# Multiple Linear Regression

**Instructions**

Please share your answers filled in this word document. Submit code files wherever applicable.

Please ensure you update all the details:

**Name:**

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: Shahina Athar**

**Batch ID: 10122020**

**Topic: Multilinear Regression**

**Grading Guidelines:**

**1. An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline.
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of problems are completed.

(OR)

* + All assignments are submitted after the deadline.
* **Grade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of the problems are completed.

(OR)

* + Less than 80% of problems in the assignments are submitted after the deadline.
* **Grade D: (>= 60 and < 70):**
  + Assignments submitted after the deadline and with 50% or less problems.
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of problems in the assignments are submitted after the deadline.

(OR)

* + Less than 30% of problems in the assignments are submitted before the deadline.
* **Grade F: (< 50):** No submission (or) malpractice.

1. **Business Problem**
   1. **What is the business objective?**
   2. **Are there any constraints?**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its data type and its relevance to the model building. And if not relevant, provide reasons and a description of the feature.**

1. **Data Pre-processing**

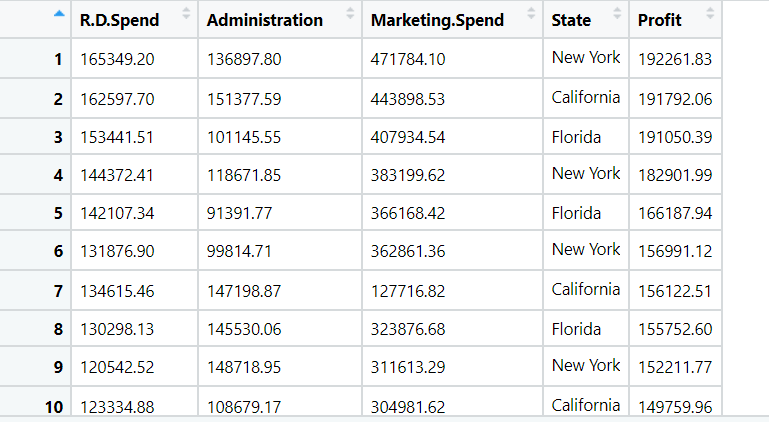
**3.1 Data Cleaning, Feature Engineering, etc.**

**3.2 Outlier Treatment.**

1. **Exploratory Data Analysis (EDA):**
   1. **Summary.**
   2. **Univariate analysis.**
   3. **Bivariate analysis.**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options).**
   2. **Perform Multi linear regression model and check for VIF, AvPlots, Influence Index Plots.**
   3. **Train and Test the data and compare RMSE values. Tabulate R-Squared and RMSE values for different models in the documentation and provide an explanation.**
   4. **Briefly explain the model output in the documentation.**
   5. **Tune the model and improve its accuracy.**
3. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Problem Statements: -**

1. An analytics company has been tasked with the crucial job of finding out what factors affect a startup company and if it will be profitable or not. For this, they have collected some historical data and would like to apply multilinear regression to derive brief insights into their data. Predict profit, given different attributes for various startup companies.



1. **Business Problem**
   1. **What is the business objective?**

**Ans:** An analytics company has been tasked with the crucial job of finding out what factors affect a startup company and if it will be profitable or not. For this, they have collected some historical data and would like to apply multilinear regression to derive brief insights into their data. Predict profit, given different attributes for various startup companies.

1. **Briefly explain the model output in the documentation.**

|  |  |
| --- | --- |
| **R-Squared** | 0.950 |
| **Train RMSE** | 7216.059795616378 |
| **Test RMSE** | 14102.686337895531 |

Equation is, y = 5.541e+04+0.8111(RandDSpend)- 0.0629(Administration)+0.0199(MarketingSpend)+608.6022(State\_Florida)+

1151.1433(State\_NewYork) -> B0+B1x+B2x+B3x+B4X+B5X

1st of all, we load the data then, convert categorical variable into dummy variables then perform all visualization including univariate, bivariate, check data is normally distributed or not, then build the model, Then apply Influence Index plots and removed high influence values so that our model will improve after that again build the model and we get the R-Squared Value which is given above and also get the equation of line Then we predict the model after that we again do  some visualizations then split into train and test and find the test RMSE  and Train RMSE.

1. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Ans:** With the help ofmultilinear regression analytics company can easily find out what factors affect a startup company and if it will be profitable or not.

1. Perform multilinear regression with price as the output variable and document the different RMSE values.



1. **Business Problem**
   1. **What is the business objective?**

**Ans:** Perform multilinear regression with price as the output variable and document the

different RMSE values.

