ECS 171: Homework 1

Hong Fan

912524085

Question1:

The auto-mpg data was downloaded from UCI Machine Learning Repository and was loaded into Matlab after removing the column—“car name” and 6 records which have missing values.

Prctile() command is used to find thresholds for low, medium and high mpg so that all samples will be divided into these three categories as equally as possible.

mpg = 18.6144 is the threshold of low and medium mpg.
Mpg = 26.9144 is the threshold of medium and high mpg.

Question2:

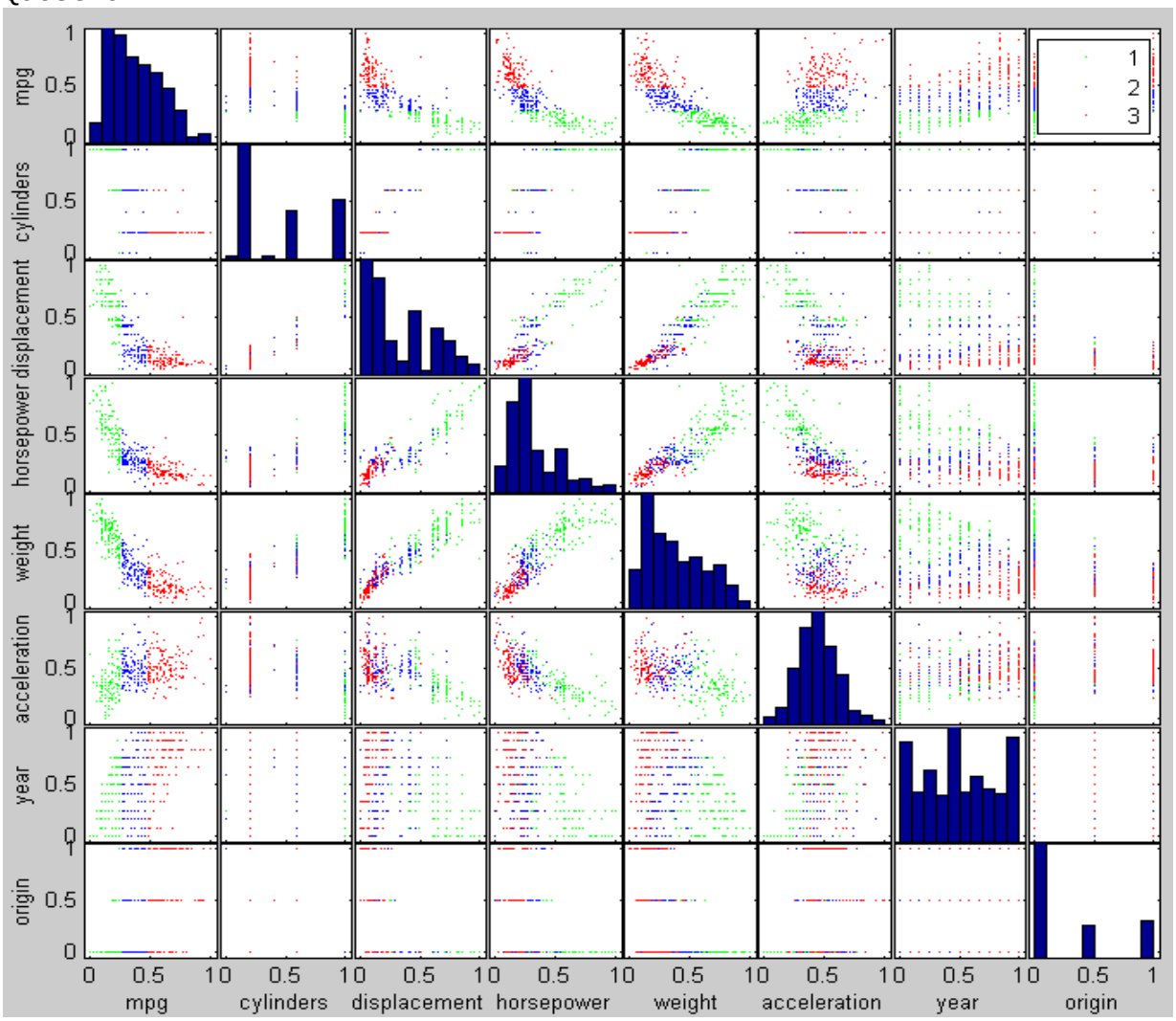

Figure 1 2-D Scatter Plot Matrix

In Figure 1, low, medium and high mpg are represented by color green, blue and red respectively.

