Journal of Methods and Measurement in the Social Sciences,
Vol. 9, No. 2, 22-31, 2018

The Pressing Need to Test for Autocorrelation: Comparison of Repeated Measures ANOVA and Interrupted Time Series Autoregressive Models

Jay Schyler Raadt
The University of North Texas

Neglecting to measure autocorrelation in longitudinal research methods such as Repeated Measures (RM) ANOVA produces invalid results. Using simulated time series data varying on autocorrelation, this paper compares the performance of repeated measures analysis of variance (RM ANOVA) to interrupted time series autoregressive integrated moving average (ITS ARIMA) models, which explicitly model autocorrelation. Results show that the number of RM ANOVA signaling an intervention effect increase as autocorrelation increases whereas this relationship is opposite using ITS ARIMA. This calls the use of RM ANOVA for longitudinal educational research into question as well as past scientific results that used this method, exhorting educational researchers to investigate the use of ITS ARIMA.

Keywords: ANOVA, ARIMA, Autocorrelation, Longitudinal Research, Method Comparison

When studying the efficacy of an educational intervention, it is important to incorporate the aspect of time. Such longitudinal research designs show how a measured phenomenon changes due to an intervention that is hypothesized to enhance learning. Generally, a longitudinal research design has the advantage of increased statistical power, which reduces the probability of erroneously failing to reject a false null hypothesis, called a Type-II error (Shadish, Cook, & Campbell, 2001, p. 267). It is well established that researchers ought to consider the statistical power of their analyses (Cohen, 1962), which increase statistical conclusion validity of an intervention (Shadish, Cook, & Campbell, 2001, p. 45). Thus, the allure of longitudinal designs is understandable. Moreover, making an inference about behavior from just one pre-test and post-test lacks internal validity in the form of the testing effect and the instrumentation effect. In other words, if an instrument is prone to low reliability (instrumentation effect) or if a test-taker benefits from mere familiarity with the instrument (testing effect), then these effects will be more apparent in a repeated measures context (Shadish, Cook, & Campbell, 2001, p. 60). Overall, incorporating the aspect of time increases the internal validity and statistical conclusion validity of a research study.

Using a traditional repeated measures analysis of variance (RM ANOVA), the change in a measured phenomenon among a sample of individuals can be split between pre-intervention and post-intervention

phases using multiple pre-tests and post-tests. However, longitudinal data analysis is complicated because the past is a good predictor of the present and future. For example, an individual with a high score in math aptitude will likely have a high score in math aptitude one week later. The relationship between a measured phenomenon and itself in the past is called autocorrelation and it is a violation of the assumption of independence.

Autocorrelation is very common in longitudinal data analysis, but Scheffé (1959) shows that there is an increase in type-I error rate when observations are autocorrelated because autocorrelation violates the assumption of independence. However, Scheffé's proof is dense and mathematically esoteric. Many if not all applied educational researchers would not be able to critically examine this mathematical proof and, if anything, accept it on the basis of an argument from authority. Therefore, as an alternative to accepting Scheffe's proof on authority alone, this study tests Scheffé's proof using simulated data at levels of autocorrelation ranging from -0.4 to 0.4 in steps of 0.1. This simulated data is analyzed using RM ANOVA, which does not control for autocorrelation, and interrupted time series autoregressive integrated moving average (ITS ARIMA), which does control for autocorrelation. Thus, the hypothesis of this study is that as the autocorrelation of the simulated samples increase, RM ANOVA will identify a higher proportion of statistically significant differences than ITS ARIMA identifies.

Important Prior Knowledge

RM ANOVA. The procedure for conducting an RM ANOVA is similar to that of other statistical tests: state a hypothesis, set an alpha level of statistical significance, compute the test statistic, and interpret results. RM ANOVA experiments require random sampling, a normally distributed outcome variable in the population, homogeneity of variance, and equal correlation coefficients between measurement pairs in the population (Hinkle, Wiersma, & Jurs, 2003, pp. 357-363). Balanced design and time-invariant covariates are also requirements of RM ANOVA (Kwok, Underhill, Luo, Elliott, & Yoon, 2008).

