Regression-based techniques for statistical decision making in single-case designs

Rumen Manolov, Jaume Arnau, Antonio Solanas and Roser Bono
Universidad de Barcelona

The present study evaluates the performance of four methods for estimating regression coefficients used to make statistical decisions about intervention effectiveness in single-case designs. Ordinary least square estimation is compared to two correction techniques dealing with general trend and a procedure that eliminates autocorrelation whenever it is present. Type I error rates and statistical power are studied for experimental conditions defined by the presence or absence of treatment effect (change in level or in slope), general trend, and serial dependence. The results show that empirical Type I error rates do not approach the nominal ones in the presence of autocorrelation or general trend when ordinary and generalized least squares are applied. The techniques controlling trend show lower false alarm rates, but prove to be insufficiently sensitive to existing treatment effects. Consequently, the use of the statistical significance of the regression coefficients for detecting treatment effects is not recommended for short data series.

Técnicas fundamentadas en la regresión para la decisión estadística en diseños de caso único. El estudio evalúa el rendimiento de cuatro métodos de estimación de los coeficientes de regresión utilizados para la toma de decisiones estadísticas sobre la efectividad de las intervenciones en diseños de caso único. La estimación por mínimos cuadrados ordinarios se compara con dos métodos que controlan la tendencia en los datos y un procedimiento que elimina la autocorrelación cuando ésta es significativa. Los resultados indican que las tasas empíricas y nominales de falsas alarmas no coinciden en presencia de dependencia serial o tendencia al aplicar mínimos cuadrados ordinarios o generalizados. Los métodos que controlan la tendencia muestran tasas más bajas de error Tipo I, pero no son suficientemente sensibles a efectos existentes (cambio de nivel o de pendiente), por lo que el uso de la significación estadística de los coeficientes de regresión para detectar efectos no se recomienda cuando se dispone de series cortas de datos.

Single-case designs allow psychologists to study the evolution of the behavior of a single experimental unit (an individual or a group taken as a whole) in different conditions. These conditions usually involve the presence or absence of psychological treatment and permit assessing its effectiveness for introducing modifications in the response of interest. However, the evaluation of the relationship between the intervention and the response rate requires baseline stability, phase alternation and replications across time and in different settings (Kazdin, 1978). Among the obstacles for single-case data analysis the shortness of data series (Huitema, 1985) and the serial dependence between the measurements (Matyas & Greenwood, 1997; Parker, 2006) have to be highlighted. The importance of the latter has been underlined due to its impact on Type I error rates on a variety of procedures (Busk & Marascuilo, 1988; Sharpley & Alavosius, 1988; Suen & Ary, 1987).

The present study centers on regression-based techniques for making statistical decisions regarding treatment effectiveness in N= 1 designs. The reason for choosing this type of procedures is that they are well-known and can easily be applied using commonly available software even by psychologists who are not experts in statistics. Previous studies (Brossart, Parker, Olson, & Mahadevan, 2006; Parker & Brossart, 2003) have shown that the R^2 values obtained by regression-based techniques for effect size estimation can be excessively low —in the case of Gorsuch's (1983) Trend analysis— or excessively high for Allison and Gorman's (1993) and White, Rusch, Kazdin, and Hartmann's (1989) models. Additionally, positive autocorrelation leads to overestimating the effect size (Beretvas & Chung, 2008; Manolov & Solanas, 2008a). The abovementioned models have also been compared versus the no effect model using an *F* test, obtaining Type I error rates greater than 10% for autocorrelation of .3 (Fisher, Kelley, & Lomas, 2003). In contrast, there is evidence that the regression coefficients estimate precisely the data parameters defined by simulation (Solanas, Manolov, & Onghena, 2010) and, thus, it is necessary to test whether the *p* values associated with these coefficients are to be recommended for hypothesis testing.

