MSc Data Science

Data Analytics and Visualisation - CI7330

Coursework

Submitted by

Hanumanth Cherukuri

K2046544

Contents

List of Figures

List of Tables

Section 1: A statistical summary of all the variables.

Stats	deprivation	spend	log spend	covid
Count	16000	16000	16000	16000
Min	4.012	4.23	1.443	-
1st Quantile	23.980	9.04	2.202	-
Median	31.817	10.74	2.374	-
Mean	31.301	11.7	2.411	-
3rd Quantile	38.514	13.18	2.579	-
Max	59.021	42.72	3.755	-
SD	10.04945	3.9914	0.30102	-
Unique Values	80	1926	16000	2
Data Type	num	num	num	factor

Table 1 statistical summary of variables

The supermarket sales dataset has 16,000 samples and 4 columns, with the names deprivation, covid, spend, and log spend. There are no null values in the dataset. The variables deprivation, spend, and log spend are numerical variables, and covid was of the integer data type, which further changed to the factor data type. The variable deprivation has a minimum value of 4, mean of 31 and maximum of 59. The mean spend is 11.7, and the minimum, maximum are 4.23 and 42.72, respectively. The covid has 2 labels from the dataset that 0 denotes the pre-covid and 1 for during lockdown. The data is balanced that pre- covid and during lockdown have 8000 samples each.

Section 2: Are they correlated?

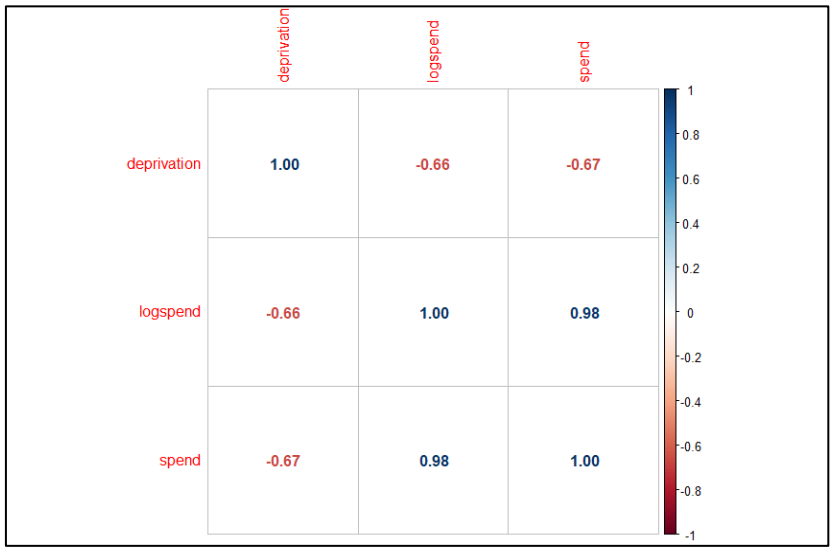

Figure 1 correlation plot

The correlation coefficient determines the linear relationship between two variables. The correlation matrix shows all the correlation coefficients between -1 to 1 for all pairs in the system. The Pearson correlation method is used to find the correlation between the variables. As the covid attribute is a logical data type, the correlation analysis only applies to the numeric variables. For in-depth analysis, the correlation test method can be applied to determine the t-test statistic value, degrees of freedom, significance level, confidence interval, and sample estimates. Here we stick to the correlation plot for better visualisation. From Figure 1 the variables deprivation and spend are negatively correlated with a score of -0.66. As we know, log spend is the log transformation of spend that is positively correlated with 0.98.

Section 3: Did the spend change between pre-COVID and lockdown?

```
        Welch Two Sample t-test

data:  spend by covid
t = 10.277, df = 15197, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 0.5231924 0.7697976
sample estimates:
mean in group 0 mean in group 1
       12.02426        11.37777
```

Figure 2 Two sample t-test

To determine the spend change between pre-covid and lockdown, a two sample t-test was performed.

