CARNEGIE MELLON UNIVERSITY
DATA, INFERENCE & APPLIED MACHINE LEARNING ([redacted])
ASSIGNMENT 4

INSTRUCTIONS

- Submissions should be made via canvas.
- **Single** Python/MATLAB code file(.ipynb or .m) **[Do not Submit checkpoints for .ipynb]**. In addition, each line of code should be documented by text. This demonstrates that the code is unique and owned by the student.
- Assignment report(.pdf) with full evidence that the assignment was completed by the student and demonstrate a full understanding of each step in the process including textual descriptions of each result (statistics, table, graph etc) represents and insights that can be gained.
- Indicate the libraries you have used in your code at the beginning of the report (After the title page).
- Using ChatGPT for any assignment is not allowed as it could lead to being flagged for plagiarism.
- Data files (as given).

Submission process:

1. Put source code **file and data files** in a single folder
2. Name of the folder should be the same as your andrew ID
3. [redacted]
4. After attaching zipped file, click on "Add Another File" from assignment submission page and **attach your report**
5. Submit your assignment

N.B. This process will allow us to compile your reports in **Turnitin** to check for plagiarism.

Specific reasons for a submission being classified as incomplete include:

- Failure to correctly name your folder with your Andrew ID
- [redacted]
- A missing report describing the steps, results, and insights
- A missing dataset required for running the code
- A missing code file such as .ipynb or .m file
- An error in the file path needed to run the code

The student is responsible for checking that their submission is complete. Students will lose 10% as for late submission even if the submission is repaired during the 24 hours after the deadline has passed, and receive 0 for the assignment if it is not repaired.

[redacted]

1. Linear regression with one explanatory variable (20 points)

Load in monthly house prices data in pounds sterling (£) from Jan 1991 to Dec 2016 from *monthlyHousePricesUK.csv* and the FTSE100 index from Yahoo Finance (ticker = ^FTSE) over the same period (01-Jan-1991 to 31-Dec-2016).

a) Using the FTSE100 index monthly returns as dependent variable and the house prices monthly returns as explanatory variable, create a regression model with MATLAB/Python including a constant and calculate the correlation coefficient **(5points)**
b) What do the results tell us? **(5points)**
c) Use a hypothesis test to back up your conclusion about the existence of a significant relationship between these two variables. **(10 points)**

2. Linear regression with multiple explanatory variables (30 points)

The *college.csv* file contains information about different US colleges and universities. We are going to use the number of applications received, the number of enrolled students, the number of out of state students, the number of admitted students who were in the top 10% and number of admitted students who were in the top 25% of their class to predict the graduation rate.

a) Calculate the correlation coefficients of the aforementioned variables. **(5 points)**
b) Considering the graduation rate as the dependent variable, use stepwise to build the linear regression model. **(5 points)**
c) Which predictor variables are useful in predicting the graduation rate? Explain how you got those variables. **(5 points)**
d) Would the set of predictor variables be useful in predicting the graduation rate if you were to use BIC to select the model? Why? **(5 points)**
e) Compare the accuracy of the model using only useful predictors with the accuracy of the model using all five predictors? **(5 points)**
f) Given a set of predictors corresponding to Carnegie Mellon University, what graduation rate value should the most accurate model predict? **(5 points)**

3. Open study (30 points)

Design and undertake a study to assess a trend in the domain of transport for one or more countries of your choice. Your study should be based on publicly available data and explained using mathematical facts. Explain assumptions, methodology and findings. An example would be to study the relationship between increase in transport and road traffic accidents. The World Health Organization has data for road traffic deaths per country in 2010 and there is a World Bank indicator for Passenger cars (per 1000 people). Can you predict the situation in 2021?

Deliverables: You should turn in a report that includes the trend you are studying, the data source **(5 points)**, your assumptions **(5points)**, the methodology used along with its implementation in MATLAB/Python **(10 points)**, and finally the findings and conclusions, which should be backed with code and figures **(10 points)**.

4. Model Fitting and Prediction (20 points)

The bank of Israel has published the data for unemployment rate (per 100 Israeli workforces) from 1980-12-31 to 2013-09-02. Download the data from Quandl (code: ODA/ISR_LUR) into MATLAB/Python. Estimate the likely rate of unemployment by the year 2020 **(10 points)**. Explain how one can evaluate the accuracy of the estimate **(5 points)** and provide the accuracy as a percentage. **(5 points)**