MULTIPLE LINEAR REGRESSION ANALYSIS

Black Friday Sales

By
Clarine Anslum-3687114
Shamini Puthooppallil Baby -3674381

9-June-2019

Table of Contents

ABSTRACT

The aim of this project was to examine the product purchase on Black Friday sales influenced by factors such as Gender, Age, Occupation, City of residence, Product, Category and etc.

To facilitate this, the data was sourced from https://www.kaggle.com/mehdidag/black-friday and started analysing the factors by fitting multiple linear regression model.

The regressors that were strongly related to purchase amount were manipulated and the evidence for the same have been proved.

INTRODUCTION

Consumers simply being tired of the too-good-to-be-true deals means that you may not get enough of sales on Black Friday to gain a positive return on investment. With November being the new December and Black Friday deals starting earlier and earlier each year, consumers are simply fatigued by deals.

With the "hype" of Black Friday, customers are exposed to a retail environment that can stimulate frustration and aggression. Black Friday is traditionally known for long lines with customers waiting outdoors in cold weather waiting for the store to open, confusion and chaos of customers once the retail doors are opened for business, heavily crowded stores, a limited amount of products available at a reduced price, long checkout lines, and the lack of availability of advertised sale products.

Here, the store wants to know better the customer purchase behavior against different products.

METHODOLOGY

The regression problem is trying to predict the response variable (the amount of purchase)is Purchase amount in dollars and the predictor variables will be verified using statistical tools and the programming language R. Also, we were interested in determining the multicollinearity among the independent variables and how they affect the dependent variable during the regression analysis.

Transaction Data

The data was sourced from the open source platform Kaggle (https://www.kaggle.com/) with all rights reserved for public access and the link for the data is https://www.kaggle.com/mehdidag/black-friday.

The dataset here is a sample of the transactions made in a retail store with 550 000 observations about the black Friday in a retail store, it contains different kinds of variables either numerical or categorical. This dataset contains missing values and need to be cleansed before the analysis.

The variables are : User_ID, Product_ID, Gender, Age, Occupation ID, City_Category, Stay_In_Current_City_Years, Marital_Status, Product_Category_1, Product_Category_2, Product_Category_3, Purchase, Purchase amount in dollars

Hypothesis

Ho : Purchase amount on Black Friday sale is not influenced by the customers' characteristics
H1 : Purchase amount on Black Friday is strongly influenced by customers' characteristics

Data preparation

Before starting the data analysis and modelling, the main duty of the data scientist is to make sure that the data provided is in correct format. If the dataset is not in proper format, the entire work needs to be repeated. The required libraries for the time series analysis is in Appendix 1.

Dataset is loaded to the R software using read.csv function and performed required pre-validations for the loaded dataset. Sample data is viewed and make sure that the data loaded correctly to the R software.

	User_ID <int>	Product_ID <fctr>	Gender <fctr>	Age <fctr>	Occupation <int>	City_Category <fctr>	Stay_In_Current_City_Years <fctr>
1	1000001	P00069042	F	0-17	10	A	2
2	1000001	P00248942	F	0-17	10	A	2
3	1000001	P00087842	F	0-17	10	A	2
4	1000001	P00085442	F	0-17	10	A	2
5	1000002	P00285442	M	55+	16	C	4+
6	1000003	P00193542	M	26-35	15	A	3

6 rows | 1-8 of 12 columns

See the code snippet in Appendix 2

The Regression analysis is performed using the R Markdown and the common statistical tools in the upcoming sessions. Structure of the dataset, column values, null values and impossible values are checked using some basic R data preprocessing packages. We have replaced the null values and the impossible values with appropriate values.

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>
User_ID	1	537577	1002991.85	1714.39	1003031	1002983.60	2145.32
Product_ID*	2	537577	1693.33	1002.58	1647	1673.93	1187.56
Gender*	3	537577	1.75	0.43	2	1.82	0.00
Age*	4	537577	3.49	1.35	3	3.35	1.48
Occupation	5	537577	8.08	6.52	7	7.69	8.90
City_Category*	6	537577	2.04	0.76	2	2.05	1.48
Stay_In_Current_City_Years*	7	537577	2.86	1.29	3	2.82	1.48
Marital_Status	8	537577	0.41	0.49	0	0.39	0.00
Product_Category_1	9	537577	5.30	3.75	5	4.85	4.45
Product_Category_2	10	537577	6.78	6.21	5	6.44	7.41

1-10 of 12 rows | 1-8 of 13 columns Previous 1 2 Next

This code snippet is available in Appendix 3

Train and Test data split

The data set was split into two (70:30) train and test data. The train set was selected for feature selection and model fitting. Once the model selection from the available candidate models was over by adequacy and accuracy test, the model was tested and validated with the help of test data set.

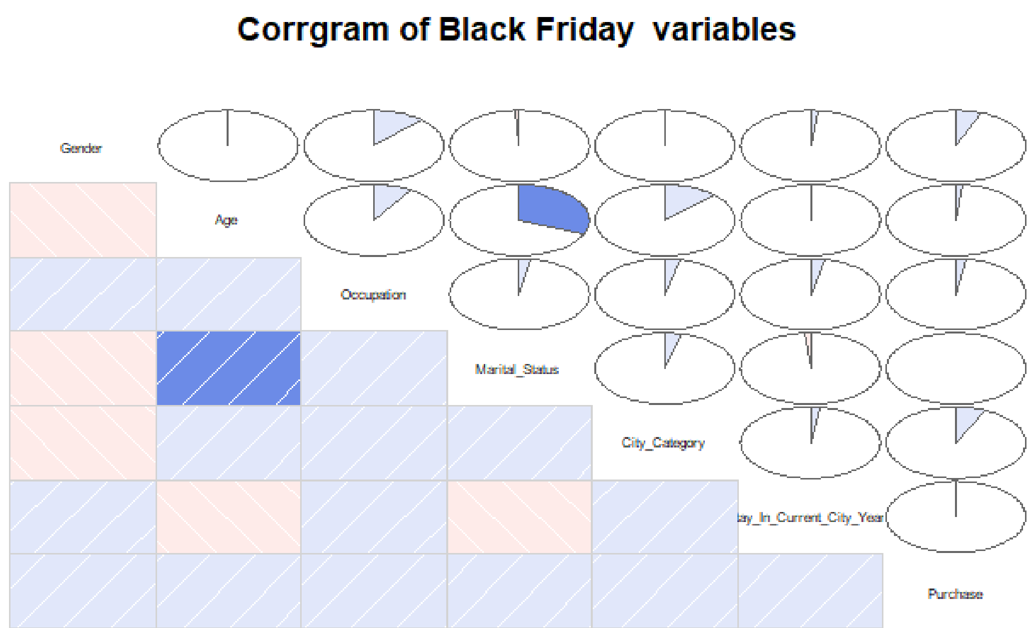
```
[1] "Number of rows in train data"
[1] 376303
[1] "Number of rows in test data"
[1] 161274
[1] "Number of rows in Raw datset "
[1] 537577
```

See Appendix 4 for the R code

Correlation of regressors and model training

	User_ID	Occupation	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
User_ID	1.00	-0.02	0.02	0.00	0.00	0.00	0.01
Occupation	-0.02	1.00	0.02	-0.01	0.01	0.01	0.02
Marital_Status	0.02	0.02	1.00	0.02	0.00	0.00	0.00
Product_Category_1	0.00	-0.01	0.02	1.00	-0.04	-0.39	-0.31
Product_Category_2	0.00	0.01	0.00	-0.04	1.00	0.09	0.04
Product_Category_3	0.00	0.01	0.00	-0.39	0.09	1.00	0.28
Purchase	0.01	0.02	0.00	-0.31	0.04	0.28	1.00

According to the above table, the correlation between the numeric variables were found to be weak with less than 0.4. The correlation was visualized as per the plot below:



See Appendix5 for R code chunks

The multiple regression model was the appropriate one for our hypothesis question. In order to achieve the best model, the feature selection was done in two ways.

