Serial Correlation

Causes, Consequences, Detection, Remedy

Often times, the data in environmental studies are collected over temporal periods. The monitoring of wind speed, animal weight or water quality over time intervals is a common procedure in environment studies. When fitting any model to data, it is always important to look at the residuals. The residuals (e) are indicative of the true errors (u), which is the net sum of all other variables not in the model that affect your dependent variable. When the error term from different time periods (or across cross-section observations) are not independent and are correlated, the error term is said to be serially correlated.

**Intuition for interpreting the error term in regression**

***Example.***

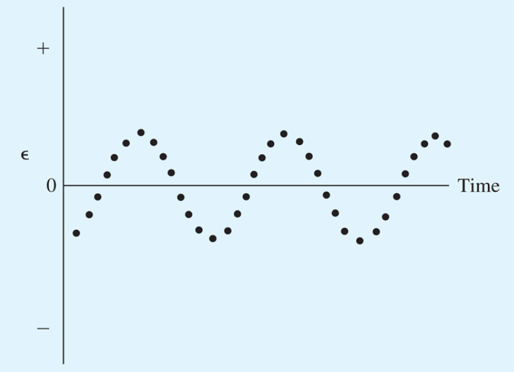
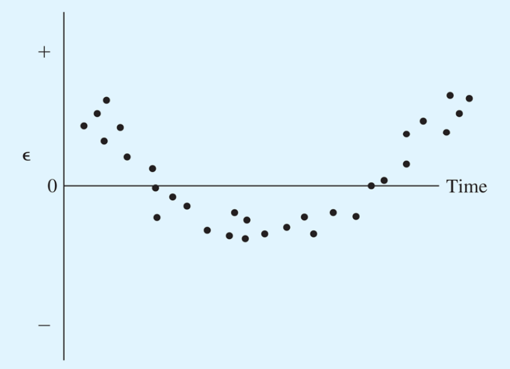
You are trying to predict average panda weight with age and sex. You collect data and design the following model:

PandaWeight = c + b\*Age + d\*Sex + e

e is the net sum of the effects of all other explanatory variables not specified in the model. One can think of panda family genealogy, diet, exposure to diseases, parasites, injuries and many other unobserved variables that can affect the weight of a panda. But with a large enough sample, or with independent and normally distributed errors, we assume that these variables are higher for some panda and lower for some, such that when we add them up, they sum to zero. Hence the net effects of Age and Sex is truly b and d, not because of some other factor.

Serial correlation poses a significant problem because it is common in time series data and is a key assumption in nearly all statistical techniques, when this assumption can be frequently violated[[1]](#footnote-1). With serial correlation, each error (or residual) is indicative of the residual before or after it. Our predictions of effects are less accurate on average. Serial correlation can also arise as a symptom of other underlying problems in the model, such as omitted variable and/or misspecification of the model, in which case the consequences are much larger, as the results will be biased[[2]](#footnote-2). Hence, serial correlation can also be a good way of detecting if your model is lacking in some relevant variables or if the fit can be made better by adjusting the functional form.

Serial correlation can be positive or negative[[3]](#footnote-3). Negative serial correlation is rare and would mean that the residuals would flip signs (from positive to negative and back again) in each time interval.



***Figure 1a & figure 1b*** Positive serial co­­rrelation[[4]](#footnote-4)

The error term tends to have the same sign as the term before. If e increases from t=1 to t=2, e also tends to increase from t=2 to t=3. (Figure from: Using Econometrics, A Practical Guide, Pearson Addison-Wesley, 2011)

The most commonly assumed kind of serial correlation is first order serial correlation, in which the current value of the error term is a function of the previous value of the error term:

et= ρet–1 + ut

e = the error term of the model

ρ = the first-order autocorrelation coefficient, AR(1) errors

u = a random shock (not serially correlated)

Causes

***Pure serial correlation***

This occurs when there is serial correlation across time in the true population. When you deal with time series analysis, this frequently happens as observations made within a shorter temporal period of each other will tend to be more similar than observations made further apart.

***Impure serial correlation***

Impure serial correlation is not serial correlation in the population but in the residuals (e) *after* you have estimated your model. If you detect serial correlation in your residuals after estimating a model, it could be indicative of one of the following problems in your model[[5]](#footnote-5).

