Received: 14 September 2021 | Revised: 9 November 2022 | Accepted: 11 January 2023

DOI: 10.1002/sim.9669

RESEARCH ARTICLE

Statistics in Medicine WILEY

In praise of Prais-Winsten: An evaluation of methods used to account for autocorrelation in interrupted time series

C Bottomley[1,2] | **M Ooko**[1,2,3] | **A Gasparrini**[4,5] | **RH Keogh**[6]

[1]London School of Tropical Medicine & Hygiene, MRC International Statistics and Epidemiology Group, London, UK

[2]Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK

[3]Department of Epidemiology and Demography, Kemri-Wellcome Trust Research Programme, Kilifi, Kenya

[4]Department of Public Health, Environments and Society, London School of Hygiene and Tropical Medicine, London, UK

[5]Centre for Statistical Methodology, London School of Hygiene and Tropical Medicine, London, UK

[6]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK

Correspondence
Bottomley C, London School of Tropical Medicine & Hygiene, MRC International Statistics and Epidemiology Group, London, UK.
Email: christian.bottomley@lshtm.ac.uk

Funding information
UK Medical Research Council (MRC), Grant/Award Number: MR/R010161/1

Interrupted time series are increasingly being used to assess the population impact of public health interventions. These data are usually correlated over time (auto correlated) and this must be accounted for in the analysis. Typically, this is done using either the Prais-Winsten method, the Newey-West method, or autoregressive-moving-average (ARMA) modeling. In this paper, we illustrate these methods via a study of pneumococcal vaccine introduction and explore their performance under 20 simulated autocorrelation scenarios with sample sizes ranging between 20 and 300. We show that in terms of mean square error, the Prais-Winsten and ARMA methods perform best, while in terms of coverage the Prais-Winsten method generally performs better than other methods. All three methods are unbiased. As well as having good statistical properties, the Prais-Winsten method is attractive because it is decision-free and produces a single measure of autocorrelation that can be compared between studies and used to guide sample size calculations. We would therefore encourage analysts to consider using this simple method to analyze interrupted time series.

KEYWORDS
autocorrelation, interrupted time series, intervention analysis

1 | INTRODUCTION

An interrupted time series (ITS) consists of observations made before and after an event of interest that are used to assess its impact. The event may be planned, such as the introduction of a vaccination program, or unplanned such as the 2008 global financial crisis or recent COVID-19 pandemic.[1] In either case, a defining feature of this study design is that the observations are made at the population level. ITS analyses therefore assess the population-level impact of an event. For example, we may be interested in estimating the difference in the mean disease incidence following the introduction of a vaccine vs what it would have been had the vaccine not been introduced.

Bottomley C and Ooko M are joint first authors.

The impact of the event can be estimated by fitting a regression model to compare the pre- and post-event periods, adjusting for confounding due to time trend and seasonality where necessary. A problem with this approach is that the usual independence assumption is difficult to justify when the residuals are correlated in time, as is often the case in ITS analyses. Typically, residuals that are close in time are more similar than those that are further apart. This so-called autocorrelation must be accounted for otherwise the analysis will produce incorrect—usually anticonservative—*p*-values and confidence intervals.[2]

Three approaches are commonly used to account for autocorrelation in ITS. The first is to assume the residuals follow a first-order autoregressive process and fit the regression model using the Prais-Winsten procedure.[3] As a form of generalized least squares (GLS), the Prais-Winsten method works by applying a linear transformation to the outcome and explanatory variables in order to decorrelate the error term. Because a first-order autoregressive error model is assumed, the appropriate linear transformation is determined by a single parameter representing the correlation between residuals at consecutive time points. The second approach is to fit several autoregressive-moving-average (ARMA) models by maximum likelihood, and then use either the autocorrelation function or a statistical criterion, such as the AIC, to choose the best-fitting model.[4,5] Finally, the third approach is to ignore autocorrelation in the estimation of the regression parameters and adjust the standard errors using the Newey-West method.[6] This approach is essentially an extension of the robust standard errors methodology that is commonly used to adjust for clustering and heteroskedasticity.

Here we conduct a simulation study to evaluate these different methods under a range of autocorrelation scenarios. In this evaluation, we assume the ARMA error model is unknown and consider the selection of an appropriate model as part of the estimation procedure. The study builds on previous simulation studies where an order-1 autoregressive model has been assumed.[7-10] Our main finding is that the Prais-Winsten method generally has coverage closer to the nominal value than other methods.

The paper is structured as follows. We begin by describing a regression model that is widely used to analyze ITS and three methods commonly used to account for autocorrelation (Sections 2 and 3). The methods are illustrated using data from a pneumococcal vaccine impact study (Section 7). We then present a simulation study to evaluate the methods in terms of bias, mean square error, and confidence interval coverage (Section 5). Results from the simulation study show marked differences in coverage, with the Prais-Winsten method generally having better coverage than other methods. In response to this finding, we explore reasons for the observed variation and approaches that can be used to bring coverage closer to the nominal level (Section 6). We also briefly explore the issue of statistical power. Finally, we conclude with a discussion of our findings (Section 7).

2 | MODELLING ITS

An ITS consists of a number of measurements, such as the number of cases of disease, made on a population before and after an event of interest. We let y_t $(t = 0, \ldots, n-1)$ denote observations of the outcome at n equally spaced times, and τ denote the time of the event, which for concreteness we assume is an intervention rather than an unplanned event. A simple model for the ITS y_t $(t = 0, \ldots, n-1)$ is:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \tag{1}$$

where x_t represents an indicator for the intervention ($x_t = 0$ for $t < \tau$ and $x_t = 1$ for $t \geq \tau$), and ε_t is a mean zero error that represents other determinants of the outcome. Assuming x_t is independent of ε_t, that is assuming no confounding, an unbiased estimate of the intervention effect, β_1, can be obtained by regressing y_t on x_t.

In practice, it is often necessary to adapt this basic model. In particular, the model must be modified when the "other determinants" include confounding factors that are correlated with x_t. If such factors are ignored in the regression, then the resulting estimate of β_1 is no longer an unbiased estimate of the intervention effect. Graphically, confounding manifests itself as a trend in y_t (unless the confounding factors are perfectly correlated with x_t). Thus, approaches for dealing with confounding often involve modeling trend rather than modeling confounder effects directly.[11] The trend can be modeled as linear, non-linear, or stochastic.[12]

The simplest approach, which is frequently used in the medical literature and is often reasonable for modeling short time series, is to assume that the confounding can be controlled via a linear trend term. This is the so-called segmented regression model:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 t + \varepsilon_t. \tag{2}$$

The model sometimes also includes an interaction between x_t and t. Usually the interaction term is interpreted as a changing intervention effect but it could equally represent a non-linear trend.

