The importance of testing residual autocorrelation in longitudinal studies

Abstract

While the theory of longitudinal data analysis (LDA) has a solid foundation, there are instances where the assumptions of the analytical model remain unverified. Failure to examine autocorrelation in residuals (ACR) can elevate the risk of committing a Type I error, leading to the rejection of a true null hypothesis. This study compares two distinct analytical models within LDA: the polynomial (straight line, quadratic…) model and the autoregressive (AR) model. Three separate studies were conducted to investigate this comparison.

In Study 1, a real dataset was analyzed using a polynomial model, during which ACR was checked. In Study 2, the same dataset was reexamined using an AR model, followed by an ACR analysis. Study 3 involved contrasting the results of Studies 1 and 2 through a confounding test. Notably, the conclusions derived from Studies 1 and 2 diverged considerably. While Study 1 yielded significant inferences concerning the variable *Gender*, this significance was not replicated in Studies 2 and 3, likely attributable to a Type I error in Study 1.

In Study 3, the core independent variables (IVs) from Study 1, specifically *Day of the week* and *Day of the week squared* failed to garner support. Simultaneously, the AR IVs from Study 2 were validated. Consequently, this study underscores the advantages of the AR model, confirming its statistical and conceptual adequacy in the realm of LDA. The implications extend to considerations of enhancing data analysis in longitudinal studies.

Keywords: intensive longitudinal designs, pooled time series, panel data, clinical trials, autocorrelation in residuals

Introduction

In any field of psychology, more and more longitudinal studies are performed, with a variety of ways in which authors analyse the data (Asparouhov & Muthén, 2020, 2023; Box et al, 1970, 2016; Lee & Yu, 2015). However, not all analysis methods used are equally suitable, depending on the type of data analysed and the procedure used for its analysis. This research reviews the main longitudinal design systems: (a) contemplating longitudinal data analysis (LDA) encompassing single case designs (SCD) and data from multiple participants over time, known as intensive longitudinal designs (ILD); (b) discussing the challenge of autocorrelation and serial dependence in analyzing these data types; (c) explaining the consequences of not addressing autocorrelation in the original data; (d) conducting analyses in Study 1 (polynomial model) and Study 2 (AR, or autoregressive, model) on real longitudinal data to observe the effects of disregarding serial dependence and autocorrelation of raw data, as well as the absence of autocorrelation of residuals (ACR); (e) presenting a corrective data analysis procedure in Study 2, AR model, to address ACR in the same dataset analyzed in Study 1; (f) contrasting the outcomes of Studies 1 and 2 in Study 3, using a confounding test to determine the best fit for the model, confirming the presence of serial dependence in the original data and the 'white noise' nature of residuals in the polynomial model, Study 1; finally, (g) revealing how Study 1's analysis led to potentially mistaken inferences about the IVs, likely due to a Type I error.

Single subject designs and intensive longitudinal data analysis

The emergence of single subject designs and statistical analyses involving SCD in psychology stemmed from two key trends, firstly, it arose from the development of behavioral research (Skinner, 1963), secondly, it evolved from the development of LDA in engineering (Box et al., 1970, 2016).

The introduction of SCD designs raised questions regarding the most appropriate analysis for this data. Initially, analyses using mean contrasts with *t* or *F* tests were conducted (Gentile et al., 1972; Shine & Bower, 1971). Criticisms arose due to inappropriate analyses, particularly with the escalation of *t* or *F* values, potentially increasing the likelihood of committing a Type I error when autocorrelation (ACR) was present (Hartmann, 1974). This problem was initially pointed out by various authors (Aitken, 1934; Cochrane & Orcutt, 1949), who suggested utilizing Box-Jenkins ARIMA models (1970) to remove ACR in the data.

ILD systems, an extension of SCD (Walls & Schafer, 2006), involve recording data from multiple individuals at various times, measuring one or more variables simultaneously with regular periodicity (hours, days, weeks, etc.). Publications labeled as ILD encompass longitudinal data, intensive data, daily diary, experience sampling, ambulatory assessment, ecological momentary assessment, panel data, etc., depending on the field under investigation. The analysis of ILD has seen increased frequency in biology, psychology, and medicine due to technological advancements in record systems (Stinson, Liu & Dallery, 2022).

Autocorrelation in longitudinal data

Autocorrelation within longitudinal data presents persisting challenges in SCD designs and continues to persist in ILD. Specifically, this relates to the autocorrelation of original variable data, or serial dependence, and ACR (Arnau & Bono, 2003; Jones et al., 1977; Tong & Dubé, 2022). While the ACR has to be reviewed in univariate time series (Box et al., 2016), in ILD researchers often overlook this aspect, despite the fact that several authors recommend its verification (Bolger & Laurenceau, 2013; Singer & Willet, 2003).

The debate surrounding the consideration of ACR in psychology has been ongoing. In a review of previously published works within a clinical psychology journal, Huitema (1985) concluded that the presence of autocorrelation in the raw data does not impede the analysis of temporal data as independent from cross-sectional research (employing *F* or *t*-tests, regression, etc.), disregarding the potential ACR. Huitema's approach has been under scrutiny due to calculating the average of correlations found in reviewed articles, which averaged close to zero. Huitema's conclusions seem to disregard three significant aspects: the varying nature of autocorrelation in each case, the limited statistical power due to a small number of observations per phase (Box et al., 1970), and the improper standardization of correlation values (Matyas & Greenwood, 1991). Huitema's stance has been questioned by various authors, prompting a reconsideration of his earlier standpoint (Kazdin, 1982; Suen & Ary, 1987); simulation studies suggest that higher autocorrelation increases the probability of Type I errors (Hibbs, 1974; Huitema et al., 1999).

The 'pre-whitening' technique (Cochrane & Orcutt, 1949), suggested for removing autocorrelation in raw variables, has been found to lead to various statistical errors (Hamed, 2008); consequently, its utilization has diminished.

The consequences of residual autocorrelation

While in psychology there was a debate about the convenience of using AR models for LDA, Kmenta (1971, pp. 274-281) showcased, in a context of statistical economy, that in cases where ACR is significant, utilizing ordinary least squares (OLS) for parameter estimation leads to underestimated errors' variances. Consequently, the variances and standard errors of the parameters, included in the estimators' denominators, are similarly underestimated. This situation causes overestimation of values in t, z, F, R^2, b_0, b_1,… statistics, leading to an increased risk of Type I errors.

Thus, if ACR is not significant, using OLS or similar procedures becomes appropriate for correct parameter estimation in temporal regression.

In Kmenta's manual (1971), a straightforward illustration is provided regarding ACR when it follows an AR1 pattern (Bolger & Laurenceau, 2013). It begins by assuming a forecast model where residuals (e_t) are autocorrelated with a value of ρ, represented as $e_t = \rho e_{t-1} + \varepsilon_t$. This formulation leads to an evaluation of the variance of e_t:

$$\text{Var}(e_t) = \text{Var}(\varepsilon_t)/(1-\rho^2), \quad (1)$$

here, several aspects should be noted. The first is that if $\rho = 0$, then $\text{Var}(e_t) = \text{Var}(\varepsilon_t)$. The second aspect is that for values of $\rho \neq 0$, $\text{Var}(e_t) > \text{Var}(\varepsilon_t)$. Also, the value of $\text{Var}(e_t)$ will be larger the greater the absolute value of ρ, making it easier to make Type I errors; this confirms simulation studies with time series.