* 1. Omitted variable

The variable of interest responds to explanatory variable(s) that is/are missing in your model. If this missing variable is serially correlated, the effect will be captured in your error term (e), which will experience periods of positive runs when it is high, and negative runs when it is low.

Y = a\*X1 + b\*X2 + u

***Example.***

You are trying to predict panda weight (X1) across time but have left out bamboo abundance (X2). Your predictions might correctly capture some of the variation in panda weight across the year but it might underestimate panda weight for periods of time when bamboo abundance increases (positive runs of residuals observed) and vice versa.

***Example.***

You are trying to predict farmer’s income in California over time, but observe a series of positive residuals for some months and a series of negative residuals for others. You may have left out rainfall in your model, which affects farmer’s income (yield and produce) and is serially correlated (El Nino in 2016 means weeks of abnormally high rainfall vs dry spells). Your model will tend to overestimate when rainfall goes down and underestimate when rainfall increases.

* 1. Incorrect functional form

Serial correlation can also be caused by misspecification of the functional form. The variable of interest might exhibit a quadratic or exponential trend, but if we fit a linear model through it, the residuals will appear to be serially correlated.

***Example.***

You are trying to predict the effect of fertilizer on crop yield, and fit a linear model to it. But more commonly, fertilizer effect is very high at low levels and almost zero at high levels (there is only so much nitrogen the soil can absorb), hence it actually follows a quadratic model. Your residuals will appear to be low at low levels of fertilizer and high at higher levels of fertilizer.

Consequences

* If it is pure serial correlation, OLS estimates are still unbiased and consistent but inefficient, meaning that chances of you getting the correct estimate on average is lower. I think of it as the target on a dartboard becoming smaller.
* Your standard errors are biased, hence your hypothesis tests are invalid (typically, it indicate significance when it is not, due to overestimation of the t-score) [[6]](#footnote-6).
* If it is impure serial correlation, which arises as a symptom of OVB or misspecified functional form, then your OLS estimates are biased and inconsistent!

Detection

First, go ahead and estimate your model. Depending on your data, you can use linear regression or any other non-linear regression such as exponential, logistic or sinusoidal functions. Since we want to examine serial correlation in the residuals (e), testing for serial correlation is inherently a diagnostic technique that comes after the model is estimated[[7]](#footnote-7). The model must first be fitted to obtain values of the coefficients and the residuals (e). Detecting serial correlation can be broadly categorized into two main ways: Informal and formal[[8]](#footnote-8).

I will be using whale abundance data from the Gulf of Alaska Data Portal to illustrate the techniques.

Date source:

Suzanne Teerlink, Terrance Quinn, Jan Straley, and Olga Von Ziegesar (2013). *Historical humpback whale abundance estimates in Prince William Sound, Alaska: 1978 - 2009*. [Data file]. Retrieved from https://goa.nceas.ucsb.edu/#view/df35d.124.19

***Estimate your model.***

whalels <- lm(Sightings ~ Year, data = whaleAbundance)

summary(whalels)

Coefficients:

Estimate Std. Error t-value Pr(>|t|)