In order to plot more easily, I used 1, 2, 3 to represent low mpg, medium mpg and high mpg (shown in the code). I also normalized each variable before ran regressions.

Regarding the three mpg categories, pair-wise combinations among mpg with displacement, horsepower and weight are more informative than others. Among these pair-wise combinations, the 2-D scatter plot above shows clear linear patterns implying mpg tends to decrease as horsepower, weight and displacement go up.

Question3:

The code for this question is saved in OLSestimate.m.

Question4:
Mean squared error is estimated by:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

Where n is the sample size. Y hat is the estimated mpg rating and Y is the actual mpg rating.

In OLSestimate(data_normalized, order_poly, one_ind_var), one_ind_var asks the user to type index(2-cylinders, 3- displament, 4- horsepower, 5 – weight, 6- acceleration, 7-model year, 8- origin).

Table1: Training MSE for different single variables with different orders:

	Polynomial Order				
Single Variable in the Model	0	1	2	3	4
Cylinders	0.0279	0.0088	0.0087	0.0078	0.0077
Displacement	0.0279	0.0077	0.0064	0.0063	0.0061
Horsepower	0.0279	0.0100	0.0075	0.0075	0.0074
Weight	0.0279	0.0060	0.0047	0.0046	0.0046
Acceleration	0.0279	0.0216	0.0210	0.0208	0.0204
Model year	0.0279	0.0256	0.0254	0.0254	0.0252
Origin	0.0279	0.0176	0.0169	0.0489	3.4420e+26

Table2: Testing MSE for different single variables with different orders:

	Polynomial Order				
Single Variable in the Model	0	1	2	3	4
Cylinders	0.0328	0.0186	0.0185	0.0173	0.0173
Displacement	0.0328	0.0141	0.0118	0.0118	0.0115
Horsepower	0.0328	0.0134	0.0125	0.0123	0.0123
Weight	0.0328	0.0110	0.0109	0.0108	0.0107
Acceleration	0.0328	0.0319	0.0318	0.0316	0.0316
Model year	0.0328	0.0305	0.0287	0.0261	3.3293e+11
Origin	0.0328	0.0240	0.0225	2.6380	1.4444e+32

The following 7 plots show testing data and polynomial model fittings with single variable and different orders. Red line represents 0 order polynomial, blue is for 1st order polynomial fitting, black for 2nd order polynomial fitting, green for 3rd order polynomial fitting, and yellow is for 4th order polynomial fitting