Model 1

The variables were converted into numeric values and trained the data sets

The model equation:

```
[1] Purchase = 7025 + 677 * Gender + 36 * Age + 9 * Occupation + -51 * Marital_Status + 442 *
City_Category + 17 * Stay_In_Current_City_Years
```

The summary statistics

All of the regressors are significance as the p-values of t-statistics is significance at 5%. Nevertheless, the r-squared value is very low with 0.008.

The summary is as follow:

```
Call:
lm(formula = Purchase ~ Gender + Age + Occupation + Marital_Status +
    City_Category + Stay_In_Current_City_Years, data = train_BF_NUMERIC)

Residuals:
   Min     1Q Median     3Q    Max
 -9937  -3527  -1164   2902  15658

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                7024.673     47.711 147.233  < 2e-16 ***
Gender                      676.742     18.915  35.778  < 2e-16 ***
Age                          35.816      6.361   5.631 1.80e-08 ***
Occupation                    8.952      1.254   7.137 9.55e-13 ***
Marital_Status              -51.277     17.327  -2.959  0.00308 **
City_Category               442.243     10.731  41.212  < 2e-16 ***
Stay_In_Current_City_Years   16.523      6.283   2.630  0.00854 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4963 on 376296 degrees of freedom
Multiple R-squared:  0.008567,	Adjusted R-squared:  0.008551
F-statistic: 541.9 on 6 and 376296 DF,  p-value: < 2.2e-16
```

Multicollinearity check

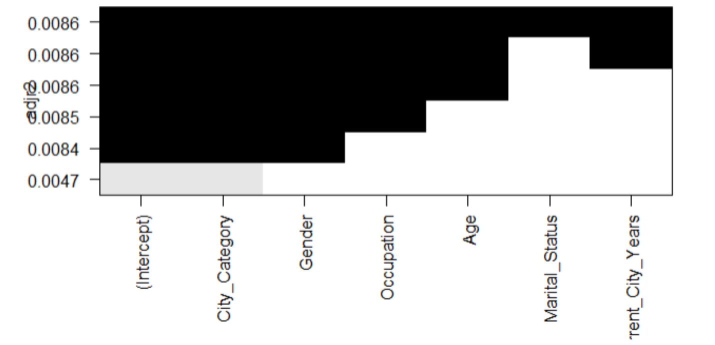
The multicollinearity is not affected in this model as all VIF values are very low (nearly 1). This is shown in the below plot.

Variables <chr>	Tolerance <dbl>	VIF <dbl>
Gender	0.9856327	1.014577
Age	0.8832722	1.132154
Occupation	0.9761341	1.024449
Marital_Status	0.9017933	1.108902
City_Category	0.9840959	1.016161
Stay_In_Current_City_Years	0.9982352	1.001768

Feature selection

All the validation matric values as per below suggested the model with all 6 variables.

Adj.R2 <int>	rsq <int>	CP <int>	BIC <int>
6	6	6	4



Residual checks

1. Homogeneity of residuals variance assumption failed as the p-value is not significant.

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 746.1348, Df = 1, p = < 2.22e-16

        studentized Breusch-Pagan test

data:  model_BF
BP = 1007, df = 6, p-value < 2.2e-16
```

2. Assumption of Normality of residual fails as the p-value is not significant.

```
        Shapiro-Wilk normality test

data:  model_BF_res$residuals
W = 0.9586, p-value < 2.2e-16
```

3. Assumption of uncorrelation is significant as the p-value is significant

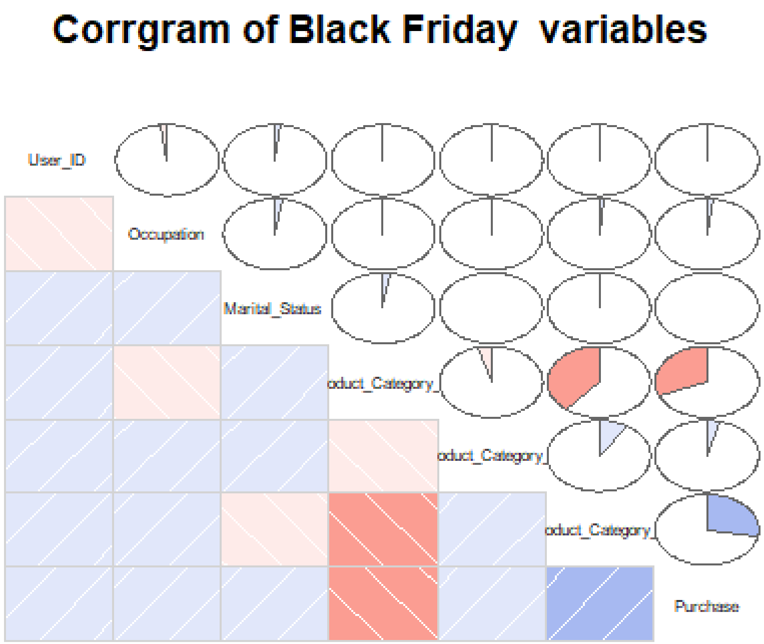
```
lag Autocorrelation D-W Statistic p-value
  1     0.009957675      1.979506   0.594
Alternative hypothesis: rho != 0
```

As per the statistical tests performed on residuals, we concluded that this model is not adequate for the Black Friday dataset.

See the appendix 6 for R chunk of model fitting and adequacy of the above model.

Model 2

The variables City and marital status were converted into factor values and trained the data sets. Chorogram of the features is as follows:



The model equation:

```
[1] Purchase = 9949 + 12 * Occupation + 164 * City_CategoryB   + 695 * City_CategoryC   + 15 * Stay_In_Current_City_Years
-318 * Product_Category_1 + 38 * Marital_Status1  + 9 * Product_Category_2 + 149 * Product_Category_3
```

The summary statistics

All of the coefficients of regressors are significance as the p-values of t-statistics is significance at 5%. Nevertheless, the r-squared value is low with 0.13.

The summary is as follow:

```
Call:
lm(formula = Purchase ~ Occupation + City_Category + Stay_In_Current_City_Years +
    Product_Category_1 + Marital_Status + Product_Category_2 +
    Product_Category_3, data = train_BF)

Residuals:
     Min       1Q   Median       3Q      Max
-11457.0  -3233.2   -576.4   2255.6  17090.3

Coefficients:
                           Estimate Std. Error  t value Pr(>|t|)
(Intercept)                9948.966     29.331  339.201  < 2e-16 ***
Occupation                   11.787      1.160   10.160  < 2e-16 ***
City_CategoryB              163.745     18.680    8.766  < 2e-16 ***
City_CategoryC              695.129     19.973   34.804  < 2e-16 ***
Stay_In_Current_City_Years   15.091      5.873    2.570   0.0102 *
Product_Category_1         -318.203      2.189 -145.370  < 2e-16 ***
Marital_Status1              37.968     15.405    2.465   0.0137 *
Product_Category_2            8.550      1.222    6.994 2.68e-12 ***
Product_Category_3          149.089      1.315  113.363  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4639 on 376294 degrees of freedom
Multiple R-squared:  0.1336,	Adjusted R-squared:  0.1336
F-statistic:  7252 on 8 and 376294 DF,  p-value: < 2.2e-16
```

The ANOVA table of the model from method 2 is as follows:

```
Analysis of Variance Table

Response: Purchase
                               Df     Sum Sq    Mean Sq   F value    Pr(>F)
Occupation                      1 4.4221e+09 4.4221e+09   205.463 < 2.2e-16 ***
City_Category                   2 4.6590e+10 2.3295e+10  1082.353 < 2.2e-16 ***
Stay_In_Current_City_Years      1 2.5272e+08 2.5272e+08    11.742 0.0006112 ***
Product_Category_1              1 9.1514e+11 9.1514e+11 42519.641 < 2.2e-16 ***
Marital_Status                  1 1.6798e+08 1.6798e+08     7.805 0.0052105 **
Product_Category_2              1 5.5006e+09 5.5006e+09   255.571 < 2.2e-16 ***
Product_Category_3              1 2.7659e+11 2.7659e+11 12851.134 < 2.2e-16 ***
Residuals                  376294 8.0989e+12 2.1523e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In our model, p-values of F-statistic are less than 0.05 for all feature variables and this is highly significant. Our model is an adequate one for the data set Black Friday sales.