(Intercept) -6484.3812 900.5255 -7.201 9.60e-08 \*\*\*

Year 3.3086 0.4514 7.329 6.96e-08 \*\*\*

# looks pretty significant, whale sightings have gone up in the past 30 years.

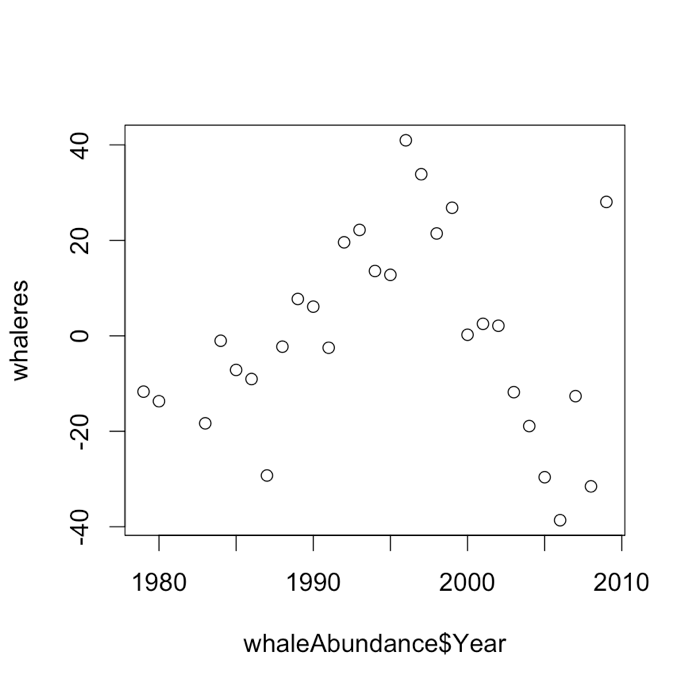
***Get your residuals.***

whaleres <- whalels$residuals

***Informal methods, mostly visual means of detecting serial correlation:***

1. Plot your residuals against your explanatory variables

plot(whaleAbundance$Year, whaleres)

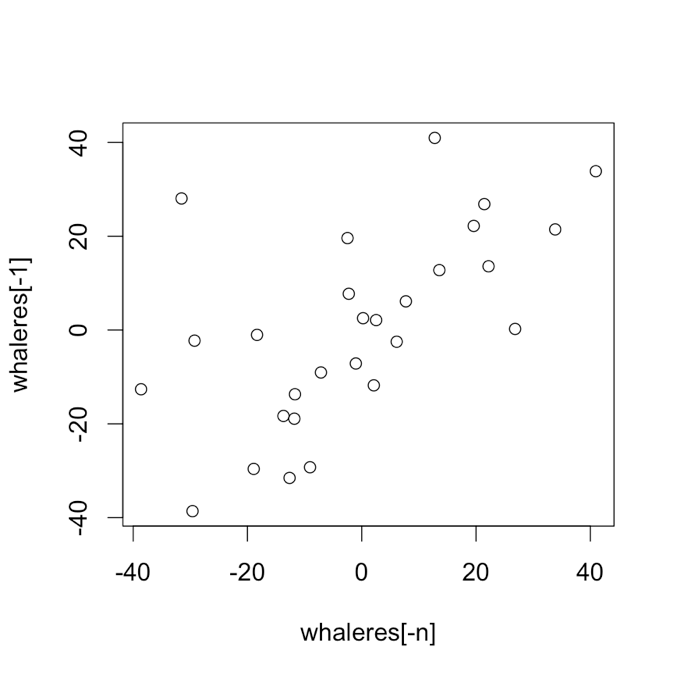


# The residuals exhibit a trend! Going up first, and then down. Since the residual is the difference between the actual observed y and the predicted y, my model is underestimating the real value in the middle years and over estimating in the early (1980s) and later years(2000s).

1. Plot your residuals against lags

n = length(whaleres)

plot(whaleres[-n], whaleres[-1])

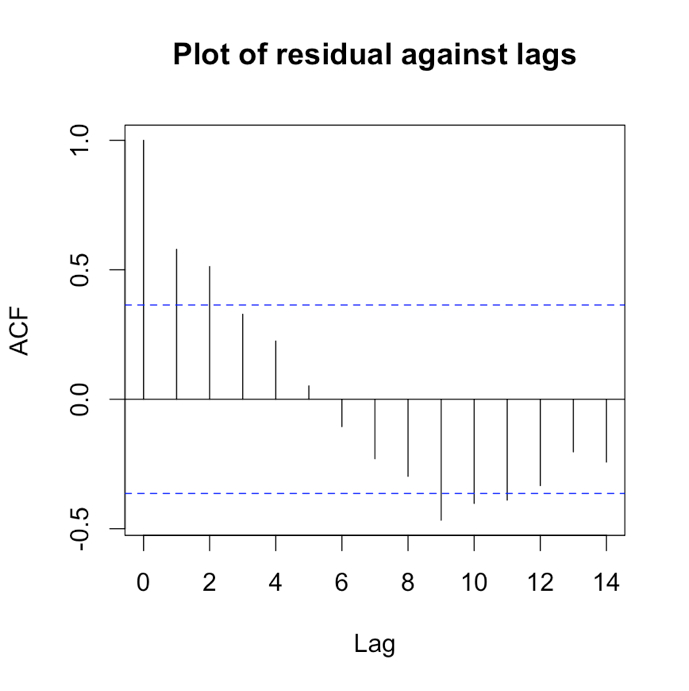


# The residuals seem to be positively correlated. What’s the value of the correlation coefficient, ρ? See bgtest() below for this.