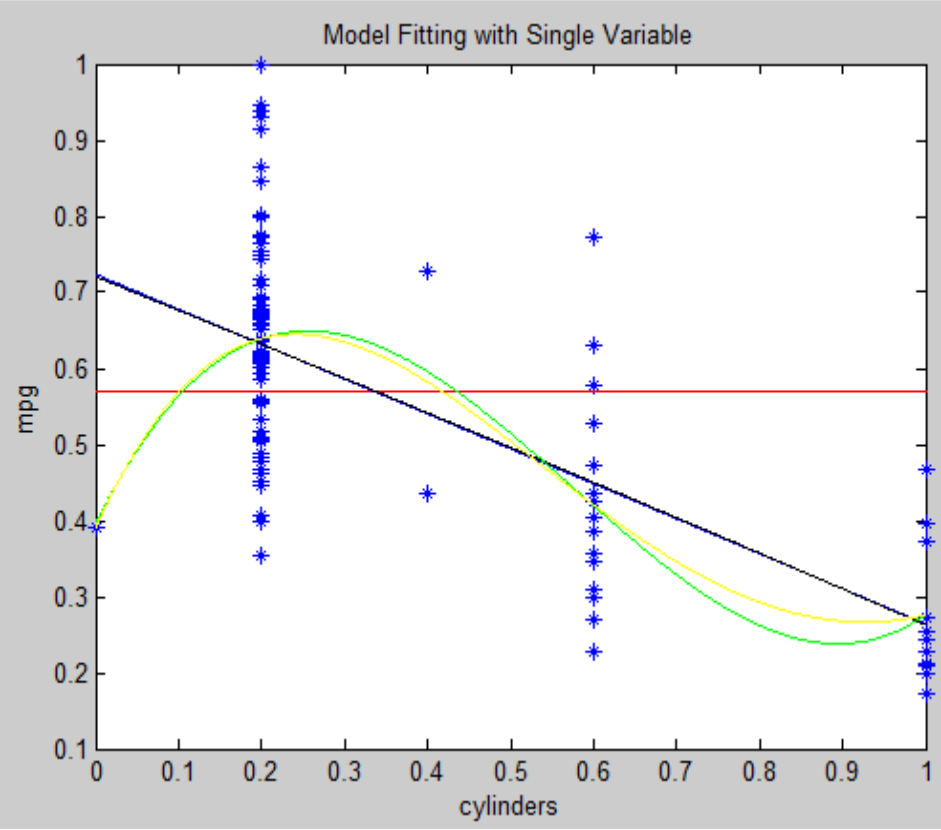

Figure2. 0th to 4th order polynomial (mpg vs cylinders)

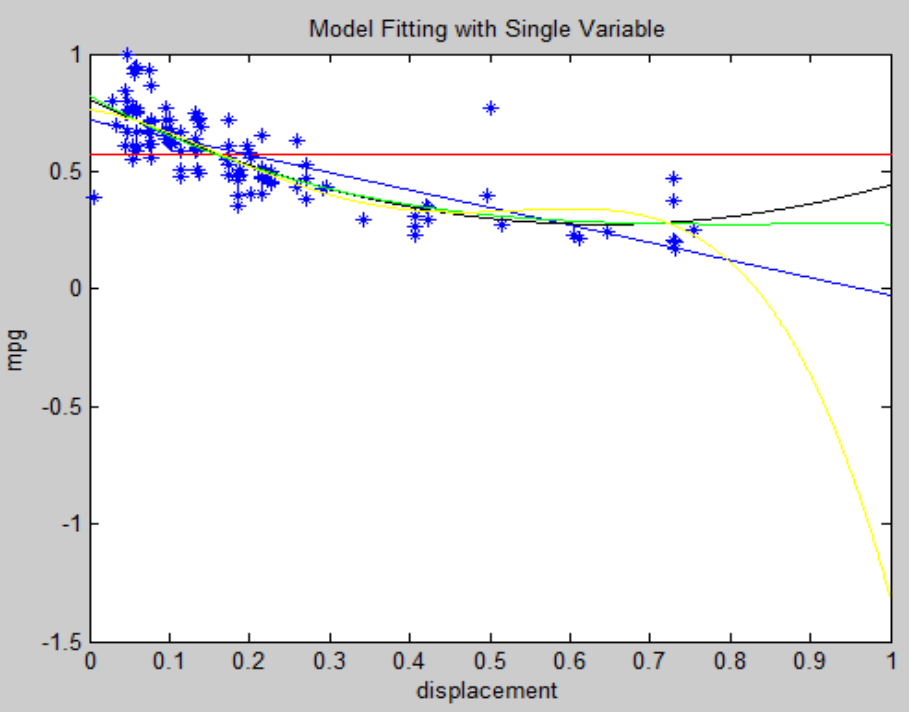

Figure3. 0^{th} to 4^{th} order polynomial (mpg vs displacement)

