Addressing Autocorrelation in Time Series Data: A Comparison of Four Analytic Methods Using Data from College Course Evaluations

Larry Ludlow **Shenira A. Perez**
Boston College

The sequential nature of observations in time series make them inherently prone to autocorrelation. This can be problematic because autocorrelation violates a major assumption associated with many conventional statistical methods. Although numerous analytic techniques address autocorrelation, the literature is generally devoid of discussions that contrast the benefits and disadvantages of various methods. This paper provides readers with a brief introduction to autocorrelation and related concepts, and uses empirical data from college course evaluations to contrast the results of four commonly used methods for adjusting autocorrelation in social science research. Implications of results and recommendations for choosing between these strategies are discussed.

A time series is a collection of sequentially ordered observations of a variable that can be used to examine longitudinal causal patterns, forecast trends, or explore the impact of a single event at a specified point in time (McLeary & Hay, 1982). Historically, this type of data has been used in the fields of finance, sociology, economics, and meteorology to study a vast array of phenomena, such as global warming, financial trade markets, and unemployment rates (e.g., Edwards & Richardson, 2004; Jackman & Layard, 1991; Loomis, Castillejos, Gold, McDonnell, & Borja-Aburto, 1999; Shahbaz, Nanthakumar, Rashid, & Talat, 2015). More recently, educational researchers have begun to draw on the broad applicability of time series designs to examine long-term, time-dependent processes within the context of learning and education (e.g., Erdem & Ucar, 2013; MacSuga-Gage, & Gage, 2015; Umble, Cervero, Yang, & Atkinson, 2000).

Ironically, this increased interest in time series data can also pose analytic challenges for educational researchers who are not accustomed to working with temporally ordered data (Box-Steffensmeier, Freeman, Hitt, & Pevehouse, 2014). Unlike cross sectional data, which are collected at one point in time and can be structured within a data file in any order, time series data are collected and analyzed sequentially, which can produce observations and error terms that are correlated across time. The correlation of residuals across time—or *autocorrelation*—is problematic because it violates a critical assumption associated with certain statistical methods, such as ordinary least squares (OLS) regression. If not corrected, autocorrelation can seriously distort the results yielded from time series analysis (Chatterjee & Price, 1991; Pindyck & Rubinfeld, 1991).

Although there are several analytic options for adjusting autocorrelation, these methods have their respective limitations and vary considerably in complexity and sophistication (Brillinger, 2001). Choosing the most effective approach can be challenging for researchers who lack training in time series analysis because it involves a consideration of the specific characteristics of both the data and the modeling techniques. Unfortunately, those who turn to the extant literature for direction will find that it is generally devoid of sources that provide procedural comparisons or contrast the limitations of various strategies (see Nicholich & Weinstein, 1981 for an exception; see Huitema & McKean, 2007 for a comparison of statistical methods for identifying autocorrelation). Moreover, in many cases, authors either provide overly technical descriptions of procedures that can alienate novice researchers (e.g., Afzal, Gagnon, & Mansell, 2015) or fail to indicate how they addressed autocorrelation in their data altogether (e.g., Chao, 2012; Clyde & Klobas, 2001). This lack of readily accessible comparative information creates a burden for those who are new to time series analysis and require additional direction when formulating their analytic strategies.

Objectives

The purpose of this paper is to illustrate a time series analysis that is accessible to researchers with varying levels of statistical expertise and, in doing so, address a gap in the literature through the comparison of four analytic procedures employed on the same data. The paper begins with a summary of pertinent background information related to the nature of autocorrelation in time series data. Following this, we outline four commonly employed procedures used to address autocorrelation in social science research and then use real data from college course evaluations to contrast the results of the procedures. Lastly, we present a general discussion regarding the implications of using one statistical method over another and provide additional recommendations for researchers who are considering working with time series data.[1]

Background Information

There are several recurring concepts in the time series literature that should be clearly understood by anyone working with this type of data. However, time series are inherently complex and the literature is associated with a cumbersome degree of jargon that is somewhat inconsistent across disciplines (Tsay, 2000). In this section, we provide a basic summary of essential background information and present it in a manner we hope will facilitate readers' understanding of the results and discussion presented later in the paper.

Time series, time series data, and time series analysis

First and foremost, we begin by describing what constitutes a *time series* and how it differs from *time series* data and *time series analysis*. This distinction is important because these terms are not synonymous and yet are often used interchangeably in the literature. *Time series* refers to a set of observations measured sequentially between equally spaced intervals of time (aka *time lags*), whereas *time series data* are a particular portion of a time series that are analyzed using various statistical methods. Finally, the umbrella of statistical techniques used to analyze time series data is referred to as *time series analysis* (McCleary & Hay, 1982).

When time series data are analyzed using an appropriate analytical strategy, they can be used to answer important empirical questions and provide valuable insight into real-world phenomena (e.g., Beck & Katz, 2011; Brown, Katz, & Murphy, 1984; Marston, 1988). However, in order to maximize the precision and efficiency of the parameter estimates yielded in time series analysis, the data should meet some basic requirements. One, the data should have a minimum of $n > 30$ observations for each variable in the series. Two, the lags between observations should be equally spaced throughout the series. Lastly, all observations in a time series should rely on the same measurement tool and be based on the same outcome (McLeary & Hay, 1982). Although there are analytical methods that can accommodate data that do not meet these requirements (e.g., *short time series analysis* for $n < 30$; Bloom, 2003), these are beyond the scope of this paper.

Autocorrelation

In cross-sectional data, observations are randomly collected and thus, assumed to be independent of one another. In contrast, observations in a time series are collected in sequential order and are, therefore, effectively related to each other through time-dependent processes. This element of time series makes this type of data more likely to contain a *chronic* dependency among the residuals, which is referred to as *autocorrelation* (aka *serial correlation*). More specifically, in any given dataset that does *not* contain autocorrelated residuals, the degree of correlation between the residuals for two or more time points (i.e., rho) is equal to zero, $\rho\left(\varepsilon_t, \varepsilon_{t-t_j}\right) = 0$, where t=1,…N observations and j=the degree of lag between observations. In contrast, in data that contain autocorrelated residuals, the error term (ε) associated with each observation shares common variance with the error in a given number of preceding residuals in the series, such that $\varepsilon_t = \rho\varepsilon_{t-j} + \omega_t$, where ω_t are assumed to be *i.i.d.* $N(0, \sigma^2)$.