1. **Briefly explain the model output in the documentation.**

|  |  |
| --- | --- |
| **R-Squared** | 0.950 |
| **Train RMSE** | 274.3714301901709 |
| **Test RMSE** | 278.52949548805515 |

Coefficients are:

Intercept 308.0040 (B0)

speed 9.3196 (B1)

hd 0.7986 (B2)

ram 47.8208 (B3)

screen 123.0266 (B4)

ads 0.6627 (B5)

trend -52.0261 (B6)

cd\_yes 60.1647 (B7)

multi\_yes 105.4382 (B8)

premium\_yes -509.7279 (B9)

1st of all, we load the data then, delete unwanted columns, convert categorical variable into dummy variables then perform all visualization including univariate, bivariate, check data is normally distributed or not, then build the model, Then apply Influence Index plots and removed high influence values so that our model will improve after that again build the model and we get the R-Squared Value which is given above and also get the equation of line, and also we do VIF for better model performance but it gives low value of R-square and adjusted R square.Then we predict the model after that we again do  some visualizations then split into train and test and find the test RMSE  and Train RMSE.

1. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Ans:** With the help ofmultilinear regression analytics we can easily find out what factors affect on price.

1. An online car sales platform would like to improve its customer base and their experience by providing them an easy way to buy and sell cars. For this, they would like an automated model which can predict the price of the car once the user inputs the required factors. Help the business achieve their objective by applying multilinear regression on the given dataset. Please use the below columns for the analysis purpose: price, age\_08\_04, KM, HP, cc, Doors, Gears, Quarterly\_Tax, and Weight.



1. **Business Problem**
   1. **What is the business objective?**

**Ans:** An online car sales platform would like to improve its customer base and their experience by providing them an easy way to buy and sell cars. For this, they would like an automated model which can predict the price of the car once the user inputs the required factors.

1. **Briefly explain the model output in the documentation.**

|  |  |
| --- | --- |
| **R-Squared** | 0.885 |
| **Train RMSE** | 1342.9698259162574 |
| **Test RMSE** | 1328.4383667274078 |

Coefficients are:

Intercept -1.474e+04 (B0)

Age\_08\_04 -112.0180 (B1)

KM -0.0170 (B2)

HP 36.6084 (B3)

cc -3.7954 (B4)

Doors -122.5437 (B5)

Gears 464.9839 (B6)

Quarterly\_Tax 5.2134 (B7)

Weight 30.6398 (B8)

1st of all, we load the data then, Select specific columns, then perform all visualization including univariate, bivariate, check data is normally distributed or not, data is not normal so we transform into normal, then build the model, Then apply Influence Index plots and removed high influence values so that our model will improve after that again build the model and we get the R-Squared Value which is given above and also get the equation of line. Then we predict the model after that we again do  some visualizations then split into train and test and find the test RMSE  and Train RMSE.

1. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Ans:** An online car sales platform would like to improve its customer base and their experience by providing them an easy way to buy and sell cars..With the help ofmultilinear regression analytics we can easily predict the price of the car once the user inputs the required factors.

1. With the growing consumption of avocados in the USA, a freelance company would like to do some analysis on the patterns of consumption in different cities and would like to come up with a prediction model for the price of avocados. For this to be implemented, build a prediction model using multilinear regression and provide your insights on it.

Snapshot of the dataset is given below: -

A close up of a piece of paper

Description automatically generated

1. **Business Problem**
   1. **What is the business objective?**

**Ans:** With the growing consumption of avocados in the USA, a freelance company would like to do some analysis on the patterns of consumption in different cities and would like to come up with a prediction model for the price of avocados.

1. **Briefly explain the model output in the documentation.**

|  |  |
| --- | --- |
| **R-Squared** | 0.502 |
| **Train RMSE** | 0.2849206746453047 |
| **Test RMSE** | 0.28373173966683357 |

Coefficients are:

Intercept -384.8289 (B0)

Total\_Volume -0.0183 (B1)

tot\_ava1 0.0052 (B2)

tot\_ava2 0.0154 (B3)

tot\_ava3 -0.0021 (B4)

Total\_Bags -0.0185 (B5)

Small\_Bags 0.0207 (B6)

Large\_Bags 0.0053 (B7)

XLarge\_Bags 0.0066 (B8)

type 0.4203 (B8)

year 28.8846 (B9)

region 3.0579 (B10)

1st of all, we load the data then, Find the relation between AveragePrice and region, Convert categorical variable into numeric, then perform all visualization including univariate, bivariate, check data is normally distributed or not, data is not normal so we transform into normal, then build the model, Then apply Influence Index plot, Applying transformations => CubeRoot(x) to the input variables and build the model and we get the R-Squared Value which is given above and also get the equation of line. Then we apply VIF, After applying VIF, we remove some varibales which has high vif value and again build the model but R-Square and Adjusted R-square decrease, Then we predict the model after that we again do  some visualizations then split into train and test and find the test RMSE  and Train RMSE.

1. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Ans:** With the growing consumption of avocados in the USA, With the help Multilinear Regression Analysis a freelance company can easily analyse the patterns of consumption in different cities and also predict the price of avocados with the help of prediction model.