The results from stepwise regression

```
Start:  AIC=6407684
Purchase ~ 1

                             Df  Sum of Sq        RSS     AIC
+ Product_Category_1          1 9.2674e+11 8.4208e+12 6368397
+ Product_Category_3          1 7.5524e+11 8.5923e+12 6375984
+ City_Category               2 4.7553e+10 9.3000e+12 6405769
+ Product_Category_2          1 1.3807e+10 9.3337e+12 6407130
+ Occupation                  1 4.4221e+09 9.3431e+12 6407508
+ Stay_In_Current_City_Years  1 4.4767e+08 9.3471e+12 6407668
<none>                                     9.3476e+12 6407684
+ Marital_Status              1 1.2563e+06 9.3475e+12 6407686

Step:  AIC=6368397
Purchase ~ Product_Category_1

                             Df  Sum of Sq        RSS     AIC
+ Product_Category_3          1 2.8717e+11 8.1336e+12 6355342
+ City_Category               2 3.6819e+10 8.3840e+12 6366752
+ Product_Category_2          1 6.1249e+09 8.4147e+12 6368125
+ Occupation                  1 3.4645e+09 8.4173e+12 6368244
+ Marital_Status              1 4.6637e+08 8.4203e+12 6368378
+ Stay_In_Current_City_Years  1 2.5490e+08 8.4206e+12 6368388
<none>                                     8.4208e+12 6368397
- Product_Category_1          1 9.2674e+11 9.3476e+12 6407684

Step:  AIC=6355342
Purchase ~ Product_Category_1 + Product_Category_3

                             Df  Sum of Sq        RSS     AIC
+ City_Category               2 3.1136e+10 8.1025e+12 6353903
+ Occupation                  1 2.9332e+09 8.1307e+12 6355209
+ Product_Category_2          1 1.2815e+09 8.1324e+12 6355285
+ Marital_Status              1 3.7244e+08 8.1333e+12 6355327
+ Stay_In_Current_City_Years  1 2.5168e+08 8.1334e+12 6355333
<none>                                     8.1336e+12 6355342
- Product_Category_3          1 2.8717e+11 8.4208e+12 6368397
- Product_Category_1          1 4.5867e+11 8.5923e+12 6375984

Step:  AIC=6353903
Purchase ~ Product_Category_1 + Product_Category_3 + City_Category
```

```
Step:  AIC=6353903
Purchase ~ Product_Category_1 + Product_Category_3 + City_Category

                             Df  Sum of Sq        RSS     AIC
+ Occupation                  1 2.3037e+09 8.1002e+12 6353798
+ Product_Category_2          1 1.0694e+09 8.1014e+12 6353855
+ Stay_In_Current_City_Years  1 1.7594e+08 8.1023e+12 6353897
+ Marital_Status              1 1.5370e+08 8.1024e+12 6353898
<none>                                     8.1025e+12 6353903
- City_Category               2 3.1136e+10 8.1336e+12 6355342
- Product_Category_3          1 2.8148e+11 8.3840e+12 6366752
- Product_Category_1          1 4.5531e+11 8.5578e+12 6374474

Step:  AIC=6353798
Purchase ~ Product_Category_1 + Product_Category_3 + City_Category +
    Occupation

                             Df  Sum of Sq        RSS     AIC
+ Product_Category_2          1 1.0526e+09 8.0992e+12 6353751
+ Stay_In_Current_City_Years  1 1.3844e+08 8.1001e+12 6353794
+ Marital_Status              1 1.2629e+08 8.1001e+12 6353794
<none>                                     8.1002e+12 6353798
- Occupation                  1 2.3037e+09 8.1025e+12 6353903
- City_Category               2 3.0507e+10 8.1307e+12 6355209
- Product_Category_3          1 2.8107e+11 8.3813e+12 6366632
- Product_Category_1          1 4.5508e+11 8.5553e+12 6374365

Step:  AIC=6353751
Purchase ~ Product_Category_1 + Product_Category_3 + City_Category +
    Occupation + Product_Category_2

                             Df  Sum of Sq        RSS     AIC
+ Stay_In_Current_City_Years  1 1.3818e+08 8.0990e+12 6353747
+ Marital_Status              1 1.2681e+08 8.0990e+12 6353747
<none>                                     8.0992e+12 6353751
- Product_Category_2          1 1.0526e+09 8.1002e+12 6353798
- Occupation                  1 2.2869e+09 8.1014e+12 6353855
- City_Category               2 3.0301e+10 8.1295e+12 6355152
- Product_Category_3          1 2.7662e+11 8.3758e+12 6366387
- Product_Category_1          1 4.5479e+11 8.5539e+12 6374308
```

Continued…….

```
Step:  AIC=6353747
Purchase ~ Product_Category_1 + Product_Category_3 + City_Category +
    Occupation + Product_Category_2 + Stay_In_Current_City_Years

                             Df  Sum of Sq        RSS     AIC
+ Marital_Status              1 1.3074e+08 8.0989e+12 6353743
<none>                                     8.0990e+12 6353747
- Stay_In_Current_City_Years  1 1.3818e+08 8.0992e+12 6353751
- Product_Category_2          1 1.0523e+09 8.1001e+12 6353794
- Occupation                  1 2.2495e+09 8.1013e+12 6353849
- City_Category               2 3.0239e+10 8.1293e+12 6355145
- Product_Category_3          1 2.7663e+11 8.3756e+12 6366383
- Product_Category_1          1 4.5471e+11 8.5537e+12 6374300

Step:  AIC=6353743
Purchase ~ Product_Category_1 + Product_Category_3 + City_Category +
    Occupation + Product_Category_2 + Stay_In_Current_City_Years +
    Marital_Status

                             Df  Sum of Sq        RSS     AIC
<none>                                     8.0989e+12 6353743
- Marital_Status              1 1.3074e+08 8.0990e+12 6353747
- Stay_In_Current_City_Years  1 1.4210e+08 8.0990e+12 6353747
- Product_Category_2          1 1.0528e+09 8.0999e+12 6353790
- Occupation                  1 2.2215e+09 8.1011e+12 6353844
- City_Category               2 3.0042e+10 8.1289e+12 6355132
- Product_Category_3          1 2.7659e+11 8.3755e+12 6366378
- Product_Category_1          1 4.5483e+11 8.5537e+12 6374301

Call:
lm(formula = Purchase ~ Product_Category_1 + Product_Category_3 +
    City_Category + Occupation + Product_Category_2 + Stay_In_Current_City_Years +
    Marital_Status, data = train_BF)

Coefficients:
               (Intercept)          Product_Category_1          Product_Category_3               City_CategoryB
                   9948.97                     -318.20                      149.09                       163.74
            City_CategoryC                  Occupation          Product_Category_2   Stay_In_Current_City_Years
                    695.13                       11.79                        8.55                        15.09
           Marital_Status1
                     37.97
```

Multicollinearity check

The multicollinearity is not affected in this model as all VIF values are very low (nearly 1). This is shown in the below plot.

Variables <chr>	Tolerance <dbl>	VIF <dbl>
Occupation	0.9970087	1.003000
City_CategoryB	0.6725253	1.486933
City_CategoryC	0.6700825	1.492353
Stay_In_Current_City_Years	0.9982561	1.001747
Product_Category_1	0.8470607	1.180553
Marital_Status1	0.9970463	1.002962
Product_Category_2	0.9916949	1.008375
Product_Category_3	0.8416879	1.188089

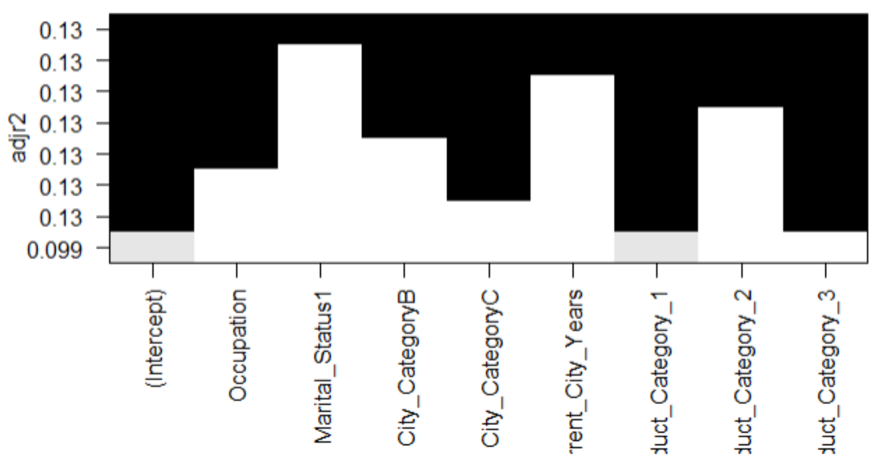
8 rows

Feature selection

All the validation matric values as per below suggested the model with all 6 variables.

