Journal of Educational and Behavioral Statistics
Winter 1996, Vol. 21, No. 4, pp. 390–404

The Runs Test for Autocorrelated Errors: Unacceptable Properties

Bradley E. Huitema
Joseph W. McKean
Jinsheng Zhao
Western Michigan University

Keywords: *autocorrelation, independence, regression assumptions, runs test, time series*

The runs test is frequently recommended as a method of testing for nonindependent errors in time-series regression models. A Monte Carlo investigation was carried out to evaluate the empirical properties of this test using (a) several intervention and nonintervention regression models, (b) sample sizes ranging from 12 to 100, (c) three levels of α, *(d) directional and nondirectional tests, and (e) 19 levels of autocorrelation among the errors. The results indicate that the runs test yields markedly asymmetrical error rates in the two tails and that neither directional nor nondirectional tests are satisfactory with respect to Type I error, even when the ratio of degrees of freedom to sample size is as high as .98. It is recommended that the test generally not be employed in evaluating the independence of the errors in time-series regression models.*

An expository article on Mood's (1940) runs test was recently published in this journal (Mogull, 1994). It described some of the more popular applications of the test and pointed out the following anomalies in the performance of the one-sample version when there are runs of two: First, departures from randomness are not identified, and, second, power decreases rather than increases with increases in sample size. The purpose of the present article is to describe the problem with employing the runs test as a diagnostic measure in evaluating the assumption of random errors in regression models.

Regression models are often a desirable choice for the analysis of time-series data, especially in the case of small N. Alternative methods of time-series analysis based on ARIMA models (Box & Jenkins, 1976) and spectral models (e.g., Granger & Hatanaka, 1964; Jenkins & Watts, 1968; Priestley, 1981) have been frequently recommended in the behavioral sciences, but the minimum sample size requirement for these approaches (50–100 for ARIMA and 100–200 for spectral analysis) and other practical considerations render

---

This material is based on work supported by the National Science Foundation under Grant Number SES-9205328. Requests for reprints should be addressed to the first author.

them of questionable value in many applications. The interrupted time-series experiment is an example of a behavioral science time-series design that is often successfully modeled using regression procedures (Huitema, 1985). In the case of a time-series design, the general linear model can be written as

$$Y_t = \beta_0 + \beta_1 X_{1_t} + \beta_2 X_{2_t} + \cdots + \beta_m X_{m_t} + \epsilon_t$$

where Y_t is the score on the response measure at time t, β_0 is the process intercept, β_1 through β_m are the process partial regression coefficients, X_{1_t} through X_{m_t} are the values of the X variables at time t, and ϵ_t is the process error at time t.

Inferential procedures associated with this regression model were derived under certain assumptions regarding the errors. It is assumed that the errors possess the following properties:

(1) Independence. The value of each error is statistically independent of all other errors. In the case of a time-series design, it is assumed that errors before and after the error at time t contain no information (beyond that provided by the mean) that is useful in predicting the error at time t.
(2) Constant variance. Regardless of the level of the X variable(s) included in the model, the associated population errors have a common variance.
(3) Normality. The population errors associated with each level of the X variable(s) are normally distributed.

Because time-series data are collected across time on a single sampling unit, it is always appropriate to question whether the errors of the time-series process are independent. There are several reasons to be concerned with large departures from independence. For example, it is well known that a high degree of positive dependency among the errors generally leads to (a) serious underestimation of standard errors for the regression coefficients and (b) prediction intervals that are excessively wide. The latter effect occurs because information regarding the dependency is not acknowledged in the prediction interval expressions that were derived under the independence assumption. Additional details on the effects of dependent errors can be found in most econometrics textbooks (e.g., Johnston, 1984).

Some method for testing the independence assumption should be considered a routine aspect of the diagnostics analysis that follows the fitting of a time-series regression model. The two most frequently recommended approaches for testing this assumption are the Durbin-Watson test (Durbin & Watson, 1950, 1951) and the runs test. A survey of recent textbooks suggests that the Durbin-Watson test has become the more popular of the two, but the runs test continues to be recommended for testing the independence assumption in time-series regression models (e.g., Chatterjee & Price, 1991; Cromwell, Labys, & Terraza, 1994; Daniel, 1990; Draper & Smith, 1981; Rawlings, 1988).

