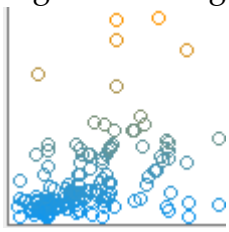
IAML - Assignment 2

s1474146

1 Simple Linear Regression

1.a

The *engine power* alone is not sufficient to predict car price. There are 2 reasons by looking at the graph below (*price* is y-axis, *engine-power* is x-axis). First, this graph suggested the presence of outlier value since most of the data points are concentrated in the bottom half of the graph. Second, the value of those 2 attributes does not seem to be strongly correlated (if they are, I those data points would be distributed along some straight line)



There are outliers value in *prices*. The *Linear Regression* model does not work well with dataset containing outliers, since those outliers with extremely large (or small) values would effect the predicted parameters a lot. To improve the performance of the model, the *log* function can be applied to the *prices* attribute.

1.b

The result of **SimpleLinearRegression** model suggests those attributes are weakly correlated (CC = 0.41)

If one more unit of engine-power is added:

- Original dataset: $0.09 * power + 3038.37$
- Modified dataset: $0.09 * power + 3038.28$
- The *intercept* is decreased by 0.09. It is just a basic arithmetic, if we add 2 unit, then the intercept will be decreased by 0.18.

By examining the magnitude of the regression coefficient, it is not possible to tell whether or not engine-power is an important influential variable on price. Since those coefficents simply shows the slope and the intercept. We have to look at the correlation coefficients in order to interpret those regression coefficients correctly.

1.c

Measurements:

- CC = 0.41. CC tells us how the *engine-power* and *price* are related linearly. If we have CC = 1, then for some increment i in *engine-power*, we will on average have $c \times i$ increment in *price*, where c is a slope of linear regression function. The bigger the $abs(CC)$, the more correlated those 2 variables will be. A way to visualize it is that if CC is equal to 1 or -1, all data points can be fitted perfectly by a straight line.
- RMSE = 6120.9. Root Mean Squared Error is one of the way to measure the error of the prediction versus the true data. RMSE is calculated by taking the root of average of the sum of squared error over all data points. By taking the square of error, RMSE place strong penalty on large error comparing to the smaller one.
- MAE = 3970. Mean Absolute Error is another way to measure error between the predicted value and true value. This one puts no penalty for the larger error. Smaller or larger error will contribute the amount according to their value.

1.d

The result of **SimpleLinearRegression** for dataset *PartA_base*:

- CC = -0.14
- RMSE = 6762
- MAE = 4934

The *simplest* baseline for predicting the price based on **linear regression** is that the model will predict the mean of the *price* for all *engine-power*'s value. Formally, it is:

$$price'_i = mean(\textbf{price}) + 0 \times enginepower_i \quad (1)$$

The function for regression line from Weka :

$$price_i = 11684.723 \quad (2)$$

And predicting the mean of *price* is what Weka did. By checking the *Preprocessing* tab, I saw that 11684.723 is the mean of attribute *price*

MULTIVARIATE LINEAR REGRESSION

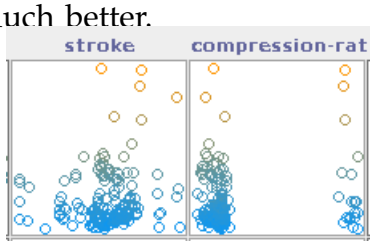
2.a

Considering each plot of each attributes against *price*, there is no single attribute that stands out from others as the strong predictor for *price*. In order to be a good predictor for price, an attribute should have strong correlation with *price*. And I could not see any pattern that suggest strong correlation.

However, there are some attributes that are more suitable for predicting price than the others. Those attributes with their distribution not skewed to one side are better predictor for price than those attributes with skewed distribution. For example:

- width
- length

Attributes useless at predicting price would possess those properties: (i) skewed distribution, (ii) the price distribution along one range of attribute's value is not different from another range of attribute's value. For example, look at the graph below, high price is encoded by orange color, low & medium price by blue color. The graph on the left plot *stroke* against *price*, and it seems like that it can predict *price* much better.



These attributes are not useful:

- torque
- compression_ratio

Significant correlation: when looking at the graph of different attributes ploted against each other, if they seem to distributed along some immaginary line with slope equal to 1 or -1.