Regarding other alternatives for single-case data analysis, none of them has been established unequivocally as suitable, especially due to the influence of autocorrelation. Historically, the first

technique applied was visual inspection (Johnston & Pennypacker, 2008), but it has been shown that, among other drawbacks, false alarm rates tend to increase in presence of positive autocorrelation (Matyas & Greenwood, 1990). The closely related split-middle method (White, 1974) has also been shown not to control Type I error rates under serial dependence (Crosbie, 1987). Another simple technique proposed is the *C* statistic (Tryon, 1982), which is rather a first-order autocorrelation estimator than an index for assessing intervention effectiveness (DeCarlo & Tryon, 1993). Analysis of variance assumes independence and nonzero autocorrelation affects the easiness of obtaining significant results (Scheffé, 1959; Toothaker, Banz, Noble, Camp, & Davis, 1983). Randomization tests, on the other hand, do not assume explicitly the lack of serial dependence (Edgington & Onghena, 2007), but positive autocorrelation has been found to distort Type I error rates in some cases (Manolov & Solanas, 2008b) and to increase the probability of omitting an existing effect (Ferron & Ware, 1995). Although interrupted time-series analysis has been designed to control serial dependence (Harrop & Velicer, 1985) it has been shown to affect Type I error rates when few data points are available (Greenwood & Matyas, 1990).

These findings encourage exploring the Type I error rates and statistical power of regression-based techniques. In order to deal with the consequences of measuring behavior longitudinally, several proposals have been developed (see Arnau & Bono, 2004, for a review). The following section presents in detail the procedures tested here.

Procedures studied

The regression models used were common to all procedures and were specified to detect separately level change or slope change (see formulae and design matrices X_1 and X_2, respectively, in the Appendix). The ordinary least squares (OLS) estimation uses the two models as presented in the Appendix, without modification.

The generalized least squares estimation (GLS; Simonton, 1977a), also referred to as Autoregressive analysis by Gorsuch (1983), starts with an OLS estimation and then tests the residuals for independence (Durbin & Watson, 1971). If there is no autocorrelation OLS and GLS concur, whereas the presence of serial dependence leads to correcting the dependent variable according to $y_1^* = y_1 \sqrt{1 - r_1^2}$ for the first value and using $y_t^* = y_t - r_1 y_{t-1}$ for the following ones. In the previous expressions r_1 is the estimate of the lag-one autocorrelation ρ_1 calculated from the Durbin-Watson statistic *d* as $r_1 = 1 - d/2$. Simmonton (1977b, 1978) adds that the independent variable should also be corrected in the same way as the dependent variable. Design matrices X_3 and X_4 presented in the Appendix show the result of using these transformations. GLS's final step consists in an OLS estimation using the corrected variables. As the aforementioned steps indicate, GLS is intended to deal with serial dependence.

Differencing analysis (DA; Gorsuch, 1983) has an initial step of differencing both the dependent and the independent variables, subtracting each value from the previous one. The transformed series has, thus, one value less than the original one. An OLS estimation is applied to the differenced series using the design matrices X_5 and X_6 from the Appendix for the level change and slope change models, respectively. This type of correction is designed for correcting linear trend.

Trend analysis (TA; Gorsuch, 1983) includes an initial OLS regression analysis using time as independent variable (see design matrix X_7 in the Appendix). The residuals of the analysis are saved and used as a dependent variable in OLS regressions using the X_1 and X_2 design matrices for applying the level change and slope change models. The purpose of this procedure is also correcting linear trend, since the influence of time on the dependent variable is eliminated.

Method

Series length selection

The present research focused on data series following the AB design structure representing an initial phase of assessment of the behavior of interest followed by treatment introduction. Since studies with few observation points are more feasible in applied psychological settings, the series (N) and phase lengths (n_A and n_B) included here were: a) N= 10 with $n_A = n_B = 5$; b) N= 15 with $n_A = 5$ and $n_B = 10$; and c) N= 20 with $n_A = n_B = 10$.

