11-1-2009

Detecting Lag-One Autocorrelation in Interrupted Time Series Experiments with Small Datasets

Clare Riviello
University of Texas at Austin, clareriviello@gmail.com

S. Natasha Beretvas
The University of Texas at Austin, tasha.beretvas@mail.utexas.edu

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Detecting Lag-One Autocorrelation in Interrupted Time Series Experiments with Small Datasets

Clare Riviello S. Natasha Beretvas
University of Texas at Austin

The power and type I error rates of eight indices for lag-one autocorrelation detection were assessed for interrupted time series experiments (ITSEs) with small numbers of data points. Performance of Huitema and McKean's (2000) z_{HM} statistic was modified and compared with the z_{HM}, five information criteria and the Durbin-Watson statistic.

Key words: Autocorrelation, information criteria, type I error, power.

Introduction

Educational research contains many examples of single-subject designs (Huitema, McKean, & McKnight, 1999). Single-subject designs, also known as interrupted time series experiments (ITSEs), are typically used to assess a treatment's effect on special populations such as children with autism or developmental disabilities (Tawney & Gast, 1984). The design consists of repeated measures on an outcome for an individual during baseline and treatment conditions (A and B phases, respectively). Use of repeated measures on an individual is designed such that the subject acts as his/her own control; this also helps rule out the possible influence of potential threats to validity including history, practice, and maturation effects.

 With ITSE data, the pattern of scores over time is compared for the A (baseline) versus the B (treatment) phases. The comparison

can lead to inferences about the effect of introducing the treatment on the trend in the outcome scores. To describe the change in trend, the effect on the level of the scores and on the possible growth pattern must be assessed. Numerical descriptors of these trends are not well estimated given the number of repeated measures is as small as is commonly found in educational single-case design research (Busk & Marascuilo, 1988; Huitema, 1985). One of the sources of these estimation problems is related to the autocorrelated structure inherent in such designs (Huitema & McKean, 1991; White, 1961; Kendall, 1954; Marriott & Pope, 1954).

 Several test statistics and indices recommended for identifying potential autocorrelation exist. Unfortunately these statistics are typically recommended only for datasets with a larger numbers of data points than are typically encountered with ITSEs. Huitema and McKean (2000) introduced a test statistic, z_{HM}, to identify lag-one autocorrelation in small datasets. The Type I error rate of the z_{HM} was within nominal levels and sufficient power was associated with this statistic. The current study introduces a modification of the z_{HM} designed to enhance further its statistical power. This study assesses the Type I error rate and power of both versions of the z_{HM}. The performance of the two z_{HM} statistics is also compared with that of other test statistics and indices that are commonly used to identify autocorrelated residuals for models used to summarize trends for small ITSE datasets.

Clare Riviello is an Engineering Scientist at Applied Research Laboratories. Email: clareriviello@gmail.com. Natasha Beretvas is an Associate Professor and chair of Quantitative Methods in the Department of Educational Psychology. Her interests are in multilevel and meta-analytic modeling techniques. Email: tasha.beretvas@mail.utexas.edu.

Autocorrelation

One of the fundamental assumptions when using ordinary least squares estimation for multiple regression is that errors are independent. When the independence assumption does not hold, this leads to inaccurate tests of the partial regression coefficients (Huitema and McKean, 2000). For data consisting of repeated measures on an individual, it is likely that a model can explain some but not all of the autocorrelation. In addition, when the residuals in a regression model are autocorrelated the model must account for this to ensure accurate and precise estimation of parameters and standard errors. Thus, it is important to be able to detect autocorrelation so that the proper methods for estimating the regression model can be employed.

This study is designed to focus solely on first-order (lag-one) autocorrelation. For a multiple regression model including k predictors, x_i, of outcome y at time t using:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + ...\beta_k x_{kt} + \varepsilon_t . \quad (1)$$

If there is a lag-one autocorrelation, ρ_1, between residuals, then ε_t, the residual at time t, is related to ε_{t-1}, the residual at time $t-1$ as follows:

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + a_t \quad (2)$$

where ρ_1 is the autocorrelation between residuals separated by one time period. It is assumed that ε_t and ε_{t-1} have the same variance, and a_t is assumed to follow a standard normal distribution.

