15.071: The Analytics Edge **Fall 2022**
Homework 1: Linear Regression

Out: September 12 Due: September 25 11:59 pm.
Please upload your write-up and code in pdf format with file name "Lastname 15071 _HW1.pdf".

Problem 1: Climate Change (20 points) *We strongly recommend that you complete this problem before doing Problem 2, as it will reinforce the* **R** *programming skills that you will need for Problem 2.*

In this problem, we will attempt to study the relationship between average global temperature and several other environmental factors that affect the climate. There have been many studies documenting that the average global temperature has been increasing over the last century. The projected consequences of a continued rise in global temperature are dire, with rising sea levels and an increased frequency of extreme weather events potentially affecting billions of people.

The file **climate_change.csv** contains climate data from May 1983 to December 2008. Below is a brief description of all the variables in this dataset:

- **Year**: the observation year.
- **Month**: the observation month.
- **Temp**: the difference in degrees Celsius between the average global temperature in that period and a reference value. This data comes from the Climatic Research Unit at the University of East Anglia.
- **CO2, N2O, CH4, CFC-11, CFC-12**: atmospheric concentrations of carbon dioxide (CO2), nitrous oxide (N2O), methane (CH4), trichlorofluoromethane (CCl3F; commonly referred to as CFC-11) and dichlorodifluoromethane (CCl2F2; commonly referred to as CFC-12), respectively. This data comes from the ESRL/NOAA Global Monitoring Division CO2, N2O and CH4 are expressed in ppmv (parts per million by volume – i.e., 397 ppmv of CO2 means that CO2 constitutes 397 millionths of the total volume of the atmosphere) CFC-11 and CFC-12 are expressed in ppbv (parts per billion by volume).
- **Aerosols**: the mean stratospheric aerosol optical depth at 550 nm. This variable is linked to volcanoes, as volcanic eruptions result in new particles being added to the atmosphere, which affect how much of the sun's energy is reflected back into space. This data is from the Godard Institute for Space Studies at NASA.
- **TSI**: the total solar irradiance (TSI) in W/m2(the rate at which the sun's energy is deposited per unit area). Due to sunspots and other solar phenomena, the amount of energy that is given off by the sun varies substantially with time. This data is from the SOLARIS-HEPPA project.
- **MEI**: multivariate El Nino Southern Oscillation index (MEI), a measure of the strength of the *El Nino/La Nina-Southern Oscillation* (a weather effect in the Pacific Ocean that affects global temperatures). This data comes from the ESRL/NOAA Physical Sciences Division.

We are interested in studying how changes in the environmental factors affect future temperatures, as well as how well these variables explain temperature changes so far. To do this, first read the dataset **climate_change.csv** into **R**.

Then, split the data into a training set, consisting of all the observations up to and including 2002, and a test set consisting of the remaining years (*hint: use subset*). A training set refers to the data that will be used to build the model (this is the data we give to the lm() function), and a test set refers to the data we will use to test the predictive ability of our model.

a) (4 points) Build a linear regression model to predict the dependent variable **Temp**, using **CO2, CH4, N2O, CFC-11, CFC-12, TSI, Aerosols** and **MEI** as independent variables (**Year** and **Month** should NOT be used in the model). Use ONLY the training set to train your model, and report both the training set R^2 and the test set OSR^2.

b) (3 points) Which variables are significant in your model? (We will consider a variable signficant only if the p-value is below 0.05).

c) (3 points) The current scientific opinion is that **N2O** and **CFC-11** are greenhouse gases: gases that are able to trap heat from the sun and contribute to the heating of the Earth. However, the regression coefficients of both the **N2O** and **CFC-11** variables are negative, indicating that, all else equal, larger atmospheric concentrations of either of these two compounds are associated with lower global temperatures.

Which of the following is the simplest correct explanation for this contradiction?

- Climate scientists are wrong that **N2O** and **CFC-11** are greenhouse gases **-** this regression analysis constitutes part of a disproof.
- There is not enough data, so the regression coefficients being estimated are not accurate.
- All of the gas concentration variables reflect human development **-** **N2O** and **CFC-11** are correlated with other variables in the data set.

d) (4 points) Compute the correlations between all the variables in the training set. Which of the independent variables is **N2O** highly correlated with (absolute correlation greater than 0.7)?

e) (3 points) Which of the independent variables is **CFC-11** highly correlated with?

f) (3 points) Given that the correlations are so high, build a model with only **N2O, TSI, Aerosols**, and **MEI** as independent variables. Remember to use the training set to build the model. Report the coefficients, the R^2 of and the OSR^2 your new reduced model.

