Regression and Time Trends Exercise: Making inference about trends in the presence of autocorrelation.

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

In the last exercise, we found that, although their appears to be an overall trend towards higher temperatures in the USA during the 20th century, the data has some properties that pose problems for assessing whether that trend was significant using linear regression. In particular, it looks like there is temporal autocorrelation in the data, which might give us false confidence that this trend is significant!

Luckily, there are a few different ways that we can modify our approach to the problem and make inference about trends even in the presence of autocorrelation. Here are two of them:

Subsampling the data
The biggest problem with temporal autocorrelation is that it "fools" us into thinking that we have more independent observations of a system than we actually do. Luckily, temporal autocorrelation tends to be strongest between observations that are close together in time, and gets less strong as we examine observations that are spaced farther apart. We can take advantage of this by using only a subset of our observations to make inference about trends. If autocorrelation is only a problem for a lag of 1 (adjacent observations), then we might be able to deal with the problem by including only half of the observations in the regression, ensuring that each observation is separated by two time-intervals, instead of one. If we still have a significant trend after subsampling the data, and the residuals from that model look OK, we can be reasonably confident that there is a long-term trend in the data.
Advantages: Simple, easy to understand and communicate.
Disadvantages: Doesn't use all of the data.

Adding a lag to the model
Another way of dealing with autocorrelation is to include values of the response in a previous time as another term in your regression model. Adding a lag changes the regression equation from this:

$$response = intercept + slope * year_t + error$$

to this:

$$response = intercept + slope_t * year_t + slope_{t-h} * year_{t-h} + error$$

In this way, we are including the correlation between years directly into our model.
Advantages: Uses more of the data.
Disadvantages: The slope associated with year no longer represents the long-term trend. Cannot use observations at the beginning of the series.

The following instructions will help you get comfortable with using these two strategies.

Instructions:

- Examine the temperature trend in the USA from 1900 to 2010 by 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. Remind yourself why we can't trust the standard errors and p-values from the model output using Ordinary Least Squares.

- Click on the tab above the trend graph that says "Subsampling". A new control will appear in the sidebar below that lets you choose how many observations will be between each one that you use in the regression model. Change the sampling interval and watch how it changes the time-trend plot, the diagnostic plots, and the model output.

- Click on the tab that says "Lagging". This will add a lagged value of the response into the regression model. This lagged value now shows up as a blue dotted line on the time-trend plot. The slider control now allows you to change the lag on the new term we've added to the regression model. By default, it's set at a lag of one, meaning that we are using the value in the previous year to predict the value in the current year. Change the value of the lag, and watch how it changes the time-trend plot, the diagnostic plots, and the model output.

Questions:

1. What is the minimum value of the subsampling interval that we need in order for the diagnostic plots to indicate we can proceed with making valid inference about trends?

2. How does subsampling the data change our estimate of the long-term trend in temperatures in the USA? How does it change our confidence that this trend is significant?

3. What's the minimum value of the lag that we need in order for the residuals to be independent?

4. How does adding a lag to the regression model change our confidence in the long-term trend?

5. If our goal was to get the most accurate estimate of the long-term trend in temperatures over this time-period, which of the three approaches (OLS, Subsampling, Lagging) would we want to use? Why?

6. If our goal was to predict what the mean annual temperature of the USA would be in 2011, which of the three approaches (OLS, Subsampling, Lagging) would we want to use? Why?