Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Adjustment for autocorrelation in multiple-group (controlled) interrupted time series analysis and its effect on power: A simulation study of the Newey-West and Prais-Winsten methods

Ariel Linden

ariel.linden@ucsf.edu

University of California, San Francisco

Research Article

Keywords: Interrupted time series analysis, multiple-group interrupted time series analysis, controlled interrupted time series analysis, power, autocorrelation

Posted Date: March 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8865851/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Adjustment for autocorrelation in multiple-group (controlled) interrupted time series analysis and its effect on power: A simulation study of the Newey-West and Prais-Winsten methods

Ariel Linden, DrPH
University of California, San Francisco
Department of Medicine
Division of Division of Clinical Informatics & Digital Transformation (DoC-IT)
ariel.linden@ucsf.edu

Abstract

Background: Multiple-group (controlled) interrupted time series analysis (MG-ITSA) is increasingly used to evaluate interventions in healthcare settings when randomized designs are infeasible. Simulation-based power analysis has been proposed to support prospective study planning in this framework, but uncertainty remains regarding how different approaches to autocorrelation adjustment affect power and inferential performance. This study evaluates how two commonly used methods (ordinary least squares with Newey-West standard errors [OLS-NW] and Prais-Winsten [PW] regression) compare when estimating power and other performance measures in MG-ITSA designs with first-order [AR(1)] autocorrelation.

Methods: A comprehensive Monte Carlo simulation study was conducted based on the standard MG-ITSA regression model. Data were generated under a wide range of design conditions, including varying numbers of time periods, numbers of control units, effect sizes, and levels of first-order autocorrelation. Treatment effects were defined as difference-in-differences in level and difference-in-differences in trend. In addition to power, several other performance measures were computed, included percentage bias, empirical standard errors, root mean squared error (RMSE), 95% confidence interval coverage, and Type I error rates. All analyses were conducted using the POWER_ITSA package for Stata.

Results: Across all scenarios, power increased with longer series, larger effect sizes, and more control units. OLS-NW consistently produced higher power than PW for both level and trend effects, with differences often exceeding 10%. Both methods yielded approximately unbiased estimates and nearly identical RMSE. However, important differences in inferential performance were observed: OLS-NW resulted in substantially lower confidence interval coverage and

markedly inflated Type I error rates, particularly in shorter series, whereas PW maintained near-nominal coverage and Type I error control across all conditions.

Conclusions: In MG-ITSA with AR(1) autocorrelation, higher power achieved using OLS-NW comes at the cost of poor finite-sample inferential validity. PW provides more reliable hypothesis testing and uncertainty quantification, albeit with lower power. These findings highlight a fundamental trade-off between power and inferential calibration and underscore the importance of aligning estimator choice with study objectives when planning prospective MG-ITSA studies.

Keywords: Interrupted time series analysis, multiple-group interrupted time series analysis, controlled interrupted time series analysis, power, autocorrelation

Background

Interrupted Time Series Analysis (ITSA) is widely applied in healthcare research to evaluate the effects of interventions, policy changes, and natural events. This design examines outcomes measured as sequential observations collected at consistent intervals, typically representing aggregated data such as morbidity or mortality rates, and average expenditures. In most cases, a single entity (e.g., hospital or region) is exposed to a "treatment" expected to produce an observable shift ("interruption") in the level and/or trend of the series following implementation [1,2]. ITSA is often regarded as having strong internal validity due to its control of regression to the mean [1-3] and can also exhibit robust external validity when conducted on population-level data or generalized across settings [1,2]. Despite these strengths, the single-group design (SG-ITSA) suffers from key validity threats. Recent evidence suggests that SG-ITSA may fail to account for extraneous influences or may produce misleading conclusions when pre-treatment trends already reflect directional change [4].

The limitations of SG-ITSA can be overcome by incorporating one or more control units that are comparable to the treated unit on baseline level, outcome trajectory, and other observable pre-treatment characteristics [5-9]. This design, known as multiple-group (or controlled) ITSA (MG-ITSA), enables the construction of a credible counterfactual for estimating treatment effects [10].

Despite the increasing application of MG-ITSA in healthcare research (for some recent examples, see [11-20]), formal power analyses remain rare because most studies are retrospective, rendering prospective power estimation largely inappropriate [21]. Among the limited number of prospective studies, methods for calculating power have differed substantially. Fanshawe et al. [22], for example, estimated power by fitting autoregressive integrated moving average (ARIMA) models to each site's pre-intervention time series to account for trends and autocorrelation, then simulating post-intervention outcomes under prespecified effect sizes for treated sites and combining these with control-site forecasts to construct counterfactuals. Yelland et al. [23] appear to have relied on cluster-level standard deviations and intra-class correlation to estimate power to detect a relative change in outcomes, although their specific procedure was not clearly described. Power et al. [24] employed a simpler approach, comparing baseline averages with end-of-study averages to compute power. This diversity of approaches highlights the absence of a standardized methodology. Although Schochet [25] derived closed-form expressions for power and minimum

detectable effects under clustered fixed-effects models, this framework requires multiple clusters and is therefore ill-suited to typical MG-ITSA settings with one treated unit and few controls. To address this gap, Linden [26] proposed a simulation-based power analysis using ordinary least squares estimation with Newey-West heteroskedasticity- and autocorrelation-consistent (HAC) standard errors [27].

The current paper extends Linden's [26] work by comparing power curves and other model performance measures derived using OLS with Newey-West standard errors (OLS-NW) to those obtained using Prais-Winsten (PW) regression [28]. In contrast to SG-ITSA in which many univariate time series models are available to evaluate treatment effects, these two models are amongst the limited array of statistical frameworks available for evaluating MG-ITSA designs. Although alternative estimators exist, such as those derived from the autoregressive conditional heteroskedasticity (ARCH) family (see [29,30] for a comprehensive treatment of econometric time series models), their methodological complexity renders them less practical for researchers lacking advanced econometric training. In contrast, the OLS-NW approach offers ease of implementation and straightforward interpretability, while the PW model, despite its additional transformation step, remains relatively accessible and methodologically transparent.

We conduct a comprehensive simulation study to compare the performance of these two methods under a large range of scenarios to address factors unique to the MG-ITSA study design when estimating power using time series regression. These factors include the total number of time periods under study, the number of control units serving as comparators, varying effect sizes, handling of the autocorrelated nature of the data, and the measure of effect which can be expressed as either differences in the pre- to post-treatment change in *level* (i.e. difference in differences in level) or differences in the pre- to post-treatment change in *trend* (i.e. difference in differences in trend) of the outcome variable. In addition to power, we compute several other performance measures in order to explore sources of observed variation between models. The results of this study, as well as the availability of a recently introduced community-contributed package for Stata called POWER_ITSA [31], are intended to assist investigators in selecting and implementing the most appropriate technique for power estimation in prospective MG-ITSA applications.