Estimating Lag-One Autocorrelation

Several formulas are available for estimating the lag-one correlation coefficient, ρ_1, for a time series consisting of N data points. The conventional estimator is calculated as follows:

$$r_1 = \frac{\sum_{t=1}^{N-1}(Y_t - \overline{Y})(Y_{t+1} - \overline{Y})}{\sum_{t=1}^{N}(Y_t - \overline{Y})^2} \quad (3)$$

where $\overline{Y}$ is the simple average of the N values of y. Unfortunately, as evidenced by its common usage, the bias of r_1 is often ignored. The expected value of a lag-1 autocorrelation coefficient for a series consisting of N data points was analytically derived by Marriott and Pope (1954) to be:

$$E(r_1) = \rho_1 - \frac{1}{N}(1 + 3\rho_1) + O(N^{-2}) . \quad (4)$$

It should be noted that the expression in Equation 2 only covers terms to order N^{-1} [thus, the term: $O(N^{-2})$]; there are additional terms for higher orders of the inverse of N. For large samples, these higher order terms tend towards zero. However, the ITSEs of interest in this study tend to involve short series where N is reasonably small and these higher order terms are thus not as negligible. Bias clearly exists in the estimation of the autocorrelation.

Huitema and McKean (1991) listed four additional, fairly common estimators designed to reduce the bias observed in r_1. However, each of these is also highly biased for small data sets. Huitema and McKean (1991) suggested correcting for the bias in r_1 by using

$$r_1^+ = r_1 + \frac{1}{N} \quad (5)$$

which, for smaller true values of ρ_1 incorporates some of the noted bias evident in Equation 2. The authors showed that their modified estimator, r_1^+, is unbiased when ρ_1 equaled zero even for sample sizes as small as $N = 6$. Additionally, the authors found that the bias was lower for positive values of ρ_1 but higher for some negative values.

When estimating the autocorrelation, it is also necessary to calculate the error variance

of the estimator because the estimator and its variance can be combined to produce a statistic that can be used to statistically test for the autocorrelation. Bartlett (1946) derived his variance formula for the variance of r_1:

$$\sigma^2_{r_1} = \frac{1-\rho_1^2}{N}.\qquad(6)$$

by ignoring terms of order N^{-2} or higher. This formula is commonly reduced to:

$$\hat{\sigma}^2_{r_1} = \frac{1}{N}\qquad(7)$$

under the assumption of the null hypothesis that $\rho_1 = 0$ (Huitema & McKean, 1991). Huitema and McKean (1991) asserted that the commonly used Bartlett variance approximation is not satisfactory for small sample sizes. Their simulation study indicated that $\hat{\sigma}^2_{r_1}$ (see Equation 7) consistently overestimated the empirical variance. This overestimation performed quite badly for values of N of less than twenty with Bartlett's variance approximation exceeding the empirical variance by 83% and 40% for $N = 6$ and $N = 10$, respectively. The authors explored the performance of Moran's variance estimate:

$$\hat{\sigma}^{2*}_{r_1} = \frac{(N-2)^2}{N^2(N-1)}\qquad(8)$$

which, under the null hypothesis ($\rho_1 = 0$), gives precise error variance estimates. After looking at the performance of an autocorrelation test statistic using $\hat{\sigma}^{2*}_{r_1}$ as the error variance estimator, the authors concluded that $\hat{\sigma}^{2*}_{r_1}$ was not adequate for small sample sizes. In tests for positive values of autocorrelation, its results were too conservative except for large values of N. They recommended using:

$$\hat{\sigma}^{2+}_{r_1} = \frac{(N-2)^2}{N^2(N-1)}\{1-[E(r_1)]^2\}\qquad(9)$$

where

$$E(r_1) \cong \rho_1 - \frac{1}{N}(1+3\rho_1).\qquad(10)$$

(Marriot & Pope, 1954) as follows from Equation 4. Use of Equation 9 yielded values close to the empirical values of the variance of ρ_1 estimates even for Ns as small as $N = 6$.

