Environmental Data Laboratory
Professor Kirchner

Laboratory 11: Detecting and accounting for serial correlation

Serial correlation is very common in environmental time series. Serial correlation (also called autocorrelation) occurs when the residuals in a time series are not independent; instead, each residual depends, in part, on the previous residual. That probably sounds pretty arcane and technical, but serial correlation has some very important practical effects on trend analyses. To quote from the serial correlation Toolkit:

> -Spurious--but visually convincing--trends may appear in your data
> -Although the regression coefficients (or other results of your analysis) will still be unbiased, you will underestimate their uncertainties, potentially by large factors.
> -Because uncertainties will be underestimated, confidence intervals and prediction intervals will be too narrow.
> -Estimates of "goodness of fit" will be exaggerated.
> -Estimates of statistical significance will be exaggerated, perhaps vastly so. Your actual false-positive rate can be *much* higher than the α-value used in your statistical tests.

You should have already received a toolkit that explains serial correlation, and details techniques to detect it and account for it. Here, we will demonstrate some of these techniques. The purpose is to give you a little bit of direct experience with serial correlation, to go along with the theoretical concepts in the toolkit.

Open the JMP data file, "Serial correlation lab.jmp". There are two 50-year time series here, labeled simply as A and B. One of these time series has a real trend in it, and the other doesn't. Your task is to figure out which is which.

To begin, first plot both time series as a function of year, using Fit Y by X. Now, before you fit a line to the data, just *look* at them. •1: Does either series look like there is a clear trend? Which one? Now, fit straight lines to both time series, and •2: note the regression slope, its standard error, and its statistical significance (p< value), for both time series. Based on these results, which data series appears to have the more convincing trend? •3: Is there any visual evidence that either data set needs to be transformed to make it linear or to make it homoscedastic (to make the scatter roughly even, from the beginning of the time series to the end)?

As the linear regression Toolkit stressed, it is vital to look at the residuals from any fitting exercise. Let's go through a few of those steps. Save the residuals from the two regression exercises, and plot them as functions of time. •4: Is there any evidence of curvature? Any evidence of heteroscedasticity (uneven scatter)? Any clear outliers? •5: Plot the distributions of the residuals. Are there drastic departures from normality?

The foregoing exercises simply check for curvature, non-normality, and heteroscedasticity. They don't check for serial correlation. To check for serial correlation, you need to plot the residuals against their lags. To do this, you need to create two columns, called "lag A residuals" and "lag B residuals", with formulas that look like this:

Lag(Residuals Series A , 1) and Lag(Residuals Series B , 1) (1)

where you get the lag function from the "Row" category in the right-hand choice window. Look to make sure that these columns are doing what you want them to: they should reproduce the two residuals columns, displaced downward by one row.

Now, plot each of the residuals against its lags. If the data are serially correlated, there should be significant correlation between the residuals and their lags. •6: Is there such correlation, for either data set? What are the correlation coefficients for set A and set B?

You should have discovered that there is serial correlation in the residuals from data set A. That raises the spectre of all of the badnesses that were spelled out in the introduction. In particular, the uncertainty in the regression trend for data set A may have been underestimated, and therefore the statistical significance may have been overestimated. So we need a way to tell what the "real" parameters hiding inside that regression relationship are. There are several practical ways to do this. Let's review the theory behind them, again from the serial correlation Toolkit:

Linear regression assumes that the "true" underlying linear relationship between X and Y is,

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad (2)$$

where αand β are the "true" slope and intercept (corresponding to the parameters *a* and *b* that you would estimate from any particular data set), and ε_i are the errors in Y. If those errors are serially correlated with a "true" correlation of ρ (roughly corresponding to the correlation coefficient *r* that you would estimate from the lagged residuals of any particular data set), then:

$$\varepsilon_i = \rho\varepsilon_{i-1} + \xi_i \qquad (3)$$

where ξ_i is random, and *not* serially correlated. Note that a fraction ρ of each ε_{i-1} is passed on to the next ε_i; you can think of $\rho\varepsilon_{i-1}$ as the redundant part of ε_i, and ξ_i as the non-redundant part. Regression assumes that the errors are uncorrelated, like the ξ_i, not like the ε_i, which are serially correlated. So the question naturally arises: can we arrange things so that our residuals are the uncorrelated ξ_i, which regression knows how to handle, rather than the serially correlated ε_i? Watch this: solve the linear relationship for ε, at both time *i* and *i*-1:

$$\varepsilon_i = Y_i - \alpha - \beta X_i \quad \textit{and} \quad \varepsilon_{i-1} = Y_{i-1} - \alpha - \beta X_{i-1} \qquad (4)$$

Now, substitute both of these ε's into equation (3) above, and rearrange terms:

$$Y_i - \alpha - \beta X_i = \rho(Y_{i-1} - \alpha - \beta X_{i-1}) + \xi_i \;\; \textit{or} \;\; Y_i - \rho Y_{i-1} = \alpha(1-\rho) + \beta(X_i - \rho X_{i-1}) + \xi_i \qquad (5)$$