Methods

The MG-ITSA model

The MG-ITSA regression model assumes the following form [7,32]:

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 X_t + \beta_3 X_t T_t + \beta_4 Z + \beta_5 Z T_t + \beta_6 Z X_t + \beta_7 Z X_t T_t + \epsilon_t \qquad (1)$$

where Y_t is the aggregated outcome variable measured at each time-point t, T_t is the time since the start of the study; X_t is a dummy variable representing the treatment (pre-treatment periods = 0, otherwise 1); Z is a dummy variable to denote the cohort assignment (treatment or control); and $X_t T_t$, $Z T_t$, $Z X_t$ and $Z X_t T_t$ are all interaction terms among these variables. Coefficients $\beta_0 - \beta_3$ represent the control group and the coefficients $\beta_4 - \beta_7$ represent values of the treatment unit. More specifically, β_0 represents the intercept, or starting level of the aggregated outcome variable of the control group; β_1 is the slope, or trajectory of the outcome variable for the control group until the introduction of the treatment; β_2 represents the intercept at the time-point immediately following the introduction of the treatment and indicates whether there was a change in the level of the outcome immediately following the introduction of the treatment for the control group; β_3 represents the change in slope or trajectory of the outcome in the control group after introduction of the treatment until the end of the study; β_4 represents the difference in the level (intercept) of the outcome variable between the treatment unit and controls prior to the treatment; β_5 represents the difference in the trajectory (slope) of the outcome variable between the treatment unit and controls prior to the treatment; β_6 indicates the difference between the treatment unit and controls in the predicted level of the outcome variable versus their counterfactuals in the period immediately following introduction of the treatment (i.e. difference in differences in level), and β_7 represents the difference between the treatment unit and controls in the slope (trend) of the outcome variable after initiation of the treatment compared with pre-treatment (i.e. difference in differences in trend) (Linden, 2015). When the random error terms follow an AR(1) process,

$$\epsilon_t = \rho \epsilon_{t-1} + u_t \qquad (2)$$

where, the autocorrelation parameter ρ is the correlation coefficient between adjacent error terms, such that $|\rho| < 1$, and the disturbances u_t are independent $N(0,\sigma^2)$ (see [33] for a detailed discussion of autocorrelation in time-series regression models).

The two parameters (β_4 and β_5) play a particularly important role in establishing whether the treatment unit and controls are balanced on both the level and trend of the outcome variable in the pre-treatment period. If these data are from a randomized controlled trial, we would expect similar levels and trends prior to the treatment. However, in an observational study where equivalence between groups cannot be ensured, any observed differences will likely raise concerns about the ability to draw causal inferences about the outcomes [7,32]. For a visual depiction of these parameters the reader is referred to [7,34].

Estimation approaches

OLS-NW:

This approach estimates the time series data using OLS, with statistical inference based on HAC standard errors calculated using the Newey-West estimator [27]. This adjustment addresses serial correlation in the time-indexed regression errors while allowing for heteroskedasticity of unknown form. The Newey-West estimator provides a consistent estimate of the asymptotic covariance matrix of the OLS estimators by approximating the long-run variance of the regression disturbances through a weighted sum of sample autocovariances up to a prespecified lag length, with higher-order autocovariances down-weighted to ensure a non-negative definite variance estimate. When the truncation lag grows slowly with the sample size, the associated *t*- and *F*-statistics are asymptotically valid under very general forms of heteroskedasticity and serial correlation, which makes the Newey-West approach a practical default for situations where the exact error process (for example, AR(1) vs higher-order dynamics) is unknown or possibly misspecified [35].

Prais-Winsten:

The Prais–Winsten method is a feasible generalized least squares (FGLS) procedure for linear time-series regressions that corrects for AR(1) serial correlation in the errors while retaining the full sample, including the first observation. Assuming the error term $u_t = \rho u_{t-1} + e_t$with $| \rho | < 1$, the method first obtains an estimate of ρ (autocorrelation coefficient) from an auxiliary regression of the OLS residuals on their lag, and then applies the corresponding quasi-differencing transformation to both the dependent and explanatory variables. OLS on this transformed equation yields estimates that are asymptotically more efficient than OLS under the AR(1) specification, and the associated standard errors and test statistics are asymptotically valid for inference in the presence of such serial correlation [35].

Appendix Table 1 provides a side-by-side comparison of the OLS-NW and PW estimation approaches across various modeling features.

The POWER_ITSA package for Stata

All simulations in the current study were conducted using the community-contributed package for Stata called POWER_ITSA [31], which computes power via simulation for ITSA designs. POWER_ITSA passes on user-specified values representing coefficients from the MG-ITSA regression model (Equation 1) to the ITSADGP [36] package, which generates artificial datasets containing one treated unit and a user-specified number of control units. Next, for each time series dataset generated (containing the treatment unit and the number of control units specified), POWER_ITSA passes on the simulated time series to the ITSA package [7] to estimate the treatment effects represented by coefficients β_6 and β_7 in Equation 1 (i.e. the difference in differences in level of the time series and difference in differences in trend, respectively). POWER_ITSA offers both OLS-NW and PW regression as estimation models, and in the current

study, all analyses were repeated using both methods. After model estimation on each dataset, POWER_ITSA performs a Wald test to determine if the coefficient β_6 or β_7 is 0 (representing the difference in differences in level and difference in differences in trend, respectively), and the associated *P*-value is saved. Summing the number of instances in which the *P*-value is < 0.05 (alpha) provides an estimate of the overall power for the given simulated scenario.

Simulation strategy

Table 1 presents the POWER_ITSA inputs implemented in the current study. The overall goal was to assess if (and to what extent) the various factors influence power differentially for OLS-NW and PW models, including the total number of time periods under study, the number of control units serving as comparators, effect size, level of AR(1) autocorrelation, and the measure of effect (i.e. difference in differences in level [β_6 of Equation 1]) or differences in the pre- to post-treatment change in *trend* (i.e. difference in differences in trend [β_7 of Equation 1]) of the outcome variable. The simulation input values replicate those employed in [26] for consistency. The number of time periods ranged from 10 to 50 and treatment was introduced at the halfway point in the time series. The number of controls ranged from 4 to 20 in increments of 4, except when evaluating autocorrelation where the number of controls was set to 12. The starting level (intercept) was set to 10 for both the treated unit and the controls' time series. A non-zero intercept ensured that simulated datasets would not produce negative values. The pre-treatment trend (slope) for both treated unit and controls was set to 1. This small period over period increase was intended to represent a small underlying secular trend in the time series. The change in the level of the controls' time series immediately following the introduction of the treatment versus the counterfactual at that time-point was set to 0, as we would not expect any factor to influence the level of the controls' time series. The change in the level of the treatment unit's time series immediately following the

introduction of the treatment versus the counterfactual at that time-point was set to increase in increments of 2, 2.5, and 3 points when evaluating change in level (and set to 0 when evaluating change in trends). These values represent a 20%, 25% and 30% immediate increase in level versus the counterfactual at the first post-treatment time point. The post-treatment trend (slope) of the controls' time series was set to 1. In the absence of a treatment, we would expect the controls' time series to continue unchanged from baseline (set to 1 to represent a small continued secular trend from baseline). The trend (slope) of the treatment unit's time series after introduction of the treatment was set to increase in increments of 1.25, 1.5, and 2 points when evaluating changes in trend (and set to 0 when evaluating changes in level). These values represent a 25%, 50% and 100% increase in the post-treatment trend. The AR(1) correlation coefficient between adjacent error terms was set to range from -0.9 to 0.9 for both treated unit and control units for specific analyses of autocorrelation, otherwise the autocorrelation was set to 0.2 for both treated and control units, representing a small positive autocorrelation. 1 standard deviation (SD) was used in generating the normally distributed random error term for both the treatment unit and controls' time series. Figures 1 and 2 illustrate the data generating process for a difference in difference in trend and difference in differences in level, respectively. Each panel presents an actual iteration of the data generating process.

Sensitivity analysis

Three different sensitivity analyses were performed to determine if changing one or more of the input values for level and trend impacts power differentially for OLS-NW and PW methods. In the first sensitivity analysis, the treatment unit's baseline trend was set to differ from that of the controls (5 versus 1 for treated unit and the controls', respectively). The treatment unit's post-treatment trend was set to 5.50 (50% increase) and the number of controls was set to 12. In the

second sensitivity analysis, the controls' baseline level was set to differ from that of the treated unit (8 versus 10, respectively) and the number of controls was set to 12. The third sensitivity analysis examined if a combination of both level and trend changes impact power curves differentially for OLS-NW and PW models. Here, the controls' baseline level was set to 8 and the treated unit's baseline level was set to 10. The treatment unit's level change was set to 2 (20% increase) and the controls had no level change. The treatment unit's post-treatment trend was set to 2 (100% increase) and the controls post-treatment trend was set to 1 (continuation from baseline) and the number of controls was set to 12.