Data generation

Data were generated to represent the main features of real behavioral data, such as serial dependence, general trend (i.e., a persistent upward or downward drift in data initiated during baseline phase and not related to treatment introduction), and different types of treatment effect. This was achieved using the model presented in Huitema and McKean (2000): $y_t = \beta_0 + \beta_1 T_t + \beta_2 LC_t + \beta_3 SC_t + \varepsilon_t$, where y_t is the value of the dependent variable at moment t; β_0 is the intercept, β_1 is the coefficient used for specifying the magnitude of trend, β_2 is the level change coefficient, and β_3 is the slope change coefficient. The values of the dummy variables can be consulted from design matrix X_8 in the Appendix. The error term, represented by ε_t, was generated by means of a first-order autoregressive model: $\varepsilon_t = \varrho_1 \varepsilon_{t-1} + u_t$ in which autocorrelation (ρ_1) ranged from $-.3$ to $.6$ in steps of $.3$, representing the degrees of serial dependence found by Parker (2006). The u_t term was used to test three different distributions: normal, negative exponential and Laplace (double exponential). Similar Monte Carlo simulation studies have generally focused on the normal distribution, but there is evidence that it may not be always an adequate model for representing behavioral data (Bradley, 1977; Micceri, 1989) and, hence, nonnormal distributions have already been used to test the statistical properties of analytical techniques (e.g., Kowalchuk, Keselman, Algina, & Wolfinger, 2004; Sawilowsky & Blair, 1992). In order to achieve the comparability of the u_t term distributions, all of them had mean set to zero and standard deviation set to one. In the case of the negative exponential distribution, data were generated with location parameter θ= 0 and scale parameter σ= 1. Afterwards 1 was subtracted from data in order to center around zero. This distribution was included, since it is highly asymmetrical ($\gamma_1 = 2$) in comparison to the symmetrical normal distribution. Data following the Laplace distribution was generated with location parameter μ= 0 and scale parameter ϕ= .7071. This distribution was included to test the relevance of kurtosis, as it is leptokurtic ($\gamma_2 = 3$) in comparison to the normal distribution. 10,000 iterations were made per each experimental condition defined by the combination of series length, level of autocorrelation, and error distribution.

There is still no consensus on what values should be used for the beta parameters in the data generation model, since the effect size benchmarks proposed by Cohen (1988) are not appropriate for single-case data (Matyas & Greenwood, 1990; Parker et al., 2005). In addition, related investigations do not specify explicitly the model used to generate the data nor the value of the effect size parameters (e.g., Keselman & Keselman, 1990; Parker & Brossart, 2003), while others select effect sizes in relation to the power they produce in statistical tests (e.g., Algina & Keselman, 1998). In the current study, the β_0 coefficient was set to zero as the pre-intervention response rate is irrelevant, while the remaining parameters were set to zero in the conditions referring to lack of trend or effect and to one for the conditions with general trend and treatment effect being present.

Data analysis

It has been emphasized that for regression analysis it is essential to fit the correct model (Huitema & McKean, 1998) and, therefore, in the present study only the cases when the regression model specified matched the known (simulated) truth were considered. In case the procedures performed well, it would only apply to the occasions when the researcher fits the right model. Conversely, an inappropriate performance even in this ideal situation would be strong evidence against the application of regression-based procedures for making statistical decisions.

The performance of the procedures was assessed in terms of Type I error rates and statistical power using a nominal significance level of .05. A Type I error occurs when the regression coefficient b for level or slope change is statistically significant (i.e., has an associated p value ≤.05) and there is no treatment effect simulated in the data. In order to assess the correspondence between nominal and empirical Type I error rates, Serlin's (2000) robustness criterion was used, specifying that Type I error is controlled when the true null hypothesis is rejected .05 ± (.025)(.05)= .025 to .075 of the times. A Type II error takes place when the regression coefficient for level or slope change has a p value > .05 in data series with intervention effect. In the current study, power —the complementary of Type II errors— was emphasized. Both in absence and in presence of treatment effect it was possible to explore the impact of trend and autocorrelation on the probability of labeling a behavioral change as statistically significant.