The model in Equation (2) can be estimated using ordinary least squares regression (OLS). However, a problem with using OLS is that the resulting p-values and confidence intervals are only valid if the ϵ_t are mutually independent. This assumption usually does not hold for time series data since the residuals tend to be positively correlated. In this situation, OLS produces a standard error SE that is downward-biased[2] and, as a result, the confidence interval and p-value for the intervention effect are anti-conservative. In the next section, we describe three methods commonly used to account for autocorrelation in ITS analyses.

3 | METHODS USED TO ACCOUNT FOR AUTOCORRELATION

3.1 | Prais-Winsten

The Prais-Winsten method involves estimating the correlation between the error at t and $t-1$, $\text{corr}(\epsilon_t, \epsilon_{t-1})$, and then using this estimate to transform the outcome and predictor variables in such a way that the correlation is removed from the error when a linear regression model is fitted to the transformed data. The key assumption behind the method is that the error follows a first-order autoregressive process; autocorrelation is therefore only fully removed if the error follows this model. The method is an example of feasible generalized least squares and, as such, produces estimates with the same asymptotic distribution as the maximum likelihood estimator—see, for example, chap. 8 in Hamilton's textbook.[13] The following outline is based on the description presented by Woodridge.[2]

We assume the no-trend model (Equation 1) in which the error follows a first-order autoregressive process. Specifically, we assume that

$$\epsilon_t = \phi_1 \epsilon_{t-1} + \eta_t, \tag{3}$$

where $|\phi_1| < 1$ and η_t are independent disturbances with zero mean and variance σ^2. The constraint on the autoregressive (AR) parameter ϕ_1 ensures the process is stationary, that is, $\text{cov}(\epsilon_t, \epsilon_{t+h})$ is independent of t, and therefore that the variance remains constant over time.

If we also assume that ϕ_1 is known, then we can remove the correlation in the errors by applying the transformation $\tilde{y}_t = y_t - \phi_1 y_{t-1}$ and $\tilde{x}_t = x_t - \phi_1 x_{t-1}$ for $t = 1, \ldots, n-1$. In terms of the transformed data, Equation (1) becomes:

$$\tilde{y}_t = \beta_0 (1 - \phi_1) + \beta_1 \tilde{x}_t + \eta_t, \tag{4}$$

where the errors, η_t, are now mutually independent. Hence the intervention effect, β_1, can be estimated by defining a constant predictor $z_t = (1 - \phi_1)$ and regressing $\tilde{y}_t$ on $\tilde{x}_t$ and z_t in a model without an intercept term.

Because ϕ_1 is usually unknown, y_t and x_t must be transformed using an estimate of this parameter, that is, $\tilde{y}_t = y_t - \hat{\phi}_1 y_{t-1}$ and $\tilde{x}_t = x_t - \hat{\phi}_1 x_{t-1}$. Typically, $\hat{\phi}_1$ is the estimated slope parameter from the regression of $\hat{\epsilon}_t$ on $\hat{\epsilon}_{t-1}$ where these residuals are obtained from the regression of y_t on x_t (Equation 1). The regression of $\hat{\epsilon}_t$ on $\hat{\epsilon}_{t-1}$ can be conducted with or without an intercept; either way, the regression coefficient for $\hat{\epsilon}_{t-1}$ provides a consistent estimate of ϕ_1.

The above method is referred to as the Cochrane-Orcutt method.[14] The Prais-Winsten method[3] is an extension of this method that includes y_0 and x_0 in the analysis and thereby increases the precision of the parameter estimates. To ensure that the error variance is independent of time, y_0 and x_0 are scaled by the factor $\sqrt{\left(1 - \hat{\phi}_1^2\right)}$, that is $\tilde{y}_0 = y_0\sqrt{\left(1 - \hat{\phi}_1^2\right)}$ and $\tilde{x}_0 = x_0\sqrt{\left(1 - \hat{\phi}_1^2\right)}$. Then, as in the Cochrane-Orcutt method, $\tilde{y}_t$ is regressed on $\tilde{x}_t$ and z_t, where $z_t = \sqrt{\left(1 - \hat{\phi}_1^2\right)}$ for $t = 0$ and $z_t = (1 - \phi_1)$ for $t > 0$. As in the Cochrane-Orcutt method, the regression is fitted without an intercept term.

Additional covariates can be handled similarly. For example, to fit the model with trend (Equation 2)—and assuming first-order autoregressive error—we would need to transform t in addition to x_t and y_t.

3.2 | Auto-regressive-moving-average

The first-order autoregressive error model described above (Equation 3) is a special case of an auto-regressive-moving-average (ARMA) model. In this more general model, the error at time t depends on the errors at

the p most recent time points (AR part of the model) and q disturbance terms (MA part of the model), that is:

$$\epsilon_t = \phi_1\epsilon_{t-1} + \cdots + \phi_p\epsilon_{t-p} + \theta_1\eta_{t-1} + \cdots + \theta_q\eta_{t-q} + \eta_t, \tag{5}$$

where the disturbances η_t (also called innovations) are assumed to be uncorrelated and normally distributed with zero mean and constant variance σ^2.

Assuming an ARMA error, the likelihood of the data can be written as $\prod_{t=0}^{n} f\left(y_t|y_{t-1}, \ldots, y_0; \zeta\right)$ and model parameters ζ estimated by maximising this likelihood. Note that ζ includes both the parameters from the ARMA error model and from the model of the mean. The two approaches most commonly used to implement maximum likelihood estimation are: (1) to use the Kalman filter to maximize the full likelihood and (2) to maximise a conditional likelihood obtained by fixing $\eta_{p-1}, \ldots, \eta_{p-q}$ at zero and $\epsilon_0, \ldots, \epsilon_{p-1}$ at their observed values (ie, the residuals at these time points). The two approaches are described in Hamilton's textbook.[13] As an illustration of the conditional likelihood approach, consider the no trend model (Equation 1) with MA(1) error, that is, $\zeta = \left(\beta_0, \beta_1, \theta_1, \sigma^2\right)$. If we set $\eta_{-1} = 0$ then $\epsilon_0 = \eta_0, \epsilon_1 = \theta_1\epsilon_0 + \eta_1$ and $\epsilon_2 = \theta_1\left(\epsilon_1 - \theta_1\epsilon_0\right) + \eta_2$. Because $\epsilon_t = y_t - \beta_0 - \beta_1 x_t$, the first three terms of the likelihood are $y_0 \sim N\left(\beta_0 + \beta_1 x_0,\ \sigma^2\right)$, $y_1|\ y_0 \sim N\left(\beta_0 + \beta_1 x_1 + \theta_1\epsilon_0,\ \sigma^2\right)$ and $y_2|y_1, y_0 \sim N\left(\beta_0 + \beta_1 x_2 + \theta_1\left(\epsilon_1 - \theta_1\epsilon_0\right),\ \sigma^2\right)$, and the subsequent terms can be derived by further iterating the error equation.