Detecting Autocorrelation

The main purpose of estimating the correlation coefficient and calculating its error variance is to detect the presence of autocorrelation in a data set. If data are known to be autocorrelated, then methods other than ordinary least squares should be used to more accurately estimate the regression coefficients and their standard errors. One of the more commonly used tests for autocorrelation in residuals is the Durbin-Watson test statistic:

$$d = \frac{\sum_{t=2}^{N}(\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{N}\varepsilon_t^2}\qquad(11)$$

where ε_t represents the residual at time t (see Equation 2).

The procedure for carrying out this test can be confusing, thus the sequence of steps for testing the non-directional $H_0 : \rho_1 = 0$ is explained here. First both d and $(4-d)$ should be compared with the upper bound d_u. If both exceed this bound, then the null hypothesis is retained; otherwise, both d and $(4-d)$ are compared with the lower bound, d_l. If either falls below d_l, then the null hypothesis is rejected and a non-zero lag one autocorrelation is inferred. If neither d nor $(4-d)$ falls below d_l, the test is inconclusive. The concept of an inconclusive region is unsettling and, although computer methods that provide exact p-values are now becoming available, most are slow or expensive (Huitema & McKean, 2000).

It is in this context, that Huitema and McKean (2000) proposed an alternative test statistic that is simple to compute,

approximately normally distributed and does not have an inconclusive region. The test statistic was evaluated for its use to test residuals from ITSE models that have one to four phases. Huitema and McKean's test statistic is defined as:

$$z_{HM} = \frac{r_1 + \dfrac{P}{N}}{\sqrt{\dfrac{(N-2)^2}{N^2(N-1)}}} \qquad (12)$$

where P is the number of parameters in the time-series regression model and N is the total number of observations in the time series. The authors found that

$$r_{1,P}^+ = r_1 + \frac{P}{N} \qquad (13)$$

provided an unbiased estimate of ρ_1 and that the denominator of the test statistic (in Equation 12) approximates the empirical variance of $r_{1,P}^+$ (see Equation 8).

The z_{HM} test statistic is a generalization of the test proposed in Huitema and McKean's (1991) earlier work was designed for a single-phase model of ITSE data. However, the authors failed to implement all of the suggestions from their previous study. Specifically, the authors did not use the corrected error variance, $\hat{\sigma}_{r_1}^{2+}$, (see Equation 9) that they had recommended. Instead they used $\hat{\sigma}_{r_1}^{2*}$ (see Equation 8). Because $\{1-[E(r_1)]^2\} \leq 1$, use of $\hat{\sigma}_{r_1}^{2+}$ should lead to a smaller variance and thus a larger value of the test statistic and increased power over $\hat{\sigma}_{r_1}^{2*}$.

Information Criteria

As an alternative to using test statistics to detect autocorrelated residuals, it is also possible to estimate a model using ordinary least squares regression, estimate the same model assuming autocorrelated residuals, and then compare the fit of the two models. A post-hoc evaluation that compares the two models' fit can be then be conducted using an information criterion such as Akaike's Information Criterion (AIC):

$$AIC = -2Log(L) + 2k \qquad (14)$$

where L is the value of the likelihood function evaluated for the parameter estimates and k is the number of estimated parameters in a given model. The model with the smallest information criterion value is considered the best fitting model.

As an alternative to the asymptotically efficient but inconsistent AIC, several more consistent model fit statistics have been proposed (Bozdogan, 1987; Hannon & Quinn, 1979; Hurvich &Tsai, 1989; Schwarz, 1978). These include Swartz's (1978) Bayesian criterion:

$$SBC = -2Log(L) + Log(N)k \qquad (15)$$

where N is the number of observations, Hannon and Quinn's (1979) information criterion

$$HQIC = -2Log(L) + 2kLog(Log(N)); \qquad (16)$$

and Bozdogan's (1987) consistent AIC

$$CAIC = -2Log(L) + (Log(N)+1)k. \qquad (17)$$

In addition, Hurvich and Tsai (1989) developed a corrected AIC specifically for small sample sizes, which deals with AIC's tendency to overfit models:

$$AICC = -2Log(L) + \frac{2kN}{N-k-1}. \qquad (18)$$

For each of these information criteria formulations, the smaller the value, the better the model fit.