The specific purpose of the runs test in the regression context is to evaluate the hypothesis that the true errors ϵ_t are determined by a random (white noise) process. The test is carried out on the N sample residuals e_t, where $e_t = Y_t - \hat{Y}_t$, $\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_{1_t} + \hat{\beta}_2 X_{2_t} + \cdots + \hat{\beta}_m X_{m_t}$, $\hat{\beta}_0$ is the sample intercept, $\hat{\beta}_1$ through $\hat{\beta}_m$ are the sample partial regression coefficients, and X_{1_t} through X_{m_t} are the sample values of the X variables at time t.

Each time-ordered residual is replaced with the sign of the residual, and the conventional runs test is applied to the sequence of positive and negative signs. If the obtained test statistic is equal to or greater than the critical value, it is concluded that the process generating the errors is not random and that autocorrelation of the errors is present. The logic of the test is that the observed number of runs of positive and negative residuals should be similar to the expected number of runs under the hypothesis of independent errors. A discrepancy between the observed and expected numbers of runs that exceeds the discrepancy reasonably attributed to sampling error is viewed as evidence that the errors are not independent and that the model is inappropriate. In this case, it is recommended that a search be undertaken for an alternative model that, hopefully, will yield independent errors. The alternative model may contain different or additional independent variables and/or autoregressive/moving average parameters to model the dependency in the errors.

A potential but largely ignored problem with the application of the runs test to the signs of the residuals is that the residuals have properties which differ somewhat from those of the errors. An understanding of this point requires that a clear distinction be made between the process errors ϵ_t and the sample residuals e_t. The process (or population) error ϵ_t is defined as

$$Y_t - [\beta_0 + \beta_1 X_{1_t} + \beta_2 X_{2_t} + \cdots + \beta_m X_{m_t}],$$

whereas the sample (or realization) residual e_t is defined as

$$Y_t - [\hat{\beta}_0 + \hat{\beta}_1 X_{1_t} + \hat{\beta}_2 X_{2_t} + \cdots + \hat{\beta}_m X_{m_t}].$$

While ϵ_t is estimated by e_t, the properties of the two are not the same. Most importantly, the residuals e_t of a fitted model will be dependent, even though the true errors ϵ_t of the model are independent. This occurs because some dependency (widely believed to be trivial) is necessarily introduced among the residuals as a result of fitting the constraints of the model to the sample data. The effects of this dependency on graphical and formal test methods used to evaluate the hypothesis of independent errors is of interest.

Anscombe and Tukey (1963) mentioned that although correlations and constraints affect distributions of functions of the residuals, "the corresponding effects on the graphical procedures can usually be ignored" (p. 144). Hence, Draper and Smith (1981) concluded, "It would appear that in general

regression situations the effect of correlation between residuals need not be considered when plots are made, except when the ratio $(N - p)/N$. . . is quite small" (p. 152). This conclusion leaves at least two questions unanswered for the researcher. First, what is the definition of a "small" $(N - p)/N$ ratio? Second, what is the effect of the $(N - p)/N$ ratio on the runs test of the independence assumption?

Even though the theory associated with the use of the runs test in various time-series applications is quite well developed (e.g., Goodman & Grunfield, 1961; Moore & Wallis, 1943; Sen, 1965; Wallis & Moore, 1941), there is need for an evaluation of the empirical performance of the test in typical applications. This is especially true in the frequently encountered case of several predictor variables and small to intermediate sample sizes.

This article evaluates the properties of the runs test with realistic sample sizes in the context of two simple linear regression models (one with a dichotomous predictor and one with predictor values equal to the time period of the observed Y value) and three complex regression models for the analysis of interrupted time-series intervention designs that contain two, three, or four phases. Multiple regression models for these intervention designs contain parameters for describing intercept, level change, slope, and slope change for various phases. Hence, analyses of two-, three-, and four-phase designs yield four, six, and eight parameter estimates, respectively. The specific design matrices associated with both the simple and multiple predictor models investigated in this study can be found in the Appendix.