Performance measures

In addition to power, the performance of the OLS-NW and PW methods was evaluated by examining percentage bias, empirical standard errors, root mean squared error (RMSE), 95% confidence interval coverage, and type I error rate [37]. For each scenario, 2000 datasets were generated. Stata™ version 19.0 was used for conducting all analyses. The ITSA_POWER code used in these analyses is located in the online Supplement.

Results

Difference in differences in trend

Power:

Figure 3 illustrates that power consistently increases as a function of an increasing number of control units, an increasing number of time periods, and increasing effect size. OLS-NW has approximately 10.4% higher power, on average, than PW at all values along the continuum of time and number of controls. There also appears to be a greater separation in the OLS-NW power curves than in the PW curves.

Figure 4 illustrates that increasing autocorrelation is associated with lower power, which remains consistent with an increasing number of time periods under study. There is a large amount of variability in the autocorrelation curves at a low effect size, and the patterns only become clear with a larger number of time periods. OLS-NW consistently has higher power than PW at all values along the continuum of time and level of autocorrelation (approximately 14.9% higher, on average). PW consistently shows larger separation between autocorrelation levels than OLS-NW.

Percentage bias:

Appendix Figure 1 illustrates that percentage bias is related to both the number of time periods and the effect size. Bias decreases as effect size increases and as the number of time periods increases. The number of controls appears to differentially affect bias, with a larger number of controls exhibiting less bias. Both OLS-NW and PW exhibit the same overall pattern of bias with both methods close to 0 percentage bias on average, but with the PW exhibiting a larger range (OLS-NW: mean = -0.09 [min= -8.42; max = 5.11], PW: mean = -0.23 [min = -10.74; max = 10.26]).

Empirical standard errors:

Appendix Figure 2 illustrates that empirical standard errors decrease over time but do not differ much by effect size. Both OLS-NW and PW exhibit the same overall curvilinear pattern except that PW is consistently higher at each time period (OLS-NW: mean = 0.11 [min = 0.04; max = 0.35], PW: mean = 0.16 [min = 0.05; max = 0.51]). OLS-NW appears to be influenced by the number of controls, with a smaller number of controls exhibiting larger standard errors.

RMSE:

Appendix Figure 3 illustrates that RMSE is not affected by time, number of controls, effect size, or method (OLS-NW: mean = 1.01 [min = 0.98; max = 1.02], PW: mean = 1.01 [min = 0.99; max = 1.02]).

95% coverage:

Appendix Figure 4 illustrates that PW consistently produces estimates around the nominal 95% level (mean = 94.3% [min = 91.6%; max = 96.3%]) while the OLS-NW method produces substantially lower estimates across all scenarios (mean = 82.1% [(min = 65%; max = 89.6%]). OLS-NW also appears to produce coverage estimates that are influenced by the number of controls, with a lower number of controls produces better coverage than with larger numbers of controls.

Difference in differences in level

Power:

Figure 5 illustrates that power consistently increases as a function of an increasing number of control units, an increasing number of time periods, and increasing effect size. OLS-NW consistently has higher power than PW at all values along the continuum of time and number of controls (approximately 14.4% higher, on average). There is a greater separation between the number of control units in the OLS-NW curves than in the PW curves. Moreover, the power curves are much more linear than the sigmoidal curves exhibited by difference in differences in trend (Figure 3).

Figure 6 illustrates that while autocorrelation is inversely related to power (with PW more clearly showing the distinction between levels of autocorrelation), both models show a nearly flat relationship between the level of autocorrelation and the number of periods in the study. In contrast

to difference in differences in trend (Figure 4), OLS-NW does not clearly demonstrate consistently higher power than PW.

Percentage bias:

Appendix Figure 5 illustrates that both approaches overall have a percentage bias close to 0 but exhibit noisy behavior in curves representing the number of controls (OLS-NW; mean = -0.01 [min = -3.40; max = 2.62], PW: mean = -0.18 [min = -5.17; max = 3.03]).

Empirical standard errors:

Appendix Figure 6 shows that OLS-NW exhibits a linear decrease in standard errors over time, with smaller numbers of control units having higher standard errors than larger groups, but is not affected by effect size. PW exhibits decreasing curvilinear trends in standard errors over time, but no discernable separation by control group size. OLS-NW has standard errors that are approximately 25% lower on average than PW (NW: mean = 0.69 [min = 0.55; max = 1.00], PW: mean = 0.93 [min = 0.68; max = 1.42]).

RMSE:

Appendix Figure 7 illustrates that both methods exhibit nearly identical RMSE across time and number of controls (NW: mean = 1.01 (min = 0.99; max = 1.02], PW: mean = 1.01 [min = 0.99; max = 1.02]).

95% coverage:

Appendix Figure 8 illustrates that PW consistently produces estimates around the nominal 95% level (mean = 94.4% [min = 92.1%; max = 96.2%]) while the OLS-NW method produces substantially lower estimates across all scenarios (mean = 83.2% [(min = 65.5%; max = 90.6%]). OLS-NW also appears to produce coverage estimates that are influenced by the number of

controls, with a lower number of controls producing better coverage than with larger numbers of controls.

Type I error

Appendix Figure 9 illustrates the type I error rate as a function of time for OLS-NW and PW models, holding the number of controls at 12 and the level of autocorrelation at 0.20. While PW consistently produce type I errors around the expected 5% level (mean = 5.3% [min = 4.8%, max = 6.5%]), the OLS-NW method exhibits high (and decreasing) type I errors over time until leveling off around 30 time periods (mean =18.1%, [min = 11.3%, max = 30.2%]).

Appendix Figure 10 illustrates the type I error rate as a function of autocorrelation for OLS-NW and PW models, holding time constant at 50 periods and the number of controls at 12. While PW consistently produce type I errors in the range around the expected 5% level (mean = 5.32% [min = 4.2%, max = 6.5%]), the OLS-NW method exhibits a curvilinear relationship (low to high) of type I errors with increasing autocorrelation achieving a type I error of 5.53% at the autocorrelation level of 0.90 (mean = 17.1% [min = 1.6%, max = 5.53%]).

Sensitivity analysis

Appendix Figures 11, 12, and 13 illustrate the power curves for OLS-NW and PW for the 3 sensitivity analyses, respectively. In all scenarios, the PW method consistently exhibit lower power than the OLS-NW method (consistent with results of the main analyses).

Discussion

This simulation study evaluated the finite-sample performance of OLS-NW and PW regression methods for estimating difference-in-differences in level and trend within a MG-ITSA framework under AR(1) autocorrelation. Whereas prior simulation studies comparing these methods have been conducted exclusively in SG-ITSA settings [38,39], the present study extends this research

to the MG-ITSA design in which inference is based on between-group contrasts. The results demonstrate clear and systematic trade-offs between the two methods that are driven primarily by variance estimation and inferential calibration, rather than bias or overall point-estimation accuracy. These findings have important implications for the design and analysis of prospective MG-ITSA studies in healthcare research.