Null Hypothesis: There is no significant difference of spend change.

Alternative Hypothesis: There is spend change between pre-covid and lockdown.

From the output, the t-statistic is 10.277, with the p- value of the test being 2.2e-16, which is less than significance level of alpha = 0.05. We can reject the null hypothesis that there is no significant change of spend during pre-covid and lockdown. The average spend during pre-covid is 12.02426, and the average spend during lockdown is 11.37777.

Section 4: Can we use covid status and deprivation to predict average spend?

```
Call:
lm(formula = spend ~ deprivation + covid, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3967 -2.0177 -0.3669  1.5347 24.1576

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.331007   0.079840  254.65   <2e-16 ***
deprivation -0.265384   0.002323 -114.26   <2e-16 ***
covid1      -0.646495   0.046679  -13.85   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.952 on 15997 degrees of freedom
Multiple R-squared:  0.453,     Adjusted R-squared:  0.4529
F-statistic:  6624 on 2 and 15997 DF,  p-value: < 2.2e-16
```

Figure 3 summary of multiple linear regression model

$y_1 = b_0+b_1x_1+b_2x_2$ (1)

spend = 20.331007 + (-0.265384*deprivation) + (-0.646495*covid) (2)

The average spend can be determined using a multiple linear regression model. From the above figure, we can see the distribution of residuals from -7.3967 to a maximum of 24.1576, and when residuals are plotted, they follow the normal distribution.

The probability t- value of the variables deprivation and covid is 2e-16 which is less than the significance level alpha = 0.05. So, we can reject the null hypothesis that the true contribution of variables deprivation and covid is equal to 0. The variables deprivation and covid makes significant contribution to predict the spend. The residual standard error is 2.952, and the adjusted r-squared is 0.4529, which means the model explains 45 percent of the variance. The average predicted spend is 11.70, equal to the average actual spend.

Section 5: Graphs for question 3, and question 4.

5.1 visualisation of spend during pre-covid and lockdown.

