Regression_HW_3

Vyom Vats

October 19, 2016

Question 4

Reading the data and extracting only complete cases

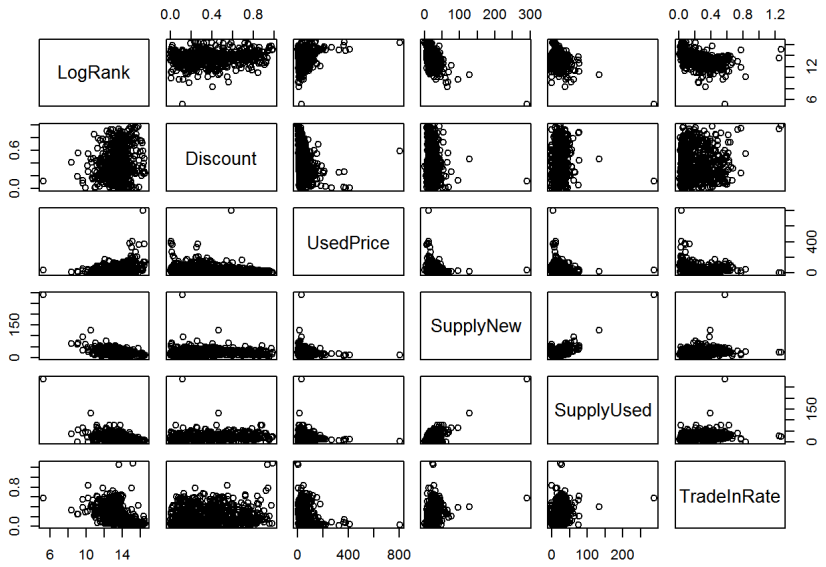
```
Input = read.csv(file="C:\\Users\\Vyom\\Documents\\Data_Amazon_NEW.csv",head=TRUE,sep=",")
NewData = Input[complete.cases(Input),]
```

Getting the variables of interest

```
NewData$PriceDiscount = 1-NewData$UsedPrice/NewData$LandedPrice

#keep only the data where the used price is lower than new price:
myData = NewData[NewData$PriceDiscount > 0,]

ModelData = data.frame(LogRank = log(myData$Rank),
                       Discount = myData$PriceDiscount,
                       UsedPrice = myData$UsedPrice,
                       SupplyNew = myData$NewCounterValue,
                       SupplyUsed = myData$UsedCounterValue,
                       TradeInRate = myData$TradeInAmount/myData$UsedPrice,
                       SubCondition = as.factor(myData$UsedSubcondition),
                       CAT=as.factor(myData$CAT))
```

$(a) Data Analysis$
$Plot the scatterplot of LogRank vs Discount, UsedPrice, SupplyNew, SupplyUsed, and TradeIn - Rate seperately. Describe the general trend($

```
pairs(ModelData[,1:6])
```



The scatter plots show that LogRank is not quite linearly related to any of the variables among Discount, UsedPrice, SupplyNew, SupplyUsed, and TradeInRate. There is no identifiable direction and shape of each of these scatter plots, except for LogRank v/s SupplyNew which does seem to have a downward slope, but even this is not very pronounced.

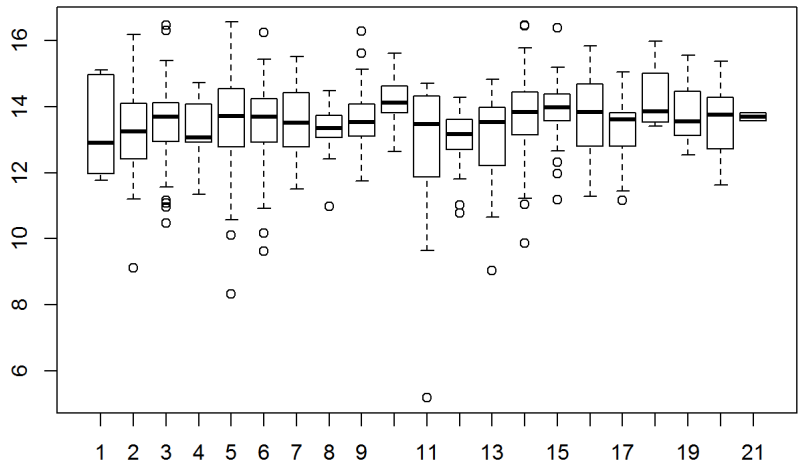
```
print(cor(ModelData[,1:6]))
```

```
##               LogRank    Discount  UsedPrice  SupplyNew  SupplyUsed
## LogRank      1.0000000  0.16444648  0.3060130 -0.5356637 -0.42822676
## Discount     0.1644465  1.00000000 -0.2250221 -0.1186229  0.08268886
## UsedPrice    0.3060130 -0.22502212  1.0000000 -0.1385574 -0.13121585
## SupplyNew   -0.5356637 -0.11862290 -0.1385574  1.0000000  0.76595333
## SupplyUsed  -0.4282268  0.08268886 -0.1312159  0.7659533  1.00000000
## TradeInRate -0.4751708  0.05384091 -0.1583217  0.2385717  0.22358708
##             TradeInRate
## LogRank     -0.47517079
## Discount     0.05384091
## UsedPrice   -0.15832165
## SupplyNew    0.23857167
## SupplyUsed   0.22358708
## TradeInRate  1.00000000
```