Results

Type I error rates

TA appears to be too conservative for all experimental conditions, including different series lengths and different error distributions, rejecting the null hypothesis practically never regardless of the values of the simulation parameters ρ_1 and β_1. Therefore, results are presented only for OLS, GLS, and DA in Tables 1, 2 and 3, respectively. All three procedures keep the Type I error rates close to nominal 5% levels in independent data series with no general trend. The presence of positive autocorrelation and/or trend is associated with excessively high false alarm rates for OLS and GLS, although for longer series GLS performs slightly better than OLS. In contrast to data series with ρ_1>0, when ρ_1<0 and there is no general trend OLS and GLS are too conservative, an expected result considering that simulating negative autocorrelation implies

more variability in the data series and makes difficult the detection of phase differences. When trend is present, GLS overcorrects for series with N≥15 and does not reject the null hypothesis almost ever. Regardless of the series length DA matches nominal alpha even when $\rho_1 \neq 0$ and/or $\beta_1 \neq 0$ for normal error series. Somewhat greater Type I error rates were observed for DA in the case of exponential and Laplace errors, although the estimates would still meet Serlin's (2000) robustness criterion.

Statistical power

OLS and GLS appear to be more sensitive to treatment effects (both level and slope changes) than DA and TA. Due to the excessively liberal Type I error rates provided by OLS and GLS in presence of trend and/or serial dependence, the interpretation of the power of the two techniques in these experimental conditions would be meaningless. The aforementioned conservativeness of TA was reflected in extremely high Type II error rates. Therefore, power estimates will be presented only for DA (see Table 4), the only procedure controlling Type I error rates. DA shows certain sensitivity to level changes and none (i.e., lower than nominal alpha) to slope changes. In fact, for the condition for which comparisons are justified (i.e., absence of autocorrelation and trend), DA shows less than one third of the power of OLS and GLS for level change and less than one twentieth for slope change. The power estimates are greater for longer data series but the improvement seems too small to be relevant. In accordance with the false alarm estimates, Type II error rates are slightly lower for exponential error data. Autocorrelation also has some influence on DA's power, observing greater sensitivity for ρ_1>0 and lower sensitivity for ρ_1<0. Conversely, the presence of trend does not affect the statistical power of DA in any direction.

Table 1
Type I error rates for ordinary least squares estimation for several experimental conditions

Data series features[a]	Series length	Exponential[b]		Normal[b]		Laplace[b]	
		LC[c]	SC[c]	LC[c]	SC[c]	LC[c]	SC[c]
$\rho_1= 0, \beta_1=0$	5+5	.0418	.0403	.0481	.0520	.0432	.0410
	5+10	.0412	.0489	.0540	.0513	.0466	.0486
	10+10	.0437	.0425	.0495	.0512	.0462	.0500
$\rho_1= .3, \beta_1=0$	5+5	.1134	.1122	.1204	.1255	.1186	.1210
	5+10	.1213	.1331	.1249	.1363	.1278	.1393
	10+10	.1324	.1370	.1353	.1437	.1365	.1417
$\rho_1= .6, \beta_1=0$	5+5	2368	.2628	.2368	.2503	.2428	.2646
	5+10	2566	.3001	.2506	.2887	.2575	.3062
	10+10	3197	.3194	.3046	.3085	.3151	.3151
$\rho_1= 0, \beta_1=1$	5+5	.9890	.9961	.9981	.9999	.9924	.9992
	5+10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	10+10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\rho_1= .3, \beta_1=1$	5+5	.9870	.9950	.9938	.9992	.9878	.9971
	5+10	.9999	1.0000	1.0000	1.0000	1.0000	1.0000
	10+10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

[a] degree of autocorrelation (ρ_1) and size of the trend parameter (β_1)
[b] distribution of the random variable u_t
[c] regression model applied, testing for level change (LC) or slope change (SC)

Discussion

The present study assesses the performance of four regression-based techniques in terms of false alarm rates and capability of detecting existing effects. The comparison takes place in the context of single-case data series generated following three error distributions and including features such as autocorrelation and general trend.