To choose the form of the ARMA error model—that is, the values of p and q —some authors recommend inspecting the autocorrelation function and partial autocorrelation functions of the residuals.[5] For example, zero autocorrelation beyond lag 1 implies an MA(1) model (ie, an ARMA model with $p = 0$ and $q = 1$). Others recommend fitting a number of different ARMA models and using a statistical criterion like the AIC, AICc or BIC to select the best fitting model.[4] This approach is appealing because it reduces subjectivity. Indeed, in an early paper on the AIC, Akaike argued for using the criteria to "relieve the time series analyst of much of the burden of making subjective judgements".[15] A drawback is that it is often unclear how many models should be used in the comparison. One strategy for dealing with this problem is to use a forward selection algorithm in which p and q are increased incrementally until there is no further improvement in model fit as measured by AIC, for example.[16]

3.3 | Newey-West

The OLS estimate of the intervention effect is unbiased provided the model for the mean is correctly specified, even in the presence of autocorrelation. Thus, another way to deal with autocorrelation is to use the OLS estimate of the intervention effect and adjust the SE. The Newey-West method does exactly this.[6] It uses the observed correlation between residuals to produce a so-called robust SE. The method is closely related to the methods proposed by White and Liang and Zeger to account for heteroskedasticity and clustering.[17,18]

The key assumption of the Newey-West method is that the error correlation is zero beyond a certain lag m. It is therefore tempting to use a large value of m to minimize the impact of this assumption. Unfortunately, however, the variance estimate is only consistent if m is small relative to the number of observations (n).[6] So how should m be chosen? One option is to use the integer part of $n^{1/4}$. This rule is motivated by the fact that in the original paper by Newey and West one of the conditions used to prove consistency was that the rate of increase in m should be slower than $n^{1/4}$.[2] Alternatively, several data-dependent strategies have also been proposed.[19,20] A simplification of one of these, assuming an AR(1) autocorrelation model with correlation parameter 0.25, leads to the rule $m = 0.75\ n^{1/3}$.[20] More recently it has been shown that size distortion in hypothesis testing may be further reduced using the rule $m = 1.3\ n^{1/2}$ in conjunction with fixed-b critical values.[21]

4 | EXAMPLE: INTRODUCTION OF A PNEUMOCOCCAL VACCINE IN KENYA

To illustrate the three different methods outlined above, we use data from an ITS study of the impact of the introduction of 10-valent pneumococcal vaccine (PCV10) on severe and very severe clinical pneumonia in Kenya.[22] The data consist of monthly hospital admissions for severe or very severe pneumonia in children <5 years collected over a period of 155 months (104 months pre-vaccine introduction and 51 months post introduction) between May 2002 and March 2015 (Figure 1A).

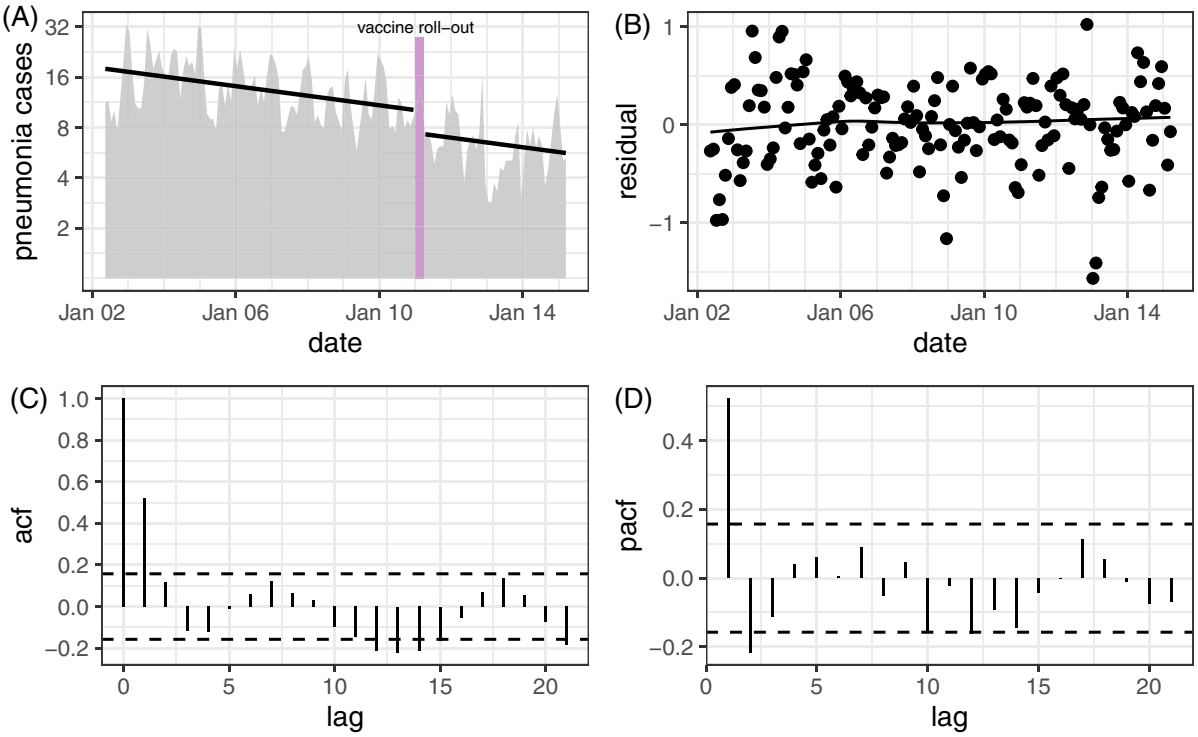

FIGURE 1 (A) Monthly incidence (per 10 000) of severe and very severe clinical pneumonia in children <5 years before and after the introduction of pneumococcal vaccination in Kilifi, Kenya. The solid black line represents the trend and vaccine impact estimated by fitting a linear regression model to the data (Prais-Winsten and ARMA model produced similar estimates). The vertical bar represents the period of vaccine roll-out (Jan - Mar 2011). (B) Residuals with loess trend line. (C) Autocorrelation function for the residuals. (D) Partial autocorrelation function for the residuals

In the original analysis, an AR(2) model was selected based on a comparison of AIC among all ARMA models with $p \leq 3$ and $q \leq 3$. Here we present a reanalysis of these data using the three methods described in Section 3. In each case, we fitted a segmented regression model (Equation 2) to the $\log_2$-transformed incidence rates. A plot of the residuals vs study month suggests that the linear trend assumption is reasonable for these data (Figure 1B). In addition to terms for trend and post-vaccine period (Jan 2011 - Mar 2015), the model included: (i) a binary indicator for health worker strikes (ii) a categorical variable for calendar month and (iii) an indicator for the vaccine roll out period (Jan - Mar 2011). Calendar month was included to account for seasonality in pneumonia incidence and the vaccine rollout period was included to exclude this period from intervention effect estimate.