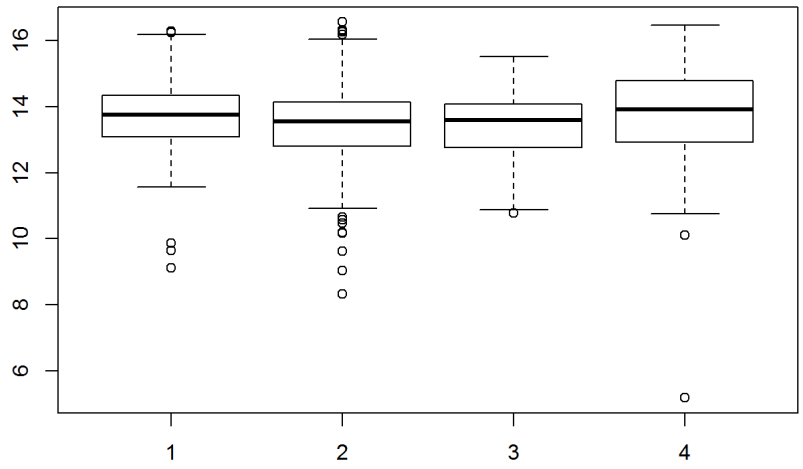
The correlation coefficients are given by the matrix above. Here also it can be observed that there is not a strong correlation between LogRank and any one of the variables among Discount, UsedPrice, SupplyNew, SupplyUsed, and TradeInRate. There is a (weak) positive correlation between LogRank and each of Discount and UsedPrice. There is a (weak) negative correlation between LogRank and each of SupplyNew, SupplyUsed, and TradeInRate.

Using the boxplot, show the relationship between LogRank and categorical variables Sub-Condition. Interpret the results.

```
boxplot(ModelData$LogRank ~ ModelData$CAT)
```




```
boxplot(ModelData$LogRank ~ ModelData$SubCondition)
```



From the boxplot of LogRank vs Cat, it can be seen that there is some variation between the spread of values of LogRank for different values of Cat. However, there is no definite pattern of this variation.

From the boxplot of LogRank vs SubCondition, it can be observed that there is very little difference in the spread of values of LogRank for different values of SubCondition. Thus SubCondition does not have much explanatory power over LogRank.

(b) Linear Regression Model Analysis Fit a Linear Regression of LogRank over input variables Discount, LogUsedPrice, LogTradeInRate

```
mdl=lm((formula = LogRank ~ Discount + UsedPrice + TradeInRate),
       data = ModelData)
summary(mdl)
```

```
##
## Call:
## lm(formula = (formula = LogRank ~ Discount + UsedPrice + TradeInRate),
##     data = ModelData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7816 -0.4396  0.1747  0.6029  4.4070
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.3841732  0.1143452 117.051  < 2e-16 ***
## Discount     1.3285239  0.1822820   7.288 1.05e-12 ***
## UsedPrice    0.0065042  0.0007826   8.311 7.00e-16 ***
## TradeInRate -3.0737610  0.2391428 -12.853  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.038 on 570 degrees of freedom
## Multiple R-squared:  0.3418, Adjusted R-squared:  0.3383
## F-statistic: 98.65 on 3 and 570 DF,  p-value: < 2.2e-16
```

What are the values of R-squared and adjusted R-squared? Interpret the model fit and discuss the reasons for these R-squared values.

Multiple R-squared: **0.3418**

Adjusted R-squared: **0.3383**

The p-value of the model is extremely small which gives evidence to reject the null hypothesis, i.e., we should conclude there is atleast one covariate which is statistically significant in predicting the dependent variable. But, the R-squared value is quite low which means that the covariates are unsuccessful in explaining the bulk of variability in the values of dependent variable.