The results obtained suggest that procedures involving OLS and GLS estimation of the regression parameters are useful only when data series are independent and present no general trend. In sequentially related data series these two procedures would reject the null hypothesis too often and may lead to inferring effectiveness even when interventions are innocuous. The fact that positive autocorrelation distorts the Type I error rates of GLS estimates indicates that the data correction performed by the procedure may need to be iterative instead of being applied only once. In contrast, the correction that takes place in TA is so strong that it eliminates both the effect of trend and serial dependence and the changes in data produced by the intervention. That is, the initial step in TA overcorrects the data and renders the procedure unable to detect treatment effects. Contrarily to the abovementioned procedures, DA shows Type I error rates approximately equal to nominal alpha even when data are autocorrelated and/or present general trend. Nevertheless, DA does not detect existing treatment effects frequently enough, particularly when the effect is a change in slope. Therefore, the differencing of the relevant variables also overcorrects, especially when the behavioral change resembles a trend. Hence, DA is not likely to label ineffective interventions as effective, but may be an excessively severe filter with high omission rates. Having more measurements available (up to twenty) does not improve sufficiently the performance of DA, as it is also the case for the remaining procedures. The results obtained did not vary greatly according to the error distribution and the conclusions regarding Type I and Type II error rates are applicable to exponential, normal and Laplace distributions. Nonetheless, the inclusion of nonnormal distributions is useful for increasing the generalizability of the findings.

In summary, the results here obtained suggest that the OLS estimation of the regression parameters cannot be advised in short

Table 2
Type I error rates for generalized least squares estimation for several experimental conditions

Data series features[a]	Series length	Exponential[b]		Normal[b]		Laplace[b]	
		LC[c]	SC[c]	LC[c]	SC[c]	LC[c]	SC[c]
$\rho_1 = 0, \beta_1 = 0$	5+5	.0458	.0398	.0509	.0523	.0454	.0408
	5+10	.0415	.0465	.0540	.0502	.0468	.0483
	10+10	.0421	.0400	.0489	.0485	.0450	.0484
$\rho_1 = .3, \beta_1 = 0$	5+5	.1188	.1060	.1204	.1238	.1195	.1177
	5+10	.1189	.1176	.1191	.1220	.1212	.1277
	10+10	.1201	.1124	.1158	.1225	.1193	.1198
$\rho_1 = .6, \beta_1 = 0$	5+5	.2255	.2309	.2237	.2279	.2273	.2385
	5+10	.2045	.2182	.1908	.2187	.1984	.2270
	10+10	.2086	.1962	.1912	.1966	.1983	.1916
$\rho_1 = 0, \beta_1 = 1$	5+5	.6404	.9698	.7010	.9705	.6654	.9767
	5+10	.0212	1.0000	.0068	1.0000	.0132	1.0000
	10+10	.0178	.9998	.0072	1.0000	.0122	1.0000
$\rho_1 = .3, \beta_1 = 1$	5+5	.5446	.9323	.5912	.9214	.5615	.9366
	5+10	.0148	.9999	.0041	1.0000	.0077	.9999
	10+10	.0177	.9965	.0102	.9984	.0126	.9988

[a] degree of autocorrelation (ρ_1) and size of the trend parameter (β_1)
[b] distribution of the random variable u_t
[c] regression model applied, testing for level change (LC) or slope change (SC)

Table 3
Type I error rates for differencing analysis for several experimental conditions

Data series features[a]	Series length	Exponential[b]		Normal[b]		Laplace[b]	
		LC[c]	SC[c]	LC[c]	SC[c]	LC[c]	SC[c]
$\rho_1 = 0, \beta_1 = 0$	5+5	.0691	.0500	.0510	.0414	.0601	.0552
	5+10	.0702	.0734	.0480	.0462	.0616	.0582
	10+10	.0706	.0606	.0477	.0432	.0607	.0561
$\rho_1 = .3, \beta_1 = 0$	5+5	.0657	.0596	.0546	.0475	.0654	.0551
	5+10	.0706	.0699	.0541	.0543	.0613	.0607
	10+10	.0690	.0647	.0542	.0481	.0590	.0593
$\rho_1 = .6, \beta_1 = 0$	5+5	.0704	.0690	.0499	.0496	.0671	.0529
	5+10	.0640	.0681	.0513	.0545	.0667	.0640
	10+10	.0616	.0592	.0507	.0495	.0700	.0595
$\rho_1 = 0, \beta_1 = 1$	5+5	.0709	.0551	.0475	.0411	.0580	.0481
	5+10	.0740	.0682	.0479	.0468	.0576	.0649
	10+10	.0715	.0690	.0452	.0446	.0640	.0582
$\rho_1 = .3, \beta_1 = 1$	5+5	.0657	.0562	.0518	.0451	.0652	.0525
	5+10	.0680	.0707	.0492	.0529	.0655	.0624
	10+10	.0661	.0649	.0471	.0472	.0639	.0586