All analyses were done using R version 3.6.1.[23] The ARMA model fitting and selection was done using the *auto.arima* function (package = *forecast*).[16] Specifically, we used *auto.arima* to implement stepwise selection based on a bias corrected version of the AIC, as recommended by Hyndman and Athanasopoulos,[4] with the constraint that $p \leq 5$ and $q \leq 5$. The Newey-West method was implemented using the *NeweyWest* function (package = *sandwich*)[24] with m chosen according to the method described by Newey and West.[19] Finally, the Prais-Winsten method was implemented using the *prais.winsten* function (package = *prais*).[25] Table 1 shows the intervention effect estimate and confidence interval generated by each method together with the OLS estimate and confidence interval (unadjusted for autocorrelation). The data and code for these analyses are available at https://github.com/christian-bottomley/ITS_Autocorrelation.

The point estimates are similar across all the methods suggesting that the vaccine reduces the incidence of severe and very severe pneumonia by about 27%. The OLS 95% confidence interval is narrower than the other confidence intervals which is unsurprising since there is strong evidence of autocorrelation (OLS: 11.5, 39.8, Prais-Winsten: −2.2, 47.0; ARMA: 5.1, 45.2; Newey-West: 6.4, 43.1). Figure 1C shows that the autocorrelation is strongest at lag 1 and Figure 1D, which shows the partial autocorrelation function, suggests an AR(2) model might be appropriate. Among the methods that account for autocorrelation, the Newey-West and ARMA confidence intervals are similar but the Prais-Winsten confidence interval is significantly wider. However, it is not obvious which is most appropriate—all of them account for significant lag-1 autocorrelation, which is the main feature of these data. This ITS analysis is not unusual in being sensitive to the choice

TABLE 1 Estimates of the trend in pneumonia incidence and vaccine impact

Method	Trend (% reduction per month)	95% CI	Vaccine impact (% reduction)	95% CI
OLS	0.55	0.35, 0.74	27.0	11.5, 39.8
Prais-Winsten	0.56	0.22, 0.89	26.4	−2.2, 47.0
ARMA[a]	0.54	0.26, 0.82	27.9	5.1, 45.2
Newey-West[b]	0.55	0.25, 0.84	27.0	6.4, 43.1

[a] ARMA model with $p = 0$ and $q = 2$.
[b] Accounting for autocorrelation up to lag 9.

of method. In an empirical evaluation of different methods for analyzing ITS, including Prais-Winsten and Newey-West, Turner et al found that statistical significance ($p < 0.05$) differed in 4 to 25% of the pair-wise comparisons.[26]

5 | SIMULATION

We assessed the performance of the different methods by simulating data from the segmented regression model (Equation 2) under 20 different autocorrelation scenarios and 4 different scenarios for ITS length (n = 20, 50, 100 and 300). To simulate from the model, we used a 1:1 ratio for the numbers of observations before and after the intervention and fixed $\beta_0 = 4$, $\beta_1 = -1$ and $\beta_2 = -1/n$ based on the relative reductions associated with intervention (25%) and trend (25%) in the pneumonia example. We note, however, that inference should be unaffected by the choice of parameter values because the SE is independent of the parameter values of the regression.[27] Our own experience of using different values and sensitivity analyses conducted in previous simulation studies also suggest that our findings are independent of the chosen regression parameter values.[7]

We assumed an MA(3) model for the error, that is we assumed $\epsilon_t = \theta_1\eta_{t-1} + \theta_2\eta_{t-2} + \theta_3\eta_{t-3} + \eta_t$ with Var $(\eta_t) = 1$. This model allows for an arbitrary correlation structure up to lag 3 but assumes zero correlation beyond this point. The 20 autocorrelation scenarios were chosen by randomly sampling θ_1 from unif(0, 1), θ_2 from unif(0, θ_1) and θ_3 from unif(0, θ_2). By selecting the parameters in this way, the autocorrelation was constrained to be positive and decreasing over time.

For each scenario, 2000 datasets were generated. Intervention effect estimates (ie, estimates of β_1) and 95% confidence intervals were obtained by implementing the methods as in the pneumococcal vaccine example, and their performance was evaluated in terms of bias, mean square error and coverage of 95% confidence intervals using the rsimsum package.[28] We also evaluated estimates obtained by fitting the true MA(3) model via maximum likelihood. The code for the simulation study and a complete table of results, including Monte Carlo error estimates, is available at https://github.com/christian-bottomley/ITS_Autocorrelation.

5.1 | Simulation results

The results from the simulation study are summarized in Supplementary Figure 1 (bias), Figure 2 (root mean square error) and Figure 3 (coverage) and in the text below.

Bias: It is well known that OLS estimates are unbiased even when errors are correlated.[13] Moreover, Prais-Winsten and ARMA estimates are also unbiased under very general conditions, including correlated errors, because they can be viewed as feasible generalized least squares estimates.[29] Our simulations results are consistent with these theoretical results. Across all sample size and autocorrelation scenarios, the mean bias was close to zero (OLS & Newey-West = −0.0039, Prais-Winsten = −0.0031, ARMA = −0.0033, MA(3) = −0.0024). Furthermore, in 76 out of the 80 scenarios, the 95% Monte Carlo confidence interval for the bias estimate included zero, irrespective of the method.

Mean square error: In scenarios where the degree of autocorrelation was low to moderate (lag-1 correlation <0.6) the different methods performed similarly in terms of MSE. The mean root MSE (across all sample size scenarios) was: OLS & Newey-West = 0.75, Prais-Winsten = 0.72, ARMA = 0.73, MA(3) = 0.77. In the high autocorrelation scenarios (lag-1 correlation ≥0.6) the Prais-Winsten and ARMA methods performed significantly better than OLS (mean root MSE: OLS & Newey-West = 1.17, Prais-Winsten = 0.97, ARMA = 0.99, MA(3) = 0.99).