Although the presence of autocorrelation is not necessarily problematic in and of itself, it can present a challenge for researchers who are interested in examining unbiased associations between two or more variables in a time series. Autocorrelation can also introduce additional analytic hurdles for novice researchers with limited statistical expertise or those who simply favor certain types of estimation methods over others. For instance, using OLS regression with autocorrelated data violates a critical assumption associated with this statistical method, such that the residual error for each observation is assumed to be random and unrelated to other errors in the data sequence $\rho\left(\varepsilon_t, \varepsilon_{t-t_j}\right) = 0$, where $\varepsilon_t = y_t - \hat{y}_t$, and $y_t - \hat{y}_t$ is the difference between the observed and predicted value for the outcome variable y (Chatterjee & Price, 1991).[2] Violating this assumption is problematic because it hinders the extent to which the parameter estimates for β_0 and β_k (where k is the number of predictors) are the best, linear and unbiased estimates. More specifically, the standard errors (*SE*) for sample estimates may be biased along with overestimated degrees of freedom, all of which would produce β_k estimates systematically varying from the true parameters, while generating excessively large t statistics and small CIs for the estimates, which then increase the likelihood of Type I statistical significance decision errors (Shumway & Stoffer, 2010).

Variability in Autocorrelation Patterns

Within the time series literature, there is an extensive degree of variability in the characteristics of analyzed data, as well as in the complexity of the time-dependent processes embedded within those data (Box-Steffensmeier et al., 2014; Box, Jenkins, & Reinsel, 1994; McCleary & Hay, 1982). Although there are dozens of ways to characterize different types of time-dependent processes, we limit our discussion by generally categorizing basic versus more complex processes. A *first-order autocorrelation* (i.e., lag-1 autocorrelation), refers to cases where the error term associated with each observation in a series is correlated with the error term of the previous observation, whereas a *higher-order autocorrelation* (i.e., lag-2….*n* autocorrelation) indicates a more complex time-dependent process. This distinction is important because, as will be demonstrated, identifying higher-order autocorrelations often requires additional probing of the residuals and the use of more sophisticated adjustment strategies (McCleary & Hay, 1982).

Diagnosing Autocorrelation

Evaluating Residual Patterns

Given that autocorrelated residuals can reflect various lagged patterns, graphical representations of the error terms are useful for identifying different time-dependent patterns in data (McCleary & Hay, 1982). For example, plotting the residuals against time is one method of doing this, such that a variable that represents time (e.g., the sequence order in which observations were measured) is plotted on the x-axis against the residuals on the y-axis.

Statistical Testing

Although there are several statistical tests for diagnosing autocorrelation, they often differ in the type of autocorrelation they are designed to assess and in the estimation methods they are compatible with (Box et al., 2014). In this paper, we limit our discussion to two widely used autocorrelation tests, the *Durbin-Watson* (D_w) and the *Ljung-Box* (Q) (McCleary & Hay, 1982).

The Durbin-Watson is designed to identify first-order autocorrelations in OLS regression models. Values for D_w can range from 0-4, with lower values reflecting a higher degree of autocorrelation (Durbin & Watson, 1950; 1951). The statistical significance of D_w is assessed by checking the value of the test statistic against the confidence intervals (CI) for the critical values specified for a given model.[3] The CIs are based on both the *k* and *n* of the specific analysis, where (*k*) refers to the number of predictors (including the constant) and (*n*) refers to the sample size. A $D_w <$ the lower limit of the CI_{dL} indicates the error terms are positively correlated and statistically significant, whereas a $D_w >$ the higher limit of the CI_{dU} indicates the positive correlation is not statistically significant.[4] Although a statistically significant D_w implies a first-order autocorrelation, this result should not be interpreted as confirming a lag-1 autocorrelation—instead, it should be interpreted as an indication that further probing of the error terms is warranted (Blattberg, 1973; Zinde-Walsh & Galbraith, 1991). One limitation of the Durbin-Watson is that it is not considered to be a reliable test for autocorrelation in models that are estimated using lagged predictor variables (Durbin & Watson, 1950; 1951)—a point we draw on later.

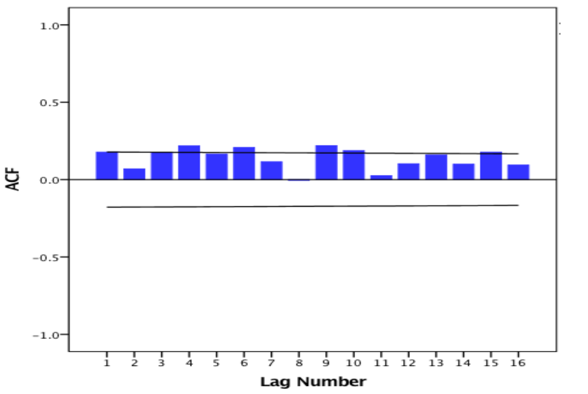

Figure 1. Autocorrelation function plot of the unstandardized residuals for the baseline model. Blue bars represent the coefficients. Lines are Upper and Lower Confidence Limits

The Ljung-Box (Q) has broader applicability than the Durbin-Watson because it can be applied to any time series and can be used in models that contain lagged predictors (Wooldridge, 2016). Another advantage of the Ljung-Box is that it can be computed for individual time lags within the same time series. For example, in most versions of SPSS, the autocorrelation function (ACF) can be used to produce a table that includes the autocorrelation coefficient and SE of each coefficient for individual lags, as well as the Ljung-Box statistic for each coefficient and its respective alpha. The ACF can also be used to produce a graphic plot of the autocorrelation coefficients for a series of lags (see Figure 1). Both the ACF table and plot are useful tools for examining and identifying patterns in residuals.

Addressing Autocorrelation

The four analytic strategies outlined in this section were primarily chosen for two reasons. First, all four strategies are commonly used to address autocorrelation in educational research and second, given their straightforward nature, they are generally accessible to researchers with varying levels of statistical expertise. The first three procedures outlined are described specifically in relation to OLS regression and can be considered relatively simple alternatives to the Autoregressive Integrated Moving Average (ARIMA) modeling technique.[5] For each method, we provide a brief statistical explanation, procedural instructions for SPSS, and a brief discussion regarding strengths and limitations.[6,7]

Modeling and Controlling for a Linear Time Trend

Oftentimes the presence of autocorrelation in time series can be attributed to an underlying time trend between the variables in the series, such that the values for both the outcome and predictor are either increasing or decreasing at a similar rate. In these cases, if the underlying trend that relates the variables is not addressed, the regression analysis may yield spurious results that include exaggerated estimates of the model's explanatory power and biased parameter estimates. To diagnose a linear time trend in time series data, a simple scatterplot of the residuals as a function of time can be evaluated for an upward or downward trend in values.