Method

A large-scale Monte Carlo experiment was designed to evaluate the performance of the runs test (normal approximation version) under many conditions. The following five factors were included in the simulation design: (a) sample size, (b) model type, (c) α level, (d) null hypothesis tested, and (e) level of autocorrelation in the first-order autoregressive error structure of the process. More specifically, there were five sample sizes (N = 12, 20, 30, 50, and 100), five nonintervention and intervention models (see the Appendix for the design matrices), three levels of α (.01, .05, and .10), three null hypotheses ($\rho_1 = 0$, $\rho_1 \leq 0$, and $\rho_1 \geq 0$), and 19 levels of autocorrelation [ρ_1 = (−.90(.10).90)] in the error structure. Hence, (5)(5)(3)(3)(19) = 4,275 combinations were investigated; 1,000 samples were generated for each of these combinations. Type I error rate and power functions were investigated for each combination in the simulation. The error structure of all five models was defined as

$$\epsilon_t = \rho_1 \epsilon_{t-1} + a_t,$$

where ϵ_t is the process error at time t; ρ_1 is the lag-1 autocorrelation of the ϵ_t; and a_t is the random shock meeting the assumptions of constant variance,

normality, and independence. Each error series was started up with a standard normal variate. The first 300 observations of each generated series were ignored in order to virtually eliminate correlation between the last observation of one sample and the first observation of the next sample. The normal variates were obtained as discussed in Marsaglia and Bray (1964) using a FORTRAN generator written by Kahaner, Moler, and Nash (1989). The simulation was performed on a Sun SPARCstation I.

Results

The results of the runs test with respect to the empirical Type I error associated with nominal $\alpha = .01, .05$, and .10 for directional tests (as well as nominal $\alpha = .02, .10$, and .20 for nondirectional tests) are presented for sample sizes 12, 20, 30, 50, and 100 in Tables 1–5, respectively. Results on power against the 18 nonzero values of ρ_1 that were investigated are not presented because the unsatisfactory Type I error outcome eliminates interest in such analyses.

TABLE 1
Empirical Type I error associated with the runs test on 2-, 4-, 6-, and 8-parameter regression models: N = *12*

	Type I error		
p	α_1	α_2	α_3
	Test for positive autocorrelation (H_0: $\rho_1 \leq 0$)		
2	.045 (.039)	.037 (.032)	.008 (.005)
4	.000	.000	.000
6	.000	.000	.000
8	.000	.000	.000
	Test for negative autocorrelation (H_0: $\rho_1 \geq 0$)		
2	.139 (.157)	.082 (.084)	.017 (.018)
4	.265	.160	.031
6	.499	.311	.071
8	1.000	1.000	.526
	Nondirectional test for autocorrelation (H_0: $\rho_1 = 0$)		
2	.184 (.196)	.119 (.116)	.025 (.023)
4	.265	.160	.031
6	.499	.311	.071
8	1.000	1.000	.526

Note. For directional tests, $\alpha_1 = .10$, $\alpha_2 = .05$, and $\alpha_3 = .01$; for nondirectional tests, $\alpha_1 = .20$, $\alpha_2 = .10$, and $\alpha_3 = .02$. The parenthetical values associated with Type I error where $p = 2$ are based on design matrices in which the levels of the predictor variable are integers 1, 2, . . . , N rather than 0 and 1.

The number of parameters in the model is indicated in Tables 1–5 under the column heading p. There are two values under each α heading associated with $p = 2$; the first, not in parentheses, is the empirical Type I error associated with the two-parameter simple linear regression model in which the predictor is a zero-one dummy variable. The second value (in parentheses) is the empirical Type I error associated with the two-parameter simple linear regression model in which the predictor variable is the observation time period, a relatively continuous variable. The performance of the runs test is very similar under these two design matrices for all N and all levels of α.

It can be seen in Table 1 that in testing for positive autocorrelation with a directional test, there are no rejections whatsoever, regardless of the level of alpha selected, unless there are only two parameters in the model. However, in testing for negative autocorrelation with the corresponding directional test, the empirical Type I error is generally far above the nominal level for all values of α. It is also apparent in this table that the error rate increases as p increases. The empirical error rate is about 2 to 3 times the nominal value

TABLE 2
Empirical Type I error associated with the runs test on 2-, 4-, 6-, and 8-parameter regression models: N = *20*

	Type I error		
p	α_1	α_2	α_3
	Test for positive autocorrelation (H_0: $\rho_1 \leq 0$)		
2	.084 (.072)	.039 (.034)	.005 (.005)
4	.027	.007	.001
6	.007	.001	.000
8	.000	.000	.000
	Test for negative autocorrelation (H_0: $\rho_1 \geq 0$)		
2	.138 (.133)	.056 (.058)	.016 (.017)
4	.206	.088	.023
6	.326	.156	.049
8	.464	.256	.093
	Nondirectional test for autocorrelation (H_0: $\rho_1 = 0$)		
2	.222 (.205)	.095 (.092)	.021 (.022)
4	.233	.095	.024
6	.333	.157	.049
8	.464	.256	.093