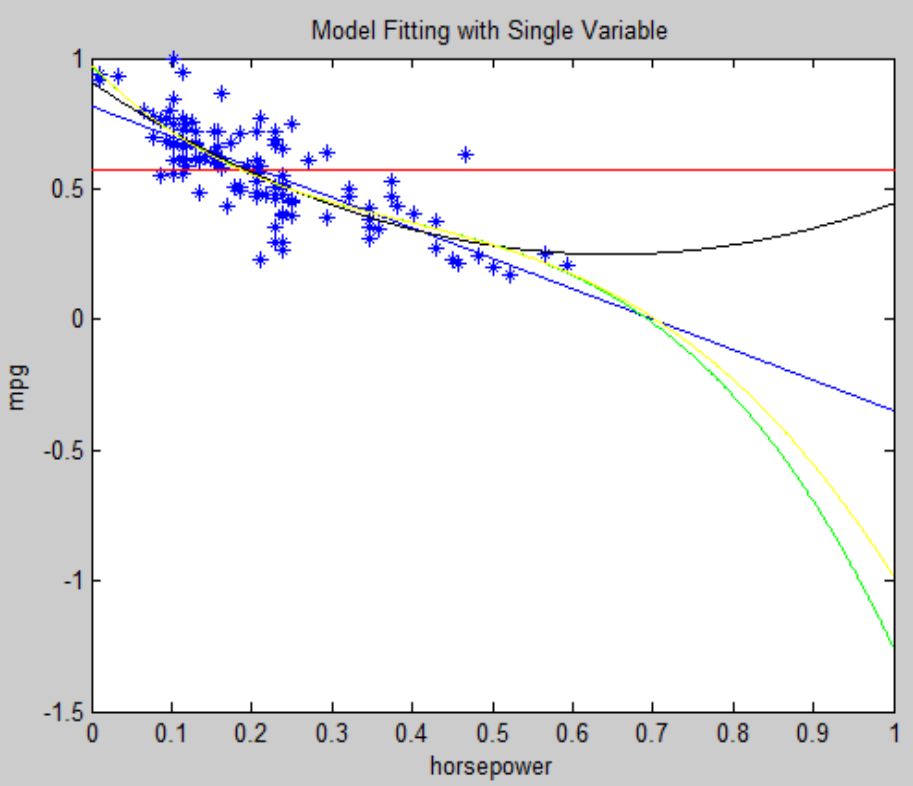

Figure3. 0^{th} to 4^{th} order polynomial (mpg vs horsepower)

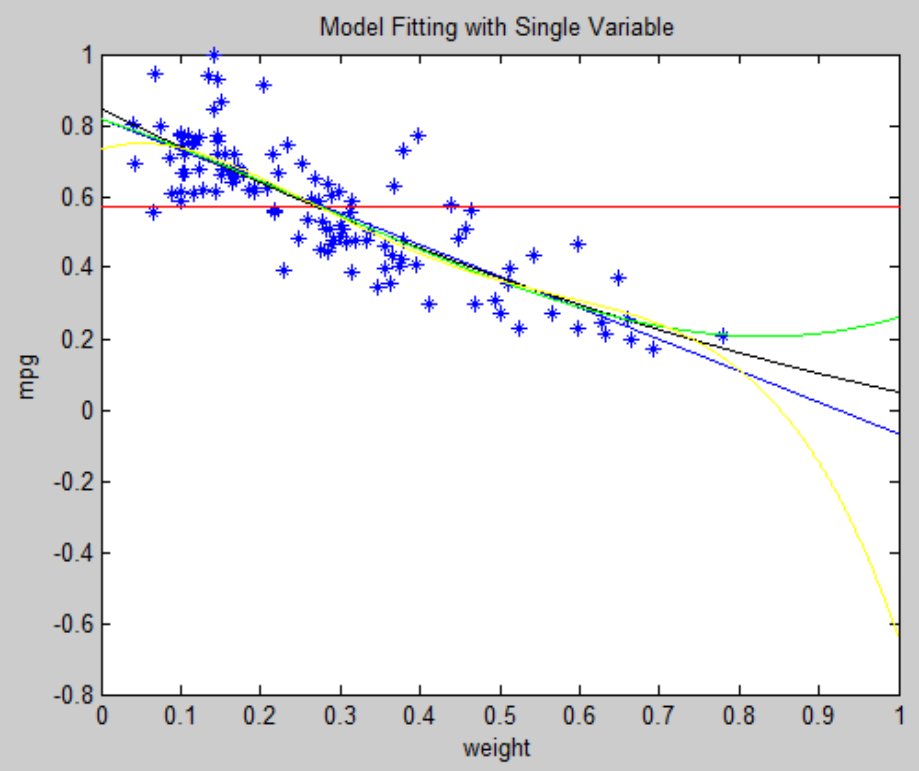

Figure4. 0th to 4th order polynomial (mpg vs weight)