Once a linear time trend has been identified, the time trend can be estimated and used as a covariate in a subsequent model, or used to eliminate the trend from each variable prior to a subsequent analysis with the detrended data (Wooldridge, 2016). A time trend can be generated by computing a variable for time (*t*), where *t=1…n,* and *t* is simply the sequence number for each observation. In the first method, the variable *t* is included in the model as a covariate in order to obtain model and parameter estimates with and without the variance explained by the time trend. In the second method (Model 2), *x* and *y* are both regressed on *t*, and the unstandardized residuals yielded in each analysis are then used in the subsequent regression as the detrended x* and y*. Although there may be negligible differences in the R^2 and β estimates, the R^2 yielded in the regression of x* on y* ($R^2_{Model\ 2}$) should be almost identical to R^2_{Change} from the regression analysis that contained the covariate with *x* added after the covariate. One benefit of these methods is that the specified models are tested on the entire sample, whereas some procedures, as we demonstrate, can negatively impact sample size.[8] However, it should be noted that this strategy may not be an effective approach for time series that contain residuals displaying non-linear trends (e.g., cubic trend; quadratic trend), because those fluctuations are indicative of more complex time-dependent processes that will not be fully captured by a linear time trend.

Lagged Covariate Predictor

Another method for adjusting autocorrelation within OLS regression is by including a *lagged* version of the outcome variable as a covariate predictor in the model (Hyndman & Athanasopoulos, 2014). When a variable is *lagged*, the values of the original variable are moved forward by the specified number of observations *lag(j).* The LAG command in SPSS can be used to create a variable with the specified *lag(j)* from an existing variable in the dataset (e.g., outcome_lag_1=LAG(outcome, 1). This new lagged outcome variable is then included as a covariate in the regression model as a way of removing the influence of the previous observation (y_{t-1}) on the subsequent observation (y_t). This strategy allows for more customization than the previous method, in that the researcher can decide the extent to which previous observations influenced each subsequent observation. However, using lagged predictor variables is also associated with a limitation, such that for every *lag(j)* specified in the model, the first *j* observations in the series are dropped from the analysis. This creates a declining power situation that can be problematic with smaller sample sizes and those lost cases themselves may contain valuable trend observations at the onset of the time series.

Cochrane-Orcutt Procedure

The Cochrane-Orcutt procedure is a transformation method, where the correlation coefficient rho (ρ) of the residuals $(e_t = \mathrm{Y}_t - \widehat{\mathrm{Y}}_t)$—which is computed using $\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_{t-1}^2}$—is used to adjust the values of the outcome y and all predictors x_1, x_2, etc., such that the values for y and x respectively, become $y_t' = y_t - \rho(y_t - 1)$ and $x_t' = x_t + \rho(x_t - 1)$ (Cochrane & Orcutt, 1949). When used to adjust for a first-order autocorrelation, the Cochrane-Orcutt has been shown to be as effective as more complex techniques, such as the ARIMA modeling procedure described in the subsequent section (Beck & Katz, 2011). This procedure can be easily executed in SPSS using the Cochrane-Orcutt command (i.e., AREG VAR=dependent with independent/ METHOD=CO) and autocorrelation should be assessed using the Ljung-Box test. However, because this transformation has a negative impact on sample size (i.e., *n-1* for each lag specified in the adjustment), it may not be the most effective approach for time series that contain higher-order autocorrelations.

The ARIMA Model

The ARIMA model (aka Box-Jenkins model) is an iterative and customizable modeling technique that can incorporate autoregressive (AR) terms *"p"*, differencing operations *"d"* (i.e., the integrated component in the model), and moving average (MA) terms *"q"*. This is a superior modeling technique that can account for multiple dynamic processes co-occurring in time series data (Box et al., 1994; Box-Steffensmeier et al., 2014; McLeary & Hay, 1982). For our purposes, we limit our discussion to AR models; that is, ARIMA models (p, d, q) that *only* specify p, the number of terms that describe the dependency among successive observations. For example, an AR model with the specification (0, 0, 0), implies there is no relationship between adjacent observations (hence an OLS regression model), whereas a specification of (1, 0, 0) indicates there is a lag-1relationship (see the Appendix for step-by-step instructions for specifying this ARIMA model in SPSS). ARIMAs are estimated using non-linear maximum likelihood estimation (MLE) that retains the first observation and simultaneously estimates the autocorrelation coefficient and the β estimates; they are considered more robust than OLS regression models that incorporate the Cochrane-Orcutt method (Shumway & Stoffer, 2010).

Method

Sample

The data were collected from end-of-semester course evaluations for 123 courses with a total of 2,648 students (15 undergraduate-level courses and 108 graduate-level courses) taught by the same instructor at a private university in the Northeast. Most of the courses were lecture-based and focused on statistical techniques, measurement, or methodology. The data are updated at the end of each semester and are input in the order that the course was taught over the 33-year span.[9] Most courses have a fixed schedule—for example, the General Linear Models course is taught every Fall, the Psychometrics course is taught every two years in the Spring, and Introductory Statistics was taught every Fall and Spring. Over time, the course offerings moved from general research methods and introductory statistics to higher-level multivariate and psychometrics courses.

Course Evaluations

The number of questions and their wording on each mandated evaluation remained stable, ranging from 28-29 questions. The data can be summarized in four categories: 1) student-level perceptions (e.g., time spent on the course compared to other courses, the extent to which the student understood principles and concepts), 2) administrative characteristics (e.g., year the course was taught, class size, degree level of students), 3) instructor-specific variables (e.g., tenure status, rank, marital status), and 4) instructor evaluation ratings (percent of students in the course who marked either excellent, very good, good, acceptable, and poor). Student-level perceptions are generally recorded as the percent of students in each class who *strongly agreed* with the statement. The instructor variables are coded as categorical variables and the administrative characteristics consist of a combination of continuous and categorical variables. Each class record is associated with a total of 42 variables, that when used in combination, provide a complex picture of the teaching and learning environment within which a class was taught, and can be used to examine a host of pedagogically meaningful relationships (Burns & Ludlow 2006; Chapman & Ludlow 2010; Ludlow & Alvarez-Salvat 2001; Ludlow & Klein, 2014).

Variables

These data were used to test a plausible substantively meaningful model to illustrate issues with autocorrelation that are often encountered in OLS regression analysis of time series. Our model used a continuous variable for class size (*size*) to predict our outcome variable *excell*, which represented the percentage of students who gave a course an overall rating of "excellent".[10] Class sizes ranged from 3-52 (M=21.5), and the percentage of students who rated a course as "excellent" ranged from 5.3% - 100% (M=42.2%). There were no missing data for either of these variables.