Note. For directional tests, $\alpha_1 = .10$, $\alpha_2 = .05$, and $\alpha_3 = .01$; for nondirectional tests, $\alpha_1 = .20$, $\alpha_2 = .10$, and $\alpha_3 = .02$. The parenthetical values associated with Type I error where $p = 2$ are based on design matrices in which the levels of the predictor variable are integers 1, 2, . . . , N rather than 0 and 1.

with $p = 2$ and 4, and 5 to 7 times the nominal value when $p = 6$. In the case of $p = 8$, the true null hypothesis was rejected in every sample when nominal 10% and 5% tests were employed, and in about 53% of the samples when the test was performed at the 1% level. It is readily apparent that the unacceptable and asymmetric empirical error rates associated with the two directional tests also lead to a nondirectional test that is unacceptable.

Tables 2–5 contain results based on $N = 20$, 30, 50, and 100, respectively. It can be seen that the empirical Type I error is inversely related to p in the case of testing for positive autocorrelation and positively related to p in the case of testing for negative autocorrelation. Regardless of sample size or p, there is a large discrepancy between the empirical error rate associated with testing for positive autocorrelation (H_0: $\rho_1 \leq 0$) and the error rate associated with testing for negative autocorrelation (H_0: $\rho_1 \geq 0$).

Interestingly, at $N = 100$ the nondirectional test of H_0: $\rho_1 = 0$ appears to be reasonably satisfactory at all levels of α. For example, with a nominal $\alpha = .10$ and $p = 8$, the empirical error rate is about .12. While this error rate

TABLE 3
Empirical Type I error associated with the runs test on 2-, 4-, 6-, and 8-parameter regression models: N = *30*

	Type I error		
p	α_1	α_2	α_3
	Test for positive autocorrelation (H_0: $\rho_1 \leq 0$)		
2	.074 (.061)	.030 (.035)	.005 (.003)
4	.036	.016	.001
6	.020	.008	.000
8	.005	.000	.000
	Test for negative autocorrelation (H_0: $\rho_1 \geq 0$)		
2	.092 (.112)	.040 (.056)	.008 (.010)
4	.155	.077	.015
6	.217	.118	.035
8	.310	.181	.051
	Nondirectional test for autocorrelation (H_0: $\rho_1 = 0$)		
2	.166 (.173)	.07 (.091)	.013 (.013)
4	.191	.093	.016
6	.237	.126	.035
8	.315	.181	.051

Note. For directional tests, $\alpha_1 = .10$, $\alpha_2 = .05$, and $\alpha_3 = .01$; for nondirectional tests, $\alpha_1 = .20$, $\alpha_2 = .10$, and $\alpha_3 = .02$. The parenthetical values associated with Type I error where $p = 2$ are based on design matrices in which the levels of the predictor variable are integers 1, 2, . . . , N rather than 0 and 1.

appears to be just moderately liberal, one should not conclude that the test is satisfactory. Notice that even in this reasonably large-N case the asymmetry associated with the two directional tests is great. The empirical error rate associated with the 5% test for negative autocorrelation is about .11, whereas the error rate for positive autocorrelation is .01.

Conclusions and Discussion

It is concluded that the small sample properties of the runs test are generally unsatisfactory in testing the independence assumption in general linear models. While this conclusion is based on the sample sizes, α levels, and models included in this study, there is reason to expect similar results in the case of other general linear models.

The asymmetrical Type I error rates associated with the runs test follow the same pattern recently found with both Bartlett's (1946) test for autocorrelation and the most frequently recommended version of the Durbin-Watson test for autocorrelated errors (Huitema & McKean, 1991; Huitema,

TABLE 4
Empirical Type I error associated with the runs test on 2-, 4-, 6-, and 8-parameter regression models: N = *50*