Interpret the estimated parameters.

The value of **Intercept** is 13.38 which means that the estimated value of LogRank is 13.38 when all the covariates are set to zero. The estimate of **Discount** can be interpreted as there is an expected increase of 1.328 in the value of LogRank associated with unit increase in the value of Discount, when UsedPrice and TradeInRate are held constant. Similarly, we can say that there is an expected increase of 0.0065 in the value of LogRank associated with unit increase in the value of **UsedPrice**, when Discount and TradeInRate are held constant. Also, there is an expected decrease of 3.074 in the value of LogRank associated with unit increase in the value of **TradeInRate**, when UsedPrice and Discount are held constant.

Write down the equation for the regression line.

$LogRank = 13.384 + 1.328 * Discount + 0.0065 * UsedPrice - 3.074 * TradeInrate$

What is the p-value for the statistical significance of each parameter? At significance level of 0.05, what these p-values indicate?

For the Intercept, p-value < $2 * 10^{-16}$

For Discount, p-value = $1.05 * 10^{-12}$

For UsedPrice, p-value = $7 * 10^{-16}$

For the TradeInRate, p-value < $2 * 10^{-16}$

These p-values are all very small than 0.05, which gives enough evidence to reject each of the null hypotheses H_0 that $\beta_i = 0$. Thus all predictors are statistically significant in predicting the response variable.

Find a 95% confidence intervals for the parameters corresponding to three predictors.

The confidence intervals are as follows:

```
confint(mdl)
```

```
##                   2.5 %       97.5 %
## (Intercept) 13.15958395 13.608762491
## Discount     0.97049757  1.686550298
## UsedPrice    0.00496696  0.008041404
## TradeInRate -3.54346953 -2.604052386
```

Add categorical variables CAT and SubCondition to the model. How does the adjusted Rsquare change in comparison to the previous model? What's your interpretation of this result?

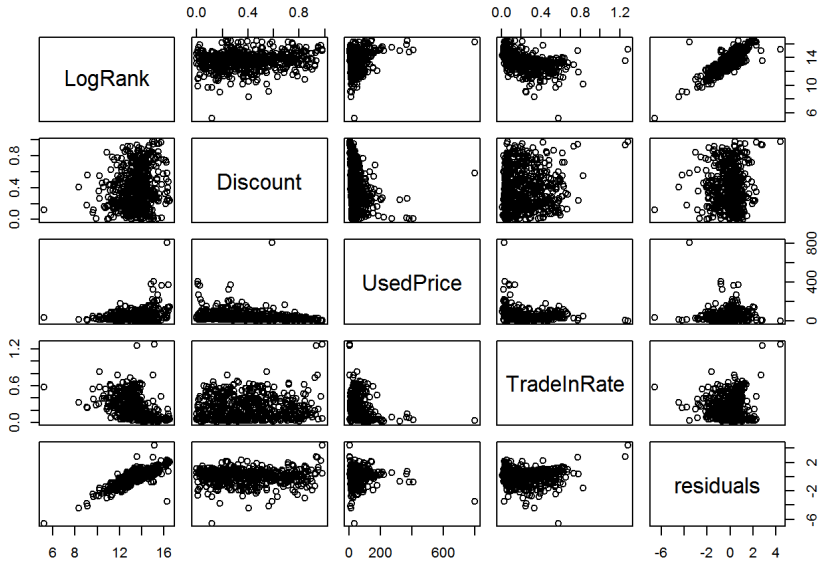
```
mdl=lm((formula = LogRank ~ Discount + UsedPrice + TradeInRate + CAT + SubCondition),
       data = ModelData)
summary(mdl)
```

```
##
## Call:
## lm(formula = (formula = LogRank ~ Discount + UsedPrice + TradeInRate +
##     CAT + SubCondition), data = ModelData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3676 -0.4111  0.1751  0.5972  4.1037
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.1487048  0.4489774  29.286  < 2e-16 ***
## Discount       1.3709365  0.1887058   7.265 1.29e-12 ***
## UsedPrice      0.0060016  0.0008238   7.286 1.12e-12 ***
## TradeInRate   -2.9952482  0.2443423 -12.258  < 2e-16 ***
## CAT2           0.0197501  0.4548817   0.043    0.965
## CAT3           0.3310070  0.4437106   0.746    0.456
## CAT4          -0.0479782  0.5142610  -0.093    0.926
## CAT5           0.2813309  0.4370549   0.644    0.520
## CAT6           0.2117482  0.4516189   0.469    0.639
## CAT7           0.5247747  0.4784101   1.097    0.273
## CAT8           0.1306072  0.5023103   0.260    0.795
## CAT9           0.4725836  0.5018878   0.942    0.347
## CAT10          0.6245685  0.5140623   1.215    0.225
## CAT11         -0.3858694  0.4831231  -0.799    0.425
## CAT12          0.0839533  0.4776872   0.176    0.861
## CAT13          0.2176615  0.5135775   0.424    0.672
## CAT14          0.3150556  0.4377989   0.720    0.472
## CAT15          0.4418481  0.4640446   0.952    0.341
## CAT16          0.5544005  0.5792880   0.957    0.339
## CAT17          0.2752904  0.4584029   0.601    0.548
## CAT18          0.6578947  0.6715090   0.980    0.328
## CAT19          0.6196345  0.5601045   1.106    0.269
## CAT20          0.2600012  0.5137827   0.506    0.613
## CAT21          0.5585473  0.8539830   0.654    0.513
## SubCondition2 -0.1276074  0.1248053  -1.022    0.307
## SubCondition3 -0.0347333  0.1622692  -0.214    0.831
## SubCondition4  0.1733981  0.1525065   1.137    0.256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.037 on 547 degrees of freedom
## Multiple R-squared:   0.37,  Adjusted R-squared:  0.3401
## F-statistic: 12.36 on 26 and 547 DF,  p-value: < 2.2e-16
```