Across both difference in difference in level and trend estimates, power increased with longer series, larger effect sizes, and greater numbers of control units. These patterns are consistent with earlier MG-ITSA power simulations using OLS–NW alone [26]. However, the present study also demonstrates that there are substantial differences in power estimates produced by the two estimation approaches in the MG-ITSA context. For difference in differences in trend, OLS–NW exhibited consistently higher power than PW across all simulated conditions, exceeding PW by approximately 10–15%, on average. This advantage persisted across varying levels of autocorrelation, time periods, numbers of control units, and effect sizes, and was robust in sensitivity analyses. Similar power advantages for OLS-NW were observed for difference in differences in level, although differences between methods were attenuated when power was examined as a function of autocorrelation. These findings are broadly consistent with Bottomley et al. [39], who reported higher power for OLS-NW relative to PW in SG-ITSA, and with Wooldridge [35] who notes that FGLS methods such as PW may have lower power than OLS-based approaches when autocorrelation is modest or imperfectly estimated.

Both OLS-NW and PW produced approximately unbiased estimates for difference in differences in level and trend parameters, with percentage bias centered near zero across scenarios. Bias decreased with increasing effect size and longer series which is consistent with those reported

in SG-ITSA studies [38,39], indicating that extending the design to multiple groups does not introduce systematic bias in point estimation.

Empirical standard errors declined with increasing series length for both methods, but PW consistently produced larger standard errors than OLS-NW, particularly in shorter time series, consistent with results reported by Bottomley et al. [39] and the finite-sample properties of FGLS estimators described by Wooldridge [35]. In contrast, OLS-NW standard errors decreased as the number of control units increased, suggesting greater sensitivity to cross-sectional information than with PW. This difference in variability largely explains the observed divergence in power, 95% coverage, and Type I error between the two estimators.

Despite the differences in standard errors, RMSE was nearly identical across methods and simulation conditions, indicating that the two approaches achieve comparable overall estimation accuracy, with differences primarily arising in inference rather than point estimation.

In contrast to its power advantages, OLS-NW exhibited poor finite-sample inferential performance. Coverage for OLS-NW was substantially below the nominal 95% level across all scenarios for both difference in differences in level and trend estimates, with average coverage near 82–83%. Coverage further deteriorated as the number of control units increased. In contrast, PW consistently achieved near-nominal coverage across all simulated conditions.

Similarly, type I error rates for PW remained close to the nominal 5% level across time and autocorrelation conditions, while OLS-NW showed severe inflation, particularly in shorter series. Although type I error for OLS-NW decreased as the number of time periods increased, it remained unacceptably high across many plausible MG-ITSA configurations. These results concur with those of Bottomley et al. [39] in SG-ITSA, but importantly show that including one or more control units does not mitigate the finite-sample inferential deficiencies of OLS-NW.

Taken together, these results highlight a fundamental trade-off in MG-ITSA between power and inferential validity under AR(1) serial correlation, which was the sole autocorrelation structure considered in the present simulations. Within this setting, OLS-NW offers clear advantages in terms of power, particularly when multiple control units are available, but at the cost of inflated Type I error and poor confidence interval coverage. Prais-Winsten regression, which is explicitly designed to address AR(1) autocorrelation, is less powerful but provides well-calibrated inference across the range of AR(1) design conditions examined. Importantly, these differences arise not from bias or point-estimation accuracy, but from how each estimator handles first-order serial correlation and variance estimation in finite samples.

These findings underscore the importance of aligning estimator choice with the inferential priorities of MG-ITSA studies when the data series include AR(1) autocorrelation. When control of false positives and reliable uncertainty quantification are paramount, PW may be preferred despite lower power. Conversely, when power considerations dominate, results obtained using OLS-NW should be interpreted cautiously, particularly in the presence of AR(1) autocorrelation. Given that OLS-NW can be implemented with higher-order lag structures, its performance under more complex forms of serial dependence may differ from the patterns observed here.

From a practical perspective, applied researchers conducting MG-ITSA studies under plausible AR(1) autocorrelation should choose between methods guided by the relative importance of power versus inferential validity. PW regression is recommended when valid hypothesis testing, confidence interval coverage, and control of type I error are primary objectives, particularly in studies with a limited number of time points. OLS-NW may be useful in exploratory analyses or settings where maximizing power to detect intervention effects is prioritized; however, its use should be accompanied by caution given the potential for substantial under-coverage and inflated

false-positive rates in finite samples. Regardless of the method selected, researchers should report key design features, such as series length, number of control units, and assumptions about the autocorrelation structure, and consider sensitivity analyses using alternative estimators to assess the robustness of substantive conclusions.

The present study has limitations. The simulation strategy developed here was designed to replicate the most common MG-ITSA study in healthcare research: a single treatment-group compared to 4 or more control units with a single treatment (treatment) period, in which OLS or FGLS time series regression is used as the statistical model and assumes AR(1) error processes for both treatment and control units. The simulations were restricted to AR(1) error processes to reflect the design of PW regression, and a common assumption in ITSA studies. Given that OLS-NW can accommodate higher-order or more complex dependence structures, the relative finite-sample performance of these estimators under AR(p) or non-AR serial correlation remains an open empirical question [35]. Additionally, several other factors were not included in the simulation design: multiple consecutive treatments, seasonality, and different outcome variable types. Generally, when there is suspected seasonality in the data and it cannot be modeled directly at the design stage, investigators should err on the side of increasing the length of the study, as the minimum required sample size increases with variability, model complexity, and seasonal period length [40]. When other factors appear to impact the time series, the investigator should attempt to model them at the design stage of the simulations to get the most accurate estimates of power. Finally, OLS-NW and PW were the two underlying statistical approaches used here for estimating power via simulation because of their common use and ease of application. Future research should investigate how power estimates for MG-ITSA studies compare when different modeling

approaches are considered, such as other time series regression models (e.g. ARCH), Bayesian frameworks [41] and machine learning algorithms [42].

Conclusion

This simulation study demonstrated that in MG-ITSA studies under AR(1) autocorrelation, the choice of autocorrelation adjustment entails a clear trade-off between statistical power and inferential validity. OLS-NW consistently yielded higher power for detecting difference-in-differences in both level and trend across a wide range of design conditions, but at the cost of substantially inflated Type I error rates and poor confidence interval coverage in finite samples. In contrast, PW regression was less powerful but provided well-calibrated inference, maintaining near-nominal coverage and Type I error control across all scenarios examined. Importantly, both methods produced approximately unbiased estimates with comparable RMSE, indicating that observed differences arise primarily from variance estimation rather than point-estimation accuracy. These findings highlight the need for investigators to align estimator choice with their inferential priorities in MG-ITSA studies: PW is preferable when valid hypothesis testing and uncertainty quantification are paramount, whereas OLS-NW may be used cautiously when maximizing power is the primary objective.

References

1. Campbell DT, Stanley JC. Experimental and quasi-experimental designs for research. Chicago: Rand McNally; 1966.
2. Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin; 2002.
3. Linden A. Assessing regression to the mean effects in health care initiatives. BMC Med Res Methodol. 2013;13:119. https://doi.org/10.1186/1471-2288-13-119
4. Linden A, Yarnold PR. Using machine learning to identify structural breaks in single-group interrupted time series designs. J Eval Clin Pract. 2016;22:855–859. doi:10.1111/jep.12544.
5. Linden A. Challenges to validity in single-group interrupted time series analysis. J Eval Clin Pract. 2017a;23:413–418. https://doi.org/10.1111/jep.12638
6. Linden A. Persistent threats to validity in single-group interrupted time series analysis with a crossover design. J Eval Clin Pract. 2017b;23:419–425. https://doi.org/10.1111/jep.12668
7. Linden A. Conducting interrupted time-series analysis for single- and multiple-group comparisons. Stata J. 2015;15(2):480–500. https://doi.org/10.1177/1536867X1501500208
8. Abadie A, Diamond A, Hainmueller J. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. J Am Stat Assoc. 2010;105(490):493–505. https://doi.org/10.1198/jasa.2009.ap08746
9. Linden A. A matching framework to improve causal inference in interrupted time-series analysis. J Eval Clin Pract. 2018;24:408–415. https://doi.org/10.1111/jep.12874
10. Rubin DB. Estimating causal effects of treatments in randomized and non-randomized studies. J Educ Psychol. 1974;66:688–701.

