Problem Set 5

Issued: Friday, November 16, 2018 **Due:** Wednesday, November 28, 2018

Problem 5.1: BPP Data Analysis

The goal of this problem is to analyze the PriceStats data from the MIT Billion Prices Project, provided by Professor Rigobon. In addition, the file `PriceStats_CPI.csv` also contains a column "CPI." The CPI (consumer price index, the price of a "market basket of consumer goods and services" - a proxy for inflation) is released monthly by the Bureau of Labor Statistics. The file `T10YIE.csv` lists (during the same time period) the break-even rate (BER), or the difference in yield between a fixed rate and inflation adjusted 10 year treasury note. This difference can be interpreted as what the market views will be the inflation rate for the next 10 years, on average. There are 122 months of data in `PriceStats_CPI.csv`. In the following questions, you may want to either work on a log scale or operate on inflation rates. For your results, report the mean squared prediction error for 1 month ahead forecasts starting September 2013. For example, to predict the CPI in May 2015, you can use all the data before May 2015. You should perform all of your model fitting on the months prior to September 2013, and use the remaining months for evaluation.

(a) First, we will try to predict the monthly CPI without using the BER or PriceStats. Fit an AR model to the CPI data (take the first CPI value of each month as that month's CPI, you may or may not want to work in log scale in order to make the model comparable to models you fit in part (c)) and report the mean squared prediction error for 1 month ahead forecasts. Which order model gives the best predictions? (**Hint:** one way to determine the proper order to use is to examine the auto-correlation and partial auto-correlation functions of the residuals. Start with a single AR term and add other terms as necessary.)

(b) How might you calculate monthly inflation rates from the CPI data and your 1 month ahead predictions? How about from PriceStats data? And BER data? (What dates would you use? Or would you use an average of many dates?) Overlay your estimates of monthly inflation rates (there should be 4 lines, one for each dataset, plus the predictions) over time (months from September 2013 onward).

(c) Next, we will include external regressors to try to improve the predictions. Include as external regressors monthly average PriceStats data and BER data to fit a new AR model to the CPI. Report your prediction error. Try instead using PriceStats data and BER data from the last day of each month as your external regressors. Fit another AR model. Which model performs better in prediction? (**Hint:** Again, in order to match the units of your predictors and responses, you'll want to either work on a log scale or work with inflation rates. Please justify your choices in which values you decide to work with.)

(d) Try to improve your model from part (c). What is the smallest prediction error you can obtain? You might consider including MA terms, adding a seasonal AR term, or adding multiple daily values (or values from different months) of PriceStats and BER data as external regressors.

(e) Consider the MA(1) model, $X_t = W_t + \theta W_{t-1}$, where $\{W_t\} \sim WN(0, \sigma^2)$. Find the autocovariance function of $\{X_t\}$.

(f) Consider the AR(1) model, $X_t = \phi X_{t-1} + W_t$, where $\{W_t\} \sim WN(0, \sigma^2)$. Suppose $|\phi| < 1$. Find the autocovariance function of $\{X_t\}$. (You may use, without proving, the fact that $\{X_t\}$ is stationary if $|\phi| < 1$.)

Problem 5.2: The Mauna Loa CO_2 concentration.

In 1958, Charles David Keeling (1928-2005) from the Scripps Institution of Oceanography began recording carbon dioxide (CO_2) concentrations in the atmosphere at an observatory located at about 3,400 m altitude on the Mauna Loa Volcano on Hawaii Island. The location was chosen because it is not influenced by changing CO_2 levels due to the local vegetation and because prevailing wind patterns on this tropical island tend to bring well-mixed air to the site. While the recordings are made near a volcano (which tends to produce CO_2), wind patterns tend to blow the volcanic CO_2 away from the recording site. Air samples are taken several times a day, and concentrations have been observed using the same measuring method for over 60 years. In addition, samples are stored in flasks and periodically reanalyzed for calibration purposes. The observational study is now run by Ralph Keeling , Charles's son. The result is a data set with very few interruptions and very few inhomogeneities. It has been called the "most important data set in modern climate research."

The data set for this problem can be found in `CO2Data.csv`. It provides the concentration of CO_2 recorded at Mauna Loa for each month starting March 1958. More description is provided in the data set file. We will be considering only the CO_2 concentration given in **column 5**. The goal of the problem is to fit the data and understand its variations. You will encounter missing data points, part of the exercise is to deal with them appropriately.

Let C_i be the average CO_2 concentration in month i ($i = 1, 2, \cdots$, counting from March 1958). We will look for a description of the form:

$$C_i = F(t_i) + P_i + R_i$$

where:

- $F : t \mapsto F(t)$ accounts for the long-term trend.
- t_i is time at the middle of the ith month, measured in fractions of years after Jan 1, 1958.
- P_i is periodic in i with period 12, accounting for the seasonal pattern.
- R_i is the remaining residual that accounts for all other influences.

The decomposition is meaningful only if the range of F is much larger than the amplitude of the P_i and this amplitude in turn is substantially larger than that of R_i.

(a) Fit the data to a linear model $F_1(t) \sim \alpha_1 + \alpha_2 t$. Plot the data and the fit. What are the values of $\hat{\alpha}_1$ and $\hat{\alpha}_2$? Plot the residual error. Comment.

(b) Fit the data to a quadratic model $F_2(t) \sim \beta_1 + \beta_2 t + \beta_3 t^2$. Plot the data and the fit. What are the values of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$? Plot the residual error. Comment.

(c) Which fit (F_1 or F_2) is better in capturing the trend in the data? Explain.

(d) Consider $F_2(t)$. We will now extract the periodic component which appears in the data. Average the residual $C_i - F_2(t_i)$ over each month. Namely, collect all the data for Jan (resp. Feb, Mar, etc) and average them to get one data point for Jan (resp. Feb, Mar, etc). The collection of those points can be interpolated to form a periodic signal P_i. Plot P_i.

(e) Plot the fit $F_2(t_i) + P_i$. What can we conclude on the variation of the CO_2 concentration since 1958?