The limitations of an RM ANOVA are well-established and stem from the requirements of the analysis. Random sampling is common to all experimental designs, but some quasi-experimental designs control enough for threats to validity such that an RM ANOVA results are acceptable (Shadish, Cook, & Campbell, 2001). A normally distributed outcome variable in the population is common to all analyses that fall under the general linear model, but if this assumption is not met, then the outcome variable can be transformed, perhaps using a Box-Cox Transformation (Osborne, 2013). Homogeneity of variance is a testable

assumption that when violated there are alternative statistical tests that still use an RM ANOVA design. Equal correlation coefficients between measurement pairs in the population, called sphericity, is also a testable assumption with corrections, namely Greenhouse-Geisser's and Huynh-Feldt's (Hinkle, Wiersma, & Jurs, 2003). A balanced design in RM ANOVA requires that all individuals in the study must have the same number of assessments and the time between these assessments must be equal. If an RM ANOVA is unbalanced, whereby some individuals have more assessments than others, these extra measurements must be dropped from the analysis. Finally, RM ANOVA requires time-invariant covariates, which are variables that do not change over time, such as place of birth, race, or sex, but it cannot accept time-variant covariates, such as level of education, marital status, or place of residence (Kwok et al, 2008).

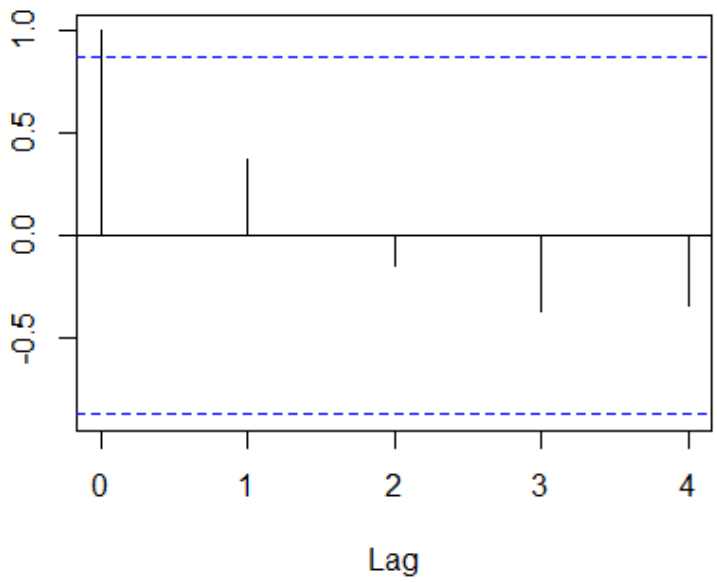
Autocorrelation. Autocorrelation is similar to other correlation coefficients, such as Pearson's r, but it is different in two ways. First, Pearson's r finds the direction and magnitude of similarity between *two different* observed variables whereas autocorrelation finds the similarity of an observed variable with *itself* in the past. Second, for any one variable, there are $n - 1$ autocorrelation coefficients because that is how many ways in which pairs of observations can be combined.

The pairs that the autocorrelation coefficient compares is defined by the 'lag,' k, of autocorrelation, which is how far in the past an observation is correlated with itself. The symbol for the autocorrelation coefficient is r_k and can be calculated using (1).

$$r_k = \frac{\sum_{t=1}^{T}(y_{k+t}-\bar{y})(y_t-\bar{y})}{\sum_{t=1}^{T}(y_t-\bar{y})^2} \quad (1)$$

For example, in a longitudinal data set $x = \{1,5,9,10,11\}$, when $k = 1$, autocorrelation measures the magnitude of similarity between the pairs (1, 5), (5, 9), (9, 10), and (10, 11), and $r_1 = 0.37$; when $k = 2$, autocorrelation measures the magnitude of similarity between the pairs (1, 9), (5, 10), and (9, 11) and $r_2 = -.15$; $k = 5$ is *null* because $5 > n_x - 1$. (Glass, Wilson, & Gottman, 1975/2008). Because there are $n - 1$ autocorrelation coefficients, the autocorrelation of a set of data is usually reported in a type of figure called a correlogram. The correlogram for the heuristic data set x appears in Figure 1.

Figure 1. A correlogram for the heuristic data set x. The blue dotted line represents the critical value for a statistically significant autocorrelation coefficient (Box, Jenkins, & Reinsel, 1994; Glass et al., 1975).