Results

Baseline Model

For the first analysis, we estimated a baseline model by regressing the outcome variable *excell* on the predictor variable *size*. As seen in Table 1, this analysis yielded a statistically significant $R^2 = .249$ and a parameter estimate of b = -1.10, indicating that on average, with every one unit increase in *size* (i.e., one student), the value of *excell* (i.e., percent of students who gave the class a rating of "excellent") decreased by 1.10 percent.

Table 1. Model Estimates for Lag-1.

Model	N	df	R^2	$Adj\ R^{2*}$	b	β	F	D_w
Baseline	125	123	0.249*	-	-1.10*	-0.499*	40.17	1.61*
Linear Trend	125	122	0.270*	0.264*	-0.79*	-0.357*	36.76	2.01
Lag-1	124	121	0.276*	0.264*	-1.11*	-0.505*	22.65	-
CO Transf.	124	122	0.270*	-	-1.01*	-0.519*	44.35	-
ARIMA *AR(1)*	125	-	0.273*	-	-1.09*	-	-	32.50*

Note. D_w is *not* provided for models that contain a lagged predictor variable (Q statistic is provided for the overall ARIMA model). * Denotes a statistically significant $R^2, Adj\ R^2, \beta, D_w, or\ Q$ $(p \leq .05)$

Figure 2[11] presents the unstandardized residuals from this model plotted across the chronological teaching sequence of the classes. The solid, flat horizontal line represents the residual value of zero and variation around this line across the full sequence of the classes should be random and uniform. The dashed, positively sloped line represents the regression of the residuals on the class sequence. The corresponding fit of this line (captured by R^2 =.15) clearly demonstrates that systematic time-related variation remains in the residuals.

The Durbin-Watson tested for autocorrelation, and at $\alpha = .05$, the test statistic was below the lower-bound threshold of the confident interval (D_W= 1.61; CI=1.69_{dL}-1.72_{dU}; k=2; n=123), indicating that the lag-1 autocorrelation in the data was statistically significant. Next, using the ACF function, we explored the values and statistical significance of the autocorrelation coefficients for the first 16 lags in the dataset. This analysis revealed statistically significant autocorrelation coefficients for lag-1, and lags-3-16, as well as a marginally significant autocorrelation for lag-2 (see Table 2). In addition, the ACF plot in Figure 1 highlighted a distinctive characteristic of the autocorrelation pattern; across various lags there was an increase in the autocorrelation coefficient for every 3rd lag, particularly in lag 1, 3, 6, and 9. Interestingly, these time lags of 1, 3, and 6 reflect the semesters in which certain graduate level classes were taught, whereas time lags 2, 4, and 5 represent the semesters in which certain undergraduate courses were taught.

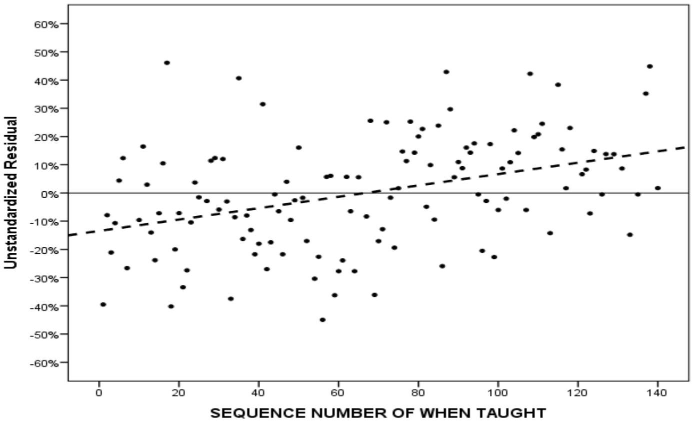

Figure 2. Unstandardized residuals plotted against time (i.e., the sequence order in which courses were taught).

Table 2. ACF Comparison for Each Model (Post Lag-1 Adjustment)

Model	*Lag-1*	2	*3*	4	5	6	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
Baseline	.178*	.069†	.175*	.218*	.164*	.207*	.115*	-.005*	.219*	.188*	.026*	.102*	.159*	.100*	.178*	.094*
Linear	-.011	-.110	.109	.082	.004	.156	-.050	-.172	.139	.034	-.116	.001	.023	-.027	.083	-.045
Lag-1	.038	-.004	.153	.170	.099	.183*	.077†	-.061†	.204*	.151*	-.034*	.068*	.134*	.080*	.122*	.036*
CO	-.012	.024	.172*	.115	.134	.201*	.055†	-.081†	.147*	.148*	.041*	.103*	.074*	.011†	.270*	-.044*
ARIMA^{1}	-.012	.008	.136*	.169	.094	.173†	.084†	-.068	.202*	.154*	-.026*	.076*	.131*	.044*	.155*	.037*

Note. Includes baseline model for comparison. * Denotes a statistically significant lag ($p \leq .05$)
† Denotes a marginally significant lag ($p \leq .10$)

Lag-1 Adjustment

In this section, we report the results from three sets of analyses using each of the procedures. For the first round of analyses, we only adjusted for a lag-1 autocorrelation.

Modeling and controlling for a linear time trend. Two models were estimated using the variable *sequence* (i.e., the order in which classes were taught). The first model was estimated using the variable *sequence* as a covariate when regressing *excell* on *size*, and the second model used the *sequence* variable to detrend the original *excell* and *size* variables. The *detrended excell* was regressed on the *detrended size* in lieu of the original variables (see *linear trend* and *detrended* models in Table 1). These analyses produced a b= -.787 for *size* and an $R^2_{Linear\ trend}$= 0.270, as well as a nonsignificant D_w= 2.01—which indicated the lag-1 autocorrelation had been successfully reduced to nonsignificance (see Table 1). As shown in Table 2, the autocorrelation coefficients for lags 1-16 were no longer statistically significant ($p > 0.05$).[12]

Lag-1 covariate. Next, we estimated a model that included a lagged covariate as a predictor. To do this, we computed a lag-1 version of our outcome variable $excell_{lag_1}$ and included this new variable as a covariate along with *size* in the regression. This produced an R^2= .276 and b= -1.11 (see Table 1). The autocorrelation was assessed using the Ljung-Box statistic provided in the ACF table, which showed that lags 1-5 had been reduced to nonsignificance. However, lag-6 and lags 9-16 remained significant ($p < 0.05$), and lags 7 and 8 remained marginally significant ($p < 0.10$; see Table 2).

