Homework 1

Due: Monday 1/27/20 by 11:59pm

Clearly, 1. and 2. do not have right or wrong answers - you will be given full credit for each as long as your responses indicate that you made an honest effort.

Course and Project Preparation

1. List up to three topics you hope we'll cover, from the most to least interesting to you. I will try to incorporate the most popular topics into the last few weeks of the course as time permits.
2. In a few sentences sentences, describe the kind of data you are interested in analyzing using the tools you gain in this course. The goal of this exercise is to get you started thinking about the kind of data you might use for your course project.

Regression and `R` Review

3. This problem will require that you work with `broc` data posted on the course website, which contains the average price of one pound of broccoli in urban areas each month, from July 1995 through December 2019.
   (a) Working with time series data often involves working with dates, which can be tricky to manipulate because they are often provided as characters. For instance, in the `broc` data the first value of the `date` variable is a string `"1995-07-01"`. This makes things like plotting difficult or extracting the year, month, or day difficult. Fortunately, it is easy to convert a string variable to something `R` calls a `Date` variable! See the Quick-R tutorial on on converting strings to `Date` variables: https://statistics.berkeley.edu/computing/r-dates-times. Create a new variable `fdate` in the `broc` data frame that is a `Date` object, and line plot of `broc$fdate` against `broc$price`.
   (b) Fit the following two regression models. Model 1 is given by $\texttt{price}_i = \mu + \sum_{j=1996}^{2019} \gamma_j(\texttt{year}_i = j) + \sum_{k=2}^{12} \alpha_k(\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2)$. Model 2 is given by $\texttt{price}_i = \mu + \beta\texttt{year}_i + \sum_{k=2}^{12} \alpha_k(\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2)$. Revisit the tutorial (https://statistics.berkeley.edu/computing/r-dates-times) for help extracting the month or year from a `Date` object in `R`. Using AIC, a z-test, or an F-test, choose between the Model 1 and Model 2. Justify your choice in at most one sentence, and provide any relevant numerical evidence, e.g. AIC values, z-value and corresponding p-value, or F-value, degrees of freedom and corresponding p-value.
   (c) Fit an additional model, which we'll call Model 3: $\texttt{price}_i = \mu + \beta\texttt{days since start}_i + \sum_{j=1996}^{2019} \gamma_j(\texttt{year}_i = j) + \sum_{k=2}^{12} \alpha_k(\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2)$. Using AIC, a z-test, or an F-test, choose between Model 1 and Model 3. Justify your choice in at most one sentence, and provide any relevant numerical evidence, e.g. AIC values, z-value and corresponding p-value, or F-value, degrees of freedom and corresponding p-value.
   (d) Fit one more additional model, which we'll call Model 4: $\texttt{price}_i = \mu + \beta\texttt{days since start}_i + \sum_{k=2}^{12} \alpha_k(\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2)$. Revisit the tutorial (https://statistics.berkeley.edu/computing/r-dates-times) for help computing the number of days since the start, July 1995. Using AIC, a z-test, or an F-test, choose between Model 3 and Model 4. Justify your choice in at most one sentence, and provide any relevant numerical evidence, e.g. AIC values, z-value and corresponding p-value, or F-value, degrees of freedom and corresponding p-value.

(e) Make one plot with four panels. In each panel, plot the prices in date order, from first to last. In the first panel, add fitted values from Model 1 along with 95% confidence intervals for each fitted value. In the second panel, add fitted values from Model 2 along with 95% confidence intervals for each fitted value. In the third panel, add fitted values from Model 3 along with 95% confidence intervals for each fitted value. In the fourth and last panel, add fitted values from Model 4 along with 95% confidence intervals for each fitted value.

(f) Plot the residuals from Model 1 in date order, from first to last. In at most one sentence, describe what (if any) evidence you observe for any remaining correlation across time.

(g) Suppose you wanted to forecast broccoli prices one month into the future, i.e. you wanted to compute $\mathbb{E}\left[y_{295}\right]$ under Model 1, Model 2, Model 3, or Model 4. Under which of the four models can you compute $\mathbb{E}\left[y_{295}\right]$ using the available data? Explain in at most one sentence.