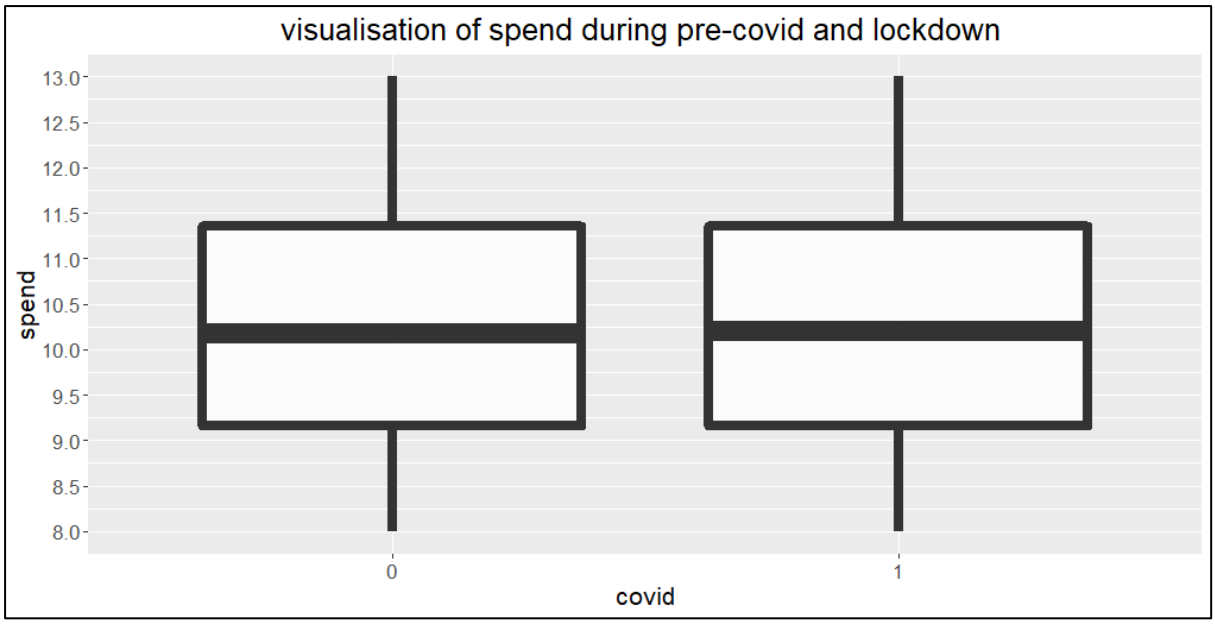

Figure 4 visualisation of spend

The Figure 4, Boxplot illustrates the spend during pre-covid, and lockdown. The first quartile is approximately 9.25, the third quartile is approximately 11.75, and the median is in the same range for both pre-covid and during covid spend of 10.25. We, can conclude that there is no major change of spend between two groups.

5.2 Predicting spend using deprivation and covid variables

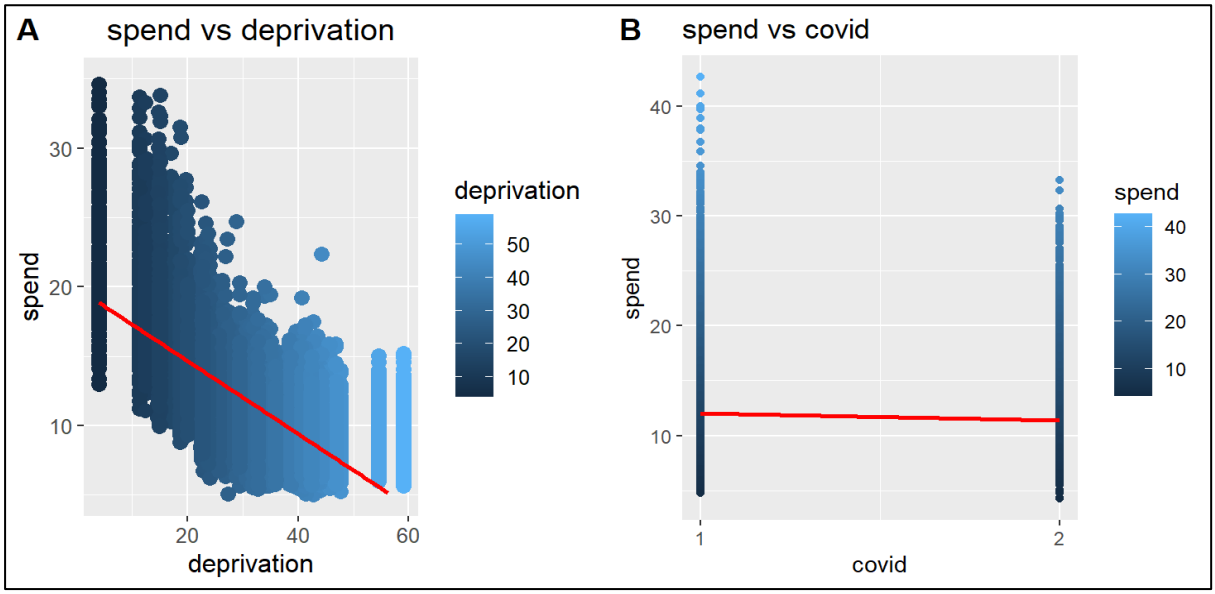

Figure 5 mlr-model regression line

The above Figure 5 illustrates the linear relationship between spend, deprivation, and covid. The independent variables deprivation and covid were negatively correlated with the dependent variable spend. As the deprivation increases spend decreases, and there is a slight decrease of spend during covid, denoted by 2 in the figure 5. The variables deprivation and covid showed significant variability in predicting spend.