ITS ARIMA. ARIMA models are part of the larger generalized linear model framework. Unlike RM ANOVA, ITS ARIMA explicitly models autocorrelation by including an autoregressive coefficient, ϕ, where ϕ is a function of r_k. In addition to autoregression, ARIMA models also include an order of differencing and an order of moving averages, usually summarized in the form $ARIMA\,(p, d, q)$, where p is the order of autoregression, d is the order of differencing, and q is the order of moving averages (Glass, Wilson, & Gottman, 1975/2008).

ARIMA models can be used to show a trend over time, or they can be used to establish evidence for an intervention effect. Such an analysis uses a pre-intervention level, L, and a post-intervention level, $L + \delta$, and these levels are compared using a t-test (Glass, Wilson, & Gottman, 1975/2008, pp. 119-150).

Like all statistical models, ITS ARIMA has assumptions, principally homogeneity of variance and homogeneity of mean, together called the assumption of stationarity. This assumption is testable using the Augmented Dickey-Fuller test or the Phillips-Perron test, part of a class of statistical tests called unit root tests, where the null hypothesis is that there is a unit root present in the time series. If there is a failure to reject this null hypothesis, then the observations are not stationary and an order of differencing is introduced.

Data that follow a first order autoregressive model would be symbolized as $ARIMA(1,0,0)$ and can be expressed in matrix form such that $y_t = Xb + a_t$ as in (2), where the double subscript k is the k^{th} phase of the quasi-experiment (Glass, Wilson, & Gottman, 1975/2008, pp. 128-133).

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_k} \\ --- \\ y_{n_k+1} \\ y_{n_k+2} \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1-\phi & 0 \\ \vdots & \vdots \\ 1-\phi & 0 \\ --- & --- \\ 1-\phi & 1 \\ 1-\phi & 1-\phi \\ \vdots & \vdots \\ 1-\phi & 1-\phi \end{bmatrix} \begin{bmatrix} L \\ \delta \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n_k} \\ --- \\ a_{n_k+1} \\ a_{n_k+2} \\ \vdots \\ a_N \end{bmatrix}$$

L and δ have standard error

$$s_a^2 = \frac{a^T a}{N-2}$$

Moreover, L and δ are distributed on t with $df=N-2$:

$$\frac{\hat{L} - L}{s_a\sqrt{c^{11}}} \sim t_{N-2}$$

$$\frac{\hat{\delta} - \delta}{s_a\sqrt{c^{22}}} \sim t_{N-2}$$

where

$$c = (X^T X)^{-1}$$

Response to Intervention

One way to measure the efficacy of an educational intervention is with the Response to Intervention (RTI) protocol. In RTI, a sample of at-risk students is identified. Then, these students are monitored for a relatively short period using a standardized measurement. After this baseline phase, an intervention is implemented, such as a multi-tiered instructional approach. The students' progress is then measured in a follow-up phase. The RTI protocol reduces educational expenditures through a more valid process of labelling students as learning disabled (LD) and it helps to address the IQ-achievement discrepancy (Fuchs & Fuchs, 2006).

Method

Sample Simulation

An RTI scenario was simulated. The number of observations was set to 90 in the pre-intervention phase and 90 in the post-intervention phase. This would approximate a one semester pre-intervention phase and a one semester post-intervention phase (Jimerson, Burns, and VanDerHeyden, 2015).

The R package *forecast* was used to simulate this data. Using the *arima.sim* function, data sets were simulated for 66 participants over 180 observations at varying levels of ϕ, ranging from -0.4 to 0.4 in steps of 0.1. All data were screened to ensure they followed an $ARIMA$ (1,0,0) model. This method was replicated 100 times. These data allowed for the analysis of 900 RM ANOVA experiments, but because ITS ARIMA is suited for single-subject research designs, the same data allowed for the analysis of 59,400 quasi-experiments. To the best of my knowledge, there is no general implementation of ITS ARIMA quasi-experiments in R. Testing the intervention effect in an ARIMA (1,0,0) is different from testing the intervention effect in ARIMA (2,0,0), ARIMA (1,0,1), or any other combination of autoregression, integration, or moving average. Thus, I created a function to test for the intervention effect in R based on the matrix algebra recreated above, as originally reported by Glass, Wilson, & Gottman (1975/2008). R syntax for this simulation, including testing for the intervention effect, is available upon request.