Adj.R2 <int>	rsq <int>	CP <int>	BIC <int>
8	8	8	6

This has been proved by sub-set selection method as well.



Residual checks

1. Homogeneity of residuals variance assumption is plausible as the p-value is significant.

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.06392559, Df = 1, p = 0.8004

         studentized Breusch-Pagan test

data:  ModelBFFit
BP = 8.2843, df = 12, p-value = 0.7625
```

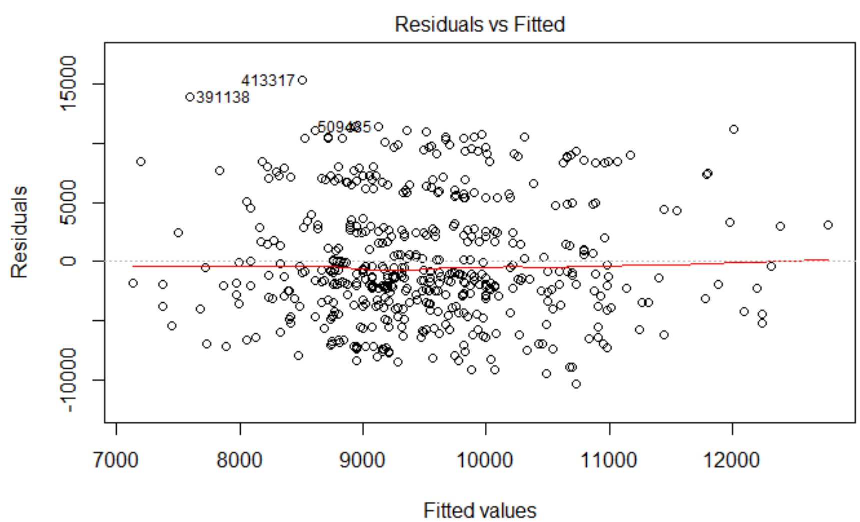
2. Assumption of Normality of residual fails as the p-value is not significant.

```
             Shapiro-Wilk normality test

data:  ModelBFFit$residuals
W = 0.95633, p-value = 5.25e-11
```

3. Assumption of uncorrelation is significant as the p-value is significant

```
 lag Autocorrelation D-W Statistic p-value
   1      0.01524002      1.969417   0.748
 Alternative hypothesis: rho != 0
```

4. Assumption of independence and identical is significant as the sample points plot is spread over the mean line and there is not pattern in the points.



As per the statistical tests performed on residuals, we concluded that this model is adequate for the Black Friday dataset.

See the Appendix7 for R chunk of model fitting and adequacy of the above model.

RESULTS

The model2 provided the RMSE and R-squared as follows on the test data:

```
[1] "Model R2 (Test Data)"
[1] 0.0081936
[1] "Model RMSE (Test Data)"
[1] 4953.563
```

The summary statistic of fitted model proved all the regressors are significant enough to predict the response variable purchase amount.

All the residual assumption analysis suggested that the assumptions are significant except for normality. This implies the suggested model is adequate to the black Friday data set.

DISCUSSION

As per the results, there are evidence to reject the null hypothesis that the response variable is not influenced by the regressors. The test hypothesis is statistically significant to support the decision that purchase price of Black Friday sale is influenced by the factors Occupation, City B, Years spent is the city, Product_Category_1, Marital_Status, Product_Category_2 and Product_Category_3.

CONCLUSION

We conclude that the purchase price is influences by the factors of customers and this has been proved by multiple linear regression.

Despite the model was adequate for the data set, the r-squared and RMSE values indicated that the scope of this project could be extended to improve these values by selecting more adequate models.

REFERENCE

1. https://www.kaggle.com/mehdidag/black-friday
2. https://www.electric-design.co.uk/the-pros-and-cons-of-black-friday
3. https://www.rdocumentation.org
4. https://thekeep.eiu.edu/cgi/viewcontent.cgi?article=1012&context=fcs_fac

Appendix

1. Required libraries

```r
#loading the libraries
library(readr)
library(tidyverse)
library(ggcorrplot)
library(car)
library(olsrr)
library(lmtest)
library(FitAR)
library(corrgram)
library(psych)
library(dplyr)
```

2. Load the data

```r
#Load the datset into R environment
BFData <- read.csv("C:/WorkingFolder/2ndyear/Regression Analysis/Project/BlackFriday.csv")
```

3. Pre-validations

```r
#check the structure of the data
str(BFData)

## 'data.frame':    537577 obs. of  12 variables:
##  $ User_ID                   : int  1000001 1000001 1000001 1000001 1000002 1000003 1000004 1000004 1000004 1000005 ...
##  $ Product_ID                : Factor w/ 3623 levels "P00000142","P00000242",..: 671 2375 851 827 2733 1830 1744 3319 3597 2630 ...
##  $ Gender                    : Factor w/ 2 levels "F","M": 1 1 1 1 2 2 2 2 2 2 ...
##  $ Age                       : Factor w/ 7 levels "0-17","18-25",..: 1 1 1 1 7 3 5 5 5 3 ...
##  $ Occupation                : int  10 10 10 10 16 15 7 7 7 20 ...
##  $ City_Category             : Factor w/ 3 levels "A","B","C": 1 1 1 1 3 1 2 2 2 1 ...
##  $ Stay_In_Current_City_Years: Factor w/ 5 levels "0","1","2","3",..: 3 3 3 3 5 4 3 3 3 2 ...
##  $ Marital_Status            : int  0 0 0 0 0 0 1 1 1 1 ...
##  $ Product_Category_1        : int  3 1 12 12 8 1 1 1 1 8 ...
##  $ Product_Category_2        : int  NA 6 NA 14 NA 2 8 15 16 NA ...
##  $ Product_Category_3        : int  NA 14 NA NA NA NA 17 NA NA NA ...
##  $ Purchase                  : int  8370 15200 1422 1057 7969 15227 19215 15854 15686 7871 ...

#check the sample
head(BFData)
```

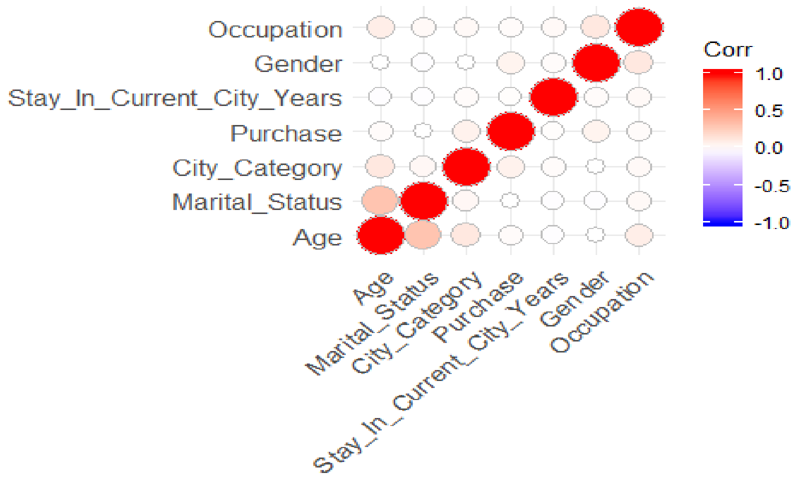
```
#REPLACE NA to 0
BFData[is.na(BFData)] <- 0

describe(BFData)
```

4. Train, test split

```
#Split the dayaset into train and test for regression analysis
set.seed(10)
split = sample(1:nrow(BFData), 0.7 * nrow(BFData))
train_BF = BFData[split,]
test_BF = BFData[-split,]
#print train data count
print("Number of rows in train data")

