DATA ANALYTICS PROJECT

Linear Regression – Automobile data

Anshul Aggarwal
101410007

Anshul Saraswat Kaul
101410008

Arpit Semwal
101410009

Amandeep Grover
101590002

Team No. 7

Submitted to:
Dr. Ashutosh Mishra

Introduction to linear regression

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

- One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.
- The other variable, denoted y, is regarded as the response, outcome, or dependent variable.

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the criterion variable and is referred to as Y. The variable we are basing our predictions on is called the predictor variable and is referred to as X. When there is only one predictor variable, the prediction method is called simple regression.

Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line.

Linear regression models are often fitted using the least squares approach, “Least squares" means that the overall solution minimizes the sum of the squares of the residuals made in the results of every single equation. The best fit in the least- squares sense minimizes the sum of squared residuals, a residual being: the difference between an observed value, and the fitted value provided by a model.

The lm() function

In R, the lm(), or “linear model,” function can be used to create a simple regression model.
The lm() function accepts a number of arguments (“formula”, data). The following list explains the two most commonly used parameters.

- **formula:** describes the model, the formula argument follows a specific format. For simple linear regression, this is “YVAR ~ XVAR” where YVAR is the dependent, or predicted, variable and XVAR is the independent, or predictor, variable.
- **data:** the variable that contains the dataset

The output of this function gives us the coefficient and the intercept through which the regression line can be generated.

Data Cleansing

Data quality is a main issue in quality information management. Data quality problems occur anywhere in information systems. These problems are solved by data cleaning. Data cleansing is the process of altering data in a given storage resource to make sure that it is accurate and correct. Data cleansing is also known as data cleaning or data scrubbing.

One of the most important task in data cleansing is handling the missing data.

Few methods to handle missing data are:

- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.
- Fill in the missing value manually
- Use a global constant to fill in the missing value
- Imputation: Use the attribute mean to fill in the missing value, or use the attribute mean for all samples belonging to the same class to fill in the missing value: smarter
- Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree

Root Mean Square Error (RMSE)

It represents the standard deviation of the differences between predicted values and observed values. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation, and are called *prediction errors* when computed out-of-sample. The RMSE serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. RMSE is a good measure of accuracy, but only to compare forecasting errors of different models for a variable and not between variables, as it is scale-dependent.

The formula of RMSE is

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

where, y_j are the original values and $\hat{y}_j$ are the predicted values.

Now, in our Linear Regression Model, we are trying to find a best-fit line minimising the error E_y for every training data we have chose.

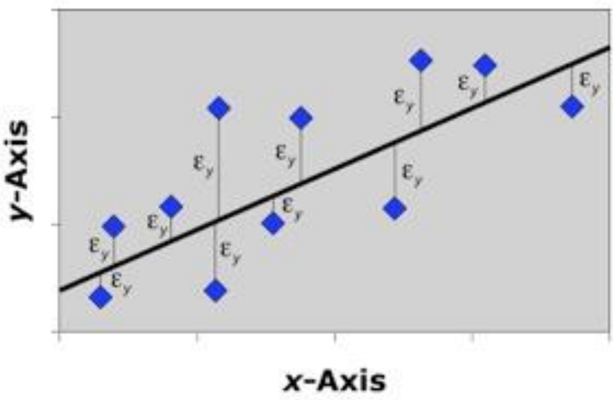

Figure 1 Scatter Plot v/s Regression line

Data Set Description

Number of Instances: 205

Number of Attributes: 26 total

- 15 continuous
- 1 integer
- 10 nominal

Attribute Information

1. **Symboling:** -3, -2, -1, 0, 1, 2, 3. -3 is safe and 3 is risky.

2. **Normalized-losses:** continuous from 65 to 256.

3. **Make:** Alfa-Romero, Audi, BMW, Chevrolet, Dodge, Honda, Isuzu, Jaguar, Mazda, Mercedes-Benz, Mercury, Mitsubishi, Nissan, Peugeot, Plymouth, Porsche, Renault, Saab, Subaru, Toyota, Volkswagen, Volvo.

4. **Fuel-Type:** Diesel, Gas.

5. **Aspiration:** Std, Turbo.

6. **Num-of-doors:** Four, Two.

7. **Body-Style:** Hardtop, Wagon, Sedan, Hatchback, Convertible.

8. **Drive-Wheels:** 4wd, fwd, rwd.

9. **Engine-Location:** Front, Rear.

10. **Wheel-Base:** Continuous from 86.6 120.9.

11. **Length:** Continuous from 141.1 to 208.1.

12. **Width:** Continuous from 60.3 to 72.3.
13. **Height:** Continuous from 47.8 to 59.8.
14. **Curb-weight:** Continuous from 1488 to 4066.
15. **Engine-Type:** dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. **Num-of-cylinders:** eight, five, four, six, three, twelve, two.
17. **Engine-Size:** Continuous from 61 to 326.
18. **Fuel-System:** 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. **Bore:** Continuous from 2.54 to 3.94.
20. **Stroke:** Continuous from 2.07 to 4.17.
21. **Compression-ratio:** Continuous from 7 to 23.
22. **Horsepower:** Continuous from 48 to 288.
23. **Peak-rpm:** Continuous from 4150 to 6600.
24. **City-mpg:** Continuous from 13 to 49.
25. **Highway-mpg:** Continuous from 16 to 54.
26. **Price:** Continuous from 5118 to 45400. - **TARGET ATTRIBUTE**

Data Pre-processing and Cleaning

The given dataset has some small inconsistencies in the form of missing data, described as under, along with how that attribute was handled:

Attribute #	**Attribute Name**	**Number of instances missing a value**	**Handling**
2	Normalized-losses	41	Mean of all available attribute values
6	Num-of-doors	2	Mode of all available attribute values
19	Bore	4	Mean of all available attribute values
20	Stroke	4	Mean of all available attribute values
22	Horsepower	2	Mean of all available attribute values
23	Peak-rpm	2	Mean of all available attribute values
26	Price	4	Mean of all available attribute values

The data now has no missing values. There was no instance of noisy data found, as given in the dataset description.

