Jin Yang

INFO 6105
Data Science Engineering Methods and Tools

Northeastern University, Fall 2019

Problem Set 1, Due: Oct 12, 2019

Problem Set Rules:

1. Each student should hand in an individual problem set at the beginning of class.
2. Discussing problem sets with other students is permitted. Copying from another person or solution set is *not* permitted.
3. Late assignments will *not* be accepted. No exceptions.

1. (Total: 50 points)

   In this question, you should use the Carseats data set to predict the sales in a new store with Price=$120, Advertising=$10000, ShelveLoc = Good, 'Urban=Yes, US=Yes.

   (a) (4 points) Fit a multiple regression model to predict Sales using Price, Advertising Urban, and US. Write out the model in equation form, being careful to handle the qualitative variables properly.

Model:

$$Sales = -0.0520\,Price + 0.1311\,Advertising - 0.0896\,Urban - 0.1038\,US + 12.8298$$

Here, Urban and US are set to 1 if Yes

$$Sales = -0.052 \times 120 + 0.1311 \times 10 - 0.0896 - 0.1038 + 12.8298 = 7.7074$$

(b) (4 points) Provide an interpretation of each coefficient in the model. Be careful–some of the variables in the model are qualitative!

-0.052 Price : Sales goes down 0.052 when price increase 1

0.1311 Advertising : Sales increase 0.1311 when Advertising increase 1

-0.0896 Urban
-0.1038 US
> Sales decrease when Urban / US change from No to Yes

(c) (8 points) Using the model from (a), predict sales in the new store and calculate 68% and 95% confidence intervals.

MSE = 5.4313 RMSE = $\sqrt{MSE}$ = 2.3305

Mean = 7.496

68% ⇒ Mean ± RMSE ⇒ (5.1658, 9.8268)

95% ⇒ Mean ± 2 × RMSE

⇒ (2.8353, 12.1573)

(d) (8 points) Using the model from (a), what is the probability that sales will be greater than 12000 units in the new store?

$$Z = \frac{(X-\mu)}{RMSE} = \frac{12-7.4863}{2.3305} = 1.93$$

Look up the z-table, we get 0.9732

probability $= (1 - 0.9732) \times 100\% = 2.68\%$

(e) (8 points) Using the model from (a), what is the probability that sales will be between 6000 and 10000 units in the new store?

$$Z_1 = \frac{(X-\mu)}{RMSE} = \frac{10-7.4863}{2.3305} = 1.074$$

$$Z_2 = \frac{7.4863-6}{2.3305} = 0.6421$$

$p1 = 85.769\%$ $p2 = 73.891\%$

$p = 1 - (1-85.769\%) - (1-73.891\%)$

$= 59.66\%$

(f) (5 points) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

predictors : ShelveLoc , Price , Advertising

Reason : $P > |t| = 0.00$

(g) (8 points) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome. Using this model, predict sales in the new store and calculate 68% and 95% confidence intervals.

Mean = 7.4963 MSE = 3.2144 RMSE = 1.7928

Sales = 11.2824 + 4.8486 ShelveLoc[T.Good]
+ 1.8593 ShelveLoc[T.Medium] - 0.0562 Price +
0.1064 Advertising

68% : $11.2824 \pm 1.7928 \Rightarrow (9.4895, 13.0753)$

95% : $11.2824 \pm 1.7928 \times 2 \Rightarrow (7.6966,$
$14.8682)$

(h) (5 points) How well do the models in (a) and (g) fit the data?

R-Squared of (a) = 0.261

R-Squared of (g) : 0.606

(a) fits not that well

(g) fits well

2. (Total: 50 points) This problem involves the sales data set for Toyota Corolla, which can be found in the file ToyotaCorolla.csv. The data set contains 1436 observations on the following 10 variables.

Price (in Dollars)

Age (in months)

Mileage

FuelType Fuel Type (diesel, petrol, CNG)

MetColor Metallic color (1=yes, 0=no)

Automatic Automatic transmission (1=yes, 0=no)

Displacement Engine displacement (in cu. inches)

Doors Number of doors

Weight (in pounds)

Horsepower Engine horsepower

(a) (3 points) Which of the predictors are quantitative, and which are qualitative?

Quantitative: Price, Age, Mileage, Displacement, Doors, weight, Horsepower

Qualitative: FuelType, MetColor, Automatic

(b) (3 points) What is the range (i.e., min and max) of each quantitative predictor?