## [1] "Number of rows in train data"

nrow(train_BF)

#print test data count
print("Number of rows in test data")
nrow(test_BF)
print("Number of rows in Raw datset ")
nrow(BFData)
```

5. Correlations

```
#display the correllation
round(cor(Filter(is.numeric, BFData)),2)

corrgram(train_BF_NUMERIC, upper.panel=panel.pie,main= "Corrgram of Black Friday  variables"
)

# Visualize the correlation matrix (2nd method)
ggcorrplot(corr_BF, method = "circle",hc.order = TRUE)
```



6. Model fitting and Adequacy (method 1)

```
#replace the unwanted +
BFData$Stay_In_Current_City_Years[BFData$Stay_In_Current_City_Years =="4+"] <- "4"

## Warning in `[<-.factor`(`*tmp*`, BFData$Stay_In_Current_City_Years ==
## "4+", : invalid factor level, NA generated

#subset the train data and select the relevant variables
train_BF_NUMERIC<- train_BF %>%
  select(Gender, Age, Occupation, Marital_Status, City_Category, Stay_In_Current_City_Years, Purcha
se)
train_BF_NUMERIC$Gender <- as.integer(train_BF_NUMERIC$Gender)
train_BF_NUMERIC$Age <- as.integer(train_BF_NUMERIC$Age)
train_BF_NUMERIC$City_Category <- as.integer(train_BF_NUMERIC$City_Category)
train_BF_NUMERIC$Stay_In_Current_City_Years <- as.integer(train_BF_NUMERIC$Stay_In_Current_
City_Years)

#check the structure of train dataset
str(train_BF_NUMERIC)

## 'data.frame':    376303 obs. of  7 variables:
##  $ Gender                    : int  1 2 2 1 2 2 2 1 2 2 ...
##  $ Age                       : int  1 6 2 2 5 3 4 3 2 3 ...
##  $ Occupation                : int  10 16 0 3 12 14 17 14 4 0 ...
##  $ Marital_Status            : int  0 1 0 0 1 0 0 1 0 1 ...
##  $ City_Category             : int  1 2 1 3 3 1 3 3 1 1 ...
##  $ Stay_In_Current_City_Years: int  3 5 3 5 4 4 2 2 4 5 ...
##  $ Purchase                  : int  9946 15601 15242 6546 8012 5450 15461 15854 7875 10619 ...
```

```r
#subset the test data and select the relevant variables
test_BF_Numeric <- test_BF %>%
  select(Gender, Age, Occupation, Marital_Status, City_Category, Stay_In_Current_City_Years, Purcha
se)
test_BF_Numeric$Gender <- as.integer(test_BF_Numeric$Gender)
test_BF_Numeric$Age <- as.integer(test_BF_Numeric$Age)
test_BF_Numeric$City_Category <- as.integer(test_BF_Numeric$City_Category)
test_BF_Numeric$Stay_In_Current_City_Years <- as.integer(test_BF_Numeric$Stay_In_Current_City_
Years)

#check the structure of test dataset  dataset
str(test_BF_Numeric)

## 'data.frame':    161274 obs. of  7 variables:
##  $ Gender                    : int  2 2 2 2 1 1 1 1 1 1 ...
##  $ Age                       : int  5 3 3 3 4 4 4 4 4 4 ...
##  $ Occupation                : int  7 20 12 12 1 1 1 1 1 1 ...
##  $ Marital_Status            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ City_Category             : int  2 1 3 3 2 2 2 2 2 2 ...
##  $ Stay_In_Current_City_Years: int  3 2 5 5 5 5 5 5 5 5 ...
##  $ Purchase                  : int  15686 3957 19614 5982 16352 8886 5875 7089 8770 15212 ...

# Compute a correlation matrix
corr_BF <- cor(train_BF_NUMERIC)
corr_BF

##                                  Gender          Age Occupation
## Gender                      1.000000000 -0.004961799 0.11769858
## Age                        -0.004961799  1.000000000 0.09216286
## Occupation                  0.117698583  0.092162859 1.00000000
## Marital_Status             -0.010128291  0.313038932 0.02541060
## City_Category              -0.004452842  0.122412295 0.03316017
## Stay_In_Current_City_Years  0.015939548 -0.005249214 0.03183773
## Purchase                    0.059646531  0.017164263 0.02175039
##                            Marital_Status City_Category
## Gender                       -0.010128291  -0.004452842
## Age                           0.313038932   0.122412295
## Occupation                    0.025410600   0.033160166
## Marital_Status                1.000000000   0.040503473
## City_Category                 0.040503473   1.000000000
## Stay_In_Current_City_Years   -0.013152051   0.019671595
## Purchase                      0.000366598   0.068631706
##                            Stay_In_Current_City_Years    Purchase
## Gender                                    0.015939548 0.059646531
## Age                                      -0.005249214 0.017164263
## Occupation                                0.031837735 0.021750388
## Marital_Status                           -0.013152051 0.000366598
## City_Category                             0.019671595 0.068631706
## Stay_In_Current_City_Years                1.000000000 0.006920346
## Purchase                                  0.006920346 1.000000000
```

```
# Compute a matrix of correlation p-values
p_corr_BF <- cor_pmat(train_BF_NUMERIC)
p_corr_BF

# Visualize the correlation matrix of full data
```

```
#Fit multiple linear regression
model_BF <- lm(Purchase ~ Gender + Age + Occupation + Marital_Status + City_Category + Stay_I
n_Current_City_Years, data = train_BF_NUMERIC)

#create the equation from the above model

equation_BF <- noquote(paste('Purchase =',
                   round(model_BF$coefficients[1],0), '+',
                   round(model_BF$coefficients[2],0), '* Gender', '+',
                   round(model_BF$coefficients[3],0), '* Age', '+',
                   round(model_BF$coefficients[4],0), '* Occupation', '+',
                   round(model_BF$coefficients[5],0), '* Marital_Status', '+',
                   round(model_BF$coefficients[6],0), '* City_Category', '+',
                   round(model_BF$coefficients[7],0), '* Stay_In_Current_City_Years'))
#Display the equation
equation_BF

## [1] Purchase = 7025 + 677 * Gender + 36 * Age + 9 * Occupation + -51 * Marital_Status + 442 *
City_Category + 17 * Stay_In_Current_City_Years

#check the summary statistics
summary(model_BF)

#Analysis of variance
anova(model_BF)

## Analysis of Variance Table
##
## Response: Purchase
##                                Df     Sum Sq    Mean Sq   F value
## Gender                          1 3.3256e+10 3.3256e+10 1350.3195
## Age                             1 2.8498e+09 2.8498e+09  115.7114
## Occupation                      1 1.6437e+09 1.6437e+09   66.7388
## Marital_Status                  1 2.0632e+08 2.0632e+08    8.3772
## City_Category                   1 4.1953e+10 4.1953e+10 1703.4542
## Stay_In_Current_City_Years      1 1.7034e+08 1.7034e+08    6.9165
## Residuals                  376296 9.2675e+12 2.4628e+07
##                               Pr(>F)
## Gender                     < 2.2e-16 ***
## Age                        < 2.2e-16 ***
## Occupation                 3.109e-16 ***
## Marital_Status              0.003800 **
```

```
## City_Category           < 2.2e-16 ***
## Stay_In_Current_City_Years  0.008541 **
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#presence of multicollinearity between variables.
vif(model_BF)

##                     Gender                        Age
##                   1.014577                   1.132154
##                 Occupation             Marital_Status
##                   1.024449                   1.108902
##              City_Category Stay_In_Current_City_Years
##                   1.016161                   1.001768

#tolerance and vif
ols_vif_tol(model_BF)

## # A tibble: 6 x 3
##   Variables                  Tolerance   VIF
##   <chr>                          <dbl> <dbl>
## 1 Gender                         0.986  1.01
## 2 Age                            0.883  1.13
## 3 Occupation                     0.976  1.02
## 4 Marital_Status                 0.902  1.11
## 5 City_Category                  0.984  1.02
## 6 Stay_In_Current_City_Years     0.998  1.00

#Stepwise regression
# Full model should contains all the variables
fullmodel=model_BF