Handling Nominal Attributes

There are 10 nominal attributes, each with a different number of possible values set. These attributes have been handled as follows:

Attribute #	Attribute Name	Handling
3	Make	Removed from consideration for regression model
4	Fuel-type	0 for Diesel, 1 for Gas
5	Aspiration	0 for Standard, 1 for Turbo
6	Num-of-doors	0 for two, 1 for four
7	Body-style	Converted into 5 dummy binary variables
8	Drive-Wheels	Converted into 3 dummy binary variables
9	Engine-location	0 for front, 1 for rear
15	Engine-type	Converted into 7 dummy binary variables
16	Num-of-cylinders	Converted string number to numeric
18	Fuel-System	Converted into 8 dummy binary variables

The linear regression equation was then generated using the lm() function in R, with 43 input attributes, and 1 target attribute (i.e. Price). 40% of the dataset was used as the training set, and remaining 60% as the testing dataset. The model was constructed on the training set, and then tested on the testing dataset. The acceptable price error was set to be $2500 to calculate accuracy, and the root mean square error (RMSE) was also calculated.

The obtained results were:

RMSE = 3804.7

Accuracy = 68.55

```
Call:
lm(formula = formula, data = trainDataset)

Coefficients:
      (Intercept)           symboling   normalized.losses           fuel.type          aspiration
       -24759.293            -224.013              10.108          -17198.396            2773.078
      num.of.doors     engine.location          wheel.base              length               width
        -1258.980           10666.268             230.798             -56.170             997.097
            height         curb.weight      num.of.cylinders         engine.size                bore
         -248.356              -3.164            2236.705             142.761            1382.976
            stroke   compression.ratio          horsepower            peak.rpm            city.mpg
        -6787.816            -776.311              10.317               2.033              16.561
       highway.mpg             hardtop               wagon               sedan           hatchback
          -17.059           -6211.798           -4858.400           -5506.242           -7523.413
       convertible                X4wd                 fwd                 rwd                dohc
               NA             679.452             335.939                  NA          -20169.270
             dohcv                   l                 ohc                ohcf                ohcv
       -41150.277          -19987.850          -14889.298          -19864.223          -22929.640
             rotor              X1bbl               X2bbl               X4bbl                 idi
               NA            3609.824            3539.198                  NA                  NA
               mfi                mpfi                spdi                spfi
          423.628            3173.405                  NA                  NA
```

Figure 2 The generated Linear Regression Coefficients

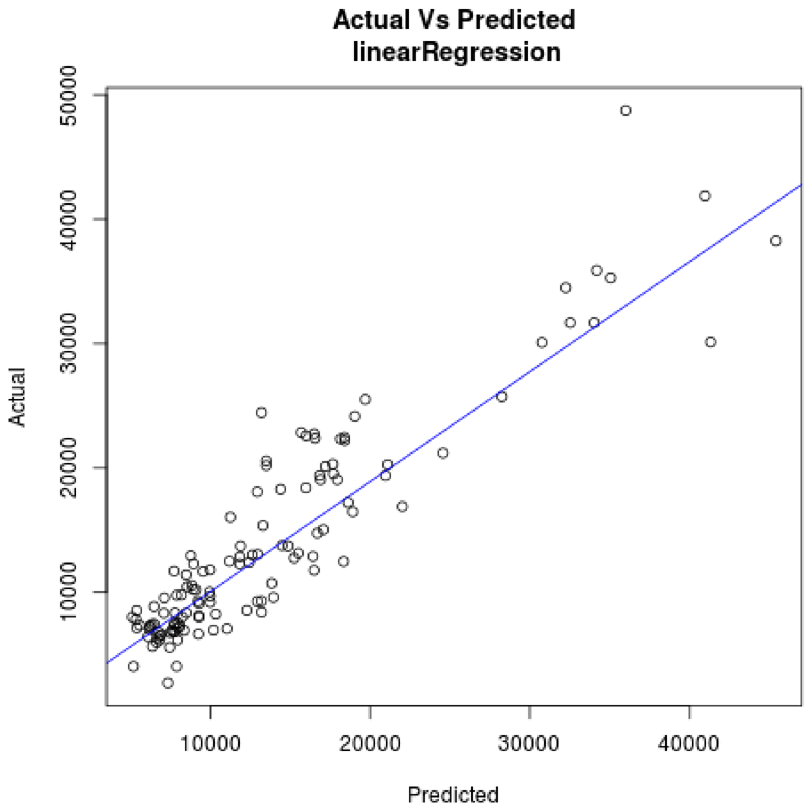

Figure 3 Actual v/s Predicted scatter plot for testing dataset