- wheel-base & length
- wheel-base & width
- city-mpg & highway-mgp

2.b

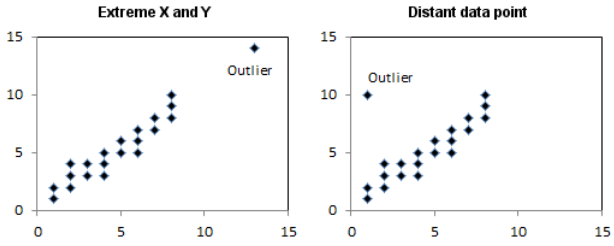
The correlation coefficient is much higher for the dataset in *2(b)* than dataset in *1(c)*. The higher value CC is, the more chance that we can fit the data with linear regression model with small error. Both the RMSE and MAE decrease by 28% and 25%.

TABLE 1

Problem number	CC	RMSE	MAE
1(c)	0.41	6120	3970
2(b)	0.73	4772	3155

2.c

Looking at the histogram of *engine-size*, there are 2 values that are completely separated from the rest (outlier). The skewed distribution caused by the outlier would affect (or would not affect) the linear regression model depending on where the outlier is according to the major data points.



The above picture is taken from (http://stattrek.com/regression/influential-points.aspx)

- Outlier would affect the model if outlier lies far away from the best fitted linear regression line (see the graph on the right)
- Outlier would not affect the model if the outlier would lie near the regression line (see the graph on the left)

TABLE 2

Dataset	CC	RMSE	MAE
Original part_B_numeric	0.73	4772	3155
Taking *log* of *engine-size*	0.81	3900	2771

To cope with outlier, I use *log* over all values of *engine-size*. The CC value increases slightly (it's good), while the errors both decreased with 27% for RSME and 18% for MAE. The performance increases because:

- Linear Regression tried to find the model that best fit the data by minimizing the squared error on training data.
- The outlier value generally are far away from the rest, and in order to accommodate outlier, the predicted model has to be shifted towward the outlier to minimize the squared error for outlier, and increasing the error for the majority of data points.

2.d

For the *width* and *height* case, the interaction term can approximately tell us the volume of the car. Since we do not have the height of the car, but based on real world observation, it can be assumed that most personal cars have the same height.

In order to find interaction terms that can improve the regression performance of the model, I ran in total 16 small experiments to test the result of combining *engine-size* with 8 other values [bore, stroke, compression-ratio, engine-power, peak-rpm, city-mpg, mean-effective-pressure, torque]. With each attributes, there will be 2 experiments:

- **Experiment A** = Both the interaction terms, and those original attributes to create that interaction terms are kept.
- **Experiment B** = The interaction terms is kept, while the original attributes of interaction terms is removed.

I left those attributes such as *wheelbase*, *width*, *height*, etc. out because I judge those value do not have any relationship at all with engine-size. Besides, the attributes *highwaympg* has strong correlation with *citympg*, therefore I think creating interaction term with one of them is sufficient. Refer to Table. 3 for the result.

TABLE 3

Interaction term with *engine-size*	CC	RMSE	MAE
compression-ratio *ExpB*	0.82	3857	2790
engine-power *ExpA*	0.79	4246	2928
city-mpg *ExpB*	0.79	4246	2928

2.e

Nominal value cannot be used in linear regression model. Because LR only works with numeric value. The output of *LR* is the weight vector (or the list of coefficients), and the value for predicted value based on input will be calculated by this equation $price_i = w_0 + w_1 a_1 + ... + w_d a_d$. The value from nominal attributes cannot be plugged into that equation.

In order to use nominal attributes for LR, a filter called **NominalToBinary** can be used to pre-process the data. For example, attribute *make* can take many text values such as *toyota*, *bmw*... By converting to binary, we can quantify how car made from different companies will contribute toward the final prices.

2.f

TABLE 4

Description	CC	RMSE	MAE
Original data - 2(b)	0.73	4772	3155
Taking *log* - 2(c)	0.81	3900	2771
Interaction terms - 2(d)	0.82	3857	2790
Complex model - 2(f) *ExpB*	0.93	2392	1612

The complex model in this problem gives the **best** result comparing with all other models. The correlation between attributes with price is highest, and the errors are smallest.

In the complex model, after converting all nominal attribute into binary, we have 76 features that can be used to predict price. While in question *2(a, b, c, d*, we only use 16 features. Disadvantages: more features means more complex model. Advantages: since we have more features, the model can fit the data better.

****

I discussed with s1474145