The AIC and SBC are supplied by default by most statistical software. For example, when using SAS's PROC AUTOREG (SAS Institute Inc., 2003) to estimate an autoregressive model, the procedure also provides results under the assumption of no

autocorrelation in residuals (i.e., using ordinary least squares, OLS, estimation). The procedure automatically provides the AIC and SBC for the OLS and autoregressive models to enable a comparison of the fit of the two models. To date, no studies have been conducted to compare use of information criteria for identification of autocorrelated residuals for ITSE data with small sample sizes.

Research Question

This study is designed to introduce and evaluate use of the variance correction suggested by Huitema and McKean (1991) in a modified version of their test statistic, z_{HM}. Specifically, the corrected test statistic being suggested and evaluated is:

$$z_{HM}^+ = \frac{r_1 + \dfrac{P}{N}}{\{1 - [E(r_1)]^2\}\sqrt{\dfrac{(N-2)^2}{N^2(N-1)}}} \quad (19)$$

Identification of lag-one autocorrelation (of residuals) was compared for the z_{HM}^+ and z_{HM} test statistics, the Durbin-Watson test statistic and the AIC, SBC, HQIC, CAIC, and AICC fit indices for conditions when $\rho_1 = 0$ and when $\rho_1 \neq 0$. This study focused only on two-phase ITSE data. This design lies at the root of commonly used single-subject designs and provides an important starting point for this investigation.

Methodology

SAS code was used to generate data, estimate models, and summarize results (Fan, Felsovalyi, Keenan, & Sivo, 2001). Several design conditions were manipulated to assess their effect on the performance of the test statistics and fit indices. These conditions included the magnitude of the treatment's effect on the level and linear growth, the degree of autocorrelation and the overall sample size of the ITSE data.

Model and Assumptions

The following two-phase, ITSE model (Huitema & McKean, 1991) was used to generate the data:

$$y_t = \beta_0 + \beta_1 t_t + \beta_2 d_t + \beta_3 [t_t - (n_A + 1)] d_t + \varepsilon_t \quad (20)$$

where n_A is the number of data points in the first phase (baseline phase A), d_t is the dummy variable coded with a zero for data points in the baseline phase and with a one for data points in the second phase, and $[t_t - (n_A + 1)]d_t$ is the centered interaction between time and treatment. The interaction term is centered in this way to provide a coefficient, β_3, that represents the treatment's effect on the slope (i.e., the difference in the linear growth between that predicted using the treatment phase data and that predicted using the baseline data). The coefficient, β_2, represents the change in the intercept from the baseline to the treatment phase (specifically, the difference in the value of y_t, when $t = n_A + 1$, predicted using treatment versus baseline phase data).

Thus, the β_2 and β_3 coefficients describe the effect of the treatment on the level and growth in y, respectively. The residuals (ε_t) were generated such that $\varepsilon_t = \rho_1 \varepsilon_{t-1} + a_t$ with ρ_1 being the true lag-one autocorrelation between residuals separated by one time unit, and a_t was randomly and independently selected from a standard normal distribution.

Because the focus in ITSE designs is on the effect of the intervention, the β_2 and β_3 coefficients (see Equation 20) are of most interest. Thus, when generating the data in this simulation study, values of β_0 (baseline data's intercept) and of β_1 (baseline data's linear growth) were not manipulated but were fixed such that β_0 was set to zero and β_1 was set to a value of 0.2 in all scenarios. This modeled data with an intercept of zero (i.e., $y_t = 0$ at $t = 0$) and a slight baseline trend. Values of β_2 and β_3, however, were varied to investigate their effect on detecting autocorrelation. Each parameter took on values 0, 0.2, and 0.4 in this fully crossed design.

In order to evaluate how the model selection criteria performed over the range of possible values for ρ_1, its value was varied to range from -0.8 up to 0.8 in increments of 0.2.

Finally, the number of data points, N, in the two phases for each scenario were varied to be 12, 20, 30, 50, or 100 with the points being divided equally between the two phases so that $n_A = n_B$ with values for each of: 6, 10, 15, 25, or 50.