[a] degree of autocorrelation (ρ_1) and size of the trend parameter (β_1)
[b] distribution of the random variable u_t
[c] regression model applied, testing for level change (LC) or slope change (SC)

Table 4
Power estimates for differencing analysis for several experimental conditions

Data series features[a]	Series length	Exponential[b]		Normal[b]		Laplace[b]	
		LC[c]	SC[c]	LC[c]	SC[c]	LC[c]	SC[c]
$\rho_1 = 0, \beta_1 = 0$	5+5	.1155	.0454	.0893	.0253	.1050	.0329
	5+10	.1249	.0556	.0960	.0299	.1069	.0407
	10+10	.1141	.0548	.1057	.0348	.1103	.0426
$\rho_1 = .3, \beta_1 = 0$	5+5	.1480	.0494	.1042	.0304	.1220	.0351
	5+10	.1356	.0617	.1116	.0337	.1215	.0414
	10+10	.1358	.0527	.1142	.0256	.1261	.0412
$\rho_1 = .6, \beta_1 = 0$	5+5	.1622	.0533	.1085	.0330	.1294	.0361
	5+10	.1572	.0511	.1233	.0309	.1303	.0435
	10+10	.1480	.0499	.1290	.0300	.1359	.0380
$\rho_1 = 0, \beta_1 = 1$	5+5	.1249	.0435	.0842	.0237	.1051	.0321
	5+10	.1262	.0600	.0976	.0328	.1116	.0470
	10+10	.1211	.0587	.0978	.0314	.1086	.0415
$\rho_1 = .3, \beta_1 = 1$	5+5	.1434	.0510	.1038	.0287	.1182	.0329
	5+10	.1358	.0554	.1111	.0337	.1190	.0416
	10+10	.1316	.0542	.1191	.0282	.1248	.0398

[a] degree of autocorrelation (ρ_1) and size of the trend parameter (β_1)
[b] distribution of the random variable u_t
[c] regression model applied, testing for level change (LC) or slope change (SC)

series, even when it is complemented by data correction methods (i.e., the trend or autocorrelation elimination intended by DA, TA, and GLS). Fitting the correct regression model and controlling trend and serial dependence whenever present does not ensure that the p values associated with the regression parameters are useful for making decisions regarding treatment effectiveness. It would be important to test whether the GLS estimation improves with the transformation of the independent variables proposed by Maddala and Lahiri (2007), prior to labeling the procedure's controlling step as ineffective. On the other hand, the proposal of McKnight, McKean, and Huitema (2000) might be a solution to the OLS problems when few serially dependent measurements are available, but it is not implemented in the most common statistical software avilable to applied psychologists.

The conclusions reached in the present study should be restricted to the procedures included and are limited by the type of effect studied (immediate and permanent) and by the two-phase design structure. Further research is needed to modify the regression-based procedures in order to improve their performance in single-case designs. These improvements need to have an effect on both p values and R^2 values, due to the utility of the latter (Cohen, 1990; 1994; Wilkinson & Task Force on Statistical Inference, 1999). Additionally, the new proposals with contrasted statistical properties ought to be implemented via programming codes in the most frequently used statistical packages.

Acknowledgements

Appendix

Design matrix $\mathbf{X_1}$ for the OLS level change model $y_t = b_0 + b_1 LC_t + e$, where LC is the dummy variable for level change, taking values as presented in the second column:

$$X_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ - & - \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}$$

Design matrix $\mathbf{X_2}$ for the OLS slope change model $y_t = b_0 + b_1 SC_t + e$, where SC is the dummy variable for slope change, taking values as presented in the second column; n_B is the phase B length:

$$X_2 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ - & - \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & n_B - 1 \end{bmatrix}$$

Design matrices for GLS estimation corresponding to the level change ($\mathbf{X_3}$) and slope change ($\mathbf{X_4}$) regression models; r_1 is the estimation of the first-order autocorrelation parameter ρ_1 obtained from the Durbin-Watson d statistic and PV represents the previous value in the series.