Price (5351, 39975)

Age (1, 80)

Mileage (1, 150993)

Displacement (79, 122)

Doors (2, 5)

weight (2205, 3560)

Horsepower (69, 192)

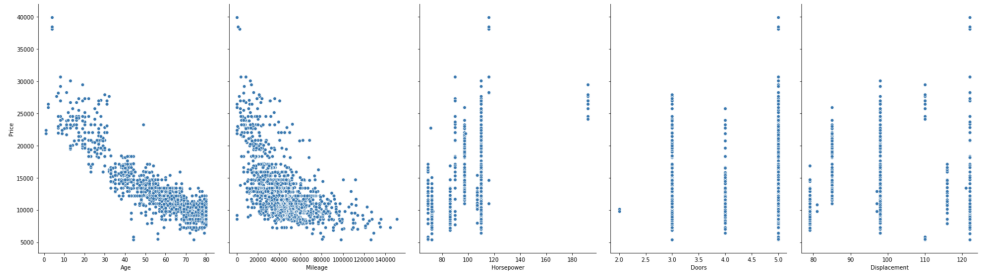
(c) (3 points) What is the mean and standard deviation of each quantitative predictor?

	Price	Age	Mileage	Displacement
mean	13199.27	55.95	42584.59	95.72
std	4461.16	18.60	22305.40	11.59

	Doors	Weight	Horsepower
mean	4.03	2364.44	101.50
std	0.95	115.95	14.98

(d) (5 points) Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

```
sns.pairplot(data, x_vars=['Age', 'Mileage', 'Horsepower', 'Doors', 'Displacement'], y_vars='Price', height=7, aspect=0
<seaborn.axisgrid.PairGrid at 0x12af10690>
```



① Price are related to Doors, Displacement, Horsepower

② As Age grows, Price goes down

(e) (4 points) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables.

Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

Horsepower, Doors, Displacement, Age, Price are all useful in predicting mpg

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.39e+04	1.32e+04	-2.560	0.011	-5.99e+04	-7922.837
Age	105.0537	55.599	1.889	0.059	-4.034	214.141
Price	-3.0630	0.269	-11.394	0.000	-3.590	-2.536
Horsepower	-250.2270	36.052	-6.941	0.000	-320.962	-179.492
MetColor	-626.3448	1014.534	-0.617	0.537	-2616.910	1364.220
Automatic	-3841.4041	2097.816	-1.831	0.067	-7957.421	274.613
Displacement	652.3795	58.038	11.241	0.000	538.506	766.253
Doors	393.0002	527.586	0.745	0.456	-642.150	1428.150
Weight	30.7931	7.074	4.353	0.000	16.914	44.673

(f) (8 points) Fit a simple linear regression with Price as the response and Age as the predictor.

(i) Is there a relationship between the predictor and the response?

Yes, there is a relationship between them.

(ii) How strong is the relationship between the predictor and the response?

The relationship is very strong.

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.488e+04	203.983	121.966	0.000	2.45e+04	2.53e+04
Age	-209.0346	3.462	-60.374	0.000	-215.828	-202.241

(iii) What is the predicted price associated for a car with an age of 48 months? What are the associated 95% confidence intervals?

predicted price = 2.488e+04 - 209.0346 * Age = 34913.66

mean = 13199.27 MSE = 3586371.96 RMSE = 1893.77

95% confidence intervals: mean +- RMSE * 2

(9411.73, 16986.81)

(g) (12 points) Fit a multiple linear regression with Price as the response and all other variables the predictors.

(i) Is there a relationship between the predictors and the response?

Yes, there is a relationship

(ii) How strong is the relationship between the predictors and the response?

	coef	std err	t	P>|t|	[0.025	0.975]
const	-5555.1340	1377.041	-4.034	0.000	-8256.957	-2853.311
Age	-152.2536	3.670	-41.489	0.000	-159.454	-145.053
Mileage	-0.0334	0.003	-11.394	0.000	-0.039	-0.028
Horsepower	44.8037	3.607	12.420	0.000	37.726	51.881
MetColor	63.6248	105.951	0.601	0.548	-144.257	271.506
Automatic	352.0347	219.154	1.606	0.108	-77.956	782.026
Displacement	-27.5497	6.336	-4.348	0.000	-39.981	-15.118
Doors	-80.6119	55.059	-1.464	0.143	-188.640	27.416
Weight	11.4258	0.663	17.221	0.000	10.124	12.728

(iii) Which predictors appear to have a statistically significant relationship to the response?

Age, Mileage, Horsepower, Automatic, displacement, weight

(iv) What does the coefficient for the age variable suggest? How accurate can you estimate the effect of age on price?

if age increse 1, the price will decrease 152.25

(v) What is the predicted price associated for a car with a mileage of 45000 miles, 48 months, diesel,

automatic transmission, 4 doors, 2568 pounds, a displacement of 122 cu. inches, a horsepower of 90, and non-metallic color? What are the associated 95% confidence intervals?

price = -5555.13 - 0.03 * 45000 - 152.2536 * 48 + 44.80 + 352.03 - 4 * 80.61
+ 2568 * 11.43 = 15213.33

interval : 13199.27 +- 2 * 1557.76 (10083.75, 16314.79)

(h) (12 points) Which predictors predictors matter most for predicting the price for a car? (Find the first and the second most important variables)

Automatic and age