The simulation study thus entailed a fully crossed design consisting of three values of β_2 crossed with three values of β_3, crossed with nine values of ρ_1, crossed with five values of N for a total of 405 combinations of conditions. One thousand datasets were generated for each of these 405 scenarios.

Analyses

After each dataset was generated, the regression model in Equation 20 was estimated using SAS's PROC AUTOREG. This procedure estimates the model using both ordinary least squares (OLS) (assuming $\rho_1 = 0$) and autoregressive methods (assuming $\rho_1 \neq 0$). The procedure provides values for the AIC and SBC for both models. HQIC, CAIC, and AICC were then calculated (see Equations 16, 17 and 18, respectively) using the log likelihood obtained from the AIC value. For each information criterion, a tally was kept describing when the autoregressive model's information criterion was lower than that of the OLS model. PROC AUTOREG additionally provides the p-value for the Durbin-Watson test statistic. As with the AIC and SBC, a tally was kept of the proportion of trials for which this p-value led to a rejection of the null hypothesis that $\rho_1 = 0$ ($p < .05$).

The z_{HM} and z_{HM}^+ were also calculated (see Equation 12 and 19, respectively) using the residuals from the OLS regression. The $E(r_1)$ in the denominator of Equation 19 was obtained by substituting $r_{1,P}^+$ for the unknown ρ_1 in Equation 6. Again, a tally was kept describing the proportion of trials for which the null hypothesis of no autocorrelation was rejected ($p < .05$). For conditions in which $\rho_1 \neq 0$, the tally by scenario for each of the eight model selection criteria provided the power to identify the correct model. For conditions in which $\rho_1 = 0$, the tally provided the type I error rate.

Results

Type I Error Rates

Table 1 contains Type I error rates by condition and criterion. Sample size appeared to have the strongest effect on type I error rates. The type I error rate was not greatly affected by the values of β_2 and β_3. Overall, the Type I error rates for z_{HM} and z_{HM}^+ were the best of the eight criteria investigated. The rates were somewhat conservative for the smallest sample size conditions ($N = 12$) with values of 0.022 and 0.035 for z_{HM} and z_{HM}^+, respectively. The z_{HM} maintained type I error rates at the nominal level across sample size conditions (with a maximum value of 0.051). The rates for z_{HM}^+ were slightly elevated (with values of 0.059) although the statistic performed much better than did the Durbin-Watson (DW) and the five information criteria (ICs) investigated.

The Type I error rates of the five ICs (SBC, AIC, HQIC, CAIC and AICC) and for the DW statistic were generally inflated across the $\rho_1 = 0$ conditions examined with the indices performing from worst to best as follows: AIC, HQIC, SBC, AICC, DW, CAIC. The Type I error rate inflation, however, decreased with increasing sample size. Only in the scenarios with the largest sample size ($N = 100$), were the CAIC and SBC's Type I error rates acceptable if somewhat conservative. The CAIC's Type I error rate performance was also acceptable (0.056) for conditions in which N was 50.

Power

Table 2 displays the power of the eight criteria used to evaluate the presence of lag-one autocorrelated residuals. In the presence of type I error inflation, the power of a criterion becomes somewhat moot. Thus, it should be kept in mind that the Type I error inflation noted for the DW and the five ICs. As would be expected, for all criteria the power was found to increase for larger sample sizes. Similarly, it was expected and found that as the magnitude of ρ_1 increased so did the power to detect the ρ_1 of the ICs and test statistics. The z_{HM} and z_{HM}^+ exhibited consistently better power levels than the SBC and DW for all positive values of ρ_1.