General hypothesis

The general hypothesis posits that in LDA (SCD, ILD, etc.), AR time series models prove more suitable than polynomial models (linear, quadratic, etc.) for accurate analysis. AR models are more adept at eliminating autocorrelation, thus minimizing Type I errors in parameter estimation. To test this hypothesis, Study 1 employs a polynomial model to analyze a dataset, followed by Study 2 which utilizes an AR model on the same data. Finally, Study 3 conducts a 'confounding test' to contrast the models from Studies 1 and 2.

Method

Procedure and data

A daily registration was completed during 45 days of lockdown for COVID-19 in Spain, from March 20th (fourth day after the start of the confinement in Spain) to May 3rd (end of the confinement) of 2020, which is 45 consecutive days, participants

were asked to respond to a daily survey comprising the MASQ-D30 and some day-to-day behaviors. On this research, we analyze the variable *Worthless*, belonging to the factor General Distress, in the Mood and Anxiety Symptom Questionnaire (MASQ-D30) Scale (Wardenaar et al., 2010). More information can be got on Flor et al. (2021). Data, input syntax and outputs are on the website repositori.uji.es/xmlui/handle/10234/204504 in SPSS format (IBM SPSS, 2022). For all analyses an $\alpha = .05$ was used.

Participants

The initial sample consisted of 319 participants recruited voluntarily through social media (web forums, WhatsApp, Twitter, and Facebook). Finally, 123 participants were selected from the total, because participants with less than 25 observations or non-consecutive registries were excluded. The research had the authorization CD/24/2020 of the university deontological commission.

Variables

Worthless: Consistent in the item 'During today, I felt worthless', an 11 points Likert scale with a possible answer between the values of 0 (I have not felt at all) and 10 (I have felt totally), we selected this variable because it is the best indicator of the factor General Distress, in MASQ-D30 Scale (Wardenaar et al., 2010), having the highest loading (.76) with its factor.

Gender: The variable, *Gender*, was coded as a categorical variable, with 0 for Male and 1 for Female, we put the option 2 for Other, but any of the participants answered this option. The sample included 40 men (32.5% of the total sample) and 83 women (67.5% of the total sample).

Age: Was measured in years, the final sample of 123 participants' mean age was 42.80 (between 21 and 75 years old), with a standard deviation of 10.35 years.

Day of the week and *Day of the week squared*: The *Day of the week* has been measured with a scale where Sunday is 1, Monday 2, and so until Saturday, that is 7; in the same way, *Day of the week squared* is for Sunday 1, for Monday is 4,… continuing until Saturday that is 49.

Study 1. Polynomial model

Hypothesis

In ILD, when we have multiple participants who are measured at different time points, one of the hypotheses of the model is multilevel (Raudenbush & Bryk, 2002; Singer & Willet, 2003), so that the temporal data, level 1, are nested within each participant, level 2. The substantive additional hypothesis at level 1 is that *Worthless* depends linearly on *Gender*, *Age*, *Day of the week* and *Day of the week squared*. At level 2, according to the data structure, there will be variation by participant in the intercept, $Var(u_0)$. Statistically (Raudenbush & Bryk; 2002), the current hypothesis will be:

$$Worthless_{jt} = (\gamma_{00} + u_{0j}) + \gamma_{10}Gender_j + \gamma_{20}Age_j + \gamma_{30}Weekday_t + \gamma_{40}Weekday^2_t + e_{jt}, \quad (2)$$

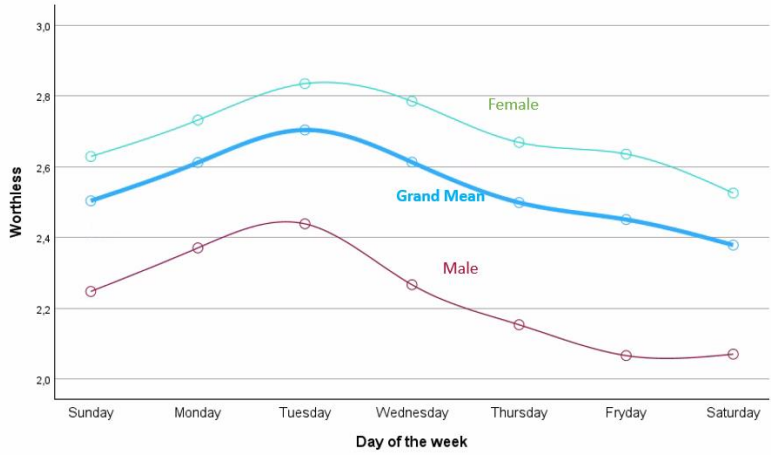
where the subscript *j* represents each participant in the study (*j*: 1, 2, … , 123), and the subscript *t* represents each measurement time point (*t*: 1, 2, … , 45). Note that the temporal IVs are *Day of the week* (*Time*: 1, 2, … , 7) and *Day of the week squared* ($Time^2$: 1, 4, … , 49).

Data analysis

The overall results for *Worthless* were *M*(*Worthless*) = 2.53 and *SD* = 1.48, ranging from 0-9. For the *Gender* variable, Male: *M*(*Worthless*) = 2.23, *SD* = 1.41, ranging 0-8; for Female: *M*(*Worthless*) = 2.68 and SD = 1.49, ranging 0-9. Female have bigger mean, standard and range. In Figure 1 we can see the means of *Worthless* by *Day of the week* and *Gender*.

Figure 1

Means of Worthless in function of Day of the week and Gender.



In Figure 1, the lines for Female are higher than for Male each day of the week, the grand mean is closer to Female because there are more women than men, and the Female and the Male lines are relatively equidistant, so there is not statistical interaction *Gender*×*Day of the week*. Note that the highest *Worthless* means are on Tuesday, and the smallest are on Saturday, but Men also on Friday.

Results

As a statistical reference, the unconditional model, Table 1 model M0, with an intercept at level 1 and also at level 2, has an AIC value of 14179.84, (σ_e^2) = .957, and its variance of the intercept at level 2, *Var*(u_0) = 1.328, so its intraclass correlation is .581 ($p < .001$), meaning approximately 58.1% of the variance of the *Worthless* variable is due to the similarity of the data within each participant.

Table 1

Statistical overall indicators for each model.

Model		−2LL[a]	Partrs[b]	Δ(-2LL); Δ(Partrs)[c]	*p*[d]	AIC	Δ(AIC)
Unconditional	M0	14175.85	3	−	−	14179.84	−
Study 1 Polynomial	M1	13851.82	8	(M1)−(M0): −324.03; 5	<.001	13857.82	(M1)−(M0): −322.02
Study 2 AR	M2	6205.81	14	(M2)−(M0): −7646.01; 11	<.001	6209.81	(M2) − (M1): −7648.01

[a] −2 Restricted Log Likelihood. [b] Number of parameters. [c] Increment in −2LL and in number the parameters. [d] *p* is the probability of the difference of models according to the difference of −2LL and of the number of parameters.

For this Study 1, we will follow the guidelines indicated by Bolger & Laurenceau (2013, 4th chapter), in the technical sections of the analysis; so we have used the option that the data structure consists of repeated measures each *Day of Lockdown* (SPSS Syntax: `REPEATED = Day of Lockdown`), that the covariance structure for each participant is AR1 (Syntax: `COVTYPE(AR1)`), and that the parameter estimation system was made by way of restricted maximum likelihood (REML). The covariance AR1 option in SPSS uses a generalized least squares (GLS) estimator of the parameters (Diggle et al., 2013; Verbeke & Molenberghs, 2000).

The statistical analysis gave the overall results of Table 1 M1[1], with an AIC of 13857.82, and −2LL = 13851.82, with 8 parameters, with a difference in −2LL compared to the unconditional model, M0: Δ(-2LL) = −322.02, Δ(*Parmtrs*) = 5, *p* <.001, indicating a good overall fit of the model to data. To compare the *goodness of fit* of two different models for the same data, when one of them is nested inside the other, we can use the Chi-squared statistical distribution, so M1 'vs' M0 can be compared with the results: [Δ(-2LL), Δ(Parameters)].