# null model contains no variable
nullmodel=lm(Purchase ~1, data=train_BF_NUMERIC)

#stepwise regression using AIC values
step(nullmodel, scope = list(upper=fullmodel), data=train_BF_NUMERIC, direction="both")

## Start:  AIC=6407684
## Purchase ~ 1
##
##                              Df  Sum of Sq        RSS     AIC
## + City_Category               1 4.4030e+10 9.3035e+12 6405909
## + Gender                      1 3.3256e+10 9.3143e+12 6406345
## + Occupation                  1 4.4221e+09 9.3431e+12 6407508
## + Age                         1 2.7539e+09 9.3448e+12 6407575
## + Stay_In_Current_City_Years  1 4.4767e+08 9.3471e+12 6407668
## <none>                                     9.3476e+12 6407684
## + Marital_Status              1 1.2563e+06 9.3475e+12 6407686
##
```

```
## Step:  AIC=6405909
## Purchase ~ City_Category
##
##                              Df  Sum of Sq        RSS     AIC
## + Gender                      1 3.3598e+10 9.2699e+12 6404550
## + Occupation                  1 3.5490e+09 9.3000e+12 6405768
## + Age                         1 7.2870e+08 9.3028e+12 6405882
## + Stay_In_Current_City_Years  1 2.9015e+08 9.3032e+12 6405900
## + Marital_Status              1 5.4526e+07 9.3035e+12 6405909
## <none>                                     9.3035e+12 6405909
## - City_Category               1 4.4030e+10 9.3476e+12 6407684
##
## Step:  AIC=6404550
## Purchase ~ City_Category + Gender
##
##                              Df  Sum of Sq        RSS     AIC
## + Occupation                  1 1.4613e+09 9.2685e+12 6404493
## + Age                         1 7.7342e+08 9.2691e+12 6404521
## + Stay_In_Current_City_Years  1 1.9873e+08 9.2697e+12 6404544
## <none>                                     9.2699e+12 6404550
## + Marital_Status              1 3.0908e+07 9.2699e+12 6404551
## - Gender                      1 3.3598e+10 9.3035e+12 6405909
## - City_Category               1 4.4372e+10 9.3143e+12 6406345
##
## Step:  AIC=6404493
## Purchase ~ City_Category + Gender + Occupation
##
##                              Df  Sum of Sq        RSS     AIC
## + Age                         1 5.9880e+08 9.2679e+12 6404470
## + Stay_In_Current_City_Years  1 1.6832e+08 9.2683e+12 6404488
## <none>                                     9.2685e+12 6404493
## + Marital_Status              1 4.2702e+07 9.2684e+12 6404493
## - Occupation                  1 1.4613e+09 9.2699e+12 6404550
## - Gender                      1 3.1510e+10 9.3000e+12 6405768
## - City_Category               1 4.3777e+10 9.3122e+12 6406264
##
## Step:  AIC=6404470
## Purchase ~ City_Category + Gender + Occupation + Age
##
##                              Df  Sum of Sq        RSS     AIC
## + Marital_Status              1 2.2032e+08 9.2676e+12 6404463
## + Stay_In_Current_City_Years  1 1.7498e+08 9.2677e+12 6404465
## <none>                                     9.2679e+12 6404470
## - Age                         1 5.9880e+08 9.2685e+12 6404493
## - Occupation                  1 1.2867e+09 9.2691e+12 6404521
## - Gender                      1 3.1635e+10 9.2995e+12 6405751
## - City_Category               1 4.1939e+10 9.3098e+12 6406167
##
## Step:  AIC=6404463
## Purchase ~ City_Category + Gender + Occupation + Age + Marital_Status
```

```
##
##                          Df  Sum of Sq        RSS     AIC
## + Stay_In_Current_City_Years  1 1.7034e+08 9.2675e+12 6404458
## <none>                                  9.2676e+12 6404463
## - Marital_Status            1 2.2032e+08 9.2679e+12 6404470
## - Age                       1 7.7641e+08 9.2684e+12 6404493
## - Occupation                1 1.2839e+09 9.2689e+12 6404513
## - Gender                    1 3.1587e+10 9.2992e+12 6405742
## - City_Category             1 4.1953e+10 9.3096e+12 6406161
##
## Step:  AIC=6404458
## Purchase ~ City_Category + Gender + Occupation + Age + Marital_Status +
##     Stay_In_Current_City_Years
##
##                          Df  Sum of Sq        RSS     AIC
## <none>                                  9.2675e+12 6404458
## - Stay_In_Current_City_Years  1 1.7034e+08 9.2676e+12 6404463
## - Marital_Status            1 2.1568e+08 9.2677e+12 6404465
## - Age                       1 7.8086e+08 9.2683e+12 6404488
## - Occupation                1 1.2545e+09 9.2687e+12 6404507
## - Gender                    1 3.1525e+10 9.2990e+12 6405734
## - City_Category             1 4.1830e+10 9.3093e+12 6406151

##
## Call:
## lm(formula = Purchase ~ City_Category + Gender + Occupation +
##     Age + Marital_Status + Stay_In_Current_City_Years, data = train_BF_NUMERIC)
##
## Coefficients:
##            (Intercept)           City_Category
##               7024.673                 442.243
##                 Gender              Occupation
##                676.742                   8.952
##                    Age          Marital_Status
##                 35.816                 -51.277
## Stay_In_Current_City_Years
##                 16.523
```

```
library(leaps)
subregmodel<-leaps::regsubsets(Purchase ~ City_Category + Gender + Occupation +
  Age + Marital_Status + Stay_In_Current_City_Years, data = train_BF_NUMERIC)
summary (subregmodel)
```

```
## Subset selection object
## Call: regsubsets.formula(Purchase ~ City_Category + Gender + Occupation +
##     Age + Marital_Status + Stay_In_Current_City_Years, data = train_BF_NUMERIC)
## 6 Variables  (and intercept)
##                            Forced in Forced out
## City_Category                  FALSE      FALSE
## Gender                         FALSE      FALSE
```

```
## Occupation                 FALSE     FALSE
## Age                        FALSE     FALSE
## Marital_Status             FALSE     FALSE
## Stay_In_Current_City_Years FALSE     FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: exhaustive
##          City_Category Gender Occupation Age Marital_Status
## 1  ( 1 ) "*"           " "    " "        " " " "
## 2  ( 1 ) "*"           "*"    " "        " " " "
## 3  ( 1 ) "*"           "*"    "*"        " " " "
## 4  ( 1 ) "*"           "*"    "*"        "*" " "
## 5  ( 1 ) "*"           "*"    "*"        "*" "*"
## 6  ( 1 ) "*"           "*"    "*"        "*" "*"
##          Stay_In_Current_City_Years
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) "*"

plot(subregmodel, scale="r2")

plot(subregmodel, scale="adjr2")

res.sum <- summary(subregmodel)
data.frame(
  Adj.R2 = which.max(res.sum$adjr2),
  rsq = which.max(res.sum$rsq),
  CP = which.min(res.sum$cp),
  BIC = which.min(res.sum$bic)
)

##   Adj.R2 rsq CP BIC
## 1      6   6  6   4

#fit the model in the sample and check the residul plot for better understanding

model_BF_res <- lm(Purchase ~ Gender + Age + Occupation + Marital_Status + City_Category + St
ay_In_Current_City_Years, data = #train_BF_NUMERIC[1:50,])
train_BF_NUMERIC[0:4000,])
#train_BF_NUMERIC[150:200,])
#plot the residuals
plot(model_BF_res)
```

Homogeneity of residuals variance

```
#statistical test
# Evaluate homoscedasticity
# non-constant error variance test
```

```r
ncvTest(model_BF)

bptest(model_BF)
```

H0: Errors have a constant variance H1: Errors have a non-constant variance

```r
#Independence of residuals error terms
acf(model_BF_res$residuals)
```

```r
LBQPlot(model_BF_res$residuals, lag.max = length(model_BF_res$residuals)-1 , StartLag = 0 + 1, k
= 0, SquaredQ = FALSE)

durbinWatsonTest(model_BF_res)

#Normality of residuals
# Test for Normally Distributed Errors
shapiro.test(model_BF_res$residuals)
```

prediction

