Regression and Time Trends Exercise: Making inference about trends in the presence of autocorrelation, part 2, Generalized Least-Squares

For this exercise, we will return to the interactive web application at http://spark.rstudio.com/statmos/mod1_regression.

Background

In the last exercise, we used subsampling and adding a lag term to attempt to account for autocorrelation in regression models and allow us to use linear regression to make inferences about trends over time.

In a way, those are both indirect ways of dealing with the problem. Remember that temporal autocorrelation is only really a problem for regression if it shows up in the **residuals,** because it violates the assumption that the errors in the model are independent. That's why we always examine plots of the residuals and compute the autocorrelation function on the residuals, not the raw data. Remember that a linear regression model has the general form:

$$Y = \beta_1 * X_1 + \sigma^2$$

Both of the strategies we used in the last exercise, subsampling and adding a lag, adjust either the response variable Y or the predictor variables (X_s). The final strategy, and the one that's used most – often in research, is to incorporate temporal autocorrelation directly into the error component of the model (σ^2), using a technique called *generalized least-squares*.

What is Generalized Least Squares?

Remember the assumption that all of the errors in an ordinary least-squares regression model are independent? We can put that statement in mathematical form by saying that the covariance of expected errors at all observations separated in time are zero, and the covariance of expected errors at all observations taken at the same time is a constant. What is covariance? It's just a measure of how strongly correlated observations are with each other. Say we had a series of residuals from a regression model that tests for trends over time. Let's call these residual values u_1, u_2, u_3 … u_T. We could write out a matrix that summarizes how much we expect these pairs of errors to vary together if all of the errors are independent:

Covariance	u_1	u_2	u_3	…	u_T
u_1	σ^2	0	0		0
u_2	0	σ^2	0		0
u_3	0	0	σ^2		0
⋮				⋱	
u_T	0	0	0		σ^2

The diagonal elements of the matrix all have a value, σ^2 indicating that observations collected in the same time interval should have a constant covariance, σ^2. Because all of the residuals are independent, all of the off-diagonal elements are zero, indicating that we expect no covariance between errors in observations collected at different times.

When we have temporal autocorrelation, then the off-diagonal elements of the matrix are no longer zero. In the simplest case, all of the observations that are separated by one time-interval (a lag of one) have an additional covariance term, *p*, and that covariance decays exponentially as the lag interval increases. This means that, instead of being all zero, the off-diagonal elements in the error-covariance matrix now look like this:

Covariance	u_1	u_2	u_3	…	u_T
u_1	σ^2	$\sigma^2 p$	$\sigma^2 p^2$		$\sigma^2 p^{T-1}$
u_2	$\sigma^2 p$	σ^2	$\sigma^2 p$		$\sigma^2 p^{T-2}$
u_3	$\sigma^2 p^2$	$\sigma^2 p$	σ^2		$\sigma^2 p^{T-3}$
⋮				⋱	
u_T	$\sigma^2 p^{T-1}$	$\sigma^2 p^{T-2}$	$\sigma^2 p^{T-3}$		σ^2

Because the autocorrelation term *p* must be less than one, then the correlation between observations at different times decays to zero as the number of time-intervals between observations increases. Because there is only one autocorrelation term, *p*, this matrix is called an AR(1) error structure.

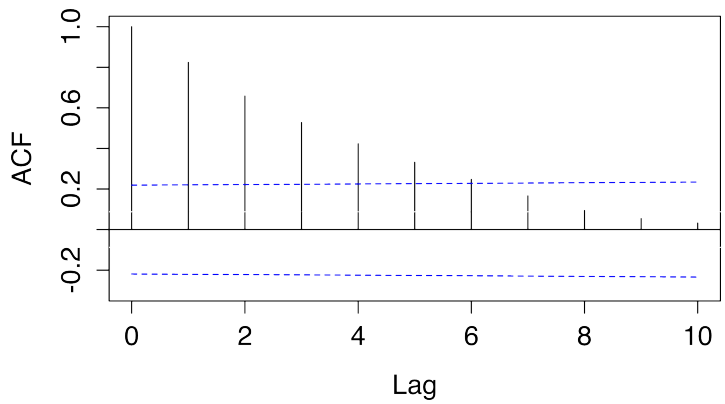
Sometimes, this is too simple an idea of how errors are correlated. We could add additional parameters that describe the relationships between errors at different time intervals. This larger family of error structures are known as AR(*p*) structures. We won't cover these in detail, but they all do basically the same thing: adjusting the error covariance matrix to account for temporal autocorrelation.

The error-covariance matrix that we discussed above is a statement about our expectations of how errors are correlated. What generalized least-squares does is estimate the unknown parameters of that error-covariance matrix from the data and incorporate it into the model fitting procedure. The result is that we get a model that should give us reliable estimates of trends over time *and the uncertainty around that trend*, something we can't get from OLS when there is strong temporal autocorrelation.

Figuring out whether GLS with an AR(p) error structure is appropriate

So how do we know whether or not the pattern of autocorrelation in our data can be represented well by an AR(*p*) error structure? Our primary clue is the plot of the Autocorrelation Function (ACF) for the residuals of a model fit by Ordinary Least

Squares. If an AR1 model for the errors is appropriate, then the ACF should decay exponentially as lag increases. This kind of ACF plot looks like this:



If this is the case, we can be reasonably confident that incorporating an AR(p) error structure into the regression using generalized least-squares is appropriate.

Instructions:

- In the online application, find data for the temperature trend in Mexico from 1900 to 2010 by entering "MEX" into the Country Code field, and changing the "from Year" and "to Year" fields to 1900 and 2010, respectively.

- Click all three check-boxes to look at the trend, diagnostic plots, and model outputs for this data. In particular, remind yourself why we can't trust the standard errors and p-values from the model output using Ordinary Least Squares.

- Click on the "GLS" tab above the trend graph. This will fit an AR(1) model to the data using Generalized Least Squares. Compare the plot of the residuals, Auto-correlation Function, and model output from the GLS model to the OLS model.

- Use these comparisons to answer the questions below.

Questions:

1. Click on the "OLS" tab and examine the plot of residual autocorrelation. Does it resemble the pattern you would expect if an AR(p) error structure was appropriate for the data? Why or why not?

2. Compare the estimates of temperature trends from the OLS and GLS analyses. How closely do the two estimates correspond?

3. Compare the standard errors and p-values for the OLS and GLS estimates of the trend. How do they differ?

4. Examine the residuals plot and the ACF plot for the GLS model. Do the residuals from the GLS analysis appear to be correlated? Based on what you see, should you trust statistical inference based on this model?

5. Compare the results of the GLS analysis to the results obtained by subsampling the data and adding a lag. How do the GLS results differ?

6. What does the GLS analysis tell you about temperature trends in Mexico over the past century?

7. Select the "Google ngrams" dataset from the dropdown menu and search for a term that you want to track over time (your choice). Try and find some trend data that would clearly not be appropriate for analysis using OLS. Which search term did you use? What time-interval?

8. Why can't you use OLS to make valid inference about trends over time using this data? Does using GLS help matters? Why or why not?