Problem 2: Forecasting Jeep Wrangler and Hyundai Elantra Sales (80 points)

Almost all companies seek accurate predictions of future sales of their products. Clearly, if a company can accurately predict sales, it can better match production with customer demand, thus reducing unnecessary inventory costs while being able to satisfy all demand.

In this exercise you are asked to predict the monthly US sales of the Jeep Wrangler (manufactured by Fiat Chrysler Automobiles (FCA)) and Elantra (manufactured by Hyundai Motor Company) automobiles, both of which are sold all over the world. The Wrangler is a compact SUV (Sports Utility Vehicle) with off-road capability made by Jeep – a subsidiary of FCA – and the Elantra is a compact sedan manufactured by Hyundai. Herein you are asked to build a linear regression model to predict monthly US sales of the Wrangler and the Elantra using economic indicators of the United States as well as Google search query volumes. The data for this problem is contained in the file **WranglerElantra2019.csv**, which you will need to download from the Canvas course site under the "Homework" section. Each observation in the file is for a single month, from January 2010 through December 2019. The variables are described in Table 1.

Table 1: Variables in the dataset WranglerElantra2019.csv.

Variable	Description
Month.Numeric	The observation month given as a numerical value (1 = January, 2 = February, 3 = March, etc.).
Month.Factor	The observation month given as the name of the month (which will be a factor variable in R).
Year	The observation year.
Wrangler.Sales	The number of units of the Jeep Wrangler sold in the United States in the given month and year.
Elantra.Sales	The number of units of the Hyundai Elantra sold in the United States in the given month and year.
Unemployment.Rate	The estimated unemployment rate (given as a percentage) in the United States in the given month and year.
Wrangler.Queries	A (normalized) approximation of the number of Google searches for "jeep wrangler" in the United States in the given month and year.
Elantra.Queries	A (normalized) approximation of the number of Google searches for "hyundai elantra" in the United States in the given month and year.

CPI.All	The consumer price index (CPI) for all products for the given month and year. This is a measure of the magnitude of the prices paid by consumer households for goods and services.
CPI.Energy	The monthly consumer price index (CPI) for the energy sector of the US economy for the given month and year.

a) Read **WranglerElantra2019.csv** into R (do not forget to navigate to the directory on your computer containing **WranglerElantra2019.csv** first). Then split the data into a training set and test set. The training set should contain all observations for 2010–2018. The test set should consist of all observations for 2019.

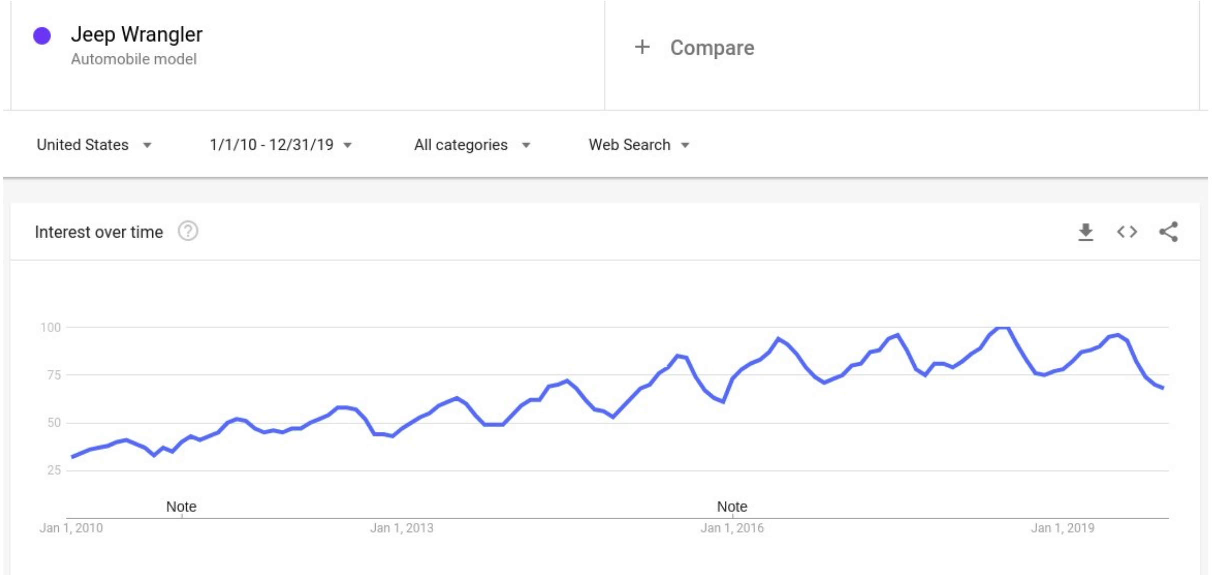
Consider the five independent variables **Year**, **Unemployment.Rate**, **Wrangler.Queries**, **CPI.Energy**, and **CPI.All.** Use the training set to build a linear regression model to predict monthly Wrangler sales, and do not add any additional variables beyond the five indicated independent variables.