$$X_3 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ - & - \\ 1 & 1-r_1PV \\ 1 & 1-r_1PV \\ \vdots & \vdots \\ 1 & 1-r_1PV \end{bmatrix} \quad X_4 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ - & - \\ 1 & 0-r_1PV \\ 1 & 1-r_1PV \\ \vdots & \vdots \\ 1 & (n_B-1)-r_1PV \end{bmatrix}$$

Design matrices for differencing analysis corresponding to the level change ($\mathbf{X_5}$) and slope change ($\mathbf{X_6}$) regression models. Since differencing is performed on the dummy variables without distinguishing between phases, each dummy variable contains N−1 values instead of N values, where N is series length.

$$X_5 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ - & - \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{bmatrix} \quad X_6 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ - & - \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}$$

Appendix (continued)

Design matrix for the regression model $y_t = b_0 + b_1 T_t + e$ used as a first step of Trend analysis, where T is the time variable taking values as presented in the second column; n_A is the phase A length:

$$X_7 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & n_A \\ - & - \\ 1 & n_A+1 \\ 1 & n_A+2 \\ \vdots & \vdots \\ 1 & n_A+n_B \end{bmatrix}$$

Design matrix for the data generation model $y_t = \beta_0 + \beta_1 T_t + \beta_2 LC_t + \beta_3 SC_t + \varepsilon_t$, where T, LC, and SC are dummy variables for trend, level change, and slope change, respectively, taking values as shown in columns two, three, and four.

$$X_8 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & n_A & 0 & 0 \\ - & - & - & - \\ 1 & n_A+1 & 1 & 0 \\ 1 & n_A+2 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & n_A+n_B & 1 & n_B-1 \end{bmatrix}$$

References

Algina, J., & Keselman, H.J. (1998). A power comparison of the Welch-James and improved general approximation tests in the split-plot design. *Journal of Educational and Behavioral Statistics, 23*, 152-169.

Allison, D.B., & Gorman, B.S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy, 31*, 621-631.

Arnau, J., & Bono, R. (2004). Evaluating effects in short time series: Alternative models of analysis. *Perceptual and Motor Skills, 98*, 419-432.

Beretvas, S.N., & Chung, H. (2008). An evaluation of modified R^2-change effect size indices for single-subject experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*, 120-128.

Bradley, J.V. (1977). A common situation conducive to bizarre distribution shapes. *American Statistician, 31*, 147-150.

Brossart, D.F., Parker, R.I., Olson, E.A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30*, 531-563.

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*, 997-1003.

Crosbie, J. (1987). The inability of the binomial test to control Type I error with single-subject data. *Behavioral Assessment, 9*, 141-150.

DeCarlo, L.T., & Tryon, W.W. (1993). Estimating and testing correlation with small samples: A comparison of the C-statistic to modified estimator. *Behaviour Research and Therapy, 31*, 781-788.

Durbin, J., & Watson, G.S. (1971). Testing for serial correlation in least squares regression. III. *Biometrika, 58*, 1-19.

Edgington, E.S., & Onghena, P. (2007). *Randomization tests* (4th ed.). London: Chapman & Hall/CRC.

Ferron, J., & Ware, W. (1995). Analyzing single-case data: The power of randomization tests. *The Journal of Experimental Education, 63*, 167-178.

Fisher, W.W., Kelley, M.E., & Lomas, J.E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis, 36*, 387-406.

Gorsuch, R.L. (1983). Three methods for analyzing limited time-series (N of 1) data. *Behavioral Assessment, 5*, 141-154.

Greenwood, K.M., & Matyas, T.A. (1990). Problems with application of interrupted time series analysis for brief single-subject data. *Behavioral Assessment, 12*, 355-370.

Harrop, J.W., & Velicer, W.F. (1985). A comparison of alternative approaches to the analysis of interrupted time-series. *Multivariate Behavioral Research, 20*, 27-44.

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

Huitema, B.E., & McKean, J.W. (1998). Irrelevant autocorrelation in least squares intervention models. *Psychological Methods, 3*, 104-116.