1. Plot a correlogram

This is a visual test of serial correlation which tells you if there exist serial correlation in your regression model, as well as an idea of the order of serial correlation.

whaleACF <- acf(whaleres, type = "correlation", main = "Plot of residual against lags")



***Formal methods.***

1. The Durbin-Watson test for serial correlation

dwtestwhale <- dwtest(Sightings ~ Year, data = whaleAbundance)

Durbin-Watson test

data: Sightings ~ Year

DW = 0.76363, p-value = 2.92e-05

alternative hypothesis: true autocorrelation is greater than 0

The Durbin-Watson test for serial correlation (Durbin and Watson, 1951) is the standard method for detecting serial correlation.

The test description:

H0: ρ = 0 (no serial correlation) and H1: ρ ≠ 0 (serial correlation)

We compare the d-statistic to the [Durbin-Watson table](mailto:http://berument.bilkent.edu.tr/DW.pdf), where n is the sample size and k, the number of variables. If testing for first order autoregression, k=1. The critical values consist of lower and upper bound values.

If d < dL Reject H0

If d > dU Do Not Reject H0

Otherwise Inconclusive

To use the Durbin-Watson test, three assumptions must be met:

* Model must include an intercept
* Serial correlation must be first order only (for more than that, see BG test below)
* The regression model does not include a lagged dependent variable, for more information, please see: http://www.ncku.edu.tw/~account/chinese/course/eco91/lecture10.pdf

1. The Breusch-Godfrey test for serial correlation

Use the bgtest() function from package lmtest. Basically, the BG test regresses residuals on their lags. You can test for as many lags as you like with the order argument.

The test description:

H0: ρ = 0 (no serial correlation) and H1: ρ ≠ 0 (serial correlation)

Below is the model for 3 lags:



The null will be ρ1=ρ2=ρ3=0 for AR(3) structure.

bgtestwhale <- bgtest(Sightings ~ Year, order = 1, type = "Chisq", data = whaleAbundance)

Breusch-Godfrey test for serial correlation of order up to 1

data: Sightings ~ Year

LM test = 10.536, df = 1, p-value = 0.001171

The LM statistic follow a chi-square distribution. You can reject the null of zero serial correlation at the 5% level (p-value = 0.00117). You can retrieve the LM test statistic with the chi-square critical value at 5% in R:

qchisq(.95, df=1)

[1] 3.841459

You can also retrieve the correlation between residuals and their lags using the coeftest() function in lmtest package.

coeftest(bgtestwhale)

z test of coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -263.12052 735.42807 -0.3578 0.7205089