```r
predTest <- predict(model_BF, newdata =test_BF_Numeric)
sseTest <- sum((predTest - test_BF_Numeric$Purchase) ^ 2)
sstTest <- sum((mean(test_BF$Purchase) - test_BF_Numeric$Purchase) ^ 2)
print ("Model R2 (Test Data)")

## [1] "Model R2 (Test Data)"

modelR2Test <- 1 - sseTest/sstTest
modelR2Test

## [1] 0.008614034

print ("Model RMSE (Test Data)")

## [1] "Model RMSE (Test Data)"

rmseTest <- sqrt(mean((predTest - test_BF_Numeric$Purchase) ^ 2))
rmseTest

## [1] 4952.513
```

7. Model fitting and Adequacy (method 2)

Include all the colums

```r
library(corrgram)
corrgram(BFData, upper.panel=panel.pie,main= "Corrgram of Black Friday  variables" )
```

```r
#change city to years
train_BF$Stay_In_Current_City_Years <- as.integer(train_BF$Stay_In_Current_City_Years)
#marital status to factor
train_BF$Marital_Status <- factor(train_BF$Marital_Status)

train_BF$Gender <- as.integer(train_BF$Gender)

#change city to years
test_BF$Stay_In_Current_City_Years <- as.integer(test_BF$Stay_In_Current_City_Years)
#marital status to factor
test_BF$Marital_Status <- factor(test_BF$Marital_Status)

ModelBFFit= lm(Purchase ~Occupation+City_Category+Stay_In_Current_City_Years+Product_Cate
gory_1+Marital_Status+Product_Category_2+Product_Category_3, data = train_BF)

#create the equation from the above model

equation_BF <- noquote(paste('Purchase =',
                  round(ModelBFFit $coefficients[1],0), '+',
                  round(ModelBFFit $coefficients[2],0), '* Occupation', '+',
                  round(ModelBFFit $coefficients[3],0), '* City_CategoryB  ', '+',
                  round(ModelBFFit $coefficients[4],0), '* City_CategoryC  ', '+',
                  round(ModelBFFit $coefficients[5],0), '* Stay_In_Current_City_Years ', '+',
                  round(ModelBFFit $coefficients[6],0), '* Product_Category_1', '+',
                  round(ModelBFFit $coefficients[7],0), '* Marital_Status1 ',
                  round(ModelBFFit $coefficients[8],0), '* Product_Category_2', '+',
                  round(ModelBFFit $coefficients[9],0), '* Product_Category_3'))
#Display the equation
equation_BF

#check the summary statistics
summary(ModelBFFit)

#Analysis of variance
anova(ModelBFFit)

## Analysis of Variance Table
##
## Response: Purchase
##                                Df     Sum Sq    Mean Sq   F value
## Occupation                      1 4.4221e+09 4.4221e+09   205.463
## City_Category                   2 4.6590e+10 2.3295e+10  1082.353
## Stay_In_Current_City_Years      1 2.5272e+08 2.5272e+08    11.742
## Product_Category_1              1 9.1514e+11 9.1514e+11 42519.641
## Marital_Status                  1 1.6798e+08 1.6798e+08     7.805
## Product_Category_2              1 5.5006e+09 5.5006e+09   255.571
## Product_Category_3              1 2.7659e+11 2.7659e+11 12851.134
## Residuals                  376294 8.0989e+12 2.1523e+07
##                              Pr(>F)
```

```
## Occupation             < 2.2e-16 ***
## City_Category          < 2.2e-16 ***
## Stay_In_Current_City_Years 0.0006112 ***
## Product_Category_1       < 2.2e-16 ***
## Marital_Status         0.0052105 **
## Product_Category_2       < 2.2e-16 ***
## Product_Category_3       < 2.2e-16 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#presence of multicollinearity between variables.
vif(ModelBFFit)
```

```
##                         GVIF Df GVIF^(1/(2*Df))
## Occupation             1.003000  1       1.001499
## City_Category          1.005348  2       1.001334
## Stay_In_Current_City_Years 1.001747  1       1.000873
## Product_Category_1       1.180553  1       1.086533
## Marital_Status         1.002962  1       1.001480
## Product_Category_2       1.008375  1       1.004179
## Product_Category_3       1.188089  1       1.089995
```

```
#tolerance and vif
ols_vif_tol(ModelBFFit)
```

```
## # A tibble: 8 x 3
##   Variables              Tolerance   VIF
##   <chr>                      <dbl> <dbl>
## 1 Occupation                 0.997  1.00
## 2 City_CategoryB             0.673  1.49
## 3 City_CategoryC             0.670  1.49
## 4 Stay_In_Current_City_Years    0.998  1.00
## 5 Product_Category_1         0.847  1.18
## 6 Marital_Status1            0.997  1.00
## 7 Product_Category_2         0.992  1.01
## 8 Product_Category_3         0.842  1.19
```

```
#Stepwise regression
# Full model should contains all the variables
fullmodel=ModelBFFit

# null model contains no variable
nullmodel=lm(Purchase ~1, data=train_BF)

#stepwise regression using AIC values
step(nullmodel, scope = list(upper=fullmodel), data=train_BF, direction="both")
```

```
## Start:  AIC=6407684
## Purchase ~ 1
##
```

```
##                          Df  Sum of Sq        RSS     AIC
## + Product_Category_1         1 9.2674e+11 8.4208e+12 6368397
## + Product_Category_3         1 7.5524e+11 8.5923e+12 6375984
## + City_Category              2 4.7553e+10 9.3000e+12 6405769
## + Product_Category_2         1 1.3807e+10 9.3337e+12 6407130
## + Occupation                 1 4.4221e+09 9.3431e+12 6407508
## + Stay_In_Current_City_Years  1 4.4767e+08 9.3471e+12 6407668
## <none>                                    9.3476e+12 6407684
## + Marital_Status             1 1.2563e+06 9.3475e+12 6407686
##
## Step:  AIC=6368397
## Purchase ~ Product_Category_1
##
##                          Df  Sum of Sq        RSS     AIC
## + Product_Category_3         1 2.8717e+11 8.1336e+12 6355342
## + City_Category              2 3.6819e+10 8.3840e+12 6366752
## + Product_Category_2         1 6.1249e+09 8.4147e+12 6368125
## + Occupation                 1 3.4645e+09 8.4173e+12 6368244
## + Marital_Status             1 4.6637e+08 8.4203e+12 6368378
## + Stay_In_Current_City_Years  1 2.5490e+08 8.4206e+12 6368388
## <none>                                    8.4208e+12 6368397
## - Product_Category_1         1 9.2674e+11 9.3476e+12 6407684
##
## Step:  AIC=6355342
## Purchase ~ Product_Category_1 + Product_Category_3
##
##                          Df  Sum of Sq        RSS     AIC
## + City_Category              2 3.1136e+10 8.1025e+12 6353903
## + Occupation                 1 2.9332e+09 8.1307e+12 6355209
## + Product_Category_2         1 1.2815e+09 8.1324e+12 6355285
## + Marital_Status             1 3.7244e+08 8.1333e+12 6355327
## + Stay_In_Current_City_Years  1 2.5168e+08 8.1334e+12 6355333
## <none>                                    8.1336e+12 6355342
## - Product_Category_3         1 2.8717e+11 8.4208e+12 6368397
## - Product_Category_1         1 4.5867e+11 8.5923e+12 6375984
##
## Step:  AIC=6353903
## Purchase ~ Product_Category_1 + Product_Category_3 + City_Category
##
##                          Df  Sum of Sq        RSS     AIC
## + Occupation                 1 2.3037e+09 8.1002e+12 6353798
## + Product_Category_2         1 1.0694e+09 8.1014e+12 6353855
## + Stay_In_Current_City_Years  1 1.7594e+08 8.1023e+12 6353897
## + Marital_Status             1 1.5370e+08 8.1024e+12 6353898
## <none>                                    8.1025e+12 6353903
## - City_Category              2 3.1136e+10 8.1336e+12 6355342
## - Product_Category_3         1 2.8148e+11 8.3840e+12 6366752
## - Product_Category_1         1 4.5531e+11 8.5578e+12 6374474
##
## Step:  AIC=6353798
```