Huitema, B.E., & McKean, J.W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*, 38-58.

Johnston, J.M., & Pennypacker, H.S. (2008). *Strategies and tactics of behavioral research* (3rd ed.). New York: Routledge.

Kazdin, A. (1978). Methodological and interpretive problems of single-case experimental designs. *Journal of Consulting and Clinical Psychology, 46*, 629-642.

Keselman, J.C., & Keselman, H.J. (1990). Analysing unbalanced repeated measures designs. *British Journal of Mathematical and Statistical Psychology, 43*, 265-282.

Kowalchuk, R.K., Keselman, H.J., Algina, J., & Wolfinger, R.D. (2004). The analysis of repeated measurements with mixed-model adjusted F tests. *Educational and Psychological Measurement, 64*, 224-241.

Maddala, G.S., & Lahiri, K. (2007). *Introduction to econometrics* (3rd ed.). London: John Wiley & Sons.

Manolov, R., & Solanas, A. (2008a). Comparing N= 1 effect size indices in presence of autocorrelation. *Behavior Modification, 32*, 860-875.

Manolov, R., & Solanas, A. (2008b). Randomization tests for ABAB designs: Comparing data-division-specific and common distributions. *Psicothema, 20*, 291-297.

Matyas, T.A., & Greenwood, K.M. (1990). Visual analysis for single-case time series: Effects of variability, serial dependence and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341-351.

Matyas, T.A., & Greenwood, K.M. (1997). Serial dependency in single-case time series. In R.D. Franklin, D.B. Allison, & B.S. Gorman (Eds.): *Design and analysis of single-case research* (pp. 215-243). Mahwah, NJ: Lawrence Erlbaum.

McKnight, S.D., McKean, J.W., & Huitema, B.E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychological Methods, 3*, 87-101.

Micceri, T. (1989). The unicorn, the normal curve and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Parker, R.I. (2006). Increased reliability for single-case research results: Is bootstrap the answer? *Behavior Therapy, 37*, 326-338.

Parker, R.I., & Brossart, D.F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy, 34*, 189-211.

Parker, R.I., Brossart, D.F., Vannest, K.J., Long, J.R., García De-Alba, R., Baugh, F.G., & Sullivan, J.R. (2005). Effect sizes in single case research: How large is large? *School Psychology Review, 34*, 116-132.

Sawilowsky, S.S., & Blair, R.C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin, 111*, 352-360.

Scheffé, H. (1959). *The analysis of variance*. New York, NJ: Wiley.

Serlin, R.C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods, 5*, 230-240.

Sharpley, C.F., & Alavosius, M.P. (1988). Autocorrelation in behavior data: An alternative perspective. *Behavioral Assessment, 10*, 243-251.

Simonton, D.K. (1977a). Cross-sectional time-series experiments: Some suggested statistical analyses. *Psychological Bulletin, 84*, 489-502.

Simonton, D.K. (1977b). Erratum to Simonton. *Psychological Bulletin, 84*, 1097.

Simonton, D.K. (1978). Erratum to Simonton. *Psychological Bulletin, 85*, 1000.

Solanas, A., Manolov, R., & Onghena, P. (2010). Estimating slope and level change in N= 1 designs. *Behavior Modification, 34*, 195-218.

Suen, H.K., & Ary, D. (1987). Autocorrelation in behavior analysis: Myth or reality? *Behavioral Assessment, 9*, 150-130.

Toothaker, L.E., Banz, M., Noble, C., Camp, J., & Davis, D. (1983). N= 1 designs: The failure of ANOVA-based tests. *Journal of Educational Statistics, 4*, 289-309.

Tryon, W.W. (1982). A simplified time-series analysis for evaluating treatment interventions. *Journal of Applied Behavior Analysis, 15*, 423-429.

White, O.R. (1974). *The «split-middle»: A «quickie» method of trend estimation*. Seattle, WA: University of Washington.

White, D.M., Rusch, F.R., Kazdin, A.E., & Hartmann, D.P. (1989). Applications of meta-analysis in individual subject research. *Behavioral Assessment, 11*, 281-296.

Wilkinson, L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 694-704.