Cochrane-Orcutt lag-1. In the next model, we used the Cochrane-Orcutt transformation to compute adjusted variables for *excell* and *size* variables. The new variables were computed following the steps outlined next. First, we computed a lag-1 version of *size* ($predictor_{lag_1}$), and then used the original variables, the lagged variables, and the ACF coefficient for the lag-1 autocorrelation $\rho = .178$ (see Table 1), to compute the new $predictor_{adjusted} = predictor - (\rho * predictor_{lag_1})$ and $outcome_{adjusted} = outcome - (\rho * outcome_{lag_1})$ variables.[13] Using these adjusted variables, we then regressed $excell_{adjusted}$ on $size_{adjusted}$. This analysis yielded an R^2 = .270 and b = -1.01 (see Table 1). The Box-Ljung Q tested the autocorrelations for lags 1-16. The ACF indicated that the autocorrelations for lags 1-5 and 8 were reduced to nonsignificance. However, the correlations for lags 9-16 remained statistically significant, and the correlations for lags 6 and 7 remained marginally significant (see Table 2).

ARIMA (1, 0, 0). This ARIMA model included a lag-1 autoregressive process AR(1), the other two parameters were set to 0, and included the original (i.e. unadjusted) *size* and *excell* variables as the predictor and outcome. Our specification included goodness of fit measures and parameter estimates (including R^2 and b), as well as the Box-Ljung Q and the ACF coefficients. This analysis yielded an R^2 = .273 and b = -1.09 (see Table 1). As seen in Table 2, the autocorrelation coefficients for lags 1-5 and 8 are no longer significant, but the correlations for lags 9-16 remained statistically significant and the correlations for lags 6 and 7 remained marginally significant. As a result, the Box-Ljung was statistically significant (see Table 1).

Discussion. In the previous sections, we reported the results of four models estimated using various methods to adjust for the presence of lag-1 autocorrelation.[14] Overall, the four analytic strategies showed varying degrees of efficacy in reducing the autocorrelation. As seen in Table 2, the *linear time trend* method was the only method that reduced the autocorrelations for lags 1-16 to nonsignificance. In comparison, the other three models estimated using a *lagged covariate*, the *CO adjustment*, and the *ARIMA model*, continued to show statistically significant autocorrelations at varying degrees. Regarding the parameter estimates yielded in these analyses, relative to the baseline model, the linear time trend model showed the largest reduction in the magnitude of the b coefficient, whereas the other three models yielded coefficient values similar to that of the baseline model.

The results underscore a couple of important points worth discussing. First, it is likely that the higher-order autocorrelations still present in the data for the models that included the lagged covariate, the Cochrane-Orcutt adjustment, and the ARIMA, may be biasing the parameter estimates, which would explain why they were similar in value to the original baseline model. The results for the linear trend model support this idea because they showed that eliminating all of the significant autocorrelations in the data weakened the association between course excellence ratings (*excell*) and class size (*size*), possibly by reducing the bias introduced by the autocorrelations.

That being said, it is also important to note that modeling a linear trend may not have been the most appropriate method for adjusting the autocorrelation in our data. When controlling for a linear time trend, the shared variance between the residuals is assumed to be progressively increasing at a constant rate. However, because our data showed both positive and negative higher-order autocorrelations, controlling for a linear time trend does not sufficiently address the pattern of the autocorrelation in our data. Moreover, the other three strategies used in this series of analyses were specified to only adjust for a lag-1 autocorrelation of $\rho = .178$, and were therefore limited in the extent to which they could reduce the strength of the higher-order correlations that were well above this value. Given that these procedures can be customized, we conducted another round of analyses using these methods to adjust for the statistically significant lag-3 autocorrelation still present in the data.

Lag-3 adjustment

In the following sections, we report the results of the lag-3 adjustments and describe procedural differences associated with these analyses.

Lag-3 covariate. We computed a new version of our outcome as a lag-3 variable, $excell_{lag_3}$, and included this as a covariate along with *size* in the regression.[15] As shown in Table 3, this yielded an R^2= .306 and b= -.873. In addition, although the autocorrelation coefficient for lags 1-3 were reduced to non-significance, the coefficients for lags 5-7 and 9-10 remained statistically significant ($p \leq .05$), while lags 4, 8, and 11-16 remained marginally significant ($p \leq .10$; see Table 4).

Table 3. Model Estimates for Lag-3

Model	N	df	R^2	$Adj\ R^2$	b	β	F	D_w
Baseline	125	123	0.249*	-	-1.10*	-.499*	40.17	1.61*
Linear Trend	125	122	0.270*	0.264*	-.787*	-.357*	36.76	2.01
Lag-3	122	119	0.306*	0.294*	-.873*	-.398*	25.82	-
CO Lag-3	122	119	0.186*		-.927*	-.431*	26.90	
ARIMA 3, 1	125		0.301*		-.898*			23.78†

Note. Includes baseline and linear trend models for comparison. D_w is *not* provided for models that contain a lagged predictor variable (Q statistic is provided for the overall ARIMA model).
* Denotes a statistically significant $R^2, Adj\ R^2, \beta, D_w, or\ Q$ $(p \leq .05)$

Table 4. *ACF Comparison for Each Model (Post Lag-3 Adjustment)*

Model	*Lag-1*	2	*3*	4	5	6	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
Baseline	.178*	.069†	.175*	.218*	.164*	.207*	.115*	-.005*	.219*	.188*	.026*	.102*	.159*	.100*	.178*	.094*
Linear	-.011	-.110	.109	.082	.004	.156	-.050	-.172	.139	.034	-.116	.001	.023	-.027	.083	-.045
Lag-3	.089	-.024	-.128	.206†	.180*	.116*	.005*	-.031†	.157*	.105*	-.009†	.014†	.122†	.073†	.133†	-.006†
COLag-3	.126	.009	.016	.196	.169†	.183*	.052*	-.022†	.197*	.138*	.013*	.076*	.135*	.108*	.163*	.030*
ARIMA^{3,1}	-.050	.010	-.081	.137	.096	.146	.007	-.091	.172	.102	-.011	.016	.087	.021	.183	.014

Note. Includes baseline and linear trend models for comparison. * Denotes a statistically significant lag ($p \leq .05$). † Denotes a marginally significant lag ($p \leq .10$)

CO lag-3. In the second iteration of the CO method, we computed a new set of adjusted variables using the autocorrelation coefficient for lag-3 ($\rho = .172$) and then followed the same steps outlined previously. Next, we regressed the new $excell_{lag3}$ on the new $size_{lag3}$ variable, which yielded an R^2= .187 and $b =$ -.929 (see Table 3). As seen in Table 4, the autocorrelations for lags 1-4 were reduced to non-significance; however, several correlations remained statistically significant (i.e., lags 6-16) and marginally significant (i.e., lag 5).