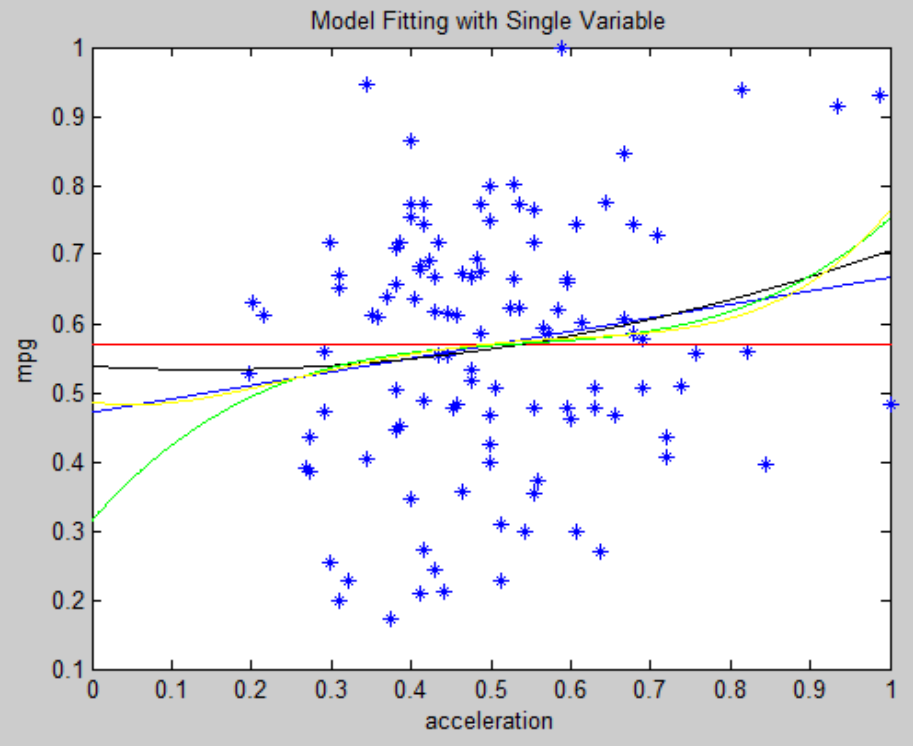

Figure5. 0th to 4th order polynomial (mpg vs acceleration)

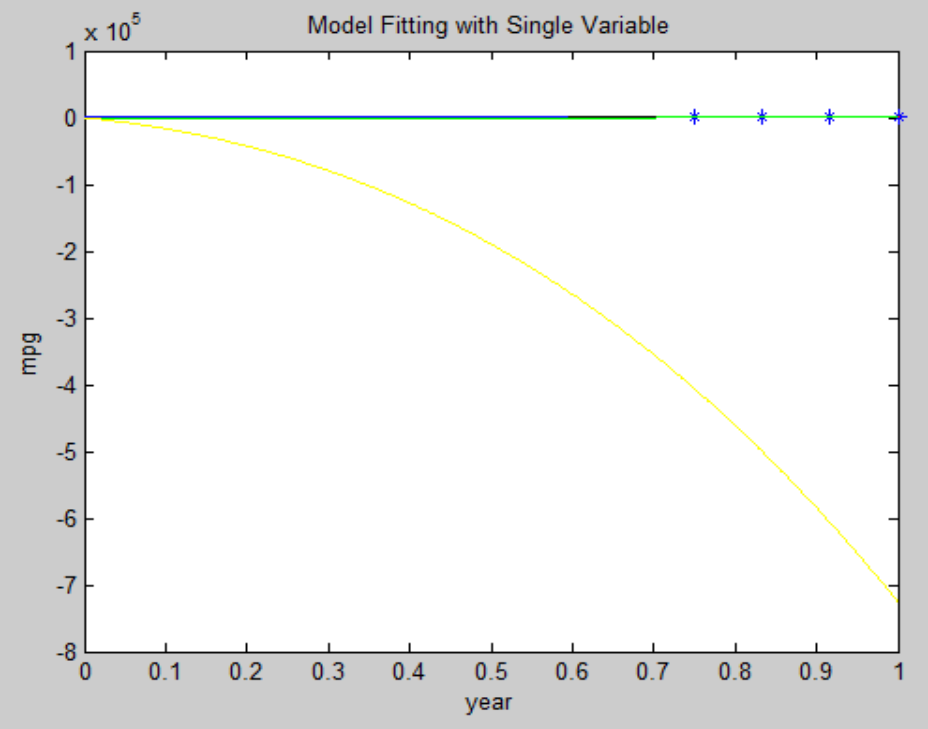

Figure6. 0th to 4th order polynomial (mpg vs year)