	Type I error		
p	α_1	α_2	α_3
	Test for positive autocorrelation (H_0: $\rho_1 \leq 0$)		
2	.091 (.085)	.042 (.044)	.007 (.006)
4	.058	.029	.003
6	.035	.018	.001
8	.028	.014	.000
	Test for negative autocorrelation (H_0: $\rho_1 \geq 0$)		
2	.113 (.109)	.065 (.063)	.014 (.011)
4	.128	.074	.020
6	.198	.131	.026
8	.237	.156	.031
	Nondirectional test for autocorrelation (H_0: $\rho_1 = 0$)		
2	.204 (.194)	.107 (.107)	.021 (.017)
4	.186	.103	.023
6	.233	.149	.027
8	.265	.170	.031

Note. For directional tests, $\alpha_1 = .10$, $\alpha_2 = .05$, and $\alpha_3 = .01$; for nondirectional tests, $\alpha_1 = .20$, $\alpha_2 = .10$, and $\alpha_3 = .02$. The parenthetical values associated with Type I error where $p = 2$ are based on design matrices in which the levels of the predictor variable are integers 1, 2, . . . , N rather than 0 and 1.

McKean, & Zhao, 1996). A common cause of the asymmetry in these tests is the dependency that is introduced among the sample residuals when the model is fit. The degree of dependency is a function primarily of the sample size N and the number of parameters p estimated. One consequence of this dependency is (generally) negative bias in autocorrelation estimates when sampling from processes having positive or moderate negative values of the lag-1 autocorrelation parameter ρ_1. In the case of small samples, the expected value of the conventional autocorrelation estimator r_1 is considerably less than the value of the underlying parameter ρ_1; the amount of bias increases dramatically as ρ_1 increases. For example, if $N = 10$, $p = 1$, and the errors are independent (yielding $\rho_1 = .00$), the expected value of r_1 is $-.10$ rather than 0; when $\rho_1 = .90$, the empirical expected value of r_1 is only .40 (Huitema & McKean, 1991). In general, bias increases as ρ_1 and p increase and sample size decreases.

It is known that there are two sources of bias in r_1. One source, associated with the fact that the sample autocovariance and the sample variance are not

TABLE 5
Empirical Type I error associated with the runs test on 2-, 4-, 6-, and 8-parameter regression models: N = *100*

	Type I error		
p	α_1	α_2	α_3
	Test for positive autocorrelation (H_0: $\rho_1 \leq 0$)		
2	.079 (.077)	.032 (.038)	.008 (.010)
4	.060	.023	.006
6	.039	.019	.000
8	.035	.010	.001
	Test for negative autocorrelation (H_0: $\rho_1 \geq 0$)		
2	.112 (.120)	.070 (.077)	.025 (.022)
4	.134	.087	.025
6	.180	.103	.029
8	.201	.114	.034
	Nondirectional test for autocorrelation (H_0: $\rho_1 = 0$)		
2	.191 (.197)	.102 (.115)	.033 (.032)
4	.194	.110	.031
6	.219	.122	.029
8	.236	.124	.035

Note. For directional tests, $\alpha_1 = .10$, $\alpha_2 = .05$, and $\alpha_3 = .01$; for nondirectional tests, $\alpha_1 = .20$, $\alpha_2 = .10$, and $\alpha_3 = .02$. The parenthetical values associated with Type I error where $p = 2$ are based on design matrices in which the levels of the predictor variable are integers 1, 2, . . . , N rather than 0 and 1.

independently distributed, biases r_1 toward 0. The other source, associated with the lack of independence of the sample mean and the sample autocovariance, biases r_1 toward -1. The bias in r_1 occurs whether this estimator is computed on the original time-series data or on the residuals of a regression model, although the amount of bias differs in these two situations. (Additional details on bias in both autocorrelation estimation and hypothesis testing can be found in Huitema and McKean, 1991, and the references contained therein; reduced bias estimators and corresponding hypothesis tests can be found in Huitema and McKean 1994a, 1994b, 1994c, 1996.)

The dependencies (mentioned above) that bias r_1 are the explanation for our earlier finding that most tests for autocorrelation are accompanied by asymmetrical one-tailed empirical Type I error values that differ greatly from the nominal values. Similarly, these dependencies explain the present findings.

It turns out that when a completely random (white noise) process is involved, the empirical expectation for the number of runs in a series of sample residuals differs systematically from the theoretical expectation. The conventional expression for the expected number of runs (given a completely random error process) is

$$\text{Expected Runs} = \left(\frac{2n_1n_2}{n_1 + n_2}\right) + 1,$$

where n_1 and n_2 are the numbers of positive and negative residuals, respectively.

