MULTIPLE LINEAR REGRSSION MODELLING CASE STUDY ASSIGNMENT

Predict monthly sales of the Hyundai Elantra in the United States

About the data

An important application of linear regression is understanding sales. Consider a company that produces and sells a product. In a given period, if the company produces more units than how many consumers will buy, the company will not earn money on the unsold units and will incur additional costs due to having to store those units in inventory before they can be sold. If it produces fewer units than how many consumers will buy, the company will earn less than it potentially could have earned. Being able to predict consumer sales, therefore, is of first order importance to the company.

In this problem, we will try to predict monthly sales of the Hyundai Elantra in the United States. The Hyundai Motor Company is a major automobile manufacturer based in South Korea. The Elantra is a car model that has been produced by Hyundai since 1990 and is sold all over the world, including the United States.

Each observation is a month, from January 2010 to February 2014. For each month, we have the following variables:

- **Month** = the month of the year for the observation (1 = January, 2 = February, 3 = March, ...).
- **Year** = the year of the observation.
- **ElantraSales** = the number of units of the Hyundai Elantra sold in the United States in the given month.
- **Unemployment** = the estimated unemployment percentage in the United States in the given month.
- **Queries** = a (normalized) approximation of the number of Google searches for "hyundai elantra" in the given month.
- **CPI_energy** = the monthly consumer price index (CPI) for energy for the given month.
- **CPI_all** = the consumer price index (CPI) for all products for the given month; this is a measure of the magnitude of the prices paid by consumer households for goods and services (e.g., food, clothing, electricity, etc.).

Problem Statement:

Build an analytical model to **predict monthly sales of the Hyundai Elantra in the United States.**

Questions:

1. Load the data set. Split the data set into training and testing sets as follows: place all observations for 2012 and earlier in the training set, and all observations for 2013 and 2014 into the testing set.How many observations are in the training set?

2. Build a linear regression model to predict monthly Elantra sales using Unemployment, CPI_all, CPI_energy and Queries as the independent variables. Use all of the training set data to do this.What is the model R-squared?

3. How many variables are significant, or have levels that are significant? Use 0.10 as your p-value cutoff.

4. What is the coefficient of the Unemployment variable? What is the interpretation of this coefficient?

5. In our problem, since our data includes the month of the year in which the units were sold, it is feasible for us to incorporate monthly seasonality. From a modeling point of view, it may be reasonable that the month plays an effect in how many Elantra units are sold.To incorporate the seasonal effect due to the month, build a new linear regression model that predicts monthly Elantra sales using Month as well as Unemployment, CPI_all, CPI_energy and Queries. Do not modify the training and testing data frames before building the model. What is the model R-Squared?

6. Re-run the regression with the Month variable modeled as a factor variable. What is the model R-Squared?

7. Which variables are significant, or have levels that are significant? Use 0.10 as your p-value cutoff. (Select all that apply.)

8. Using the model from train data set (best fit), make predictions on the test set. What is the sum of squared errors of the model on the test set?

9. What is the test set R-Squared?

10. What is the largest absolute error that we make in our test set predictions? In which period (Month, Year pair) do we make the largest absolute error in our prediction?