```
## Purchase ~ Product_Category_1 + Product_Category_3 + City_Category +
##     Occupation
##
##                             Df  Sum of Sq        RSS     AIC
## + Product_Category_2         1 1.0526e+09 8.0992e+12 6353751
## + Stay_In_Current_City_Years  1 1.3844e+08 8.1001e+12 6353794
## + Marital_Status             1 1.2629e+08 8.1001e+12 6353794
## <none>                                  8.1002e+12 6353798
## - Occupation                 1 2.3037e+09 8.1025e+12 6353903
## - City_Category              2 3.0507e+10 8.1307e+12 6355209
## - Product_Category_3         1 2.8107e+11 8.3813e+12 6366632
## - Product_Category_1         1 4.5508e+11 8.5553e+12 6374365
##
## Step:  AIC=6353751
## Purchase ~ Product_Category_1 + Product_Category_3 + City_Category +
##     Occupation + Product_Category_2
##
##                             Df  Sum of Sq        RSS     AIC
## + Stay_In_Current_City_Years  1 1.3818e+08 8.0990e+12 6353747
## + Marital_Status             1 1.2681e+08 8.0990e+12 6353747
## <none>                                  8.0992e+12 6353751
## - Product_Category_2         1 1.0526e+09 8.1002e+12 6353798
## - Occupation                 1 2.2869e+09 8.1014e+12 6353855
## - City_Category              2 3.0301e+10 8.1295e+12 6355152
## - Product_Category_3         1 2.7662e+11 8.3758e+12 6366387
## - Product_Category_1         1 4.5479e+11 8.5539e+12 6374308
##
## Step:  AIC=6353747
## Purchase ~ Product_Category_1 + Product_Category_3 + City_Category +
##     Occupation + Product_Category_2 + Stay_In_Current_City_Years
##
##                             Df  Sum of Sq        RSS     AIC
## + Marital_Status             1 1.3074e+08 8.0989e+12 6353743
## <none>                                  8.0990e+12 6353747
## - Stay_In_Current_City_Years  1 1.3818e+08 8.0992e+12 6353751
## - Product_Category_2         1 1.0523e+09 8.1001e+12 6353794
## - Occupation                 1 2.2495e+09 8.1013e+12 6353849
## - City_Category              2 3.0239e+10 8.1293e+12 6355145
## - Product_Category_3         1 2.7663e+11 8.3756e+12 6366383
## - Product_Category_1         1 4.5471e+11 8.5537e+12 6374300
##
## Step:  AIC=6353743
## Purchase ~ Product_Category_1 + Product_Category_3 + City_Category +
##     Occupation + Product_Category_2 + Stay_In_Current_City_Years +
##     Marital_Status
##
##                             Df  Sum of Sq        RSS     AIC
## <none>                                  8.0989e+12 6353743
## - Marital_Status             1 1.3074e+08 8.0990e+12 6353747
## - Stay_In_Current_City_Years  1 1.4210e+08 8.0990e+12 6353747
```

```
## - Product_Category_2       1 1.0528e+09 8.0999e+12 6353790
## - Occupation               1 2.2215e+09 8.1011e+12 6353844
## - City_Category            2 3.0042e+10 8.1289e+12 6355132
## - Product_Category_3       1 2.7659e+11 8.3755e+12 6366378
## - Product_Category_1       1 4.5483e+11 8.5537e+12 6374301

##
## Call:
## lm(formula = Purchase ~ Product_Category_1 + Product_Category_3 +
##     City_Category + Occupation + Product_Category_2 + Stay_In_Current_City_Years +
##     Marital_Status, data = train_BF)
##
## Coefficients:
##            (Intercept)       Product_Category_1
##                9948.97                  -318.20
##     Product_Category_3           City_CategoryB
##                 149.09                   163.74
##         City_CategoryC               Occupation
##                 695.13                    11.79
##     Product_Category_2  Stay_In_Current_City_Years
##                   8.55                    15.09
##        Marital_Status1
##                  37.97
```

```r
library(leaps)
subregmodel<-leaps::regsubsets(Purchase ~ Occupation + Marital_Status +
  City_Category + Stay_In_Current_City_Years + Product_Category_1 +
  Product_Category_2 + Product_Category_3, data = train_BF)
summary (subregmodel)
```

```
## Subset selection object
## Call: regsubsets.formula(Purchase ~ Occupation + Marital_Status + City_Category +
##     Stay_In_Current_City_Years + Product_Category_1 + Product_Category_2 +
##     Product_Category_3, data = train_BF)
## 8 Variables  (and intercept)
##                            Forced in Forced out
## Occupation                     FALSE      FALSE
## Marital_Status1                FALSE      FALSE
## City_CategoryB                 FALSE      FALSE
## City_CategoryC                 FALSE      FALSE
## Stay_In_Current_City_Years     FALSE      FALSE
## Product_Category_1             FALSE      FALSE
## Product_Category_2             FALSE      FALSE
## Product_Category_3             FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          Occupation Marital_Status1 City_CategoryB City_CategoryC
## 1  ( 1 ) " "        " "             " "            " "
## 2  ( 1 ) " "        " "             " "            " "
## 3  ( 1 ) " "        " "             " "            "*"
```

```
## 4  ( 1 ) "*"      " "         " "         "*"
## 5  ( 1 ) "*"      " "         "*"         "*"
## 6  ( 1 ) "*"      " "         "*"         "*"
## 7  ( 1 ) "*"      " "         "*"         "*"
## 8  ( 1 ) "*"      "*"         "*"         "*"
##          Stay_In_Current_City_Years Product_Category_1 Product_Category_2
## 1  ( 1 ) " "                        "*"                " "
## 2  ( 1 ) " "                        "*"                " "
## 3  ( 1 ) " "                        "*"                " "
## 4  ( 1 ) " "                        "*"                " "
## 5  ( 1 ) " "                        "*"                " "
## 6  ( 1 ) " "                        "*"                "*"
## 7  ( 1 ) "*"                        "*"                "*"
## 8  ( 1 ) "*"                        "*"                "*"
##          Product_Category_3
## 1  ( 1 ) " "
## 2  ( 1 ) "*"
## 3  ( 1 ) "*"
## 4  ( 1 ) "*"
## 5  ( 1 ) "*"
## 6  ( 1 ) "*"
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"

plot(subregmodel, scale="r2") ;  plot(subregmodel, scale="adjr2")

res.sum <- summary(subregmodel)
data.frame(
 Adj.R2 = which.max(res.sum$adjr2),
 rsq = which.max(res.sum$rsq),
 CP = which.min(res.sum$cp),
 BIC = which.min(res.sum$bic)
)

#fit the model in the sample and check the residul plot for better understanding
#length(train_BF$User_ID)
ModelBFFit <- lm(Purchase ~ Gender + Age + Occupation + Marital_Status + City_Category + Stay
_In_Current_City_Years, data = train_BF[0:4000,])
#train_BF])
#train_BF_Numeric2[150:200,])
#plot the residuals
plot(ModelBFFit)
```

#Homogeneity of residuals variance

```
#statistical test
# Evaluate homoscedasticity
# non-constant error variance test

ncvTest(ModelBFFit)
```

```
bptest(ModelBFFit)
```

H0: Errors have a constant variance H1: Errors have a non-constant variance

```
#Independence of residuals error terms
acf(ModelBFFit$residuals)

LBQPlot(ModelBFFit$residuals, lag.max = length(ModelBFFit$residuals)-1 , StartLag = 0 + 1, k = 0,
SquaredQ = FALSE)

durbinWatsonTest(ModelBFFit)

#Normality of residuals
# Test for Normally Distributed Errors
shapiro.test(ModelBFFit$residuals)

# predTest <- predict(ModelBFFit, newdata =test_BF)

sseTest <- sum((predTest - test_BF$Purchase) ^ 2)

sstTest <- sum((mean(test_BF$Purchase) - test_BF$Purchase) ^ 2)
print ("Model R2 (Test Data)")
modelR2Test <- 1 - sseTest/sstTest
modelR2Test
print ("Model RMSE (Test Data)")
rmseTest <- sqrt(mean((predTest - test_BF_Numeric$Purchase) ^ 2))
rmseTest
```