Appendix

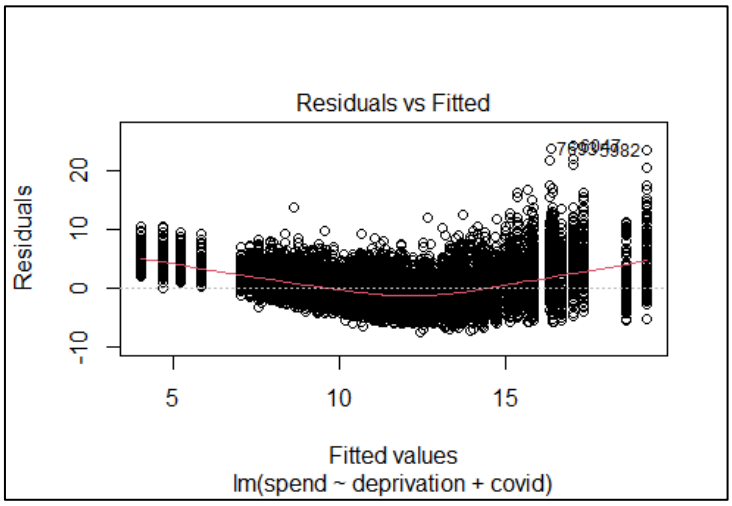

Figure 6 mlr-model residuals vs fitted

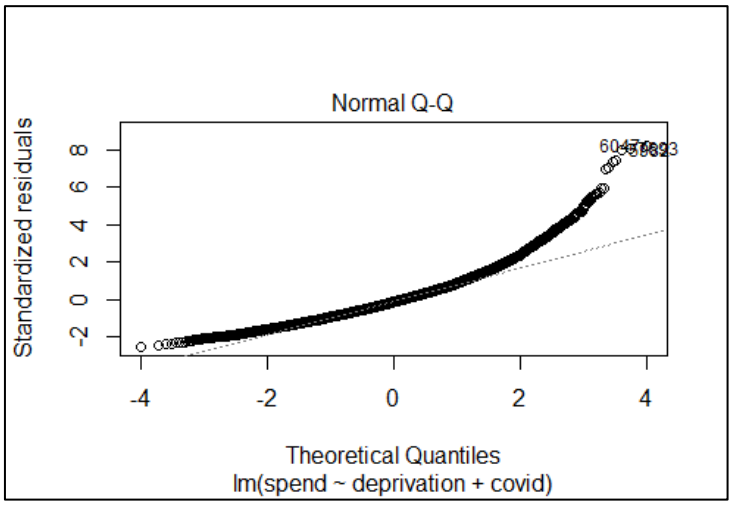

Figure 7 mlr-model normal Q-Q of mlr model

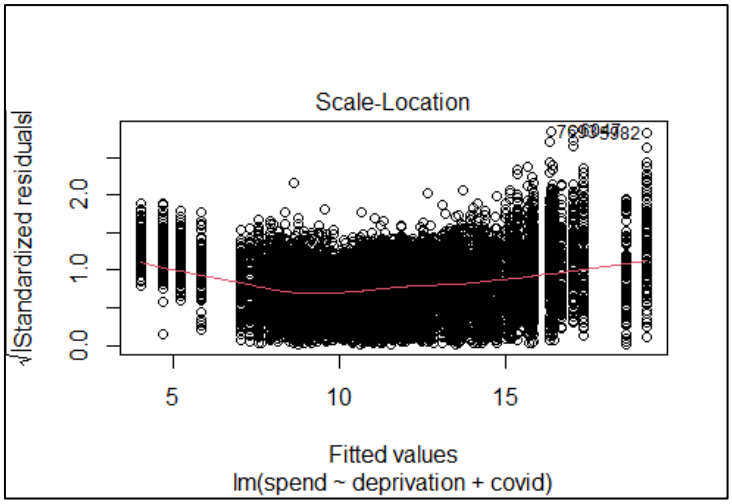

Figure 8 mlr-model scale location