Table 1: Type I Error Rates (False Detection) of Lag-One Autocorrelation
by Criterion and Condition

Condition		Information Criterion					Test Statistics ($p < .05$)		
Parm*	True Value	SBC	AIC	HQIC	CAIC	AICC	DW	z_{HM}	z_{HM}^+
ρ_1	0	0.185	0.304	0.264	0.129	0.168	0.146	0.043	0.053
β_2	0.4	0.185	0.303	0.265	0.128	0.172	0.143	0.044	0.054
	0.2	0.185	0.303	0.262	0.129	0.167	0.145	0.043	0.053
	0	0.185	0.305	0.264	0.129	0.166	0.149	0.042	0.052
β_3	0.4	0.188	0.305	0.266	0.131	0.172	0.147	0.044	0.055
	0.2	0.187	0.306	0.264	0.129	0.165	0.147	0.043	0.052
	0	0.180	0.300	0.262	0.127	0.168	0.143	0.042	0.051
N	12	0.424	0.490	0.523	0.316	0.131	0.173	0.022	0.035
	20	0.228	0.343	0.316	0.155	0.182	0.164	0.047	0.059
	30	0.146	0.272	0.221	0.092	0.183	0.149	0.047	0.059
	50	0.087	0.225	0.157	0.056	0.178	0.132	0.051	0.058
	100	0.038	0.190	0.103	0.024	0.167	0.110	0.049	0.052

*Parm. = Parameter

Both of these test statistics had better power than all other indices when $\rho_1 \geq 0.6$. These results also supported the theoretical conclusion mentioned earlier that z_{HM}^+ will always have more power than z_{HM}. For negative values of ρ_1, the ICs and DW statistic exhibited better power than the z_{HM} and z_{HM}^+. And the ICs that performed worst in terms of type I error control performed best in terms of power.

The power was also unaffected by the true values of β_2 and β_3. The power of z_{HM} and z_{HM}^+ was quite low (0.089 and 0.133, respectively) for the $N = 12$ conditions but the power levels become more comparable to those of the other criteria for larger N. However, only z_{HM} and z_{HM}^+ had exhibited acceptable type I error rates.

Conclusion

The results of the simulation study support use of the z_{HM} and z_{HM}^+ for identification of lag-one autocorrelation in small ITSE datasets. Both statistics maintain nominal rates of type I error control although z_{HM}^+'s rates seemed slightly inflated in the larger sample size conditions. Concomitant with the type I error control were found somewhat lower empirical power levels. However the type I error inflation of the five ICs and the DW prohibit their use for detection of autocorrelation in the conditions examined here and especially with ITSE data consisting of a small number of data points.

A type I error in the current context means that an autoregressive model will be estimated unnecessarily. While this should have minimal effect on the estimation of the β coefficients in Equation 20, it will likely affect the standard error (*SE*) estimates used to test the

Table 2: Power to Detect Lag-One Autocorrelation by Criterion and Condition

Condition		Information Criterion					Test Statistic ($p < .05$)		
Parm*	True Value	SBC	AIC	HQIC	CAIC	AICC	DW	z_{HM}	z_{HM}^+
ρ_1	0.8	0.614	0.672	0.661	0.569	0.585	0.574	0.689	0.699
	0.6	0.530	0.609	0.594	0.480	0.516	0.500	0.621	0.633
	0.4	0.380	0.494	0.462	0.320	0.392	0.370	0.476	0.492
	0.2	0.169	0.299	0.258	0.120	0.194	0.164	0.204	0.218
	-0.2	0.499	0.670	0.616	0.399	0.503	0.473	0.188	0.212
	-0.4	0.830	0.894	0.883	0.765	0.769	0.765	0.489	0.526
	-0.6	0.952	0.970	0.968	0.926	0.896	0.904	0.697	0.734
	-0.8	0.988	0.992	0.993	0.981	0.963	0.968	0.830	0.865
β_2	0.4	0.622	0.702	0.679	0.571	0.603	0.591	0.526	0.549
	0.2	0.619	0.699	0.681	0.569	0.601	0.590	0.523	0.546
	0	0.619	0.699	0.678	0.570	0.602	0.588	0.523	0.546
β_3	0.4	0.622	0.701	0.680	0.570	0.602	0.590	0.524	0.547
	0.2	0.620	0.700	0.680	0.570	0.603	0.590	0.524	0.548
	0	0.618	0.699	0.679	0.570	0.602	0.589	0.525	0.547
N	12	0.515	0.560	0.579	0.440	0.287	0.323	0.089	0.133
	20	0.461	0.544	0.524	0.404	0.424	0.415	0.377	0.412
	30	0.571	0.670	0.636	0.515	0.605	0.570	0.564	0.585
	50	0.717	0.812	0.775	0.678	0.788	0.754	0.732	0.743
	100	0.836	0.914	0.883	0.813	0.908	0.887	0.860	0.863