Descriptive and Inferential Statistics

Although RM ANOVA intervention effects can be analyzed using a Cohen's d effect size, there is no such effect size measurement for ITS ARIMA. Instead, this study uses F-tests to analyze RM ANOVA results and t-tests to analyze ITS ARIMA (Glass, Wilson, & Gottman, 1975/2008; Hinkle, Wiersma, & Jurs, 2003). A statistically significant result is defined with $\alpha_{crit} = 0.05$. Pearson's r is used to analyze the relationship between statistically significant results and the level of ϕ. Because the only variable in the analyses is the level of ϕ, a simple regression does not add information, for in a simple regression $\beta = r$ and $R^2 = r^2$.

Results

When analyzing the simulated data using RM ANOVA there was an average of 6.56% ($SD = 5.85\%$) statistically significant results across levels of ϕ and an average of 44.47% ($SD = 0.03\%$) statistically significant ITS ARIMA quasi-experiments (Tables 1 and 2). The number of statistically

significant results according to RM ANOVA was strongly positively correlated with the level of ϕ, $r = 0.96, t(7) = 10, p < 0.01, 95\%\ CI[0.955, 0.965]$. This relationship was opposite with ITS ARIMA, where the number of statistically significant results was strongly negatively correlated with levels of ϕ, $r = -0.92, t(7) = -6.28, p < 0.01, 95\%\ CI[-0.919, -0.921]$. Thus, there is a positive relationship between the statistically significant results identified by RM ANOVA and the level of ϕ and the 95% confidence intervals of the two correlation coefficients do not overlap.

Table 1
Descriptive Statistics for Statistically Significant RM ANOVA Experiments

Level of ϕ	Statistically Significant Results (n)	Statistically Significant Results (%)	*SD*
-0.4	0	0%	0%
-0.3	2	2%	14%
-0.2	1	1%	10%
-0.1	2	2%	14%
0	6	6%	24%
0.1	8	8%	27%
0.2	10	10%	30%
0.3	14	14%	35%
0.4	16	16%	30%

Table 2
Descriptive Statistics for Statistically Significant ITS ARIMA Quasi-Experiments

Level of ϕ	Statistically Significant Results (n)	Statistically Significant Results (%)	*SD*
-0.4	3166	47.97%	50%
-0.3	3072	46.55%	50%
-0.2	3056	46.30%	50%
-0.1	3109	47.11%	50%
0	3083	46.71%	50%
0.1	2914	44.15%	50%
0.2	2798	42.39%	49%
0.3	2670	40.45%	49%
0.4	2550	38.64%	48%

Discussion

The number of statistically significant ITS ARIMA was much higher than those of the RM ANOVA. This may seem contradictory to the hypothesis that RM ANOVA produce more statistically significant results than ITS ARIMA. However, an important pattern emerges from Table A2. Notice that the number of statistically significant results using ITS ARIMA

across replications and across levels of ϕ has a small range, from 38% to 48%, and small range of variances. These variances were consistent, averaging 49% with a standard deviation of less than 1%. The large fluctuation around the number of statistically significant results is explained by the method of randomly generating autocorrelated data. In other words, when using ITS ARIMA, there was somewhere between 0% and 100% statistically significant results, which is consistent with the random chance of simulated data. For RM ANOVA though, the range of statistically significant results was larger, from 0% to 16%, and larger range of variances. These variances grew as the level of ϕ increased.

The positive correlation between levels of ϕ and number of statistically significant results for RM ANOVA is expected. Scheffe's proof shows that as autocorrelation increases, the number of type-I errors will also increase. However, the negative relationship between levels of ϕ and number of statistically significant results in ITS ARIMA is unexpected. This relationship means that there are more type-I errors in ITS ARIMA at lower levels of autocorrelation than at higher levels of autocorrelation. This highlights the fact that ITS ARIMA is not infallible. Thus, as in any analytical decision, the researcher ought to justify their use of analysis. Specifically, these results show that if the researcher is analyzing longitudinal data, before choosing either RM ANOVA or ITS ARIMA researchers must examine the level of autocorrelation. If no autocorrelation is present in the longitudinal data, then an RM ANOVA would have lower type-I error than ITS ARIMA. However, in the presence of autocorrelation, it is best practice to choose ITS ARIMA over RM ANOVA.