This expectation is incorrect in the case of sample residuals because dependency is introduced among them by fitting the model. The actual (empirical) expectation is greater than the value suggested by this formula. In addition, the standard error of the runs, generally defined as

$$\sigma_{SE} = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}},$$

is likewise based on the false assumption that the sequence of residuals is random when the sequence of errors is random. The normal approximation for the runs test employs both the biased estimate of the expected number of runs and the biased standard error estimate; the approximation (without correction for continuity) is simply

$$z = \frac{\text{Observed Runs} - \text{Expected Runs}}{\sigma_{SE}}.$$

Hence, it can be seen that both the numerator and denominator of the runs test contain terms that were derived under the assumption that random

residuals will occur when the process errors are random. Because this assumption is false, the runs test is biased in the finite sample case; the degree of bias is quite severe in the case of multiple predictors and typical sample sizes.

It can be concluded that the runs test is generally inappropriate as a test for nonindependent errors in regression models unless few predictors and several hundred observations are available. The ratio $(N - p)/N$ must be greater than .98 for reasonable Type I error test performance. This conclusion holds for both the normal approximation and the exact critical value versions of the test.

APPENDIX

Matrices associated with 2-, 4-, 6-, and 8-parameter models

Matrices for 2-parameter intervention model

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_{n_1} \\ \hline Y_{n_1+1} \\ Y_{n_1+2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{n_1+n_2} \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 0 \\ \hline 1 & 1 \\ 1 & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 1 \end{bmatrix}$$

Matrices for 2-parameter nonintervention model

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & \cdot \\ 1 & n \end{bmatrix}$$

Matrices for 4-parameter intervention model

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_{n_1} \\ \hline Y_{n_1+1} \\ Y_{n_1+2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{n_1+n_2} \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & n_1 & 0 & 0 \\ \hline 1 & n_1 + 1 & 1 & 0 \\ 1 & n_2 + 2 & 1 & 1 \\ \cdot & \cdot & \cdot & 2 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & n_1 + n_2 & 1 & n_2 - 1 \end{bmatrix}$$

where n_1 and n_2 are the numbers of observations in the preintervention phases of the time series, respectively.

Matrices for 6-parameter model

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_{n_1} \\ \hline Y_{n_1+1} \\ Y_{n_1+2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{n_1+n_2} \\ \hline Y_{n_1+n_2+1} \\ Y_{n_1+n_2+2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{n_1+n_2+n_3} \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & n_1 & 0 & 0 & 0 & 0 \\ \hline 1 & n_1 + 1 & 1 & 0 & 0 & 0 \\ 1 & n_2 + 2 & 1 & 1 & 0 & 0 \\ \cdot & \cdot & \cdot & 2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & n_1 + n_2 & 1 & n_2 - 1 & 0 & 0 \\ \hline 1 & n_1 + n_2 + 1 & 1 & n_2 & 1 & 0 \\ 1 & n_1 + n_2 + 2 & 1 & n_2 + 1 & 1 & 1 \\ \cdot & \cdot & \cdot & n_2 + 2 & \cdot & 2 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & n_1 + n_2 + n_3 & 1 & n_2 + n_3 - 1 & 1 & n_3 - 1 \end{bmatrix}$$

n_1, n_2, n_3 are the numbers of observations in Phases 1–3, respectively.

Matrices for 8-parameter intervention model

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_{n_1} \\ \hline Y_{n_1+1} \\ Y_{n_1+2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{n_1+n_2} \\ \hline Y_{n_1+n_2+1} \\ Y_{n_1+n_2+2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{n_1+n_2+n_3} \\ \hline Y_{n_1+n_2+n_3+1} \\ Y_{n_1+n_2+n_3+2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{n_1+n_2+n_3+n_4} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & n_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & n_1+1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & n_2+2 & 1 & 1 & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & 2 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & n_1+n_2 & 1 & n_2-1 & 0 & 0 & 0 & 0 \\ \hline 1 & n_1+n_2+1 & 1 & n_2 & 1 & 0 & 0 & 0 \\ 1 & n_1+n_2+2 & 1 & n_2+1 & 1 & 1 & 0 & 0 \\ \cdot & \cdot & \cdot & n_2+2 & \cdot & 2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & n_1+n_2+n_3 & 1 & n_2+n_3-1 & 1 & n_3-1 & 0 & 0 \\ \hline 1 & n_1+n_2+n_3+1 & 1 & n_2+n_3 & 1 & n_3 & 1 & 0 \\ 1 & n_1+n_2+n_3+2 & 1 & n_2+n_3+1 & 1 & n_3+1 & 1 & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & n_1+n_2+n_3+n_4 & 1 & n_2+n_3+n_4-1 & 1 & n_3+n_4-1 & 1 & n_4-1 \end{bmatrix}$$

n_1, n_2, n_3, and n_4 are the numbers of observations in Phases 1–4, respectively.