ARIMA lag-3. The second ARIMA model was specified with an AR(3). There is one important difference to note between the specification of ARIMA models with higher-order versus first-order AR processes, such that ARIMA models with higher-order AR processes should be specified to include any previous lags that are of interest. In this instance, we estimated a model that built on the previous analysis and adjusted for both lag-1 and lag-3, with a syntax specification of [3;1, 0, 0].[16,17] This analysis yielded an R^2 = .301 and b = -.898, and showed that the autocorrelation coefficients for lags 1-16 had been reduced to nonsignificance (see Tables 3 and 4).

Discussion. The analyses reported in the previous sections addressed the lag-3 autocorrelation that remained statistically significant in three out of the four analyses previously reported. Relative to those analyses, all three strategies showed greater efficacy in reducing the autocorrelation in the data. As a result, the regression coefficients yielded for *size* in all three models were smaller than those in the previous set of analyses, which further supports our prior argument, that the strength of the association between class size and students' excellence ratings in the baseline model was biased upward, as a result of the autocorrelation. However, given the statistically significant lag-6 autocorrelation still remaining in two of the three models, we conducted one final round of analyses aimed at adjusting for the lag-6.

Lag-6 adjustment

The series of analyses were conducted to address the lag-6 autocorrelation. When appropriate, we also describe procedural differences required for a specific analysis.

Lag-6 covariate. We computed a lagged version of *excell* as a lag-6 variable, $excell_{lag_6}$, and included this as a covariate along with *size* in the regression analysis. As shown in Tables 5 and 6, this yielded an R^2= .360 and b= -.860, and reduced the autocorrelation coefficients for lags 1-16 to nonsignificance.

Table 5. Model Estimates for Lag-6

Model	N	df	R^2	$Adj\ R^{2*}$	b	β	F	D_w
Baseline	125	123	0.249*	-	-1.10*	-.499*	40.17	1.61*
Linear Trend	125	122	0.270*	0.264*	-.787*	-.357*	36.76	2.01
Lag-6	119	116	0.360*	0.349*	-.860*	-.384*	32.10	-
CO Lag-6	119	117	0.161*	-	-848*	-.401*	22.10	-
ARIMA 6, 3, 1	125		0.340*		-.656*			18.62

Note. Includes baseline and linear trend models for comparison. D_w is *not* provided for models that contain a lagged predictor variable (Q statistic is provided for the overall ARIMA model).
* Denotes a statistically significant $R^2, Adj\ R^2, \beta, D_w, or\ Q$ ($p \leq .05$)

Table 6. ACF Comparison for Each Model (Post Lag-6 Adjustment)

Model	*Lag-1*	2	*3*	4	5	6	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
Baseline	.178*	.069†	.175*	.218*	.164*	.207*	.115*	-.005*	.219*	.188*	.026*	.102*	.159*	.100*	.178*	.094*
Linear	-.011	-.110	.109	.082	.004	.156	-.050	-.172	.139	.034	-.116	.001	.023	-.027	.083	-.045
Lag-6	.080	-.002	.021	.156	.156	-.097	.018	-.086	.079	.078	-.045	.024	.090	.084	.116	-.021
COLag-6	.121	.025	.141	.188†	.156*	.073†	.060†	-.035	.178*	.131*	.011†	.102†	.141*	.124*	.142*	.025*
$ARIMA^{6,3,1}$	-.042	-.031	-.080	.125	.107	-.049	-.041	-.121	.131	.057	-.056	-.001	.063	.032	.156	-.041

Note. Includes baseline and linear trend models for comparison.
* Denotes a statistically significant lag ($p \leq .05$). † Denotes a marginally significant lag ($p \leq .10$)

CO lag-6. For the third iteration of the CO method, we computed a new set of adjusted variables and used the autocorrelation coefficient for the lag-6 ($\rho = .207$) from the original baseline model, in lieu of the coefficient for the lag-3. Next, we regressed the new $excell_{lag6}$ on the new $size_{lag6}$ variable, which yielded an R^2= .161 and $b =$ -.848 (see Table 5). As seen in Table 6, with the exception of lags 1-3, the remaining autocorrelation coefficients were either statistically significant at $p \leq .05$ (lags 5, 9-10, and 13-16) or $p \leq .10$ (lags 4, 6-7, and 11-12).

ARIMA lag-6. The next ARIMA model was specified with an AR(6). As shown in Table 5, this yielded R^2 = .340 and b = -.656. In addition, the ACF analysis showed that all of the autocorrelations for lags 1-16 had been reduced to nonsignificance (see Table 6).

Discussion. The results of the third round of analyses yielded several interesting findings. First, the Cochrane-Orcutt transformation for the lag-6 was the least effective of all three methods in reducing autocorrelation in these data. Second, the models estimated with the lag-6 covariate predictor and the ARIMA lag-6 model reduced the autocorrelations for lags 1-16 to nonsignificance. However, the parameter estimate for the ARIMA lag-6 model ($i.e., b = -.656$) was considerably weaker than the estimate from the lag-6 covariate model ($i.e., b = -.860$). One potential explanation for this difference in estimates is that the R^2 of the ARIMA indicated that this model had greater explanatory power than the lag-6 covariate model (.360 *v.* .306, respectively). Thus, the weaker ARIMA estimate but greater R^2 is likely due to a greater reduction in the bias due to the correlated residuals. The broader implications of these differences in parameter estimates are discussed further in the next section.

In general, the overall reductions in autocorrelation observed after these analyses indicated that no further adjustments were warranted. Based on these results, the ARIMA technique was the most effective of the four procedures in reducing the autocorrelation in our data (see Figure 3) and increasing the explained variance of the model. This is likely because the estimation process used in ARIMA modeling (i.e., Maximum Likelihood Estimation) benefits from a full sample size and the additional data variability contributed by the first cases in the time series records. In contrast, the Cochrane-Orcutt method reduced the effective sample size and eliminated potentially influential observations from the earliest records in the data sequence. Although the impact on sample size may be less concerning for very large datasets and/or those containing only first-order autocorrelations, the removal of influential points can negatively impact the results. This was evident in our example, where the loss of the first six observations led to both a decrease in the explained variance of the model and parameter estimates that were biased upwards.

Conclusions

The goal of this paper was to employ a relatively simple set of time series analysis procedures to illustrate different adjustments to handle the autocorrelation in the residuals that is characteristic of longitudinal data. We provided background information related to the nature of correlated residuals in time series data, as well as diagnosis procedures and subsequent addressing of autocorrelation. We then employed 25 years of college course evaluation data to illustrate the differences in procedures and results for four strategies commonly used to address autocorrelation.

Overall, the results highlight several analytic points. First, when working with time series data, it is imperative to probe for autocorrelation in a comprehensive manner that includes both an evaluation of the residual assumptions and statistical testing of the parameter estimates. This two-step process can help researchers identify time-dependent patterns in their residuals and facilitate the process of choosing an analytic strategy that is appropriate for both the magnitude and pattern of the autocorrelation in their data. Moreover, as demonstrated, when left uncorrected, autocorrelation can have a considerable impact on the accuracy of statistical estimates.