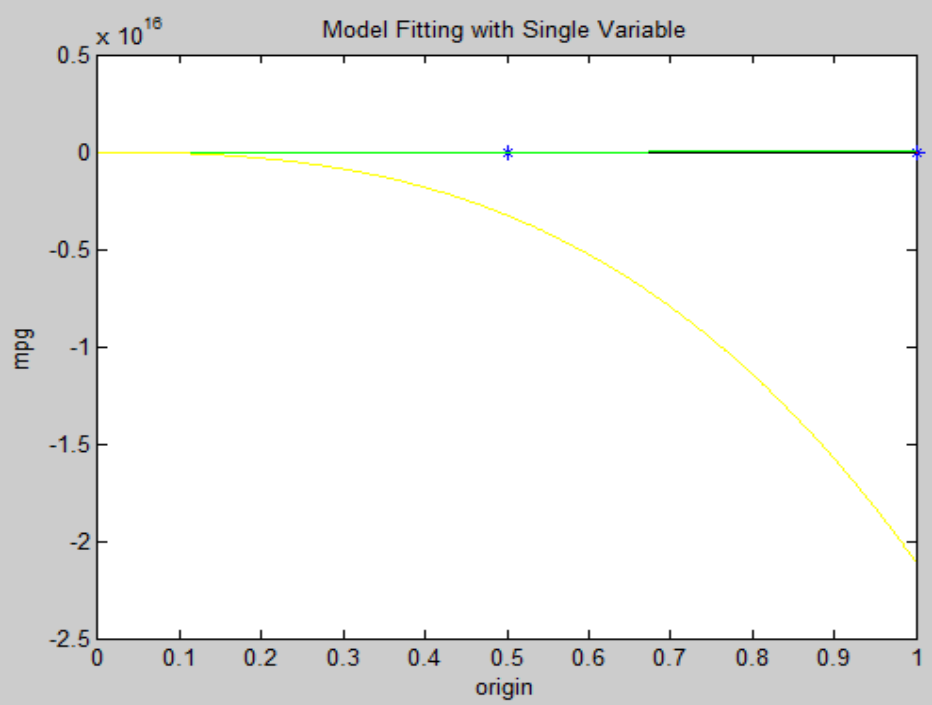

Figure7. 0th to 4th order polynomial (mpg vs origin)

The above plots show that the second order polynomial performs the best in the test set. Compared to blue lines(1^{st} order polynomial), black lines (2^{nd} order polynomial) successfully capture the “level off” trend of mpg vs horsepower and displacement (Figure2, Figure3, Figure4).

The above plots also show horsepower and weight are the most informative variables regarding mpg consumption as the plots between them and mpg consumption show clear pattern.

Question 5:

Please check the code in this file: OLSestimate_all7var.m

MSEs in training and test data sets when fitting 0^{th}, 1^{st}, 2^{nd} polynomial models using all 7 independent variables:

	Polynomial Model Order		
	0^{th} order	1^{st} order	2^{nd} order
Training data	0.0279	0.0049	0.0031
Test data	0.0328	0.0087	0.0071

Question6:

MSEs when fitting first order logistic regression using stochastic gradient descent method:

	First Order Logistic Regression
Training data	0.0037
Test data	0.0086

Before min_max normalization, mpg = 18.6144 is the boundary separating low mpg and median mpg and mpg = 26.9144 is the boundary separating median mpg and high mpg. After normalization, p1_normalized = 0.2557, p2_normalized = 0.4764 which are used as threshold when fitting logistic model.

Question7:

Before predicting its mpg rating, all its 7 features were normalized (shown in code).

After normalization, threshold of low and medium mpg = 0.2557 and threshold of medium and high mpg =0.4764 (shown in the code)
Use second-order, multi_variate polynomial:

mpg_pred = X*theta = 0.3280
According to the threshold, its mpg is medium.

Use logistic regression and set alpha = 0.05 when using gradient descent method:
w = [-0.0489; 0.0168;0.5651;-1.0329;-2.8594;0.0663;1.2035;0.2445];
X_d = X(1:8)
mpg_pred = (1 + exp(- X_d* w)).^(-1) = 0.2689
Thus its mpg is in the medium category.