Table 2

Parameter estimates for polynomial model, Study 1, of Worthless as a function of Day of the week.

[1] Gender and Age are Level 2 variables, but in SPSS is not necessary to specify this aspect, because if within a higher level, a variable always repeats its value (i.e., *Age*) coinciding with the value of the higher level identifier (*Participant*), SPSS assumes that it is a level 2 variable, and estimates it as such. Compared to Mplus, the Age variable must be indicated as level 2.

A. Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	*df*	*t*	*p*	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept (γ_{00})	2.561	.476	123	5.378	<.001	1.618	3.503
Gender (γ_{10})	.485	.221	119	2.194	.030	.047	.923
Age (γ_{20})	−.011	.010	119	−1.054	.294	−.031	.009
Weekday (γ_{30})	.135	.036	3416	3.729	<.001	.064	.205
Weekday^2 (γ_{40})	−.020	.004	3341	−4.603	<.001	−.029	−.012

[a] Dependent Variable: *Worthless*.

B. Estimates of Random Effects, Variance Parameters[a]

	Parameter	Estimate	*SE*	Wald's *z*	*p*	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Level 1	Residual (σ_e^2)	.966	.022	44.830	<.001	.924	1.009
Repeated Measures	*AR1 rho*	.255	.015	17.294	<.001	.226	.284
Level 2	Variance, V*ar*(u_{0j})	1.269	.170	7.483	<.001	.977	1.649

[a] Dependent Variable: *Worthless*.

In Table 2A, the fixed effects estimation confirms that the mean intercept of *Worthless* differs from the value of 'zero' (γ_{00} = 2.561, p < .001), as does the effect of *Gender* (γ_{10} = .485, p = .030) on *Worthless,* but the effect of *Age* (γ_{20} = −5.92, p = .294) is not significant; we will include this variable with non-significant effects in the model. The coefficients of *Day of the week*, γ_{30}=.135, p <.001, and *Day of the week squared*, γ_{40}=−.020, p <.001, are both significant, showing that data fit a squared shape.

Regarding the level 2 results in Table 2B, we observ that the variance of the intercept is significant (*Var*(u_0)=1.269, p=<.001), indicating that the general intercept at level 1 (γ_{00}=2.561) has also a significant inter-subject variability. As for the other two results in Table 2B, the residual variance (σ_e^2=.966, p <.001) is the variance corresponding to the e_{jt} errors of the statistical model of Equation 2. The forecast error autocorrelation, *AR1 rho*, is .255 (p <.001), indicating that the average correlation of the forecast errors, per participant, has that value and is significant.

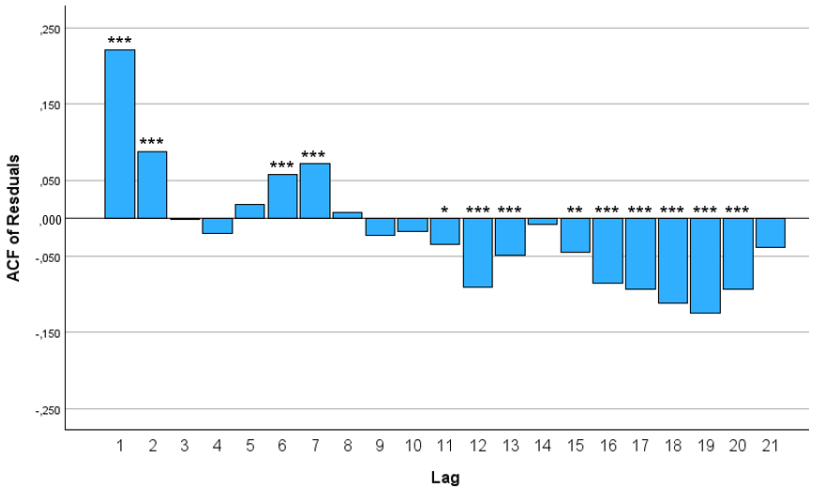
Residual analysis

To ensure accurate modeling of longitudinal data, it is essential that the residuals are "white noise", i.e., that they do not exhibit significant autocorrelation in their lags. We advise against an automatic analysis of autocorrelation function (ACF) and partial ACF (PACF) using statistical software on pooled grouped data residuals, as this method intermingles subjects' values, generating spurious correlations (e.g., merging the last value of one participant with the first value of another, and so on). Automatic ACF and PACF calculations are tailored for individual participant data, not aggregated or pooled data across different participants. A correct approach with SPSS involves manually calculated ACF and PACF as instructed in the syntax[i] for accurate analysis.

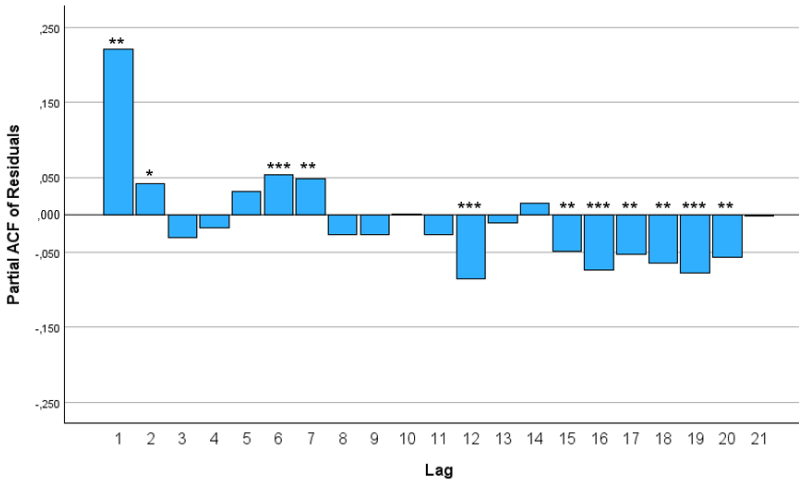
We will perform a residual analysis of Table 2 estimated values using pooled ACF and PACF of the first 21 residuals; the results are in Figure 2.

Figure 2

A. *Pooled ACF of the Study 2 residuals, Polynomial model*



B. *Pooled Partial ACF of the Study 2 residuals, Polynomial model*



*** $p < .001$, two tailed. ** $p < .01$, two tailed. * $p < .05$, two tailed.

A temporary delay is significant when both the ACF and the PACF are significant inside the same lag, note that the 11th delay is not, because the ACF is significant, but not its PACF; while that the residuals for the 1st, 2nd, 6th, 7th, 12th, and from the 15th to the 20th lags in ACF and PACF are both significant, indicating that the ACR are not "white noise", and Type I errors are likely to occur in this polynomial model, Study 1. We will analyse the same data with an AR model.