Second, the results suggest that addressing autocorrelation should be considered an iterative process that may require more than one round of adjustments. At the very least, our example underscores the importance of testing for autocorrelation even after an adjustment has been made to the data or model, because it may take more than one iteration of a procedure to successfully address the full impact of autocorrelation when it is present.

Third, because most analytic methods used to address autocorrelation differ in the types of adjustments they make, each procedure has a distinct impact on the results. This is important for two

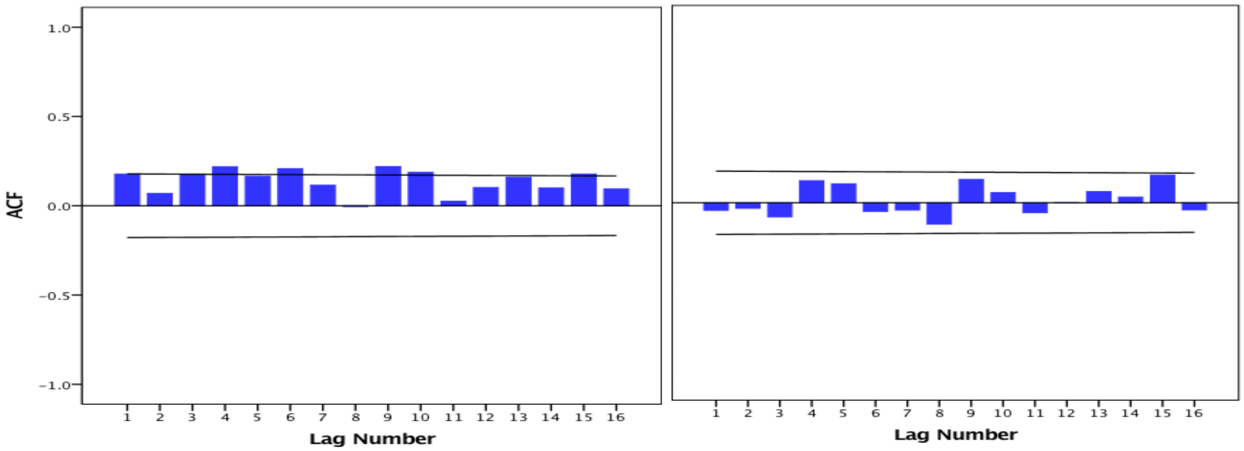

Figure 3. Comparison of autocorrelation coefficient plots for lags 1-16: Baseline model (left) v. ARIMA lag-6 model right). Blue bars represent the coefficients. Lines are Upper and Lower Confidence Limits

reasons. One, certain strategies may be better suited for certain types of autocorrelational patterns than others. As our example illustrated, given the higher-order autocorrelations in our data, estimating and controlling for a linear time trend would not have been the most appropriate strategy to address these autocorrelations. Two, considering that strategies vary in how they model the data, the resulting estimates can differ quite substantially from one strategy to another. Our advice to researchers would be to avoid using any one analytic method as a "one size fits all" solution, and to instead, rely on the unique characteristics of the data and research questions to guide the choice of analytic strategy.

Finally, even though this last point is specific to the present data context, the presence of a lag-1 autocorrelation in the time series for one instructor's professional teaching career is not particularly surprising. This is because the autocorrelation reflects (a) teaching practices that carry over from one class to the next (regardless of the content) and (b) students take multiple courses with this instructor and their perceptions about practice carry with them into the next course(s). It is somewhat more surprising, however, that these data contained lag-3 and lag-6 autocorrelations. These findings were initially considered "noise" until it was realized that they reflected a cyclical pattern due to when certain courses have been taught over the past 20 or so years. Some courses are offered only every third semester (e.g. fall, spring, summer, then fall again, spring again, and so on), others are offered every 6th semester (on a two-year cycle). Those patterns, similar to the lag-1 pattern, reflect an effect due to carried-over student perceptions and experiences, carried-over instructor teaching practices, and the very specialized content of those cycled courses. The point here is that in the analysis of an instructor's teaching record (using student ratings of instruction) it is not sufficient to focus on the immediate courses taught during a given academic review year—it is necessary to take into account what the specific courses were that were taught and when they were last taught in order to frame an appropriate baseline for evaluation and analysis.

References

Afzal, M., Gagnon, A. S., & Mansell, M. G. (2015). Changes in the variability and periodicity of precipitation in Scotland. *Theoretical and applied climatology*, *119*, 135-159.

Beck, N., & Katz, J. N. (2011). Modeling dynamics in time-series–cross-section political economy data. *Annual Review of Political Science*, *14*, 331-352.

Blattberg, R. C. (1973). Evaluation of the power of the Durbin-Watson statistic for non-first order serial correlation alternatives. *The Review of Economics and Statistics*, *75*, 508-515.

Bloom, H. S. (2003). Using "short" interrupted time-series analysis to measure the impacts of whole-school reforms: With applications to a study of accelerated schools. *Evaluation Review*, *27*, 3-49.

Box, G., Jenkins, G., & Reinsel, G. (1994). *Time series analysis, forecasting and control* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Box-Steffensmeier, J. M., Freeman, J. R., Hitt, M. P., & Pevehouse, J. C. (2014). *Time series analysis for the social sciences*. New York: Cambridge University Press.

Brillinger, D. R. (2001). Time series: General (N. Smelser & P. Baltes, Eds.). In *International encyclopedia of the social & behavioral sciences* (1st ed., Vol. 23, pp. 15724-15731). Oxford: Elsevier.

Brown, B. G., Katz, R. W., & Murphy, A. H. (1984). Time series models to simulate and forecast wind speed and wind power. *Journal of Climate and Applied Meteorology*, *23*, 1184-1195.

Burns S., & Ludlow, L. H. (2006). Understanding student evaluations of teaching quality: The unique contributions of class attendance. *Journal of Personnel Evaluation in Education*, *18,* 127-138.

Chao, R. Y. (2012). Intra-nationalization of higher education: The Hong Kong case. *Frontiers of Education in China, 7,* 508-533.

Chapman, L., & Ludlow, L. H. (2010). Can downsizing college class sizes augment student outcomes: An investigation of the effects of class size on student learning. *Journal of General Education*, 59, 105-123.

Chatterjee, S., & Price, B. (1991). *Regression analysis by example* (2nd ed.), New York: Wiley & Sons.

Clyde, L. A., & Klobas, J. E. (2001). The first internet course: Implications of increased prior participant experience. *Internet Research, 11*, 235-245.