As can be seen from the summary, the p-values of categorical variables do not indicate that they are statistically significant in expliaining LogRank. The p-values of the other variables are still very low which indicates that as before, they do a good job of predicting LogRank. However, the R-squared value is still low suggesting that that this model does not explain the majority of variability in the values of LogRank.

(c) Checking the Assumptions of the Model. Plot the relevant residual plots for the first model in part (b) and check the model assumptions. Comment on whether there are any apparent departures from the assumptions of linear regression model. If there is any, would you consider any transformation? Also please explain if there are any extreme outliers in the data/residuals? Enumerate what graphical techniques you used.

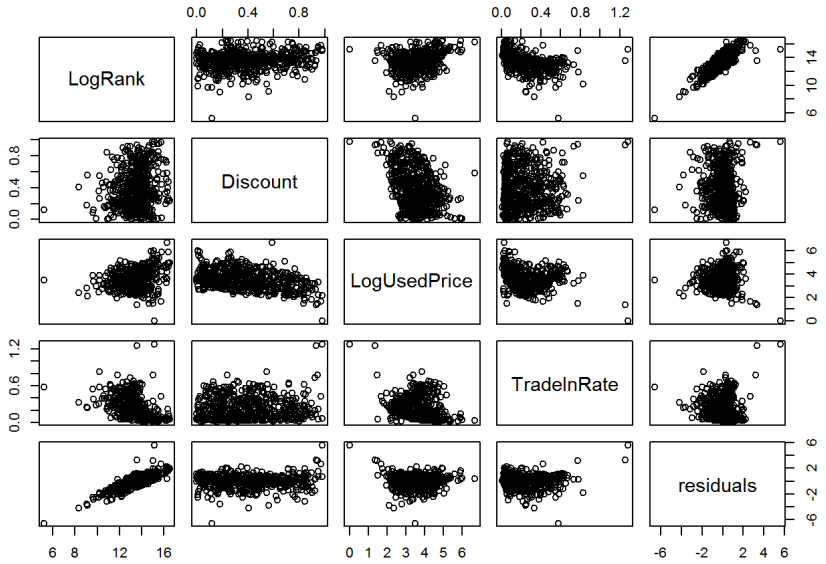
```
# preparing the model again
mdl=lm((formula = LogRank ~ Discount + UsedPrice + TradeInRate),
       data = ModelData)

#appending the residuals to the dataset to aid plotting
ModelData$residuals = rstandard(mdl)

#plotting
pairs(ModelData[,c(1,2,3,6,9)])
```



As can be seen from the scatter plots, there does seem to be non-constant variance among Residuals when plotted against UsedPrice. This is in violation with one of the assumptions of limear regression.