Study 2. Autoregressive model

The data analysed are the same as in Study 1. Our hypothesis is that the data follows an AR structure, with *Worthless* as a function of its own previous values ($Worthless_{jt} = f(Worthless_{jt-1}, Worthless_{jt-2}, \ldots, Worthless_{jt-7}, \ldots, Worthless_{jt-14}, Worthless_{jt-P7}, \ldots)$); this means that *Worthless* is function of immediate previous values, $Worthless_{jt-1}$, $Worthless_{jt-2}$, …, and of lagged 'seasonal' weekly values, $Worthless_{jt-7}$, $Worthless_{jt-14}$, … , $Worthless_{jt-P7}$, … (Flor et al., 2021; Rosel et al., 2019), being a model $AR(p)(P)_S$, where p is the number of immediate lags influencing the dependent variable (DV), and there will a seasonality of 7 days, or S=7, with a number of P lagged seasons. Note that in Study 1, $Worthless_{jt} = f(Weekday, Weekday^2)$, being *Weekday and* $Weekday^2$ the temporal IVs; but in Study 2, the temporal IVs are the auto-regressed values of

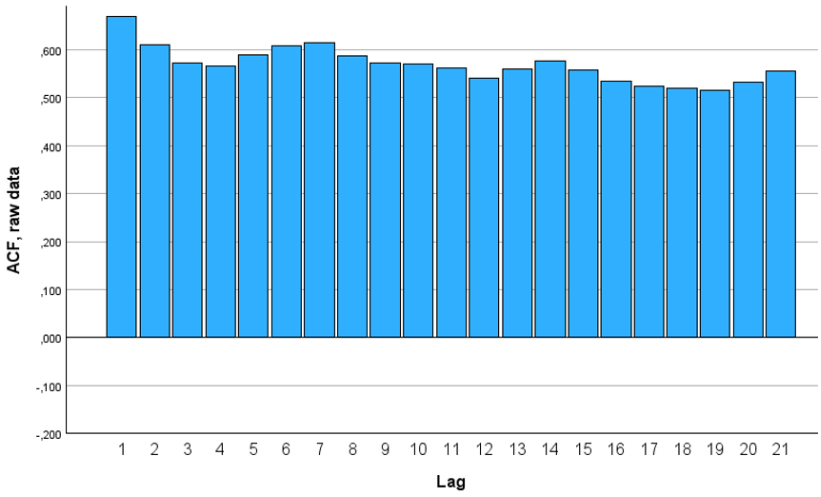
Worthless, or $Worthless_{jt} = f(Worthless_{jt-1}, Worthless_{jt-2,\ldots})$. In order to mirror the Study 1, *Worthless* will vary according to *Gender* and *Age*.

Similarly, at the between-subject level, or level 2, the intercept will vary according to the participant, i.e., $Var(u_0)$; being the general equation of the AR general hypothesis:

$$Worthless_{jt} = (\gamma_{00} + u_{0j}) + \gamma_{10}Gender_j + \gamma_{20}Age_j + \gamma_{30}Worthless_{jt-1} + \gamma_{40}\,Worthless_{jt-2}\ldots + \gamma_{70}Worthless_{jt-7} + \gamma_{140}Worthless_{jt-14} + \ldots + \gamma_{P70}Worthless_{jt-P7} + e_{jt} \quad (3)$$

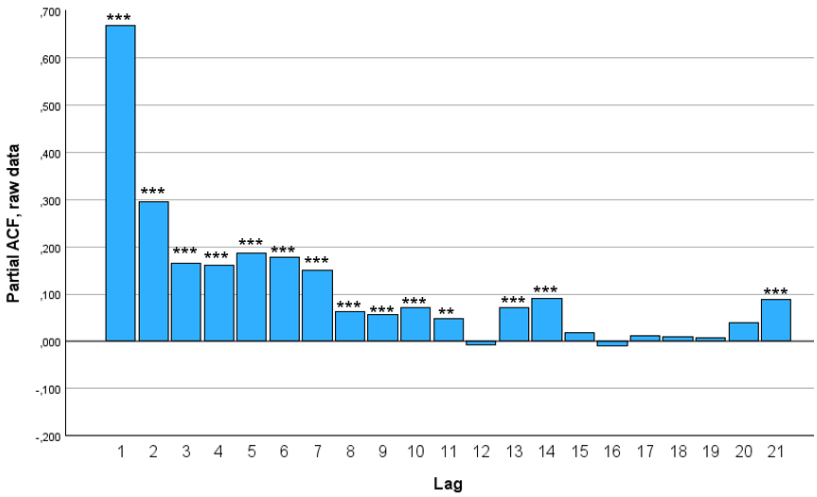
Figure 3

A. *Exploratory pooled ACF of the raw data*



Note. All the correlations are significant, having p <.001.

B. *Exploratory pooled Partial ACF of the raw data*



*** p < .001, two tailed. ** p < .01, two tailed.

To check the serial dependence of *Worthless*, we conducted an exploratory analysis of pooled ACF and PACF of the original raw data, with the results shown in Figure 3, which confirms that the data follow a long AR structure, observe that most correlations are significant and, in relation to the hypothesis about the seasonality of 7 days, it can be observed that lags 7^{th}, 14^{th} and 21^{st} are significant in the ACF and in the PACF.

Results

The same system as in Study 1, REML, was used for parameter estimation. It was not indicated that the data were repeated measures, nor that the covariance structure was AR, since these aspects were already explicitly included in Equation 3 and the corresponding analysis model.

The overall results are in Table 1; the −2LL is 6205.81 for the regression of this AR M2, which compared to the unconditional M0 model: Δ(-2LL) = −7646.01, $\Delta(df) = 11$, p <.001, indicating a significant overall fit; and the AIC value, model is 6209.81. M2 'vs' M0 can be compared with a Chi-squared test throw the −2LL values because M0 is nested inside M2. When they are not nested models, which is our case for comparing M2 'vs' M1, because the IVs are different in each model, we will use the Burnham and Anderson procedure (Burnham et al., 2011), based on the Akaike information criterion (AIC), which establishes that if $\Delta(AIC) = AIC_A - AIC_B$, when $\Delta(AIC) > |7|$, then the model with the largest value is not supported; so, we can assess that AR M2 model of Study 2 is much better than polynomial M1 model, being $\Delta(AIC) = |-7648.01|$.

Table 3

Parameter estimates for AR model, Study 2.

A. Estimates of Fixed Effects[a]

					95% Confidence Interval	
Parameter	Estimate	*SE*	*t*	*p*	Lower Bound	Upper Bound

Intercept (γ_{00})	.287	.096	2.999	.003	.099	.475
Gender (γ_{10})	.000	.041	−.009	.993	−.081	.080
Age (γ_{20})	−.003	.002	−1.557	.120	−.007	.001
$Worthless_{jt-1}$ (γ_{30})	.284	.020	13.907	<.001	.244	.324
$Worthless_{jt-2}$ (γ_{40})	.068	.021	3.226	.001	.027	.109
$Worthless_{jt-3}$ (γ_{50})	.076	.021	3.567	<.001	.034	.117
$Worthless_{jt-4}$ (γ_{60})	.030	.021	1.420	.156	−.012	.073
$Worthless_{jt-5}$ (γ_{70})	.116	.022	5.365	<.001	.074	.158
$Worthless_{jt-6}$ (γ_{80})	.083	.022	3.801	<.001	.040	.126
$Worthless_{jt-7}$ (γ_{90})	.115	.021	5.410	<.001	.073	.157
$Worthless_{jt-14}$ (γ_{100})	.072	.020	3.682	<.001	.034	.111
$Worthless_{jt-21}$ (γ_{110})	.095	.018	5.231	<.001	.059	.130

[a] Dependent Variable: *Worthless*.

B. Estimates of Random Effects, Variance Parameters[a]

					95% Confidence Interval	
Parameter	Estimate	*SE*	*Wald Z*	*p*	Lower Bound	Upper Bound
Level 1, Residual (σ_e^2)	.833	.025	33.919	<.001	.786	.882
Level 2, Intercept Var(u_0)	.000[b]	.000	−	−	−	−

[a] Dependent Variable: *Worthless*. [b] This covariance parameter is redundant; the test statistic and confidence interval cannot be computed.