Cochrane, D., & Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto-correlated residual terms. *Journal of the American Statistical Association*, *44,* 32–61.

Durbin, J., & Watson, G. (1950). Testing for serial correlation in least squares regression: I. *Biometrika, 37,* 409-428.

Durbin, J., & Watson, G. (1951). Testing for serial correlation in least squares regression: II. *Biometrika, 38,* 159-177.

Edwards, M., & Richardson, A. J. (2004). Impact of climate change on marine pelagic phenology and trophic mismatch. *Nature, 430,* 881-4.

Erdem, M., & Ucar, I. H. (2013). Learning organization perceptions in elementary education in terms of teachers and the effect of learning organization on organizational commitment. *Educational Sciences: Theory and Practice, 13,* 1527-1534.

Huitema, B. E., & McKean, J. W. (2007). Identifying autocorrelation generated by various error processes in interrupted time-series regression designs: A comparison of AR1 and portmanteau tests. *Educational and Psychological Measurement*, *67,* 447-459.

Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: Principles and practice*. Melbourne, Australia: OTexts.

Jackman, R., & Layard, R. (1991). Does long-term unemployment reduce a person’s chance of a job: A time series test. *Economica, 58*, 93-106.

Loomis, D., Castillejos, M., Gold, D., McDonnell, W., & Borja-Aburto, V. (1999). Air pollution and infant mortality in Mexico City. *Epidemiology, 10,* 118-123.

Ludlow, L. H., & Alvarez-Salvat, R. (2001). Spillover in the academy: Marriage stability and faculty evaluations. *Journal of Personnel Evaluation in Education*, *15*, 111-119.

Ludlow, L. H., & Klein, K. (2014). Suppressor variables: The difference between “is” versus “acting as”. *Journal of Statistics Education, 22,* 1-28.

MacSuga-Gage, A., & Gage, N. A. (2015). Student-level effects of increased teacher-directed opportunities to respond. *Journal of Behavioral Education, 24,* 273-288.

McCleary, R., & Hay, R. (1982). *Applied time series analysis for the social sciences* (2nd ed.). Beverly Hills, CA: Sage Publications.

Marston, D. (1988). The effectiveness of special education: A time series analysis of reading performance in regular and special education settings. *Journal of Special Education, 21*, 13-26.

Nicholich, M. J., & Weinstein, C. S. (1981). The use of time series analysis and "t" tests withserially correlated data tests. *The Journal of Experimental Education, 50,* 25.

Pindyck, R., & Rubinfeld, D. L. (1991). *Econometric models and economic forecasts* (3rd ed.). New York: McGraw-Hill.

Shahbaz, S., Nanthakumar, L., Rashid, S., & Talat, A. (2015). The effect of urbanization, affluence and trade openness on energy consumption: A time series analysis in Malaysia. *Renewable and Sustainable Energy Reviews, 47,* 683-693.

Shumway, R. H., & Stoffer, D. S. (2010). *Time series analysis and its applications: With R examples* (3rd ed.). New York: Springer Science & Business Media.

Tsay, R. (2000). Time series and forecasting: Brief history and future research. *Journal of the American Statistical Association, 95,* 638-643.

Umble, K., Cervero, R., Yang, B., & Atkinson, W. (2000). Effects of traditional classroom and distance continuing education: A theory-driven evaluation of a vaccine-preventable diseases course. *American Journal of Public Health, 90,* 1218-24.

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (6th ed.). Boston: Cengage Learning.

Zinde-Walsh, V., & Galbraith, J. W. (1991). Estimation of a linear regression model with stationary ARMA (p, q) errors. *Journal of Econometrics, 47*, 333-357.

Send correspondence to: Larry Ludlow
Boston College
Email: ludlow@bc.edu

Endnotes

1. It is important to note that our interest was in comparing the manner in which each of the procedures addressed autocorrelation and parameter estimates—and *not* in specifying the ideal model for our data.
2. Assuming the model does not have issues with multicollinearity and that it has been properly specified. (i.e., includes all relevant predictor variables).
3. The critical values tables for the Durbin-Watson can be found here: web.stanford.edu/~clint/bench/dwcrit.htm.
4. A D_w > the CI_{dL} but <CI_{dU} is considered an inconclusive test result.
5. This is not to say that the adjustment procedures described are exclusively applicable to OLS regression, but that we are focusing on these applications for this paper.
6. These instructions are specifically for SPSS v. 22.0.0.1 and may vary for other versions of this software.
7. It is also important to note that we chose these procedures because they would be appropriate for researchers with limited statistical training, and not based on how appropriate they were for our data.
8. Here we refer to the Cochrane-Orcutt adjustment and methods that use lagged variables, although the specific number of cases dropped from analyses is determined by the *lag(j)* specified in the model.
9. The courses in the data were taught between 1984-2017.
10. Each course was rated on a Likert-scale from 1= “Poor” to 5= “Excellent”.
11. The unstandardized residuals yielded by regressing *x* on *y* should be plotted (*y*-axis) against the sequence in which the measurements in the data were taken (*x*-axis).
12. The parameter estimates and the ACF coefficients for lags 1-16 were identical for both detrended models.
13. There was no need to compute an additional lag-1 outcome variable, because we had already done so for the previous model and it was saved in our dataset.
14. The *detrended* model is only discussed in the first section of the lag-1 adjustment results, as it was only estimated as a means of demonstrating an alternative method to modeling.
15. For SPSS v. 22 and higher, this can be executed using the same syntax command described in the description of this method by changing the lag specification from a 1 to a 3.
16. We recommend reviewing the syntax prior to running this analysis, as the default specification for higher-order AR processes vary across different versions of SPSS.
17. This model was specified to include the same goodness of fit measures, parameter estimates, and statistics as the prior ARIMA (1, 0, 0).

APPENDIX

Procedure for Specifying ARIMA Model with an Autoregressive Order of (*1*) in SPSS

1) ANALYZE → FORECASTING → CREATE MODELS → METHOD →ARIMA →SET CRITERIA TO (1,0,0).
2) VARIABLES TAB → SELECT OUTCOME AN PREDICTOR(S).
3) STATISTICS TAB → SELECT DISPLAY FIT MEASURES, R2, PARAMETER ESTIMATES, RESIDUAL AUTOCORRELATION FUNCTION FOR INDIVIDUAL MODELS.
4) PLOTS TAB → SELECT RESIDUAL AUTOCORRELATION FUNCTION FOR INDIVIDUAL MODELS.
5) OUTPUT FILTER TAB → SELECT INDLUDE OUTPUTS FOR ALL MODELS.
6) SAVE TAB → SELECT NOISE RESIDUALS → OK → RUN SELECTION.