To correct this, we could use log transformation of UsedPrice as shown below:

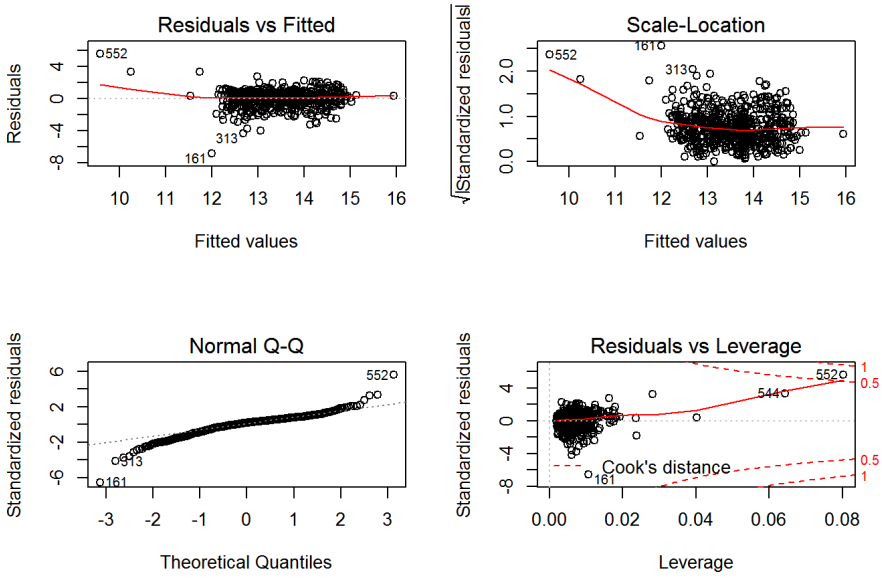
```
# preparing the model again; this time log transforming UsedPrice
mdl=lm((formula = LogRank ~ Discount + log(UsedPrice) + TradeInRate),
       data = ModelData)

#appending the residuals to the dataset to aid plotting
ModelData$residuals = rstandard(mdl)

#adding log of UsedPrice to the dataset
ModelData$LogUsedPrice = log(ModelData$UsedPrice)

#plotting
pairs(ModelData[,c(1,2,10,6,9)])
```



Now it can be observed that the variance of residuals is also constant when plotted against log of UsedPrice.

Checking if the normality assumption holds:

```
layout(matrix(c(1,2,3,4),2,2))
plot(mdl)
```



The plot above shows the QQ plot of standardized residuals with normal distribution. We can infer that the normality assumption holds true.

Checking for multicollinearity:

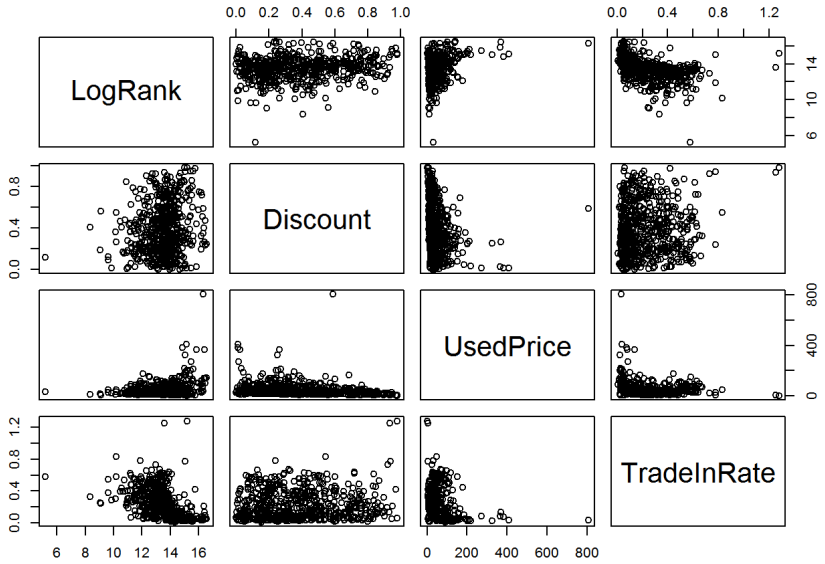
```
cor(ModelData[,c(2,3,5)])
```

```
##               Discount  UsedPrice  SupplyUsed
## Discount    1.00000000 -0.2250221  0.08268886
## UsedPrice  -0.22502212  1.0000000 -0.13121585
## SupplyUsed  0.08268886 -0.1312159  1.00000000
```

From the valuesabove we can infer that the predictors are not highly corelated.

Looking for outliers:

```
# plotting the variables in the model
pairs(ModelData[,c(1,2,3,6)])
```



It can be seen from the plot of LogRank against TradeInRate that there are 3 outliers in our dataset. The graphical techniques used to identify outliers was simple scatter plotting.