*Parm. = Parameter

statistical significance of these coefficients. The current evaluation could be extended further by comparing estimation of the OLS versus autoregressive model coefficients and their *SE*s for different levels of autocorrelation. This could help inform the current study's type I error and power results by indicating the magnitude of the effect of incorrect modeling of autocorrelation. For example, if only a small degree of accuracy and precision is gained by modeling the autocorrelation for a certain value of ρ_1, then it may not matter that the model selection criteria has low power at that value. Similarly, if an insubstantial degree of accuracy and precision results from false identification of autocorrelation, then the type I error inflation noted in this study might be of minimal importance.

As with most simulation studies, results are limited by the conditions investigated: the values of the β_2 and β_3 coefficients (see Equation 20) do not seem to have much effect on identification of ρ_1, but it should be investigated whether this is really the case or whether it just appears that way from the limited range of values of β_2 and β_3 that were chosen in this study. One of the main limitations of this study is that it considers only the two-phase ITSE data and only investigated first-order autocorrelation. Another important limitation is that performance was evaluated only for a small subset of possible data trends. All conditions included a slight positive linear trend in

baseline. In addition, the only model misspecification assessed was whether the residuals were autocorrelated.

Future research should investigate use of the z_{HM} and z_{HM}^+ for further misspecified models including when a true non-linear trend is ignored to mimic asymptotic trends resulting from ceiling or floor effects. The performance of these statistics could also be assessed for ITSEs with more than two phases (e.g., for ABAB designs) as investigated by Huitema and McKean (2000). This study also only investigated conditions in which the treatment and baseline phases had equal numbers of data points ($n_B = n_A$). Single-subject studies frequently entail unequal sample sizes per phase and the effect of uneven n should be investigated.

Based on the results of this study, researchers interested in modeling linear growth in ITSE data with a small number of data points should use z_{HM}^+ or z_{HM} to test for the presence of lag-one autocorrelation. Researchers are cautioned against using the Durbin-Watson test statistic and the various information criteria evaluated here including the AIC, HQIC, SBC, AICC, DW and the CAIC for two-phase ITSEs with Ns less than 50.

References

Bartlett, M. S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Journal of the Royal Statistical Society*, *8*, 27-41.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345-370.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment*, *10*, 229-242.

Fan, X., Felsovalyi, A., Keenan, S. C., & Sivo, S. (2001). *SAS for Monte Carlo studies: A guide for quantitative researchers*. Cary, NC: SAS Institute, Inc.

Hannon, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, *Series B*, *41*, 190-195.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment*, *7*, 107-118

Huitema, B. E., & McKean, J. W. (1991). Autocorrelation estimation and inference with small samples. *Psychological Bulletin*, *110*, 291-304.

Huitema, B. E., & McKean, J. W. (2000). A simple and powerful test for autocorrelated errors in OLS intervention models. *Psychological Reports*, *87*, 3-20.

Huitema, B. E., Mckean, J. W., & McKnight S. (1999). Autocorrelation effects on least-squares intervention analysis of short time series. *Educational and Psychological Measurement*, *59*, 767-786.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*, 297-307.

Kendall, M. G. (1954). Note on bias in the estimation of autocorrelation. *Biometrika*, *41*, 403-404.

Marriott, F. H. C., & Pope, J. A. (1954). Bias in the estimation of autocorrelations. *Biometrika*, *41*, 390-402.

SAS Institute, Inc. (2003). SAS (Version 9.1) [Computer Software]. Cary, NC: SAS Institute, Inc.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.

Tawney, L., & Gast, D. (1984). *Single-subject research in special education*. Columbus, OR: Merrill.

White, J. S. (1961). Asymptotic expansions for the mean and variance of the serial correlation coefficient. *Biometrika*, *48*, 85-94.