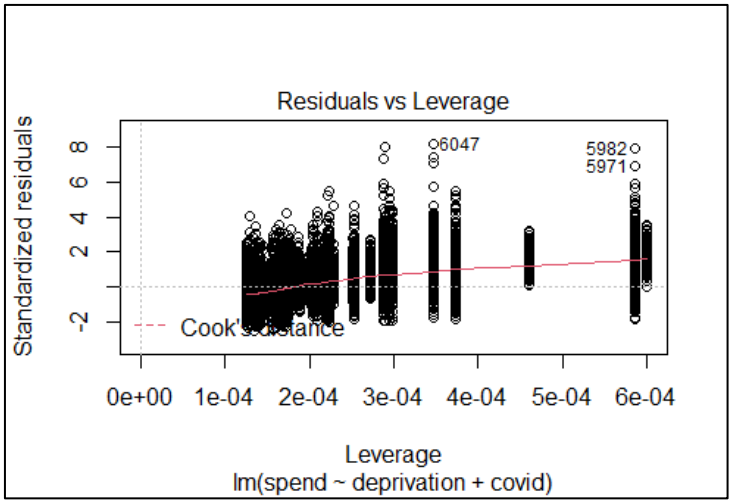

Figure 9 mlr-model residuals vs leverage

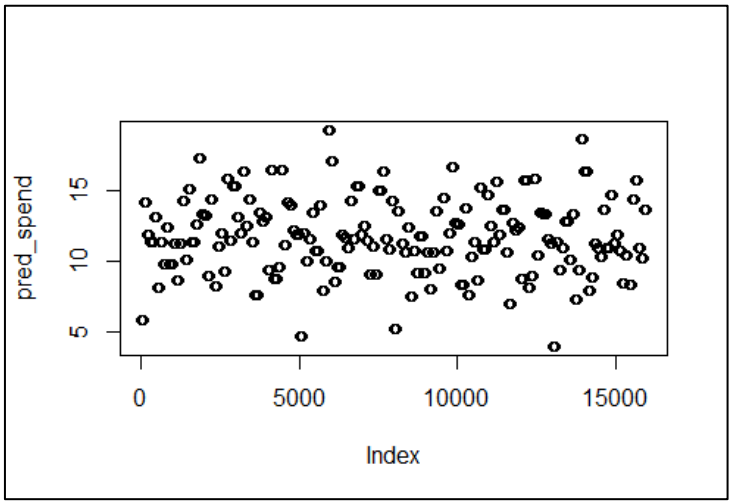

Figure 10 distribution of predicted spend of mlr-model

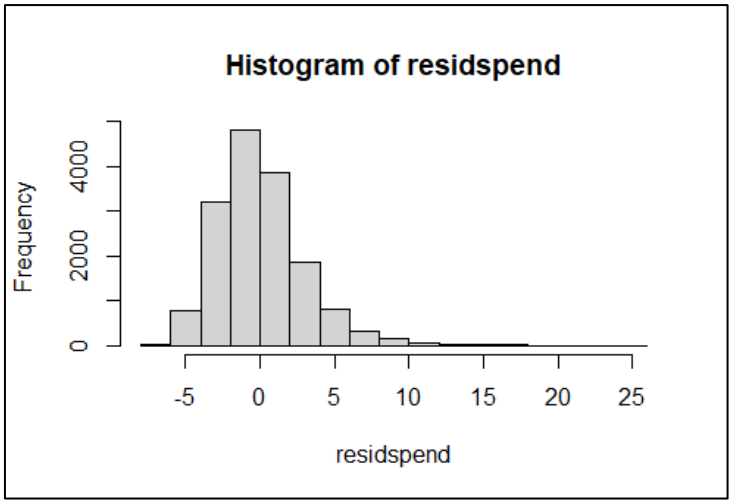

Figure 11 histogram residual spend