Assignment #5: Multiple Regression
Your name: Abhinav Kumar

The project requires you to perform, report, and interpret a multiple regression using R. (You must use R or RStudio to do the analyses.) Use this MS Word file to record your answers. You can paste the code and graphs from R into the spaces provided below.

IMPORTANT: In a separate .txt file, submit the code listing of your program for the entire project.

You should upload two files: an MS Word file with your answers and a .txt file with your code.

The data set is Fund.csv. There are approximately 1,500 observations.

The variable definitions are as follows:

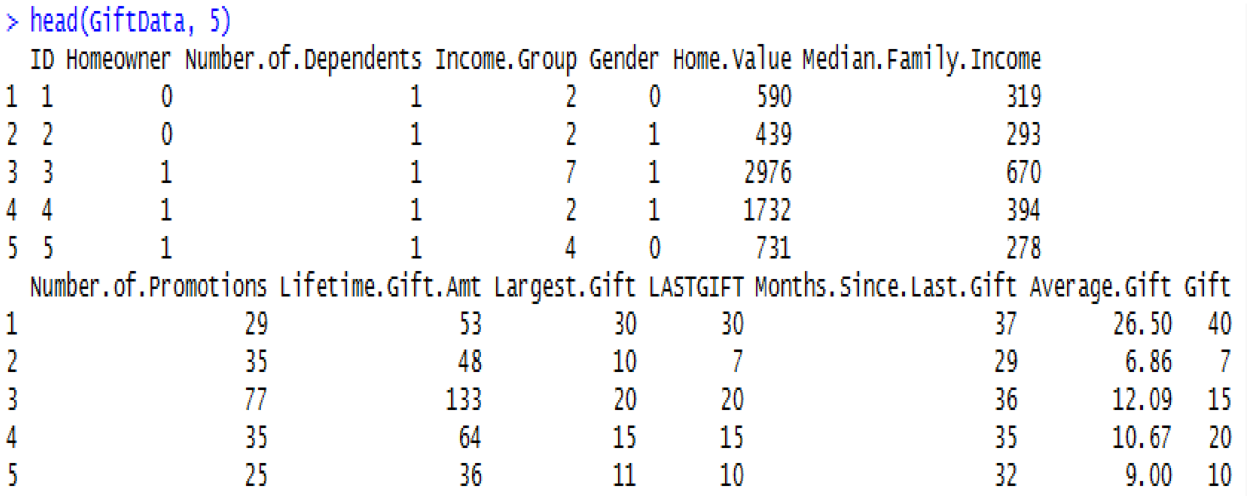
Variable	**Definition**
ID	Observation number
Homeowner	1=”Yes”, 0=”No”
Number of dependents	1 to 4
Income Group	1 to 7, 7=highest
Gender	0=Male, 1=Female
Home Value	Neighborhood home values in 000’s, from U.S. Census
Median Family Income	Median income in neighborhood in 000’s, from U.S. Census
Number of Promotions	Lifetime number of promotions
Lifetime Gift Amt	Total of all gifts to date
Largest Gift	Largest gift to date
LASTGIFT	Amount in dollars of last previous gift
Months Since Last Gift	Number of months since previous gift
Average Gift	Average amount of all past gifts
Gift	Amount in dollars of gift from current campaign

INSTRUCTIONS:

1. Read the Fund.csv file into RStudio.

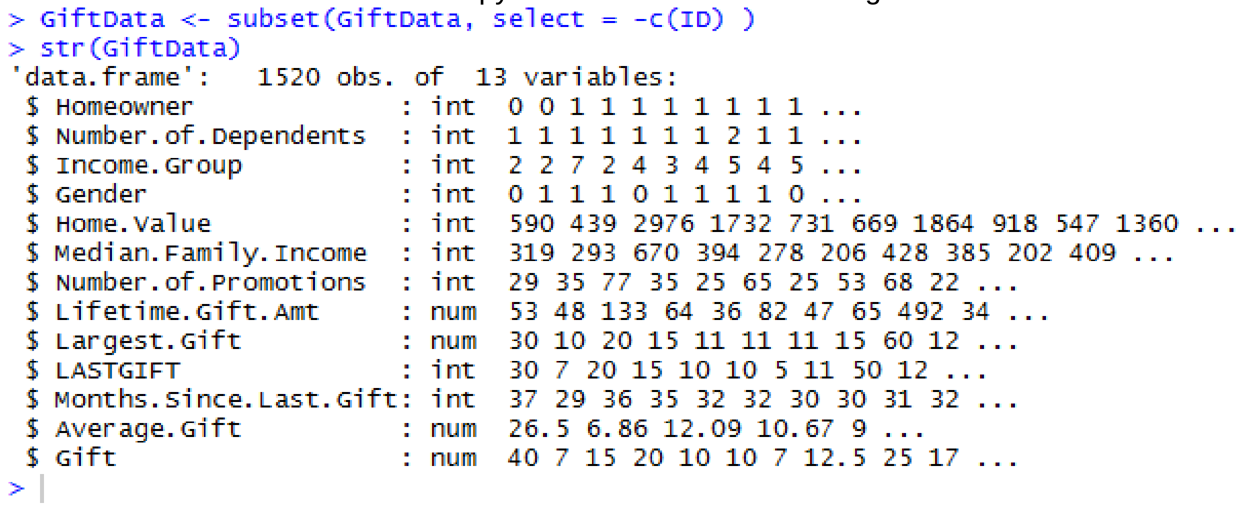
```
# The uploaded dataset name is ProjectData.csv not Fund.csv
setwd("C:/R")
GiftData <- read.csv("./ProjectData.csv", header=TRUE)
```