References

Anscombe, F. J., & Tukey, J. W. (1963). The examination and analysis of residuals. *Technometrics, 5*, 141–160.

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

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis, forecasting and control.* San Francisco: Holden Day.

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

Cromwell, J. B., Labys, W. C., & Terraza, M. (1994). *Univariate tests for time series models.* Thousand Oaks, CA: Sage.

Daniel, W. W. (1990). *Applied nonparametric statistics.* Boston: PWS-Kent.

Draper, N. R., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: Wiley.

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

Durbin, J., & Watson, G. S. (1951). Testing for serial correlation in least squares regression II. *Biometrika, 38*, 159–178.

Goodman, L. A., & Grunfield, Y. (1961). Some nonparametric tests for comovements between time series. *Journal of the American Statistical Association, 56*, 11–26.

Granger, C. W. J., & Hatanaka, M. (1964). *Spectral analysis of economic time series*. Princeton, NJ: Princeton University Press.

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. (1994a). Reduced-bias autocorrelation estimation: Three jackknife methods. *Educational and Psychological Measurement, 54*, 654–665.

Huitema, B. E., & McKean, J. W. (1994b). Tests of H_0: $\rho_1 = 0$ for autocorrelation estimators r_{F1} and r_{F2}. *Perceptual and Motor Skills, 78*, 331–336.

Huitema, B. E., & McKean, J. W. (1994c). Two reduced-bias autocorrelation estimators: r_{F1} and r_{F2}. *Perceptual and Motor Skills, 78*, 323–330.

Huitema, B. E., & McKean, J. W. (1996). Tests for the jackknife autocorrelation estimator r_{Q2}. *Educational and Psychological Measurement, 56*, 232–240.

Huitema, B. E., McKean, J. W., & Zhao, J. (1996). *Durbin-Watson test of autocorrelated errors in time-series regression: Some problems*. Manuscript submitted for publication.

Jenkins, G. M., & Watts, D. J. (1968). *Spectral analysis and its applications*. San Francisco: Holden Day.

Johnston, J. (1984). *Econometric methods* (2nd ed.). New York: McGraw-Hill.

Mogull, R. G. (1994). The one-sample runs test: A category of exception. *Journal of Educational and Behavioral Statistics, 19*, 296–303.

Mood, A. M. (1940). The distribution theory of runs. *Annals of Mathematical Statistics, 11*, 367–392.

Moore, G. H., & Wallis, W. A. (1943). Time series significance test based on signs of differences. *Journal of the American Statistical Association, 38*, 153–164.

Priestley, M. B. (1981). *Spectral analysis and time series*. New York: Academic Press.

Rawlings, J. O. (1988). *Applied regression analysis: A research tool*. Belmont, CA: Wadsworth & Brooks/Cole.

Sen, P. K. (1965). Some nonparametric tests for *m*-dependent time series. *Journal of the American Statistical Association, 60*, 134–147.

Wallis, W. A., & Moore, G. H. (1941). A significance test for time series. *Journal of the American Statistical Association, 36*, 401–409.

Authors

BRADLEY E. HUITEMA is Professor, Department of Psychology, Western Michigan University, Kalamazoo, MI 49008-3899; brad.huitema@wmich.edu. He specializes in statistical methods for quasi-experiments and single-subject designs.

JOSEPH W. MCKEAN is Professor, Department of Mathematics and Statistics,

Western Michigan University, Kalamazoo, MI 49008-5152; joe@stat.wmich.edu. He specializes in linear models, nonparametrics, and robust statistics.

JINSHENG ZHAO is a doctoral student, Department of Mathematics and Statistics, Western Michigan University, Kalamazoo, MI 49008-5152; zhao@stat.wmich.edu. He specializes in linear models, nonparametrics, and robust statistics.

Received November 9, 1994
Revision received May 22, 1995
July 20, 1995