The parameters results are shown in Table 3. Table 3A shows that the AR coefficients are significant, indicating a strong serial dependence of 21 days, or three weeks. Only the fourth lag, corresponding to $Worthless_{jt-4}$, is not significant, but we have preferred to include it, because if there is a subsequent significant simple coefficient, the fifth, $Worthless_{jt-5}$, it is more correct to leave the previous ones although they are not significant (Box & Jenkins, 1970), being a model $AR(6,3_7)$. Doing $Worthless_{jt}$ equivalent to Y_{jt} for saving space, the general equation in Table 3 will be:

$$Y_{jt} = (\gamma_{00} + u_{0j}) + \gamma_{10}Gender_j + \gamma_{20}Age_j + \gamma_{30}Y_{jt-1} + \gamma_{40}Y_{jt-2} + \ldots + \gamma_{80}Y_{jt-6} + \gamma_{90}Y_{jt-7} + \gamma_{100}Y_{jt-14} + \gamma_{110}Y_{jt-21} + e_{jt},$$

or, because $\mathrm{Var}(u_0) = 0$, Table 3B:

$$Y_{jt} = .287 + .000Gender_j - .003Age_j + .284Y_{jt-1} + .068Y_{jt-2} + .076Y_{jt-3} + .030Y_{jt-4} + .116Y_{jt-5} + .083Y_{jt-6} + .115Y_{jt-7} + .072Y_{jt-14} + .095Y_{jt-21} + e_{jt} \quad (4)$$

In Table 3A, we see that the intercept is significant, differing from the value of 'zero', γ_{00} =.287, p=.003, that the coefficient of *Gender* is practically 'zero' and not

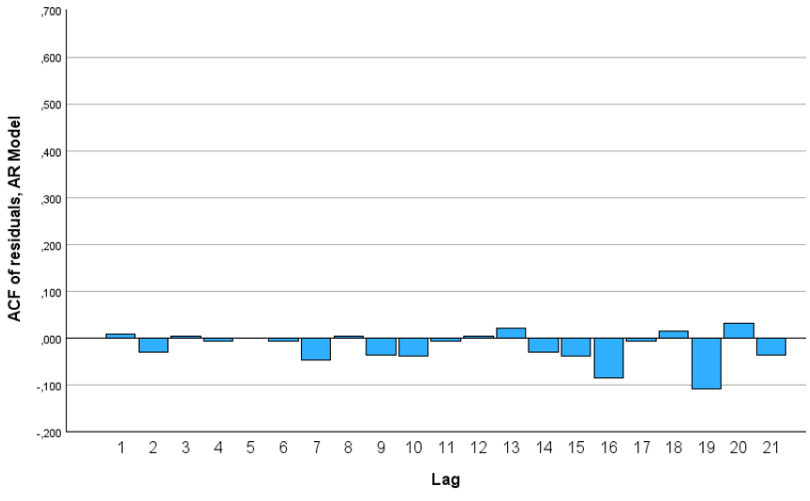
significant $\gamma_{10} < .000$, $p=.993$, and the variable *Age* neither is significant, $\gamma_{10} = -.003$, $p=.120$. Remember that in Study 1, Table 2A, the effect of Gender was significant, but here no.

Each participant has their starting level, although Equation 4 did not result in a multilevel model in the intercept, with its level 2 variance equal to zero, Var(u_0) = 0 (Table 3B), which is probably due to the fact that the AR model, Equation 4, estimates the starting intercept based on the previous values of $Worthless_{jt}$ in Equation 4 ($Worthless_{jt-1}$, $Worthless_{jt-2}$, $Worthless_{jt-3}$, …, $Worthless_{jt-7}$, $Worthless_{jt-14}$, $Worthless_{jt-21}$,), thus the possible multilevel intercept effect has already been included in the initial values of the AR model's predictions.

Residual analysis

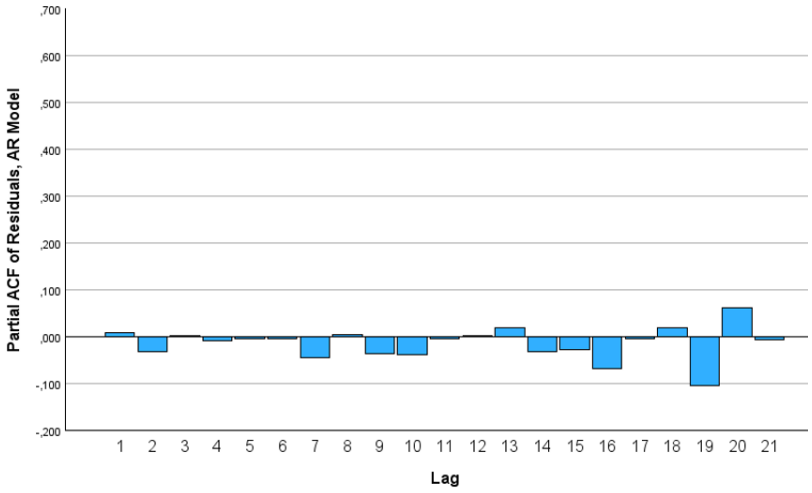
Regarding the residuals of Study 2, the AR model, the values of the ACF and PACF[2] are presented in Figure 4. An important aspect to note is that none of the ACF or PACF values of the AR model residuals are significant, being 'white noise'. Therefore, the parameters obtained in Table 3, AR Study 2, are more reliable than the parameters of polynomial Study 1, Table 2, preventing the AR model against the risk of committing Type I errors. **Figure 4**

A. *Pooled ACF of residuals from the AR model, Study 2*



B. *Pooled Partial ACF of residuals, AR model*

[2] We have used the same scale in Figures 3 and 4 (−.200 to +.700), with the aim of making them directly comparable, being the real ACF and PACF values smaller in Figure 4 than in Figure 3.



Note. All the correlations in Figure 4A and 4B are not significant, the residuals are 'white noise'.

Study 3. Confounding test

We had the doubt about what is the best model describing our longitudinal DV *Worthless*, for resolving this question, we will do a 'confounding test', initially used in epidemiology (Maldonado & Greenland, 1993; Clayton & Hills, 2013), the confounding test has been related to causation in statistical theory (Imbens & Rubin, 2015).

There are different versions of the test, but when in regression there exists two sets of competing IVs, polynomial versus AR, it is necessary to run each set of 'crude' IVs separately, getting a 'crude' regression model for each of them, and finally it is necessary to run the two sets of IVs altogether, the 'adjusted' test; the IVs that in de adjusted and the crude model regression are very similar, are the real IVs of the DV to explain; if the crude and adjusted IVs estimators are very different, they are confounding IVs (Rolf et al., 2013; VanderWeele et al., 2021); importantly, this verification process does not necessitate formal statistical tests but entails a careful comparison of estimated parameters, ensuring robust causal inference.

The 'confounding test' holds a pivotal role in the realm of causal inference. From a causal perspective, it addresses the fundamental question of whether an observed relationship between an independent variable (IV) and a dependent variable (DV)

reflects true causation or if it's distorted by the presence of confounding variables. Confounders are lurking variables that obscure the causal pathway, leading to potentially erroneous conclusions (Westreich, 2020). The test identifies true causal IVs by revealing their consistency across both models, while confounding IVs manifest as significant discrepancies. In doing so, this approach enhances our ability to establish robust causal links and ensures that our regression models accurately represent the underlying causal mechanisms, a critical aspect of rigorous scientific inquiry (Pearl, 2009; Wysocki et al., 2022).

The data analysed are the same as in Studies 1 and 2. In our context, the two 'crude' models are the polynomial of Study 1 and the AR of Study 2, we will run the 'adjusted' test on this Study 3, constituted by all the level 1 and level 2 parameters and variables included in the 'crude' models of Study 1 and Study 2. The results of this confounding test Study 3 are in Table 4.