11. Ansermino JM, Pillay Y, Tagoola A, Zhang C, Dunsmuir D, Kamau S, et al. Implementation of Smart Triage combined with a quality improvement program for children presenting to facilities in Kenya and Uganda: an interrupted time series analysis. PLoS Digit Health. 2025;4(3):e0000466. https://doi.org/10.1371/journal.pdig.0000466
12. Aye TT, Nguyen HT, Petitfour L, Ridde V, Amberg F, Bonnet E, et al. How do free healthcare policies impact utilization of maternal and child health services in fragile settings? Evidence from a controlled interrupted time series analysis in Burkina Faso. Health Policy Plan. 2025;39(9):891–901. https://doi.org/10.1093/heapol/czae077
13. Barbalat G, De Rozario R, Franck N. Reducing mental health emergency department visits through community-based assessment services: a controlled time-series analysis in the city of Lyon, France. Health Policy. 2025;129:105419. https://doi.org/10.1016/j.healthpol.2025.105419
14. Du S, Liu Y, Yang C, Yang Y, Yang Y. How DRG payment reform shapes inpatient neurological care in an underdeveloped region: evidence from a controlled interrupted time series study in Yunnan, China. Risk Manag Healthc Policy. 2025;18:2575–2590. https://doi.org/10.2147/RMHP.S530693
15. Gressler L, Lewis K. Changes in maternal morbidity and infant outcomes following state-level abortion bans post-Dobbs: a comparative interrupted time series study. BMC Public Health. 2025;25(1):2265. https://doi.org/10.1186/s12889-025-23468-8
16. Luo D, Ye X, Zhao H, Yao B, Liu W, Xu X. Reducing proton pump inhibitors overuse with an advisory, risk-based, context-aware electronic alert system: a controlled interrupted time series analysis. Int J Med Inform. 2025;185:105960. https://doi.org/doi:10.1016/j.ijmedinf.2025.105960

17. Tsunemitsu A, Shin JH, Hamada O, Tsutsumi T, Sasaki N, Kunisawa S, et al. Effects of protocol-driven care by internists on adherence to clinical practice guidelines for hip fracture surgery patients: an interrupted time series study using a nationwide inpatient database. Intern Med. 2025;64:4358–4364. https://doi.org/10.2169/internalmedicine.4358-24
18. Walter SR, Salisbury C, Scott LJ, de Vocht F, Macleod J, Ben-Shlomo Y, et al. The impact of COVID-19 lockdowns on primary care contact among vulnerable populations in England: a controlled interrupted time series study. BJGP Open. 2025;9:BJGPO.2025.0017. https://doi.org/10.3399/BJGPO.2025.0017
19. Wanigaratne S, Stukel T, Lu H, Guan J, Hynie M, Saunders NR, et al. Morbidity among resettled refugees at arrival in Ontario, Canada (1994–2017): a controlled interrupted time series study examining the effect of the Immigration and Refugee Protection Act, 2002. J Epidemiol Community Health. 2025;79(4):249–256. https://doi.org/10.1136/jech-2024-222947
20. Wyper G, Mackay DF, Fraser C, Lewsey J, Robinson M, Beeston C, et al. Evaluating the impact of alcohol minimum unit pricing on deaths and hospitalisations in Scotland: a controlled interrupted time series study. Lancet. 2023;401:1361–1370. https://doi.org/10.1016/S0140-6736(23)00497-X
21. Dziak JJ, Dierker LC, Abar B. The interpretation of statistical power after the data have been gathered. Curr Psychol. 2020;39:870–877. https://doi.org/10.1007/s12144-018-0018-1
22. Fanshawe TR, Turner PJ, Gillespie MM, Hayward GN. The comparative interrupted time series design for assessment of diagnostic impact: methodological considerations and an example using point-of-care C-reactive protein testing. Diagnostic and Prognostic Research. 2022 Mar 2;6(1):3. https://doi.org/10.1186/s41512-022-00118-w

23. Yelland J, Riggs E, Szwarc J, Casey S, Dawson W, Vanpraag D, et al. Bridging the gap: using an interrupted time series design to evaluate systems reform addressing refugee maternal and child health inequalities. Implement Sci. 2015;10:62. https://doi.org/10.1186/s13012-015-0251-z

24. Power M, Tyrrell PJ, Rudd AG, Tully MP, Dalton D, Marshall M, et al. Did a quality improvement collaborative make stroke care better? A cluster randomized trial. Implement Sci. 2014;9:40. https://doi.org/10.1186/1748-5908-9-40

25. Schochet PZ. Statistical power for estimating treatment effects using difference-in-differences and comparative interrupted time series estimators with variation in treatment timing. J Educ Behav Stat. 2022;47:367–405. https://doi.org/10.3102/10769986211070625

26. Linden A. Power considerations for multiple-group (controlled) interrupted time series analysis: a comprehensive simulation study. Eval Health Prof. 2026. In press.

27. Newey WK, West KD. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica. 1987;55:703–708. https://doi.org/10.2307/1913610

28. Prais SJ, Winsten CB. Trend estimators and serial correlation. Cowles Commission Discussion Paper No. 383; 1954. https://cowles.yale.edu/sites/default/files/files/pub/cdp/s-0383.pdf

29. Harvey AC. Forecasting, structural time series models and the Kalman filter. Cambridge: Cambridge University Press; 1989.

30. Enders W. Applied econometric time series. 2nd ed. Hoboken (NJ): Wiley; 2004.

31. Linden A. POWER_ITSA: Stata module to compute power for single and multiple-group interrupted time series analysis. Statistical Software Components S459461, Boston College Department of Economics. 2025.
32. Linden A. A comprehensive set of postestimation measures to enrich interrupted time-series analysis. Stata J. 2017c;17:73–88. https://doi.org/10.1177/1536867X1501500208
33. Kutner MH, Nachtsheim CJ, Neter J, Li W. Applied linear statistical models. 5th ed. New York: McGraw-Hill Irwin; 2005.
34. Linden A. Erratum: a comprehensive set of postestimation measures to enrich interrupted time-series analysis. Stata J. 2022;22:231–233. https://doi.org/10.1177/1536867X221083929
35. Wooldridge JM. Introductory econometrics: a modern approach. 7th ed. Boston: Cengage; 2020.
36. Linden A. ITSADGP: Stata module to generate artificial data for interrupted time-series analysis. Statistical Software Components S459403, Boston College Department of Economics. 2024.
37. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med. 2006;25:4279–4292. https://doi.org/10.1002/sim.2673
38. Turner SL, Forbes AB, Karahalios A, Taljaard M, McKenzie JE. Evaluation of statistical methods used in the analysis of interrupted time series studies: a simulation study. BMC Med Res Methodol. 2021;21:181. https://doi.org/10.1186/s12874-021-01393-1
39. Bottomley C, Ooko M, Gasparrini A, Keogh RH. In praise of Prais–Winsten: an evaluation of methods used to account for autocorrelation in interrupted time series. Stat Med. 2023;42(8):1277–1288. https://doi.org/10.1002/sim.9547

40. Hyndman RJ, Kostenko AV. Minimum sample size requirements for seasonal forecasting models. Foresight. 2007;6:12–15.

41. Ma Y, Benmarhnia T. Interrupted time series analysis in environmental epidemiology: a review of traditional and novel modeling approaches. Curr Environ Health Rep. 2025;12:50. https://doi.org/10.1007/s40572-025-00517-3

42. Linden A, Yarnold PR. Using machine learning to evaluate treatment effects in multiple-group interrupted time series analysis. J Eval Clin Pract. 2018;24:740–744. https://doi.org/10.1111/jep.12966

Abbreviations

ARCH: Autoregressive conditional heteroskedasticity regression; ARIMA: Autoregressive integrated moving average; AR(1): First-order autocorrelation; FGLS: Feasible generalized least squares regression; HAC: Heteroskedasticity- and autocorrelation-consistent standard errors; ITSA: Interrupted time series analysis; MG-ITSA: Multiple-group interrupted time series analysis; OLS: Ordinary least squares regression; OLS-NW: Ordinary least squares regression with Newey-West standard errors; PW: Prais-Winsten regression; RMSE: Root mean squared error; SD: Standard deviation; SG-ITSA: Single-group interrupted time series analysis

Supplementary Information

The online version contains supplementary material available at XXXXXX

Authors' contributions

AL conceived the study and its design, conducted all analyses, wrote the manuscript and takes public responsibility for its content.

