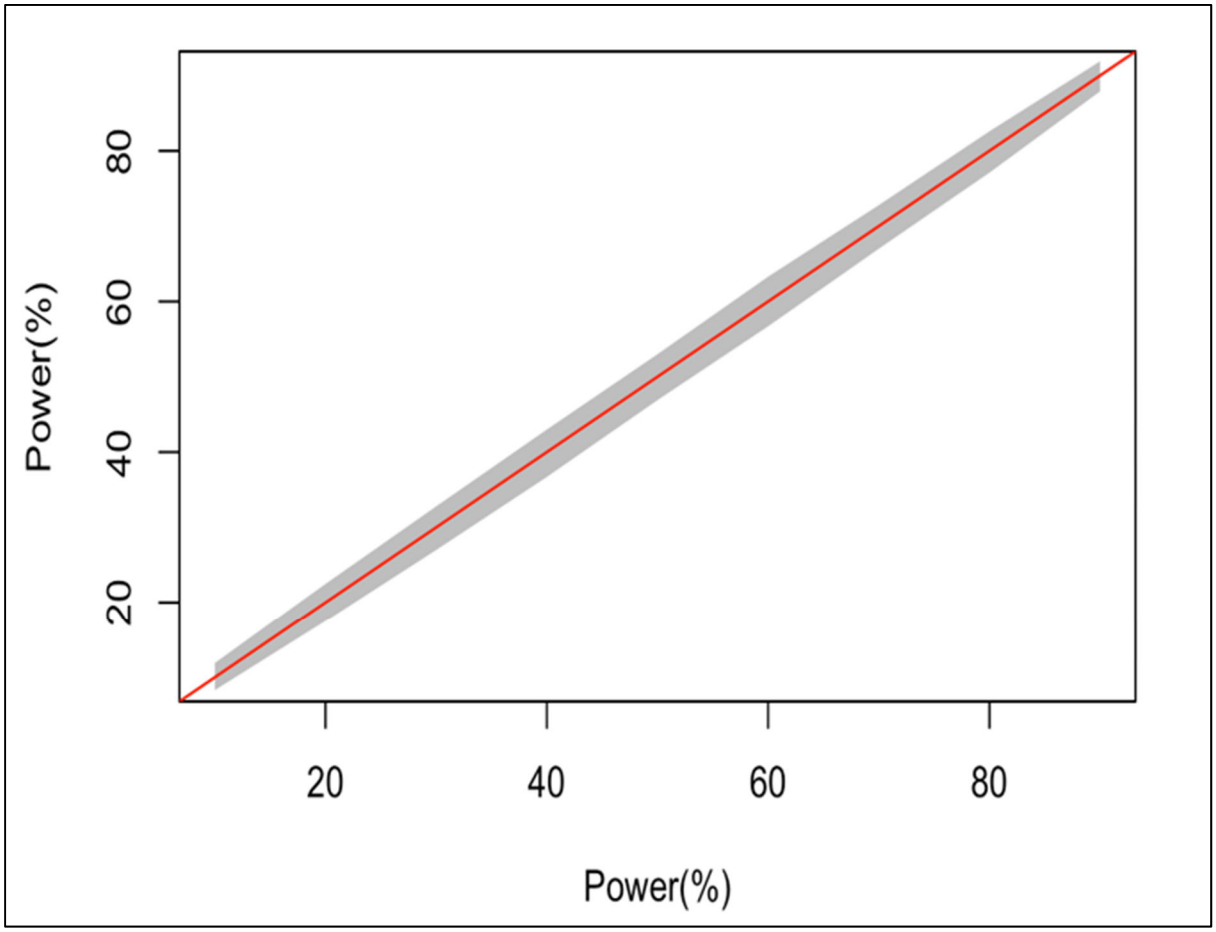


FIGURE 2. Schematic Diagram of BLW[a] Bootstrapping Method



NOTE: For BLW[a], there are always 190 observations.

FIGURE 3
Distribution of Estimated Power Values as a Function of True Power Using BLW's Method



NOTE: The grey area shows the 95% sample intervals of estimated *Power* values from Monte Carlo experiments using BLW's method (cf. Section 6 in the text). Note that sample intervals only exist for *Power* values of 10%, 20%, 30%, …, 80%, and 90%. The intervening areas are filled in to facilitate legibility.