Homework Assignment 6

Max H. Farrell – Chicago Booth
BUS41100 Applied Regression Analysis

Due at the beginning of class of week 9

1 Furniture Sales

The data set `furniture.csv` on the class website contains data from 1992–2001 on monthly furniture `sales` (in millions of dollars). If you didn't know any context this would look like an ordinary data set (rows, columns, numbers, etc), but two things make it different. First, the *order* matters. These rows are in time order, and have time labels. What we care about, and are going to try to capture, is how sales itself evolves over time.

(a) Create a variable called `time` that simply counts the months starting with the first, consecutively to the end (i.e. January, 1992 = 1; February, 1992 = 1; ..., December, 2001 = 120).

Plot the time series data for monthly furniture sales. This is just our usual plot with the outcome, `sales`, on the y axis and `time` on the x axis, but now `time` has extra meaning. Comment on what pattern(s) you see in the context of linear regression.

(b) Run a regression of `sales` against `time`, and add the regression line to the plot you made in part **(a)**. That is, fit the linear model

$$\texttt{sales}_i = \beta_0 + \beta_1 \times \texttt{time}_i + \varepsilon_i.$$

(c) Plot the residuals of this regression against time. Does this plot look how you expect? Explain.

(d) Again plot the residuals against time, but this time make all the points for December a different color. Also, plot the data and the regression line, but this time make all the points for December a different color. What do these two plots tell you? What would you change about the regression after seeing these plots?

(e) Let's turn to prediction. Remember that the idea behind prediction is to try to form a good guess for an outcome that you have not seen, Y_f, based on the data you have and a newly observed X_f. (In our favorite example: Y_f is the price of a house that has not yet sold, X_f would be the square footage, which can measure right now.) But here we don't have an X variable, just time. What we really want to do, then, is try to predict a future outcome.

Pretend you did not have the data for 2001. Re-run the regression above using only the data up to, and including, December 2000. Use this regression to predict furniture sales for each month of 2001. Comment on your predictions.

(f) Create an ACF plot of the series of `sales`. Given this plot, is the regression specification from part **(b)** the appropriate first step? Why or why not? What regression model do you recommend to capture this time series, and why?

2 United States Gas Prices

In this question we will attempt to capture the time series dependence of weekly US gas prices. The data file `USGasPrice.csv` contains the `year`, `week`, and the `price` of gas from the eighth week of 1990 through the twenty-sixth week of 2003.

In each part below, we will use one or more of the tools we learned in class. For each part, plot the **(1)** fitted values along with the original values as well as **(2)** the ACF (see the slides for week 8).

(a) As a first attempt, use annual frequencies to try to capture the idea that driving (and hence gasoline demand) is highly seasonal. Use the two plots asked for above to describe what goes wrong with this approach intuitively.

(b) There appear to be several structural changes in the data (e.g. more recent years behave differently). Identify a reasonable set of time points when the series pattern changes dramatically and permanently and then fit a constant between each time point. What does this capture?

(c) For each period you identified above, fit a separate annual cycle. What do you learn from the two plots ask for in the question?

(d) Experiment with adding further cycles, time trends, and/or lagged values, in each case allowing each period you identified in **(d)** to have a separate fit. Try a "kitchen-sink" that includes cycles, trends, and lags.

(e) Finally, what of the "kitchen-sink" model can you do without? That is, give the simplest model you can that still does nearly as well at capturing the time series dependence. What do you learn from this?

3 United Kingdom Gas Consumption

The goals of this question are to develop the best possible model for prediction of quarterly UK gas consumption (file `UKGasConsumption.csv`). Since one would expect gas consumption to increase with both population (a measure of personal consumption) and GDP (a measure of commercial production), the data consist of quarterly UK gas consumption (in millions of therms), inflation adjusted GDP, and population estimates for the years 1960 to 1986 (1987 for `GDP` and `pop`).

(a) Find a regression model that best explains the time series for gas consumption (transform to log scale). You may incorporate the effect of GDP and population into your time series (again, consider a possible transformation of these variables).

(b) Comment on your chosen model. For example, is there evidence of either mean reversion or a linear time trend in your series? What is the effect of the covariates on gas consumption?

Modeling Note: This data exhibits a common trait for quarterly data: autoregression on an annual basis. That means you want to include Y_{t-4} as your AR term, rather than the usual Y_{t-1}. For example,

```
gasdata <- read.csv("UKGasConsumption.csv")
QR <- 5:108
cos4 <- cos(QR*pi/2)
sin4 <- sin(QR*pi/2)
loggas <- log(gasdata$gas[QR]
loggaslast <- log(gasdata$gas[QR-4])
```

This provides periodic, AR, and linear effect variables.