Funding

There was no funding associated with this work.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The author declares that he has no competing interests.

Table 1 Simulation inputs

Parameter	Values related to Equation (1)	Input Values
Controls' baseline level	β_0	10
Treated unit's baseline level	$\beta_0 + \beta_4$	10
Controls' baseline trend	β_1	1
Treated unit's baseline trend	$\beta_1 + \beta_5$	1
Controls' level change post-treatment versus counterfactual	β_2	0
Treated unit's level change post-treatment versus counterfactual	$\beta_2 + \beta_6$	2, 2.5, 3 (representing 20%, 25%, and 30%)
Controls' post-treatment trend	$\beta_1 + \beta_2$	1
Treated unit's post-treatment trend	$\beta_1 + \beta_3 + \beta_5 + \beta_7$	1.25, 1.50, 2 (representing 25%, 50% and 100%)
Number of time periods		10 – 50
Initiation of the treatment		Halfway point
Number of controls		4 – 20
Autocorrelation coefficient		-0.90 – 0.90
Standard deviation for random error term		1

Notes:

(1) Autocorrelation coefficient for both groups is held constant at 0.20 for all simulations other than that specifically examining the effect of autocorrelation.

(2) The number of controls is held constant at 12 for simulations examining the effect of autocorrelation.

(3) For sensitivity analysis 1 (different baseline trends), the treatment unit's baseline trend is set to 5 and the treatment unit's post-treatment trend is set to 5.25, 5.50, 6 (25%, 50%, 100%, respectively). Controls' values are unchanged from above.

(4) For sensitivity analysis 2 (different baseline levels), the controls' baseline level is set to 8 (the treatment unit's values are unchanged from above).

(5) For sensitivity analysis 3 (combination of different level and trend changes), the controls' baseline level is set to 8, the treatment unit's level change is set to 2 (20% increase), the treatment unit's post-treatment trend is set to 2 (100% increase). All other values are unchanged from above.

Table 2. Summary of simulation results comparing OLS with Newey-West standard errors (OLS-NW) and Prais-Winsten (PW) regression in MG-ITSA

Performance measure	Simulation conditions	OLS–NW	Prais–Winsten (PW)	Interpretation
Power	Number of time periods, effect size, number of controls	Increased with more controls, longer series, and larger effect sizes; higher across all conditions	Increased with controls, time, and effect size, but consistently lower than OLS–NW	OLS–NW yielded higher power across all design configurations
Power (autocorrelation)	Autocorrelation, number of time periods	Decreased with increasing autocorrelation but remained higher across all conditions	Decreased with increasing autocorrelation	Power was uniformly higher for OLS–NW across levels of autocorrelation
Percentage bias	Number of time periods, effect size, number of controls	Decreased with increasing time periods and effect size; fewer controls associated with greater bias	Exhibited the same pattern as OLS–NW	Both methods showed similar bias patterns
Empirical standard errors	Number of time periods, effect size, number of controls	Decreased over time; larger with fewer controls	Decreased over time; consistently larger than OLS–NW	PW produced larger empirical standard errors
Root mean squared error (RMSE)	Number of time periods, effect size, number of controls	Approximately constant across conditions	Approximately constant across conditions	RMSE was not meaningfully affected by model or design factors
95% confidence interval coverage	Number of time periods, effect size, number of controls	Below the nominal 95% level	Close to the nominal 95% level	PW achieved near-nominal coverage; OLS–NW under-covered
Type I error	Time periods and autocorrelation	Inflated across conditions	Close to the nominal 5% level	PW maintained appropriate Type I error control
Sensitivity analyses	Alternative design specifications	Higher power across all scenarios	Lower power across all scenarios	Sensitivity analyses were consistent with main findings

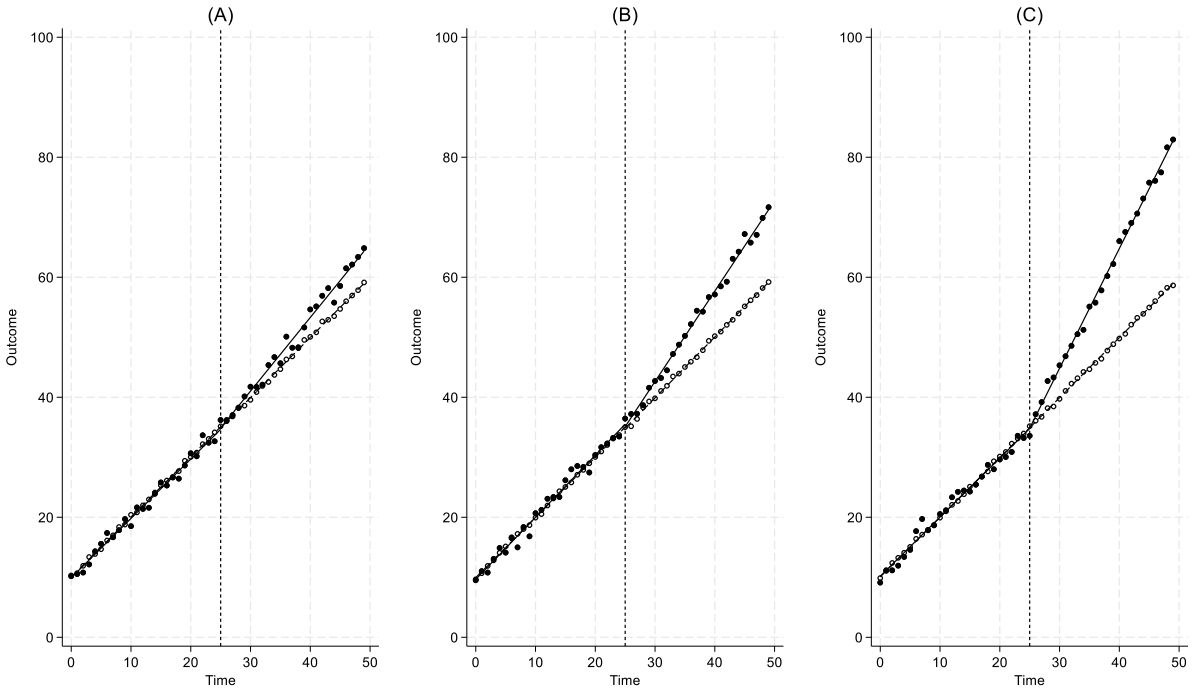


Figure 1: Single iteration of the actual data generating process for evaluating difference in differences in trend. (A) 25% increase in treatment unit's post-treatment trend, (B) 50% increase in treatment unit's post-treatment trend, and (C) 100% increase in treatment unit's post-treatment trend