Future Research

Since the inception of the ANOVA framework at the turn of the 20th century many advances have been made. For example, we know that all ANOVA-type inferential statistics can be modelled in terms of regression (Cohen, 1968) and that most statistical analyses can be modelled using structural equation models if just the covariance matrix is available (Thompson, 2015; Zientek & Thompson, 2009). There are yet more advances to be made in understanding the general linear model (GLM) and generalizations of the GLM, of which ITS ARIMA is an example.

In comparing RM ANOVA to ITS ARIMA, just one viable alternative to the older RM ANOVA framework was considered. In addition to ITS ARIMA, which does have its disadvantages (Glass, Wilson, & Gottman, 1978/2005), there are other alternatives, including hierarchical linear models with autoregressive level-I error, called an HLM with AR(1) covariance structure, (Kwok, Underhill, Luo, Elliott & Yoon, 2008; Raudenbush & Bryk, 2002) or autoregressive latent trajectories (Bollen &

Curran, 2004), called ALT. ALT is more promising than HLM with AR(1) covariance structure because, like ITS ARIMA, it directly models the autocorrelation of a series rather than modelling the autocorrelation of error.

Limitations

This study used a large number of observations in order to ensure proper estimation of the AR parameter. Such a large number of observations is uncommon in educational research. However, the inflation of type-I error is a constant for any "large N" (Scheffé, 1959). The problem is that the estimated AR parameter is uncertain as the number of observations decreases and may not be statistically significantly different from zero. Future research may replicate this study but instead iterate the number of observations in addition to iterating levels of ϕ. Additionally, it may be of analytical importance that there were 59,400 ITS ARIMA quasi-experiments versus 900 RM ANOVA experiments. Future research should disentangle this issue, perhaps through bootstrapped sampling of RM ANOVA experiments and ITS ARIMA quasi-experiments to achieve equal groups.

Conclusions

The case against using RM ANOVA for certain cases of longitudinal research is strong. This study shows that RM ANOVA have inflated type-I error due to autocorrelation. However, the general ANOVA framework is still very common and many researchers feel comfortable using this framework. Thus, educational researchers who feel most content under the ANOVA framework must at least test for autocorrelation in longitudinal studies and if autocorrelation is present, this must be cited as a limitation in their published research. On the other hand, the comfortable researcher could consider learning, practicing, and mastering the use of models that control for autocorrelation, such as ITS ARIMA, HLM with AR(1) covariance structure, or ALT. Considering these alternatives to RM ANOVA will advance educational research.

Author Note: Jay S. Raadt, University of North Texas, Department of Educational Psychology, 1300 W. Highland St., Denton, TX 76203, 940.565.4646, Jay.Raadt@unt.edu

References

Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. (1994). *Time series analysis, forecasting and control*. Prentice-Hall, Englewood Cliffs.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*, 145–153. https://doi.org/10.1037/h0045186

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*, 426-433.

Dickey, D. A. and Fuller, W. A. (1979). of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association 74*, 427-431.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly*, *41*, 93–99. https://doi.org/10.1598/RRQ.41.1.4

Glass, G. V., Willson, V. L., & Gottman, J. M. (2008). *Design and analysis of time-series experiments*. Charlotte, NC: Information Age Publishing. Original work published 1975.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston, MA: Houton-Mifflin.

Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (2015). *Handbook of response to intervention: The science and practice of multi-tiered systems of support* (2nd ed.). New York: Springer.

Kwok, O. M., Underhill, A. T., Luo, W., Elliott, T. R. & Yoon, M. (2008). Analyzing longitudinal data with multilevel models: An example with individuals living with lower extremity intra-articular fractures. *Rehabilitation Psychology, 53*, 370-386.

Osborne, J. W. (2013). *Best practices in data cleaning*. London, UK: Sage.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical Linear Models Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage.

Scheffé, H. (1959). The Effects of Departures from the Underlying Assumptions in *The analysis of variance* (pp. 334-345). New York: Wiley and Sons.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Cengage Learning.

Thompson, B. (2015). The case for using the general linear model as a unifying conceptual framework for teaching statistics and psychometric theory. *Journal of Methods and Measurement in the Social Sciences, 6*, 30-41.

Zientek, L.R., & Thompson, B. (2009). Matrix summaries improve research reports: Secondary analyses using published literature. *Educational Researcher*, *38*, 343-352.