Table 4

Parameter estimates for Study 3, adjusted model, integrated by the polynomial Study 1, and the AR Study 2 models.

A. Estimates of Fixed Effects[a]

					95% Confidence Interval	
Parameter	Estimate	*SE*	*t*	*p*	Lower Bound	Upper Bound
Intercept (γ_{00})	.374	.127	2.948	.004	.124	.624
Gender (γ_{10})	.004	.044	.099	.921	−.082	.091
Age (γ_{20})	−.003	.002	−1.565	.120	−.007	.001
$Worthless_{jt-1}$ (γ_{30})	.223	.020	10.915	<.001	.183	.263
$Worthless_{jt-2}$ (γ_{40})	.084	.021	4.039	<.001	.043	.124
$Worthless_{jt-3}$ (γ_{50})	.088	.021	4.230	<.001	.047	.129
$Worthless_{jt-4}$ (γ_{60})	.044	.021	2.076	.038	.002	.086
$Worthless_{jt-5}$ (γ_{70})	.119	.021	5.589	<.001	.077	.161
$Worthless_{jt-6}$ (γ_{80})	.088	.022	4.109	<.001	.046	.131
$Worthless_{jt-7}$ (γ_{90})	.119	.021	5.597	<.001	.077	.160
$Worthless_{jt-14}$ (γ_{100})	.072	.020	3.642	<.001	.033	.112
$Worthless_{jt-21}$ (γ_{110})	.094	.018	5.114	<.001	.058	.130
Weekday (γ_{120})	.015	.047	.320	.749	−.077	.107
$Weekday^2$ (γ_{130})	−.006	.006	−1.053	.293	−.017	.005

[a] Dependent Variable: *Worthless*.

B. Estimates of Random Effects, Variance Parameters[a]

Parameter		Estimate	*SE*	Wald's *z*	*p*	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Level 1	Residual (σ_e^2)	.832	.025	32.636	<.001	.783	.883
Repeated Measures	*AR1 rho*	.065	.065	1.011	.312	-.062	.191
Level 2	Variance, V*ar(u₀)*	.000[b]	.000	–	–	–	–

[a] Dependent Variable: *Worthless*.
[b] This covariance parameter is redundant. The test statistic and confidence interval cannot be computed.

Results

Doing a comparison between the results in Table 2, the polynomial model, and the adjusted model in Table 4, we see that the next variables in Level 1: *Gender, Day of the week,* and *Day of the week squared,* that were significant in Table 2A, are not significant in the adjusted model, or confounding test; at level 2, Tables 2B and 4B, we can observe that two significant variables in Table 2B, the *AR1 rho* and the variance of the intercept, V*ar(u₀)*, now in Table 4B are not significant.

Comparison of results

Likewise, comparing the AR model in Table 3A, with the confounding test in Table 4A, we see that *Age, Gender,* and the AR variables have almost the same values, being significant or not in both tables, almost as if there were a copy form one to the other table. In Table 4B we see that the *AR1 rho* and the intercept variance, V*ar(u₀)*, are both not significant; in Table 3B, the *AR1 rho* parameter has not been included by hypothesis and V*ar(u₀)* is also not significant.

In Study 3, we have not included the ACF or the PACF of the residuals; this decision was made as our primary objective was to compare polynomial and AR models, and the focus of this analysis did not involve these specific measures. Also, we

have not included these figures in order to save space; but it is worth noting that the residuals in Study 3 closely resemble those of the AR model in Study 2, this similarity can be observed by comparing Tables 2 and 3 results, including the residual variances (σ_e^2) of models 2 and 3.

Another peculiarity of the polynomial model, in the case of longitudinal data, is that a more complex GLS estimator could be used. The correct way would be to use an AR matrix with the significant structure of ACF and PACF for obtaining the correct GLS regressor estimators. So, we can conclude that the statistical and confounding tests verify that the AR model is more adequate to our longitudinal data than the polynomial model, despite being the most widely used model in psychology.

Discussion

In tracing the brief historical trajectory of LDA in psychology from the 1960s and 1970s, the diverse array of methods employed highlights that the commonly utilized polynomial model is not always optimal. Our comparison between polynomial and AR models underlines the superiority of the AR model in eliminating ACR, resulting in more precise and reliable IVs parameters for VD forecasting.

Based on our data, several statistical aspects favor the AR model's suitability. Despite its increased complexity and additional variables, the AR model significantly exhibits a lower AIC compared to the polynomial model. Notably, the AIC calculation penalizes larger regression models due to increased variables, making the AR model's lower AIC particularly significant.

Another advantageous aspect favoring the AR model is revealed through a ‘confounding test’. This test involved an 'adjusted' regression, encompassing both polynomial and AR IVs as explanatory factors for the DV at levels 1 and 2. Notably, only the AR variables displayed significant parameters, indicating that the IVs from the

polynomial model acted as confounding variables, exerting negligible influence on the DV. In summary, it is evident that a standalone polynomial regression (or any cross-sectional analysis, such as *t*-test, *F*-test, polynomial regression, etc.) rarely fully eradicates autocorrelation when the original data displays such tendencies. It is more likely that the IVs in the AR model of Study 2 are trustworthy for the daily variability of *Worthless* than the IVs in the polynomial model of Study 1 (Pearl, 2009; VanderWeele et al., 2021).

The previous statistical insights hold significant implications for results interpretation and application. For instance, in Study 1 (polynomial model), the variable *Gender* was significant, while in Studies 2 and 3, it did not. To illustrate the impact, let's consider a clinical pharmacological trial testing the effectiveness of a *Drug X* on a *Pathology Y*, and in our data, being *Drug X* the variable *Gender*, as well as the variable *Pathology Y* being the DV *Distress*. If the data were analyzed solely using a polynomial regression model (Study 1), the results would suggest a significant improvement in *Pathology Y* due to the *Drug X*, although in reality, this improvement does not exist, as confirmed by the AR model (Table 3A) and the confounding test (Table 4A) in Studies 2 and 3. These misleading statistical findings could have substantial consequences, potentially leading researchers in Study 1 to recommend the *Drug X* as an effective treatment for *Pathology Y* when, in reality, it is not.

In the same vein, applying the Study 1 inter-individual level 2 results suggests participant differences in intercept levels, which appear statistically significant ($Var(u_0)$ = 1.269, $p < .001$, Table 2B). However, Study 2's findings indicate no such variance ($Var(u_0) = 0$, Table 3B). This discrepancy is linked to the "inertia" effect stemming from the prior *Worthless* levels over the preceding 21 days (3 weeks!).

The bias of underestimating forecast error variance in polynomial models does not just affect an auxiliary variable like the IV Gender; it extends to the 'core' IVs, such as *Day of the week* and *Day of the week squared*. While both variables appeared significant in Study 1, their significance faded in Studies 2 and 3. It is highly probable that these IVs are indeed not significant, and their apparent significance is likely a Type I statistical error, rendering *Gender*, *Day of the week*, and *Day of the week squared* as spurious confounding IVs. These findings echo a previous applied investigation wherein, within the same day, *Hour* and *Hour squared* were employed as polynomial IVs with *Salivary alpha*-amylase as the DV, suggesting that the AR model is more fitting (Rosel et al., 2019).