Note that the serially correlated errors ε_i have disappeared, and only the well-behaved error ξ_i remains. There are several ways to fit equation (5) to data. Here we will use the Hildreth-Lu procedure, which is conceptually quite simple (and relatively easy to implement on JMP). The Hildreth-Lu procedure rewrites equation (5) such that Y_i is a function of Y_{i-1}, X_i, and X_{i-1}:

$$Y_i = \rho Y_{i-1} + \alpha(1-\rho) + \beta(X_i - \rho X_{i-1}) + \xi_i \qquad (6)$$

The Hildreth-Lu procedure turns ρ into *a parameter that is fitted to the data*, just like α and β are. This approach is simple, straightforward, and appealing. There is only one drawback: equation (18) is *nonlinear* in the parameters, so it can't be solved by linear regression. Instead, it must be fitted by nonlinear regression. In JMP, this is done through the "Nonlinear Fit" platform. This is explained in detail in "Nonlinear Regression" (Chapter 16 in the Statistics Guide in the help features of JMPIN); you can refer there if the instructions below aren't sufficiently clear.

The first step is to create a new column, called "Y" (don't make this into a formula, just keep it as a data column), and copy Data Series A into it.

The next step is to create another column and build a formula (call it "fit function") that contains the right hand side of this equation:

$$Y_i = rY_{i-1} + a(1-r) + b(X_i - rX_{i-1}) \qquad \text{(where X, obviously, is time)} \qquad (7)$$

which JMP will then try to fit to the "Y" data by choosing appropriate values for the trend *b*, the y-intercept *a*, and the serial correlation coefficient *r* (as best-fit approximations to the true values α, β, and ρ). These three constants are "parameters", and you need to define them. Here's how you do it. In the formula window, and click on the little triangle to the right of the heading of the left-hand window, click "Parameters", and then "New Parameter", then type "r" as the name, then enter an initial value for r (pick something reasonable!) and click "OK". Create the parameters "a" and "b" in the same fashion. Now, using your formidable formula-building skills, create the following formula:

r•Lag(Y , 1) + **a**•(1-**r**) + **b**•(Year - **r**•Lag(Year , 1)) (8)

You will need to select Y and Year from the "Table Columns" in the left-hand window, and a, b, and r the "Parameters" in the same left-hand window, and use the small triangle at the top of that window to toggle back and forth between displaying the parameters and displaying the table columns.

When the formula looks like (8) above, close the formula window.

Now, from the "analyze" menu, select "Nonlinear Fit". Assign "Y" as the Y, or "response", variable, and "fit function" as the X, or "predictor" variable. Then click "OK". The fitting algorithm will try to adjust the constants in "fit function" so that it looks as much like "Y" as possible. All you need to do is click "Go" (there are lots of other goodies here; if you're curious, see the help features). Now, inspect the "Solution" table for the parameter estimates and their approximate standard errors. •7: What are the estimates and standard errors for r, a, and b?

•8: What is the ratio between the standard error of the slope (b) when serial correlation is taken into account, compared to the standard error that you calculated above? (In other words, by how many times did your simple linear regression underestimate the uncertainty in the slope?) •9: From your improved estimate of the slope and its standard error, calculate a value of t that expresses by how many standard errors the slope differs from zero. What is the approximate statistical significance of this value of t? What is the ratio between this statistical significance and the value you obtained earlier? (In other words, by how many times did your simple linear regression overstate the statistical significance of the slope)?

For the sake of completeness, copy data set B into the Y column and re-run your nonlinear fitting algorithm (you don't need to re-enter the formula; just click "reset" and then "go"). •10: Does correcting for serial correlation substantially alter the standard error of set B's slope, or its statistical significance?

•11: Once serial correlation has been taken into account, which data set, A or B, do you think contains the real trend? Which trend is known more precisely? How does this compare with the results you got before you accounted for the effects of serial correlation?

----> Don't discard your data set; keep it open for the next section <----

Serial correlation in global temperature records

Now let's look at some environmental data. In particular, let's look at the global average temperature record compiled by Jones et al. Open the JMP data set, "Jones Global Temp.jmp" This is the last 40 years of a data set that stretches back into the mid-1800's. These are not temperatures in degrees C, instead, they are so-called temperature "anomalies". In order to combine data from many different sites, Jones et al. calculated the average temperature for each site during a particular "reference" period (here, 1950-1975), then they subtracted this average from the readings for each year. If the temperature in (say) October 1999 were 2°C warmer than the average for all the Octobers during the "reference" period, the temperature "anomaly" for October 1999 would be 2°. As I explained in lecture a while ago, this subtracts out a lot of the site-to-site variability, and clarifies the year-to-year changes.

It has been alleged that because there is serial correlation in long-term climate records, no reliable trends can be observed. Test this claim with the Jones time series for 1960-1999, following the same steps above. Regress the temperature anomaly against time, inspect the residuals, and plot the residuals against their lags to determine whether there's serial correlation. Then use the Hildreth-Lu procedure to correct for the serial correlation. To save yourself the trouble of building a whole new formula, (a) extend your previous data sheet to 480 rows, (b) copy the temperature anomaly data into the "Y" column and the years into the "year" column, and (c) re-use the formula you used before.

•12: Write a brief summary of what you found. How much serial correlation was there? Did it significantly affect your results? Is the global temperature trend statistically significant, even after serial correlation is taken into account? What is your best estimate of the global warming trend, in degrees C per century?

We hope that this lab has taught you to respect serial correlation, but not to fear it. It's a beast, but one that can be tamed--or at least subdued.