Year 0.13220 0.36867 0.3586 0.7198942

lag(resid)\_1 0.62740 0.16289 3.8517 0.0001173 \*\*\*

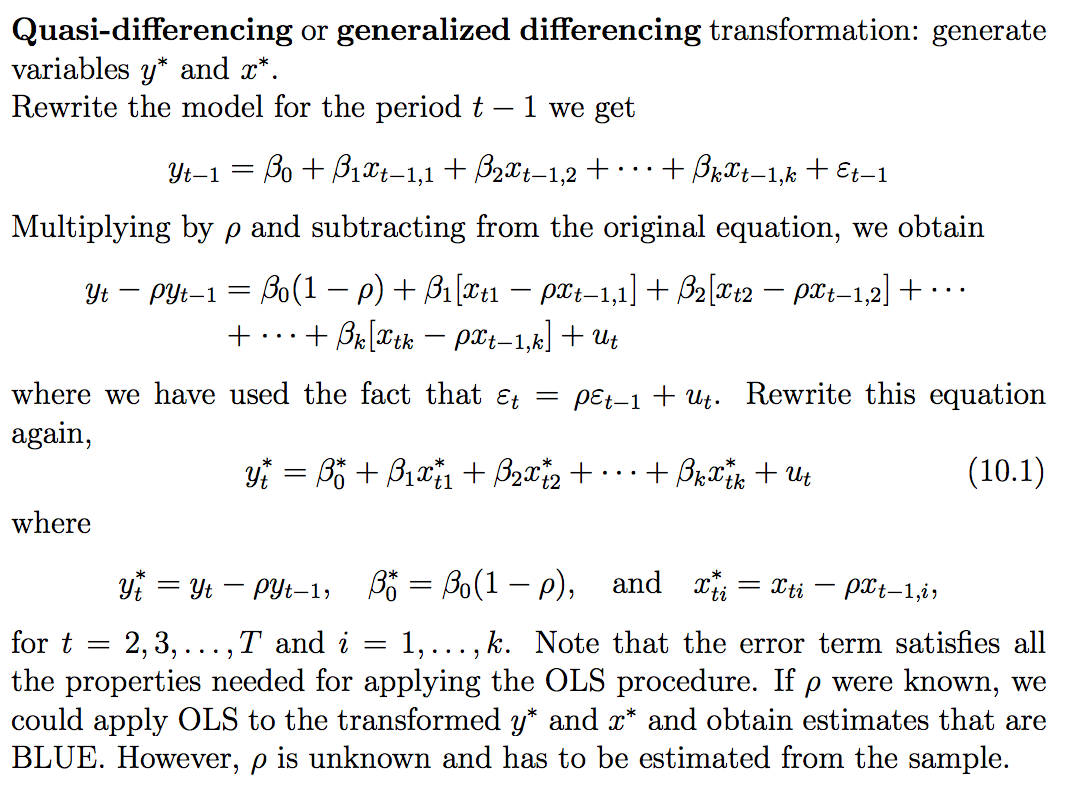
Note that the estimate of the lag (0.627) is similar to the value we get in the correlogram.

Remedy

When you have detected serial correlation, start by looking at the specification of the equation using your conceptual knowledge for possible errors that might be causing impure serial correlation[[9]](#footnote-9):

* + - Is the functional form correct?
    - Are you sure that there are no omitted variables[[10]](#footnote-10)? Especially if it is any other seasonal or temporal variation that could be correlated with the passing of time if your model contains time as one of your dependent variables.

Once you have reviewed the model carefully and decided that the serial correlation cannot be explained in any other way, you can account for serial correlation in your model using generalized least squares estimation. It is a means of transforming your model such that the errors are no longer serially correlated and the F and T statistics are robust.



**Figure 2.** Derivation of the GLS model using estimate of the error correlation coefficient by regressing residual against its lag. (Figure from: Serial Correlation, Yin-Feng Gau (2002). Retrieved from: http://www.ncku.edu.tw/~account/chinese/course/eco91/lecture10.pdf)

[Note: The steps described below draw heavily from Time-Series Regression and Generalized Least Squares in R by John Fox & Sanford Weisberg, 2010]

Use the gls() function in the nlme package (already in R base package).

* The correlation argument can be used to specify a model for error autocorrelation
* The method argument selects the method of estimation — method="ML" for maximum likelihood[[11]](#footnote-11).

For first order serial correlation,

library(nlme)

whaleGLS <- gls(Sightings ~ Year, data = whaleAbundance, correlation=corARMA(p=1), method="ML")

Generalized least squares fit by maximum likelihood

Model: Sightings ~ Year

Data: whaleAbundance

Log-likelihood: -121.8499

Coefficients:

(Intercept) Year

-7176.936652 3.656161 **<- coefficient for year is higher than in lm**

Correlation Structure: AR(1)

Formula: ~1

Parameter estimate(s):

Phi

0.6119577 **<- this is the estimation of the correlation between residual and first lag**

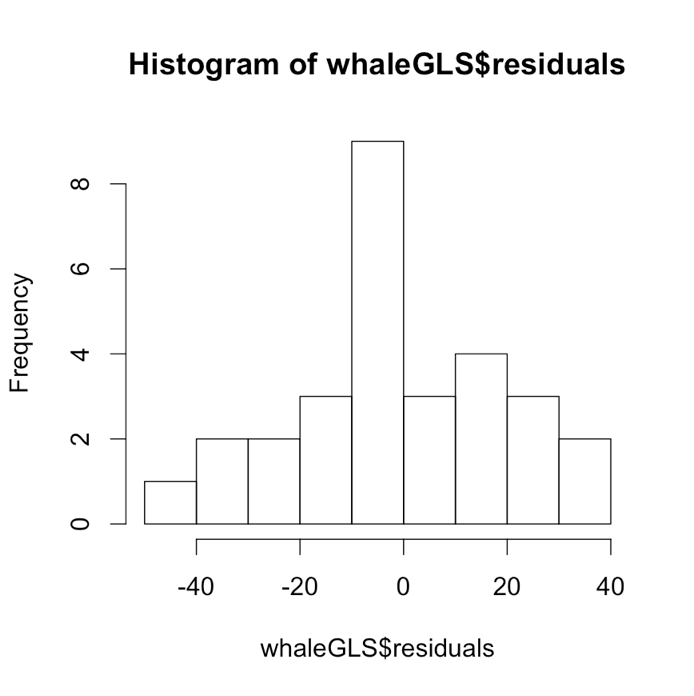
Degrees of freedom: 29 total; 27 residual