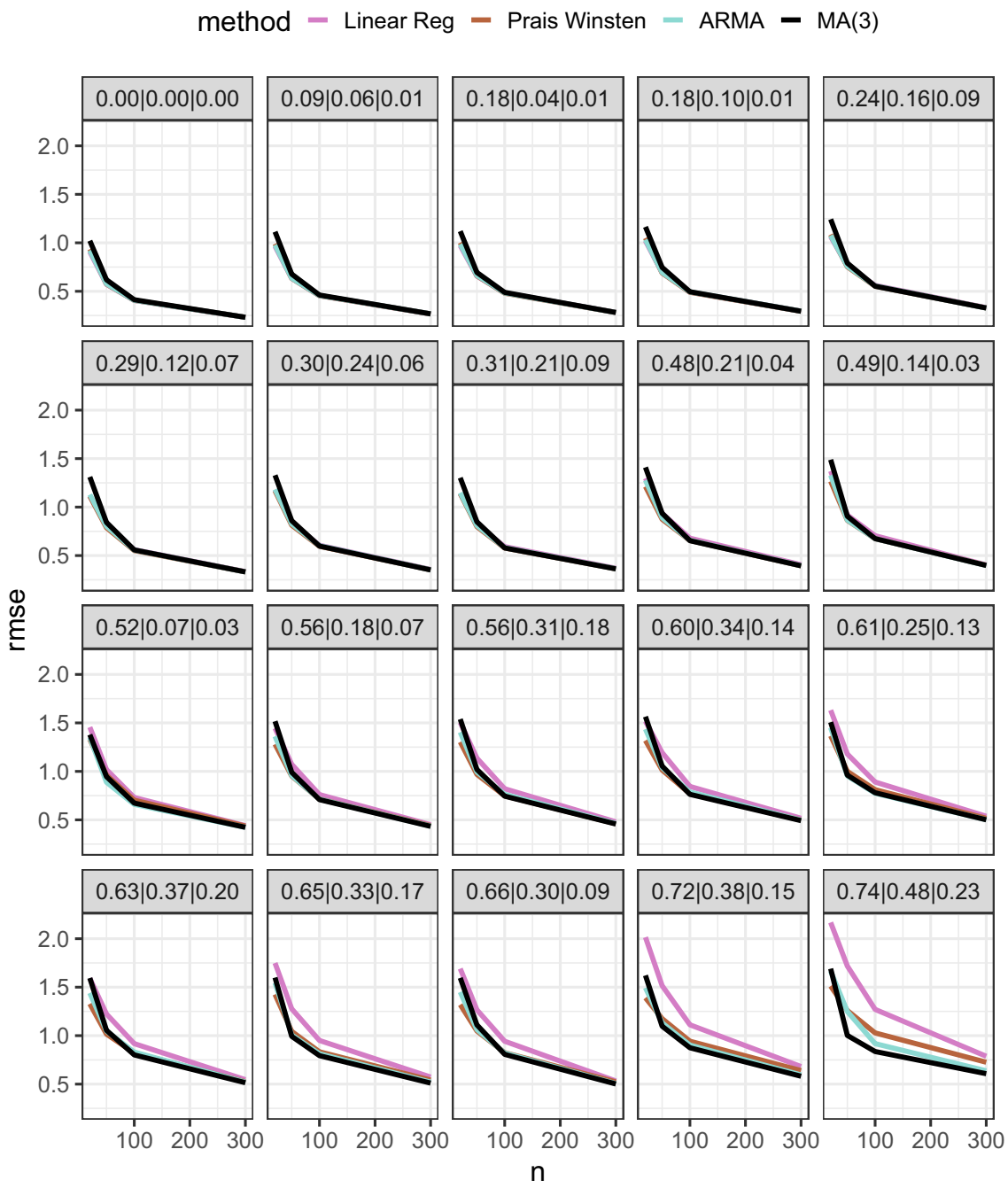

FIGURE 2 Root mean square error (RMSE) as a function of ITS length (n) in 20 autocorrelation scenarios. The scenarios range from lowest correlation in the top left (lag-1, lag-2 and lag-3 correlations of 0.06, 0.02 and 0.01 respectively) to highest correlation in bottom right (lag-1, lag-2 and lag-3 correlations of 0.74, 0.48 and 0.23 respectively)

Coverage: In general, coverage was below the nominal 95% level; however, there was significant variation between the methods. Across all scenarios, the mean coverage was: OLS = 81.3%, Prais-Winsten = 92.2%, ARMA = 88.3%, Newey-West = 82.9% and MA(3) = 82.9%. The variation in performance was particularly apparent in scenarios with $n \leq 50$ (mean coverage: OLS = 84.3%, Prais-Winsten = 88.0%, ARMA = 80.8%, Newey-West = 74.3%, MA(3) = 65.2%). In these scenarios, coverage was generally worst when an MA(3) model (the true model) was fitted and best when the Prais-Winsten method was used, suggesting an inverse relationship between coverage and number of parameters included in the error model. In a supplementary analysis exploring the coverage of different MA models, we also observed an inverse relationship between coverage and number of parameters (Supplementary Figure 2). In scenarios with $n \leq 50$ and low levels of