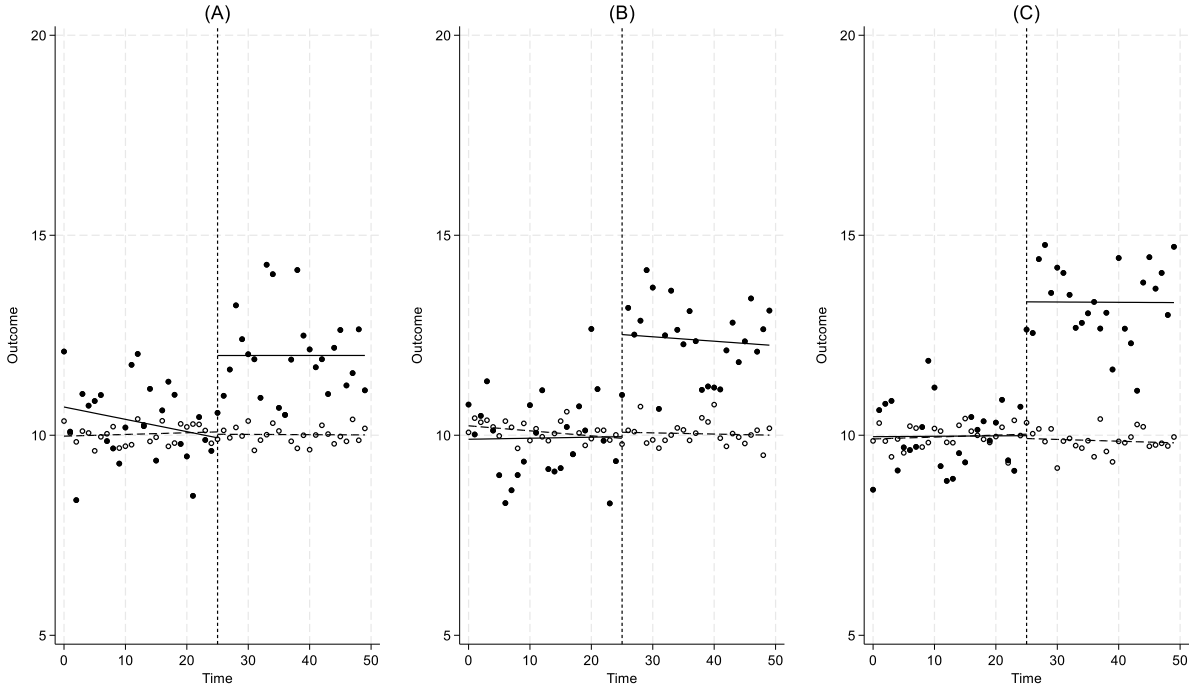


Figure 2: Single iteration of the actual data generating process for evaluating difference in differences in level. (A) 20% increase in treatment unit's post-treatment level, (B) 25% increase in treatment unit's post-treatment level, and (C) 100% increase in treatment unit's post-treatment level

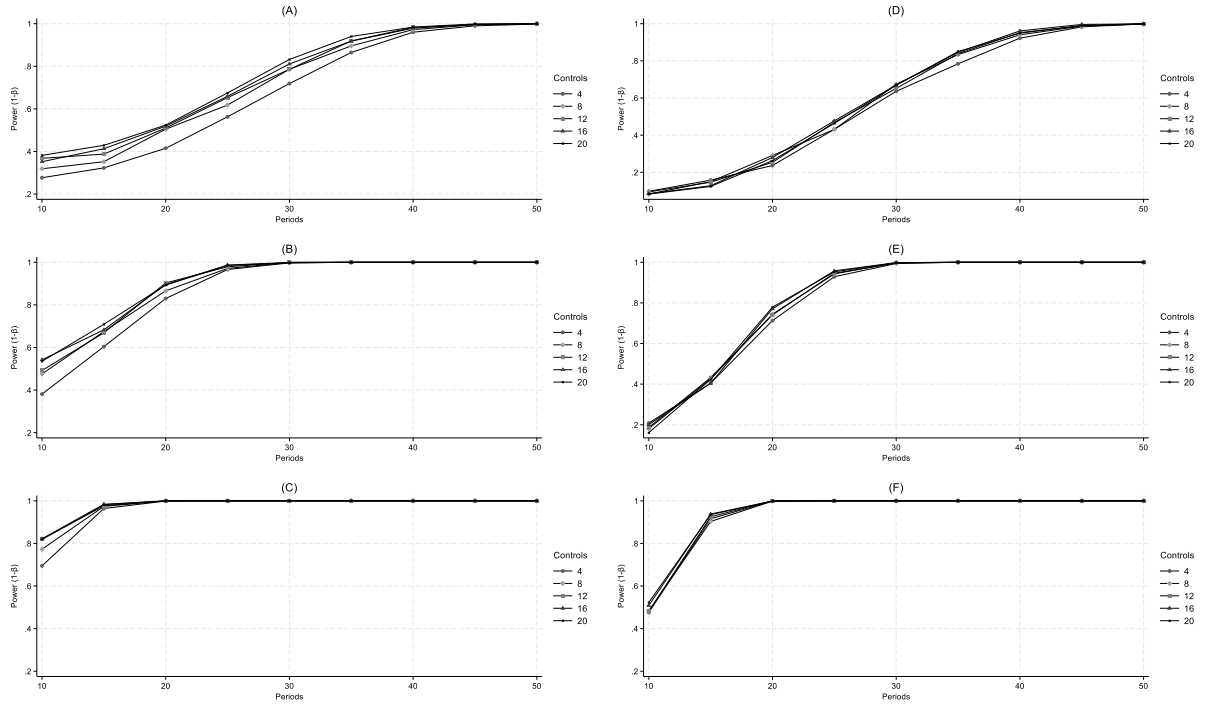


Figure 3: The impact of the number of control units and the number of time periods on power for a difference in differences in trend for OLS-NW (Panels A-C) and PW (Panels D-F) models. Panels **A** and **D** represent a 25% effect, panels **B** and **E** represent a 50% effect, and panel **C** and **F** represent a 100% effect.

