Homework_2

September 11, 2019

For this homework we will be using the "AutoData". The relevent information for the dataset has been provided below (Data can be found in blackboard). Please, read through carefully to get an understading of the data.

Data Description:

The data concerns city-cycle fuel consumption in miles per gallon,to be predicted in terms of 3 multivalued discrete and 5 continuous attributes." (Quinlan, 1993)

1. Number of Instances: 392
2. Number of Attributes: 8 including the class attribute
3. Attribute Information:
    1. mpg: continuous (Milage per gallon)
    2. cylinders: multi-valued discrete
    3. displacement: continuous
    4. horsepower: continuous
    5. weight: continuous
    6. acceleration: continuous
    7. model year: multi-valued discrete
    8. origin: multi-valued discrete

[1]:
```
# import libraries
import pandas as pd
import numpy as np
from scipy import stats
import statsmodels.formula.api as smf

# Plotting Libraries
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline
plt.style.use('seaborn-white')

## Setting Random Seeds for reproducibility
np.random.seed(1234)
import os
os.environ['PYTHONHASHSEED']='1234'
import random as rn
```

```
rn.seed(1234)
```

[2]:
```
# Read Data
df = pd.read_csv('AutoData.csv')                    # Reads Data
df = df.dropna()                                    # Drops
df.head(5)                                          # Prints a few observations from␣
 ↪the data
```

[2]:

	mpg	cylinder	displacement	horsepower	weight	accelaration	model_year	\
0	18.0	8	307.0	130	3504	12.0	70	
1	15.0	8	350.0	165	3693	11.5	70	
2	18.0	8	318.0	150	3436	11.0	70	
3	16.0	8	304.0	150	3433	12.0	70	
4	17.0	8	302.0	140	3449	10.5	70	

	origin	car_name
0	1	chevrolet chevelle malibu
1	1	buick skylark 320
2	1	plymouth satellite
3	1	amc rebel sst
4	1	ford torino

[3]:
```
print('Number of Observations:',df.shape[0])
print('Number of Variables:',df.shape[1])
```

```
Number of Observations: 392
Number of Variables: 9
```

1 Part 1: Gradient Descent Optimization (30%)

In this part your task is to create a regression model to predic ['mpg'] using the variables ['displacement', 'horsepower', 'weight', 'accelaration'].

Perform the following tasks:

1. (5%) Split the data into Train-Test [70%-30%] (Note: Don't randomize the data.).
2. (20%) Build a Gradient Descent Optimizer to Calculate the Co-efficient of the Regression Model. Validate both results by using the closed form formula ($\hat{\beta} = (X^T X)^{-1} X^T y$).
3. (5%) Calculate the Pearson Correlation for the model for the test data (Hint: Check Multiple_Linear_Regression.ipynb sample file for example).

2 Part 2: Exploring the regression Model for separate variables (70%)

In this part you are allowed to use libraries like statsmodel/scipy etc. In this part your task is to create a regression model to predic ['mpg'] using all the variables except ['car_name'].

Perform the following tasks:

1. (10%) Perform a correlation analysis of the data. What variables do you think effect 'mpg' the most (Hint: Check the correlation coefficients)?

2. (45%) Fit a regression model to predic ['mpg'] using all the variables except ['car_name']. Answer the following question:

    a. What variables/predictors do you think best suited to build/fit the model to predict 'mpg' the most (Hint: Check the p-values after fitting the model.)?

    b. Based on your observation update the model for reduced number of variables?

3. (15%) Fit a model with interaction between ['horsepower' & 'weight'] and check the pearson correlation for the predicted model (Hint: This correlation has to be calculated using the test data). Did the interaction term improved the model? if yes, why do you think so? Explain