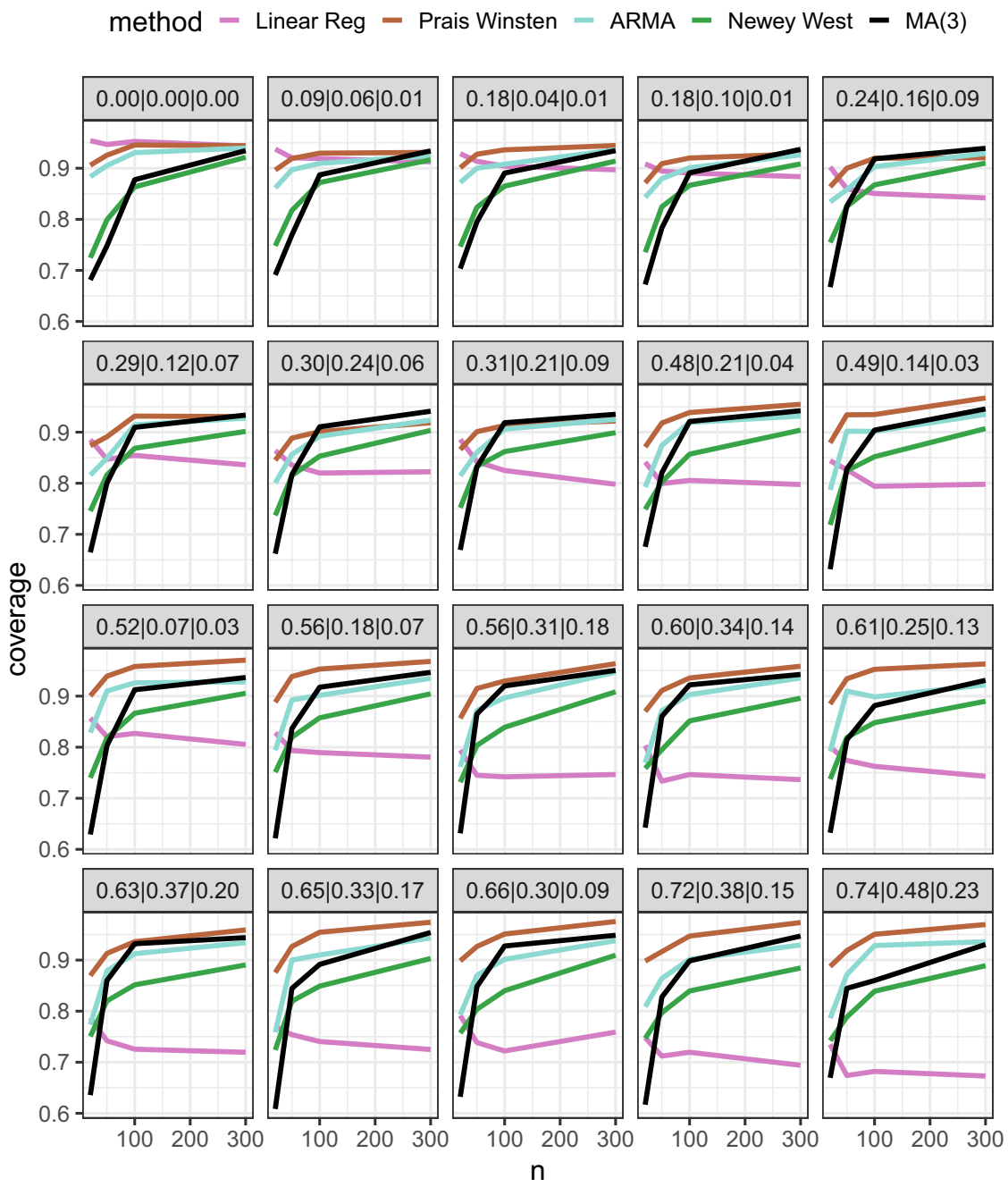

FIGURE 3 Confidence interval coverage (nominal value = 0.95) as a function of ITS length in 20 autocorrelation scenarios. The scenarios range from lowest correlation in the top left (lag-1, lag-2 and lag-3 correlations of 0.06, 0.02 and 0.01 respectively) to highest correlation in bottom right (lag-1, lag-2 and lag-3 correlations of 0.74, 0.48 and 0.23 respectively)

autocorrelation (lag-1 correlation <0.3), OLS produced coverage that was slightly closer to the nominal level than the Prais-Winsten method (mean coverage: OLS = 90.8%, Prais-Winsten = 89.9%).

6 | COVERAGE AND POWER

In our simulation study, we found that under coverage was pervasive, particularly when error models with multiple autocorrelation parameters were fitted. Here we briefly discuss the issue and possible solutions. We also show that complex error models are associated with reduced statistical power.

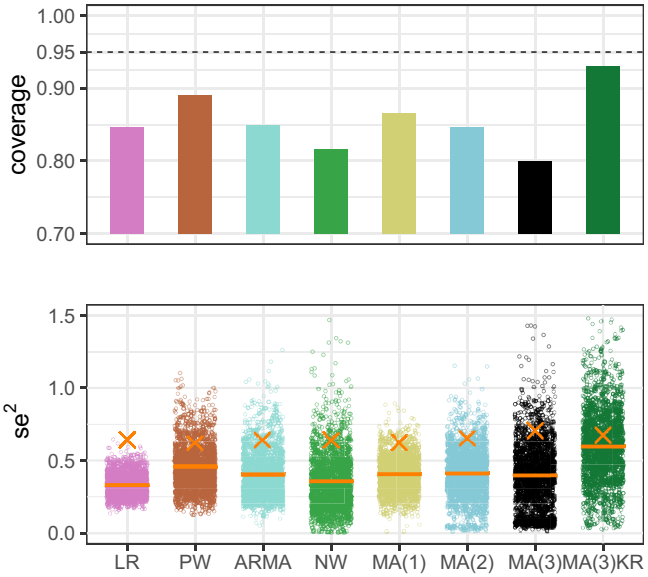

FIGURE 4 Distribution of estimates of $\text{se}\left(\hat{\beta}_1\right)^2$ and coverage under moderate autocorrelation (lag 1 = 0.29, lag 2 = 0.12, lag 3 = 0.07) and $n = 50$. The true value is denoted by a cross and horizontal bars represent the median of the distribution. Downward bias and variability in these estimates reduce the coverage of confidence intervals based on a standard normal distribution. The amount of bias differs between methods, as does the amount of variability. In particular the variability increases as the number of parameters included in the MA error model increases from 1 to 3. The coverage can be brought close to the nominal value by using the Kenward Roger (KR) method, which reduces bias and accounts for variability by basing confidence intervals on a t-distribution rather than the standard normal

6.1 | Coverage

Confidence intervals for the intervention effect estimate, $\hat{\beta}_1$, are based on the distribution of the test statistic

$$\text{T} = \frac{\hat{\beta}_1 - \beta_1}{\hat{\text{se}}\left(\hat{\beta}_1\right)}$$

where $\hat{\text{se}}\left(\hat{\beta}_1\right)$ is an estimate of the SE of $\hat{\beta}_1$. Typically, it is assumed that the distribution of T can be approximated by a standard normal distribution and 95% confidence intervals take the form $\hat{\beta}_1 \pm 1.96\ \hat{\text{se}}\left(\hat{\beta}_1\right)$.

In the hypothetical scenario where the error follows an ARMA model with known parameters, the estimated SE can be replaced by the *known* SE and T follows a standard normal distribution exactly. In this situation, confidence intervals based on this distribution are guaranteed to have correct coverage.

In practice, however, the parameters of the error model are unknown and an estimate of the SE must be used. This leads to under coverage for two reasons: (1) the SE is underestimated because of bias in the error model parameter estimates due to overfitting (2) variability in the SE estimate is unaccounted for. Variation and bias in the SE estimates are illustrated in Figure 4. In this figure, we see that differences between methods in these characteristics translate into differences in coverage. In particular, the Prais-Winsten method achieves coverage closest to the nominal value (among the standard methods) because both bias and variation are kept low. We use the MA models in the figure to illustrate the effect of increasing the number of model parameters. Consistent with our earlier observation that coverage decreases with increasing model complexity, here we see that both bias and variability increase with increasing model complexity.