The AR model outcomes ($AR(p,P_S)$ or $AR(6,3_7)$ model in our case), as illustrated in Table 3 or Equation 4, signify that a participant's feelings of *Worthless* on a specific day, for instance, a Tuesday ($Worthless_{jt}$), are influenced by various time-lagged *Worthless* values. These include the *Worthless* from the preceding day ($Worthless_{jt-1}$), two days prior ($Worthless_{jt-2}$), extending up to six days earlier ($Worthless_{jt-6}$). Additionally, it factors in the *Worthless* experienced seven days before ($Worthless_{jt-7}$), corresponding to the same weekday of the prior week. Furthermore, it considers *Worthless* from 14 days earlier ($Worthless_{jt-14}$), reflecting the Tuesday two weeks back, and the *Worthless* from 21 days before ($Worthless_{jt-21}$), indicating the Tuesday three weeks ago. Notably, the last lag, the 21st is derived from the product of the number of periods (P) and the number of days of the week (S), equaling 3 periods (periods or weeks) multiplied by 7 (days), totaling 21 days. These results confirm that human mood behavior exhibit a prolonged inertia, sometimes persisting not just for days but across several weeks (Flor et al., 2021).

In summary, we have established the superiority of the AR model over the polynomial model in ILD analysis. The AR model offers enhanced flexibility, adapting to diverse temporal patterns without imposing a fixed form on the data, accommodating not only linear or quadratic trends but also capturing periodic influences, such as weekly variations. In contrast, the conventional polynomial model remains pervasive in psychology (Cohen et al., 2021; Øverup et al., 2020), very frequent in epidemiology (Amar et al., 2020; Waterfield, 2023) and pharmacology (Hill et al., 2023; Keenan et al., 2023) studies. Surprisingly, AR studies are infrequent in these fields, and even rarer are investigations examining the ACF and the PACF of the residuals.

In our research, we have emphasized the significance of assessing the ACF and PACF) of the raw data (see Figure 3). However, we believe it is of greater importance to focus on the ACF and PACF of the residuals (see Figure 2 for the polynomial model and Figure 4 for the AR model). There are a couple of reasons for this emphasis; firstly, the ACF and PACF of the raw data represent a preliminary exploratory analysis of the potential AR model of the data. This analysis may be relatively misleading if additional IVs, such as Gender and Age in our case, are introduced. The exploratory ACF and PACF of raw data may not account for these other IVs, potentially influencing the temporal data; secondly, in psychological and social studies, unlike many physical or biological sciences, there may not be a clear behavioral regularity. As a result, the ACF and PACF can sometimes appear more complex in psychology compared to the final model. Therefore, to ensure the validity of the results derived from the proposed model, it is imperative to investigate the ACF and PACF of the residuals. This helps confirm that the residuals exhibit a white noise pattern, thereby preventing autocorrelation and the risk of biased parameter estimation, with the possibility of Type I errors.

For precise model estimation, researchers require in-depth knowledge of the field under study, essential for formulating hypotheses regarding immediate and seasonal AR temporal effects; and the ACF and PACF of the residuals should ideally reflect a statistically not significant departure from the value of 'zero' in each lag.

In statistical analyses of SCD or ILD, a crucial issue arises when the number of records is limited, with fewer than 50 records leading to reduced statistical power to reject the null hypothesis; pooled data, on the other hand, offers higher statistical power. Only when the observations per subject in pooled data fall below ten data is there a need for more specialized estimation methods (Arellano & Bond, 1991; Jin & Lee, 2012; Lee & Yu, 2015).

In summary, the preference for an AR model in LDA emerges from several compelling reasons: (a) human behavior's reliance on past values signifies an AR nature, especially in cognitive, physiological, affective, and habitual aspects; traditional statistical models, based on the independence of measurements, lose validity in LDA, where linear dependence among data points is observed; (b) our LDA indicates that the statistical fit of the AR model (Study 2) is substantially better than that of the polynomial model (Study 1); (c) polynomial models (linear, quadratic, etc.) used in LDA are prone to Type I errors due to the non-white noise nature of the residuals; (d) furthermore, the confounding test (Study 3) shows the AR model's superior causal adequacy over the polynomial model, ultimately; (e) once again, it is recommended to perform ACF and PACF of the residuals obtained, in order to check that they are 'white noise', and thus avoid Type I errors. Therefore, caution is warranted regarding results from LDA that overlook data serial dependence or fail to account for autocorrelation. Such oversights might lead to the discovery of statistically significant effects that are not objectively present in reality.

Future years are expected to see a rise in longitudinal research studies. However, critical areas need attention: (a) implement specialized training programs in longitudinal data analysis for methodologists, covering time series and temporal data analyses; (b) research groups should involve data analysis experts, fostering collaboration between researchers and specialists; (c) scientific publications should establish standards for the review of longitudinal data analysis and incorporate expert reviewers in this methodology (Hardwicke et al., 2019). It is to be expected that more ILD will be published in the coming years, but also that the quality standards of the publications will be improved.

References

Aitken, A.C. (1936). On least-squares and linear combinations of observations. *Proceedings of the Royal Society of Edinburgh*, *55*, 42–48. doi:10.1017/S0370164600014346

Amar, L.A., Taha, A.A., & Mohamed, M.Y. (2020). Prediction of the final size for COVID-19 epidemic using machine learning: A case study of Egypt. *Infectious Disease Modelling*, *5*, Pages 622-634. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7446670/

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies*, *58*(2), 277–297. https://doi.org/10.2307/2297968

Arnau, J., & Bono, R. (2003). Autocorrelation problems in short time series. *Psychological Reports*, *92*(2), 355–364. DOI: 10.2466/pr0.2003.92.2.355

Asparouhov, T., & Muthén, B. (2020). Comparison of models for the analysis of intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(2), 275-297. DOI: 10.1080/10705511.2019.1626733

Asparouhov, T., & Muthén, B. (2023). Residual structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(1), 1-31. DOI: 10.1080/10705511.2022.2074422

Bolger, N., & Laurenceau, J.P. (2013). *Intensive longitudinal methods*. New York, NY: The Guilford Press.

544 Box, G.E.P., & Jenkins, P.M. (1970). *Time series analysis: forecasting and control*. San
545 Francisco: Holden-Day.
546 Box, G.E.P., Jenkins, G.M., Reinsel, G.C., & Ljung, G.M. (2016). *Time series analysis:*
547 *Forecasting and control* (5th ed). Hoboken, NJ: Wiley.
548 Burnham, K.P., Anderson, D.R., & Huyvaert, K.P. (2011). AIC model selection and
549 multimodel inference in behavioral ecology: Some background, observations, and
550 comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35.
551 https://doi.org/10.1007/s00265-010-1029-6
552 Clayton, D., & Hills, M. (2013). *Statistical models in epidemiology*. Oxford: Oxford
553 U.P.
554 Cochrane, D., & Orcutt, G. H. (1949). Application of least squares regression to
555 relationships containing auto-correlated error terms. *Journal of the American*
556 *Statistical Association*, *44*(245), 32–61. doi:10.1080/01621459.1949.10483290
557 Cohen, K., Ramseyer, F.T., Tal, S., & Zilcha-Mano, S. (2021). Nonverbal synchrony
558 and the alliance in psychotherapy for major depression: Disentangling state-like and
559 trait-like effects. *Clinical Psychological Science*, *9*(4), 634–648.
560 https://doi.org/10.1177/2167702620985294
561 Diggle, P.J., Heagerty, P., Liang, K-Y., & Zeger, S.L. (2013). *Analysis of longitudinal*
562 *data* (2nd ed.). Oxford: Oxford U.P.
563 Flor, P., Rosel, J.F., Ferrer, E., Barrós, A. & Machancoses, F.H. (2021). Longitudinal
564 effects of distress and its management during COVID-19 lockdown in
565 Spain. *Frontiers in Psychology*, *12*:772040.
566 https://doi.org/10.3389/fpsyg.2021.772040
567 Gentile, J.R., Roden, A.H., & Klein, R.D. (1972). An analysis-of-variance model for
568 the intrasubject replication design. *Journal of Applied Behavior Analysis*, *5*, 193–
569 198.
570 VanderWeele, T.J., Rothman, K.J., & Lash, T.L. (2021). Confounding and confounders.
571 In T.L. Lash, S. Haneuse & K.J. Rothman (Eds.), *Modern epidemiology* (4th. ed.,
572 pp. 389−424). Wolters Kluwer.
573 Hamed, K.H. (2008). To prewhiten or not to prewhiten in trend analysis? *Hydrological*
574 *Sciences Journal*, *53*(3): 667–668. https://doi.org/10.1623/hysj.52.4.611
575 Hardwicke, T.E., Frank, M.C., Vazire, S., & Goodman, S.N. (2019). Should psychology
576 journals adopt specialized statistical review? *Advances in Methods and Practices in*
577 *Psychological Science*, *2*(3):240-249. doi:10.1177/2515245919858428