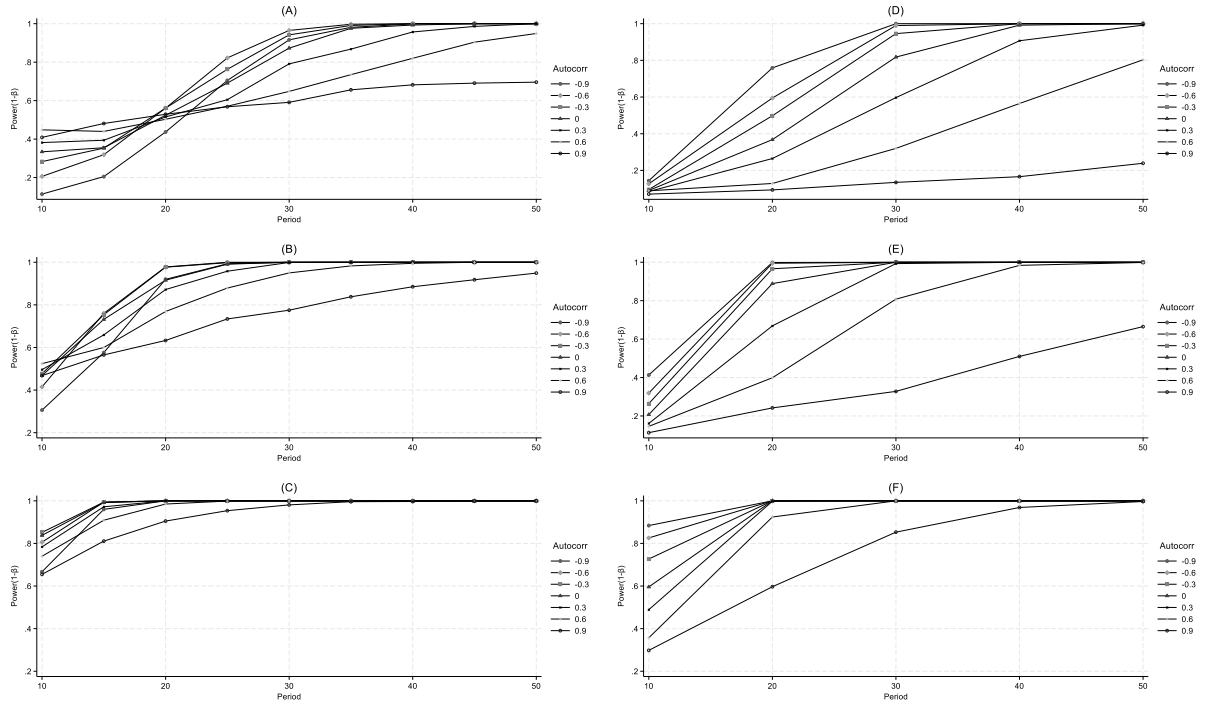


Figure 4: The impact of autocorrelation and the number of time periods on power for a difference in differences in trend for OLS-NW (Panels A-C) and PW (Panels D-F) models. Panels **A** and **D** represent a 25% effect, panels **B** and **E** represent a 50% effect, and panel **C** and **F** represent a 100% effect.

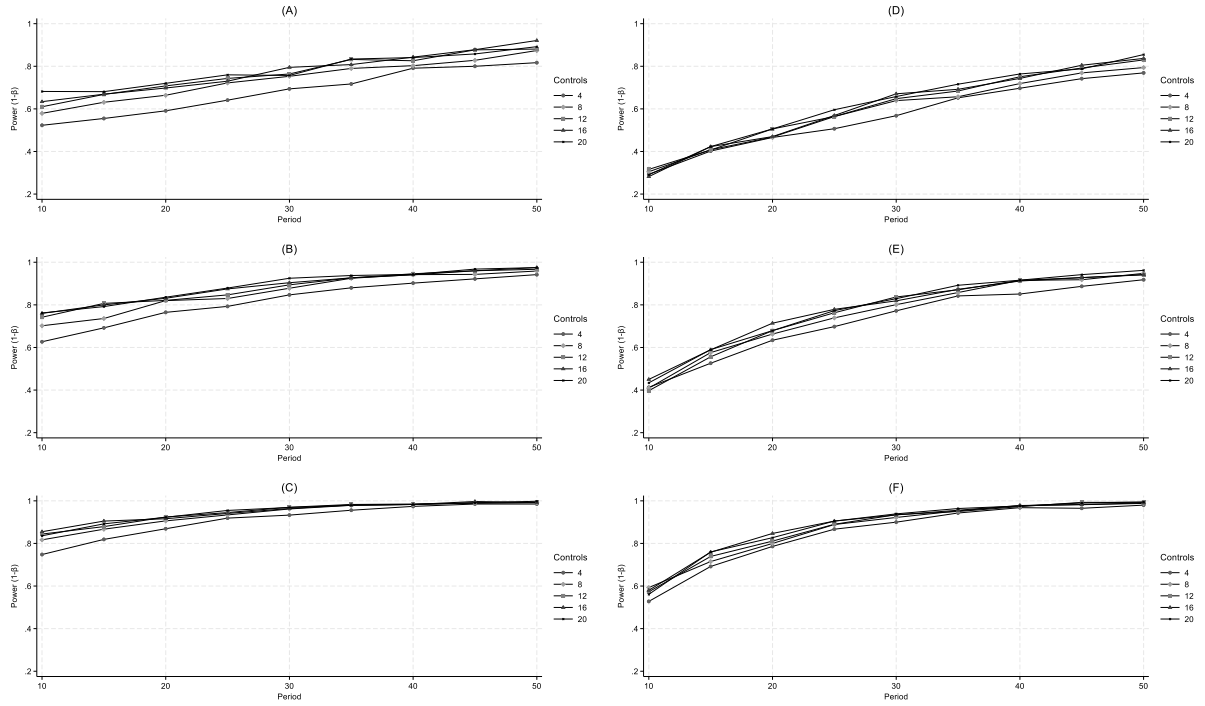


Figure 5: The impact of the number of control units and the number of time periods on power for a difference in differences in level for OLS-NW (Panels A-C) and PW (Panels D-F) models. Panels **A** and **D** represent a 20% effect, panels **B** and **E** represent a 25% effect, and panel **C** and **F** represent a 30% effect.

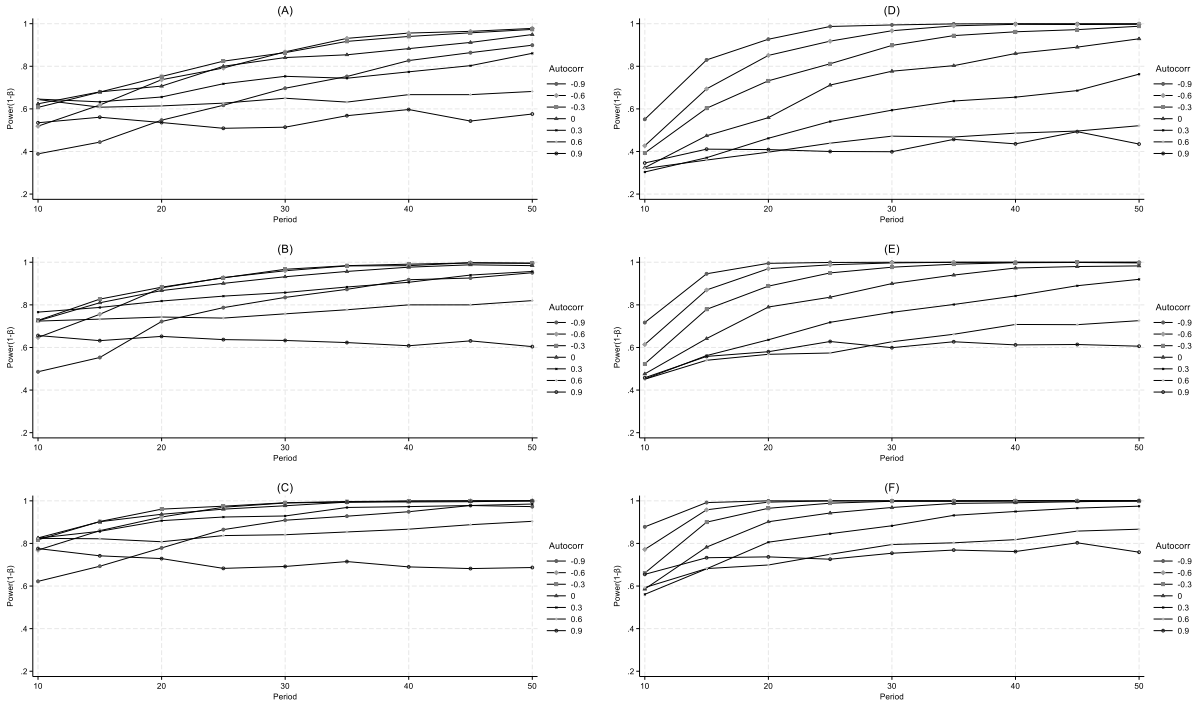


Figure 6: The impact of autocorrelation and the number of time periods on power for a difference in differences in level for OLS-NW (Panels A-C) and PW (Panels D-F) models. Panels **A** and **D** represent a 25% effect, panels **B** and **E** represent a 50% effect, and panel **C** and **F** represent a 100% effect.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SUPPLEMENTALMATERIALS.docx