Bias in the SE can be reduced by using restricted maximum likelihood estimation (REML) instead of maximum likelihood estimation, an approach commonly used in mixed effects modeling.[30,31] The idea behind REML is to linearly transform the outcome vector so that the likelihood is a function of the error model parameters only. Once the error model parameters have been estimated by maximizing this restricted likelihood, the regression parameters can be estimated by maximizing the full likelihood with the error model parameters fixed at their estimated values. Computationally the estimation of the regression parameters is equivalent to generalized least squares. By making the error model estimation

independent of the parameters of the regression model, REML reduces bias in the estimation of the error model, which, in turn, reduces bias in the SE estimates (Supplementary Figure 3). At present, REML is rarely used to analyze time series, though several authors have argued that it is useful in this setting.[10,26,27,32]

To further improve coverage, several authors[33-35] have proposed assuming $\widehat{\text{var}}\left(\hat{\beta}_1\right)$ follows a scaled chi-square distribution and using Satterthwaite's approximation[36] to compute the appropriate degrees of freedom. By making this assumption, T is approximated by a t-distribution rather than the standard normal. Specifically, it is assumed $\left(\frac{d}{\phi}\right)\hat{\phi} \sim \chi^2_d$, where $\phi = \text{var}\left(\hat{\beta}_1\right)$ and $\hat{\phi}$ is the estimate of ϕ. The degrees of freedom are estimated from the expression $d = \frac{2\phi^2}{\text{var}(\hat{\phi})}$, which is derived by matching variances, and plugging in appropriate estimates of ϕ and $\text{var}(\hat{\phi})$ ($\hat{\phi}$ is used to estimate ϕ and an estimate for $\text{var}(\hat{\phi})$ can be obtained via a Taylor series approximation). The Satterthwaite approximation can be used with either maximum likelihood or REML but since the aim is to improve coverage it is generally used with REML. Kenward and Roger suggest making a further correction to the SE, which slightly improves coverage.[31,35]

In Supplementary Figure 2, the coverage of MA(3) fitted via maximum likelihood is compared with coverage of MA(3) fitted via REML with the Kenward-Roger adjustments. It can be seen that the Kenward-Roger method largely solves the problem of low coverage.

6.2 | Power

To achieve correct coverage, the degrees of freedom must be significantly reduced when an MA(3) model is fitted even when there is no autocorrelation. This suggests that fitting an over-parameterized error model comes at a cost in terms of statistical power. For example, in the zero-autocorrelation scenario with $n = 50$, on average 9.4 degrees of freedom are used in the Kenward-Roger adjustment compared with 47 in OLS, which translates into a difference in power of 42% vs 33%. Unfortunately, it is difficult to compare power more widely because of differences in coverage. In Supplementary Figure 4, we limit the influence of coverage by restricting the comparison to methods with coverage >85%. In this analysis, MA(3) with Kenward-Roger adjustment has lower power than other methods, though at high levels of autocorrelation the comparison is only between Prais-Winsten and MA(3) with Kenward-Roger adjustment because other methods have coverage <85%.

In summary, low coverage is caused by bias and variability in the SE estimate. REML can be used to reduce bias and the Satterthwaite approximation can be used to account for uncertainty in the SE. These methods improve coverage when complex error models are used. However, such models are still often not desirable because they come at a cost of reduced statistical power.

7 | DISCUSSION

In our simulation, study we explored the performance of the most commonly used methods for handling autocorrelation in terms of bias, MSE, and confidence interval coverage. Consistent with theoretical results, we found that all methods are unbiased, and at large sample sizes ($n > 100$), there was also little to distinguish the methods—all methods were associated with similar MSE and coverage close to the nominal value. Differences were more apparent at small sample sizes. Here Prais-Winsten and ARMA were the most efficient methods, particularly at high levels of autocorrelation, and Prais-Winsten generally had coverage closest to the nominal level, though all methods, including Prais-Winsten, were associated with some under coverage.

Our findings on efficiency (MSE) are in keeping with asymptotic results and previous simulation studies. It is well known that feasible GLS estimates, which include Prais-Winsten and maximum likelihood estimates, are asymptotically efficient provided that the autocorrelation structure is correctly specified.[13] Furthermore, asymptotic efficiency is maintained even when the model is unknown if model selection is done via the AIC.[37] Surprisingly theoretical results also show that OLS — which does not account for autocorrelation in the estimation of regression parameters — is also efficient asymptotically.[38] In finite samples, a number of simulation studies have shown that accounting for autocorrelation can improve efficiency,[39,40] though OLS is probably more efficient at low levels of autocorrelation.[40]

Our simulations suggest that coverage will often be of greater concern than efficiency. An important determinant of coverage is the number of parameters included in the error model. The Prais-Winsten method is able to achieve coverage close to the nominal value because it is based on an AR(1) error model, that is, there is a single autocorrelation

parameter. In ARMA modeling good coverage is facilitated by keeping the number of parameters to a minimum. To a certain degree, this is achieved by using a model selection criterion such as AIC. However, these criteria are designed to minimize out-of-sample prediction error not to ensure correct coverage. Furthermore, model selection is data-dependent and this too can negatively impact on coverage.[41] Although ARMA modeling offers greater flexibility, our simulation study suggests that there is little advantage in terms of MSE and coverage over assuming an AR(1) model. A further advantage of assuming an AR(1) model is that it produces a produces a single measure of autocorrelation that can be used to compare between studies and guide sample size calculations for future studies.[42]

Although our simulation results suggest that the Prais-Winsten method achieves reasonable levels of coverage when analyzing time series with as few as $n = 20$ observations, alternative methods may be necessary when analyzing shorter time series. Several simulation studies have shown that estimation via REML rather than maximum likelihood can help to maintain good coverage in short time series, particularly when confidence intervals are based on a t-distribution with degrees of freedom estimated using the Satterthwaite method.[7,10,27] These methods can be implemented using software to fit mixed models—for example, the mixed command in Stata. Parametric bootstrapping offers an alternative approach.[8] Our exploration of the Kenward Roger method, which is similar to REML in conjunction with Satterthwaite degrees of freedom, suggests that this approach can help to improve coverage but that it is important to keep the number of parameters to a minimum (eg, by fitting an AR(1) model) to avoid low power. Unfortunately, REML + Satterthwaite does not maintain good coverage when time series are very short ($n < 12$) and REML often fails to coverage. In these circumstances, the best strategy appears to be to avoid any adjustment for autocorrelation and use OLS.[7,9]

In conclusion, we recommend the Prais-Winsten method over ARMA modeling and the Newey-West method for analyzing ITS. For short time series, ($n < 20$) analysts should consider using REML combined with a Satterthwaite degrees of freedom correction or OLS without any adjustment for autocorrelation.