Hartmann, D.P. (1974). Forcing square pegs into round holes: Some comments on an analysis of variance model for the intrasubject design. *Journal of Applied Behavior Analysis, 7*, 635–638.

Hibbs, D.A. (1974). Problems of statistical estimation and causal inference in time-series regression models. In H. L. Costner (Ed.), *Sociological methodology* (pp. 252-308). San Francisco: Jossey-Bass.

Hill, R., Sanchez, J., Lemel, L., Antonijevic, M., Hosking, Y., Mistry, S. N., Kruegel, A. C., Javitch, J. A., Lane, J. R., & Canals, M. (2023). Assessment of the potential of novel and classical opioids to induce respiratory depression in mice. *British Journal of Pharmacology*, 1–15. https://doi.org/10.1111/bph.16199

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment, 7*(2), 107–118.

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

IBM Corp. (2022). *IBM SPSS Statistics for Windows, v29*. Armonk, NY: IBM Corp.

Imbens, G.W., & Rubin, D.B. (2015). *Causal inference for statistics, social, and biomedical sciences*. Cambridge U.P.

Jin, F., & Lee, L. F. (2012). Approximated likelihood and root estimators for spatial interaction in spatial autoregressive models. *Regional Science and Urban Economics, 42*, 446–458

Jones, R.R., Vaught, R.S., & Weinrott, M. (1977). Time-series analysis in operant research. *Journal of Applied Behavior Analysis, 10*, 151-166. https://doi.org/10.1901/jaba.1977.10-151

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford U.P.

Keenan, R. J., Daykin, H., Metha, J., Cornthwaite-Duncan, L., Wright, D. K., Clarke, K., Oberrauch, S., Brian, M., Stephenson, S., Nowell, C. J., Allocca, G., Barnham, K. J., Hoyer, D., & Jacobson, L. H. (2023). Orexin 2 receptor antagonism sex-dependently improves sleep/wakefulness and cognitive performance in tau transgenic mice. *British Journal of Pharmacology, 1–20*. https://doi.org/10.1111/bph.16212

Kmenta, J. (1971). *Elements of econometrics*. New York, NY: Macmillan.

Lee, L. F., & Yu, J. (2015). Spatial panel data models. In B. H. Baltagi (Ed.). *The Oxford handbook of panel data* (pp. 363–401). Oxford: Oxford University Press.

Liu, X., & Wang, L. (2021). The impact of measurement error and omitting confounders on statistical inference of mediation effects and tools for sensitivity analysis. *Psychological Methods, 26*(3), 327–342. https://doi.org/10.1037/met0000345

Maldonado, G., & Greenland, S. (1993). Simulation study of confounder-selection strategies. *American Journal of Epidemiology, 138*(11), 923-936. https://doi.org/10.1093/oxfordjournals.aje.a116813

Matyas, T.A., & Greenwood, K.M. (1991). Problems in the estimation of autocorrelation in brief time series and some implications for behavioral data. *Behavioral Assessment, 13*, 137-157.

Øverup, C.S., Ciprić, A., Gad Kjeld, S., Strizzi, J.M., Sander, S., Lange, T., & Hald, G.M. (2020). Cooperation after divorce: A randomized controlled trial of an online divorce intervention on hostility. *Psychology of Violence, 10*(6), 604–614. https://doi.org/10.1037/vio0000288

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys, 3*, 96 - 146. https://doi.org/10.1214/09-SS057

Raudenbush, S.W., & Bryk , A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rolf, H.H., Groenwold, O.H., Klungel, Altman, D.G., van der Graaf, Y., Hoes, A.W., & Moons, K.G.M. (2013). Adjustment for continuous confounders: an example of how to prevent residual confounding. *Canadian Medical Association Journal, 185*(5), 401-406. DOI: https://doi.org/10.1503/cmaj.120592

Rosel, J.F., Elipe, M., Elósegui, E., Flor, P., Machancoses, F.H., Pallarés, J., Puchol, S., & Canales, J.J. (2020). Pooled time series modeling reveals smoking habit memory pattern. *Frontiers in Psychiatry, 11*: 49. https://www.frontiersin.org/article/10.3389/fpsyt.2020.00049

Rosel, J.F., Maldonado, E.F., Jara P., Machancoses, F.H., Pallarés, J., Torrente, P., Puchol, S., & Canales, J.J. (2019). Intensive longitudinal modelling predicts diurnal activity of salivary alpha-amylase. *Plos One, 14*(1): e0209475. https://doi.org/10.1371/journal.pone.0209475

Shine, L.C., & Bower, S.M. (1971). A one-way analysis of variance for single-subject designs. *Educational and Psychological Measurement, 31*, 105–113.

Singer, J.D., & Willett, J.B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford U.P.

Skinner, B.F. (1963). Operant behavior. *American Psychologist, 18*(8), 503–515. https://doi.org/10.1037/h0045185

Stinson, L., Liu, Y., & Dallery, J. (2022). Ecological momentary assessment: A systematic review of validity research. *Perspectives on Behavior Science, 45*, 469–493. https://doi.org/10.1007/s40614-022-00339-w

Suen, H.K., & Ary, D. (1987). Autocorrelation in applied behavior analysis. *Behavioral Assessment, 7*, 125-130.

Tong, K., & Dubé, C. A. (2022). Tale of two literatures: A fidelity-based integration account of central tendency bias and serial dependency. *Computational Brain & Behavior, 5*, 103–123. https://doi.org/10.1007/s42113-021-00123-0

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer.

Walls, T.A., & Schafer, J.L. (2006). *Models for intensive longitudinal data*. New York, NY: Oxford U.P.

Wardenaar, K. J., van Veen, T., Giltay, E. J., de Beurs, E., Penninx, B. W. J. H., & Zitman, F. G. (2010). Development and validation of a 30-item short adaptation of the mood and anxiety symptoms questionnaire (MASQ). *Psychiatry Research, 179*(1), 101–106. DOI: 10.1016/j.psychres.2009.03.005

Waterfield, S., Richardson, T.G., Smith, G.D., O'Keeffe, L.M., & Bell, J.A. (2023). Life course effects of genetic susceptibility to higher body size on body fat and lean mass: Prospective cohort study. *International Journal of Epidemiology*, dyad029. https://doi.org/10.1093/ije/dyad029

Westreich, D. (2020). *Epidemiology by design: A causal approach to the health sciences*. New York, NY: Oxford U.P.

Wysocki, A.C., Lawson, K.M., & Rhemtulla, M. (2022). Statistical control requires causal justification. *Advances in Methods and Practices in Psychological Science;5*(2). doi:10.1177/25152459221095823