i) (8 points) Build an initial linear model with all five independent variables. Based on model output, which variables are significant, i.e. have at least one "star" in the summary output (or more mathematically, have a p-value less than 0.05)?

ii) Using your understanding of variable significance, choose a subset of these five variables and construct a new regression model to predict monthly Wrangler sales (**Wrangler.Sales**).

1) (4 points) Justify your choice of variables.
2) (4 points) What is the linear regression equation produced by your new model, and what is your interpretation of the coefficients for the independent variables?
3) (4 points) Do the signs of the model's coefficients make sense?

iii) (6 points) How well does the model predict training-set observations, as captured, for instance, by the R^2 value of the model? In a similar spirit, how well does the model predict test-set observations, as captured, for instance, by the OSR^2 value of the model?



Jeep Wrangler
Automobile model

United States 1/1/10 - 12/31/19 All categories Web Search

Figure 1: “Jeep Wrangler” search queries (normalized) from Jan. 1, 2010 to Dec. 31, 2019.

b) **Seasonality.** One of our independent variables, **Wrangler.Queries**, was obtained from publicly available data in Google Trends. In Figure 1, we show a time-series plot of search queries for “Jeep Wrangler” over the time span from 2010–2019.

i) (6 points) Analyze Figure 1 for seasonal fluctuations. What trend do you observe? Does this trend make an intuitive business sense?

ii) (6 points) Drawing on intuition from Figure 1, let us now try to further improve the linear regression model by modeling seasonality. Construct a new linear regression model by including **Month.Factor** variable in your set of independent variables. Describe your new regression model and interpret of the coefficients of each of the **Month.Factor** categorical variables.

iii) (6 points) Which variables are significant? What is the training set R^2? Test set OSR^2?

iv) (6 points) Do you think adding the independent variable **Month.Factor** has improved the quality of your model? Why or why not?

v) (6 points) Can you think of a different way that you might use the given data to model seasonality? Do you think your new way would improve on the best model you have constructed so far? (*Note:* You don’t need to train your proposed model to answer this question)

c) **Elantra Sales.** Now, on the same training data set, construct a linear regression model to predict the outcome variable **Elantra.Sales** (i.e., monthly US Elantra sales) using a subset of the independent variables **Year**, **Unemployment.Rate**, **Elantra.Queries**, **CPI.Energy**, and **CPI.All**.

i) (6 points) Report both of your training set R^2 and the test set OSR^2. (*Note:* Your model will probably not look very good.)

ii) (8 points) Plot and compare the Sales of "Jeep Wrangler" and "Hyundai Elantra" from the time period January 1, 2010 to December 31, 2018. What do you observe? Do you think that adding seasonality would improve your model to predict Elantra sales? Why or why not?

d) Linear regression is a modeling tool that is designed to be used to generate high-quality *predictions*. But prediction is different than *recommendation/optimization*, which are tasks that managers also need to do well in the context of data analytics. To illustrate, suppose we must decide how many automobiles to produce in March in order to satisfy uncertain demand in the month of April. Suppose for simplicity that inventory of vehicles at the start of April is zero vehicles. Suppose our linear regression model predicts that the expected demand in April will be 15,855 vehicles with a standard deviation of 1,118 vehicles. Should we decide to produce 15,855 vehicles? More than 15,855 vehicles? Fewer than 15,855 vehicles?

To address these questions, please answer the following:

i) (4 points) What might be some of the various costs associated with producing <u>more</u> vehicles than the actual demand during a month? Similarly, can you think of some of the various costs associated with producing <u>fewer</u> vehicles than the actual demand?

ii) (4 points) Suppose as above that vehicle inventory at the start of April is zero. In the context of automobile production, do you think the cost of overproduction for April demand is larger or smaller than the costs of underproduction for April? Given that our linear regression model predicts that the expected value of demand in April will be 15,855 vehicles with a standard deviation of 1,118 vehicles, should we decide to produce 15,855 vehicles/more than 15,855 vehicles/fewer than 15,855 vehicles?

iii) (2 points) Can you think of ways to do prediction that might explicitly account for the differences between costs when the prediction is higher versus lower than actual demand?