DATA AVAILABILITY STATEMENT

The data and code for these analyses are available at https://github.com/christian-bottomley/ITS_Autocorrelation.

ORCID

C Bottomley https://orcid.org/0000-0002-5241-8412
RH Keogh https://orcid.org/0000-0001-6504-3253

REFERENCES

1. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *Int J Epidemiol.* 2017;46(1):348-355.
2. Wooldridge JM. Chapter 12: serial correlation and heteroskedasticity in time series regressions. *Introductory Econometrics: A Modern Approach*. fifth ed. Mason, OH: South-Western; 2009.
3. Prais SJ, Winsten CB. Trend estimators and serial correlation. *Cowles Commission Discussion Paper No. 383*. Chicago; 1954.
4. Hyndman RJ, Athanasopoulos G. *Forecasting: Principles and Practice*. second ed. Melbourne, Australia. OTexts.com/fpp2: OTexts; 2018.
5. Box GEP, Tiao GC. Intervention analysis with applications to economic and environmental problems. *J Am Stat Assoc.* 1975;70(349):70-79.
6. Newey WK, West KD. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econom.* 1987;55(3):703-708.
7. 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(1):181.
8. McKnight SD, McKean JW, Huitema BE. A double bootstrap method to analyze linear models with autoregressive error terms. *Psychol Methods.* 2000;5(1):87-101.
9. Bence JR. Analysis of short time series: correcting for autocorrelation. *Ecology.* 1995;76(2):628-639.
10. Alpargu G, Dutilleul P. Efficiency and validity analyses of two-stage estimation procedures and derived testing procedures in quantitative linear models with AR(1) errors. *Commun Stat Part B: Simul Comput*. 2003;32(3):799-833.
11. Bottomley C, Scott JAG, Isham V. Analysing interrupted time series with a control. *Epidemiol Methods.* 2019;8(1):20180010.
12. Harvey A, Koopman SJ. Structural time series models in medicine. *Stat Methods Med Res.* 1996;5(1):23-49.
13. Hamilton JD. *Time Series Analysis*. Princeton, NJ: Princeton University Press; 1994.
14. Cochran WG. Some methods for strengthening the common $\chi 2$ tests. *Biometrics.* 1954;10(4):417-451.
15. Akaike H. On the likelihood of a time series model. *J R Stat Soc Series D.* 1978;27(3/4):217-235.
16. Hyndman RJ, Khandakar Y. Automatic time series forecasting: the forecast package for R. *J Stat Softw.* 2008;27(3):1-22.
17. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica.* 1980;48(4):817-838.
18. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika.* 1986;73(1):13-22.
19. Newey WK, West KD. Automatic lag selection in covariance matrix estimation. *Rev Econ Stud.* 1994;61(4):631-653.

20. Andrews DWK. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica.* 1991;59(3):817-858.
21. Lazarus E, Lewis DJ, Stock JH, Watson MW. HAR inference: recommendations for practice. *J Bus Econ Stat.* 2018;36(4):541-559.
22. Silaba M, Ooko M, Bottomley C, et al. Effect of 10-valent pneumococcal conjugate vaccine on the incidence of radiologically-confirmed pneumonia and clinically-defined pneumonia in Kenyan children: an interrupted time-series analysis. *Lancet Glob Health.* 2019;7(3):e337-e346.
23. R Core Team. R: a language and environment for statistical computing. Published Online; 2021.
24. Zeileis A. Econometric computing with HC and HAC covariance matrix estimators. *J Stat Softw.* 2004;11(10):1-17.
25. Mohr F. prais: Prais-Winsten estimator for AR(1) serial correlation. R Package Version 1.1.1; 2019.
26. Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, McKenzie JE. Comparison of six statistical methods for interrupted time series studies: empirical evaluation of 190 published series. *BMC Med Res Methodol.* 2021;21(1):134.
27. Cheang WK, Reinsel GC. Bias reduction of autoregressive estimates in time series regression model through restricted maximum likelihood. *J Am Stat Assoc.* 2000;95(452):1173-1184.
28. Gasparini A. rsimsum: summarise results from Monte Carlo simulation studies. *J Open Source Softw.* 2018;3(26):739.
29. Kakwani NC. The unbiasedness of Zellner's seemingly unrelated regression equations estimators. *J Am Stat Assoc.* 1967;62(317):141-142.
30. Diggle PJ, Heagerty P, Liang KY, Zeger SL. *Analysis of Longitudinal Data*. second ed. Oxford, UK: Oxford University Press; 2002.
31. Spilke J, Piepho HP, Hu X. A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data. *J Agric Biol Environ Stat.* 2005;10(3):374-389.
32. Tunnicliffe WG. On the use of marginal likelihood in time series model estimation. *J R Stat Soc Series B.* 1989;51(1):15-27.
33. Giesbrecht FG, Burns JC. Two-stage analysis based on a mixed model: large-sample asymptotic theory and small-sample simulation results. *Biometrics.* 1985;41(2):477-486.
34. Fai AHT, Cornelius PL. Approximate f-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *J Stat Comput Simul.* 1996;54(4):363-378.
35. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics.* 1997;53(3):983-997.
36. Satterthwhaite FE. An approximate distribution of estimates of variance components. *Biometrics.* 1946;2(6):110-114.
37. Vrieze SI. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychol Methods.* 2012;17(2):228-243.
38. Grenander U. On the estimation of regression coefficients in the case of an autocorrelated disturbance. *Ann Math Stat.* 1954;25(2):252-272.
39. Rao P, Griliches Z. Small-sample properties of several two-stage regression methods in the context of auto-correlated errors. *J Am Stat Assoc.* 1969;64(325):253-272.
40. Spitzer JJ. Small-sample properties of nonlinear least squares and maximum likelihood estimators in the context of autocorrelated errors. *J Am Stat Assoc.* 1979;74(365):41-47.
41. Chatfield C. Model uncertainty, data mining and statistical inference. *J R Stat Soc Series A.* 1995;158(3):419-466.
42. Turner SL, Karahalios A, Forbes AB, et al. Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: a review. *J Clin Epidemiol.* 2020;122:1-11.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Bottomley C, Ooko M, Gasparrini A, Keogh R. In praise of Prais-Winsten: An evaluation of methods used to account for autocorrelation in interrupted time series. *Statistics in Medicine*. 2023;1-12. doi: 10.1002/sim.9669