Residual standard error: 20.27297

hist(whaleGLS$residuals)

qqnorm(whaleGLS$residuals)

# errors looks normally distributed now



Specifying the correlation structure as correlation=corARMA(p=2) fits an AR(2) process for the errors.

whaleGLS2 <- gls(Sightings ~ Year, data = whaleAbundance, correlation=corARMA(p=2), method="ML")

Generalized least squares fit by maximum likelihood

Model: Sightings ~ Year

Data: whaleAbundance

Log-likelihood: -121.1574 **<- log likelihood did not increase by that much**

Coefficients:

(Intercept) Year

-7201.478607 3.668038 **<- not much change from whaleGLS results**

Correlation Structure: ARMA(2,0)

Formula: ~1

Parameter estimate(s):

Phi1 Phi2

0.4218886 0.2666139

Degrees of freedom: 29 total; 27 residual

Residual standard error: 19.83065

There is not much change to the coefficient (3.656 for AR(1) vs 3.668 for AR(2)), hence I conclude that the AR(2) term is redundant and will stick with the AR(1) model.

I chose the GLS method in part because we already covered ARIMA in class. The GLS method is also slightly more versatile in the sense that you do not have to coerce your data into a time series. Many of the arguments for the gls() function are the same as for lm and in the formula argument you can fit as many variables as you want, i.e. Y ~ A + A2 + B +….

For further reading and testing of the GLS model using log-likelihood test, do visit <https://socserv.socsci.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Timeseries-Regression.pdf>

References

Greene, W. H., 2003, *Econometric Analysis, 5th ed.*, Prentice Hall.

A.H. Studenmund, *Using Econometrics, A Practical Guide*, 2011, Pearson Addison-Wesley.

John Fox & Sanford Weisberg, 2010, *Time-Series Regression and Generalized Least Squares in R*. Retrieved from <https://socserv.socsci.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Timeseries-Regression.pdf>

James Kirchner, 2001, *Environmental Data Analysis Toolkit #11: Serial Correlation*. Retrived from <http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_11.pdf>

Durbin, J. and G. S. Watson, *Testing for serial correlation in least squares regression II*, Biometrika, 38, 159-178, 1951.

1. Data Analysis Toolkit, James Kirchner, 2001 [↑](#footnote-ref-1)
2. Using Econometrics, A Practical Guide, Pearson Addison-Wesley, 2011 [↑](#footnote-ref-2)
3. Using Econometrics, A Practical Guide, Pearson Addison-Wesley, 2011 [↑](#footnote-ref-3)
4. Using Econometrics, A Practical Guide, Pearson Addison-Wesley, 2011 [↑](#footnote-ref-4)
5. Econometric Analysis, 5th ed, Greene, W. H., 2003 [↑](#footnote-ref-5)
6. Using Econometrics, A Practical Guide, Pearson Addison-Wesley, 2011 [↑](#footnote-ref-6)
7. Data Analysis Toolkit, James Kirchner, 2001 [↑](#footnote-ref-7)
8. Using Econometrics, A Practical Guide, Pearson Addison-Wesley, 2011 [↑](#footnote-ref-8)
9. Using Econometrics, A Practical Guide, Pearson Addison-Wesley, 2011 [↑](#footnote-ref-9)
10. Using Econometrics, A Practical Guide, Pearson Addison-Wesley, 2011 [↑](#footnote-ref-10)
11. Time-Series Regression and Generalized Least Squares in R, John Fox & Sanford Weisberg, 2010 [↑](#footnote-ref-11)