a. Print and the first 5 lines of the file here:



```
> head(GiftData, 5)
  ID Homeowner Number.of.Dependents Income.Group Gender Home.Value Median.Family.Income
1  1         0                    1            2      0        590                  319
2  2         0                    1            2      1        439                  293
3  3         1                    1            7      1       2976                  670
4  4         1                    1            2      1       1732                  394
5  5         1                    1            4      0        731                  278
  Number.of.Promotions Lifetime.Gift.Amt Largest.Gift LASTGIFT Months.Since.Last.Gift Average.Gift Gift
1                   29                53           30       30                     37        26.50   40
2                   35                48           10        7                     29         6.86    7
3                   77               133           20       20                     36        12.09   15
4                   35                64           15       15                     35        10.67   20
5                   25                36           11       10                     32         9.00   10
```

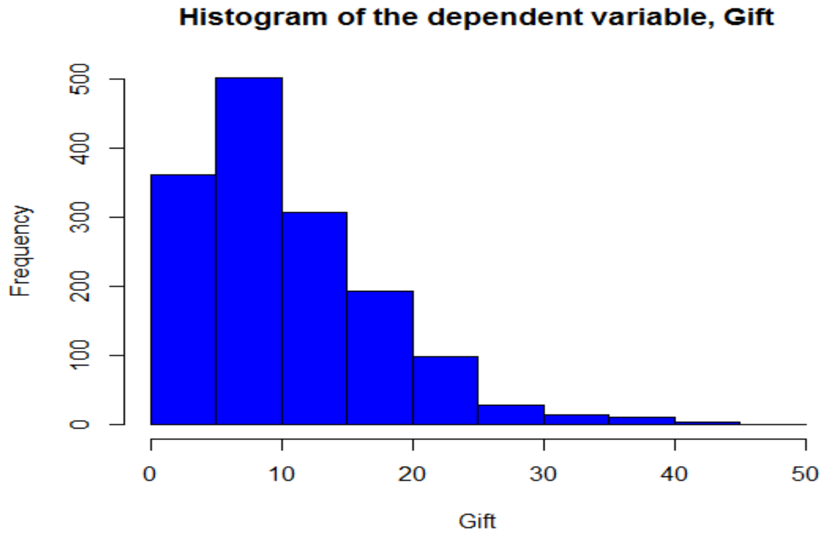
b. Remove the ID variable. Print and copy the structure of the resulting file here:



```
> GiftData <- subset(GiftData, select = -c(ID) )
> str(GiftData)
'data.frame':   1520 obs. of  13 variables:
 $ Homeowner            : int  0 0 1 1 1 1 1 1 1 1 ...
 $ Number.of.Dependents : int  1 1 1 1 1 1 1 2 1 1 ...
 $ Income.Group         : int  2 2 7 2 4 3 4 5 4 5 ...
 $ Gender               : int  0 1 1 1 0 1 1 1 1 0 ...
 $ Home.Value           : int  590 439 2976 1732 731 669 1864 918 547 1360 ...
 $ Median.Family.Income : int  319 293 670 394 278 206 428 385 202 409 ...
 $ Number.of.Promotions : int  29 35 77 35 25 65 25 53 68 22 ...
 $ Lifetime.Gift.Amt    : num  53 48 133 64 36 82 47 65 492 34 ...
 $ Largest.Gift         : num  30 10 20 15 11 11 11 15 60 12 ...
 $ LASTGIFT             : int  30 7 20 15 10 10 5 11 50 12 ...
 $ Months.Since.Last.Gift: int  37 29 36 35 32 32 30 30 31 32 ...
 $ Average.Gift         : num  26.5 6.86 12.09 10.67 9 ...
 $ Gift                 : num  40 7 15 20 10 10 7 12.5 25 17 ...
> 
```

2.

a. Produce a histogram of the dependent variable, Gift, and paste it here:

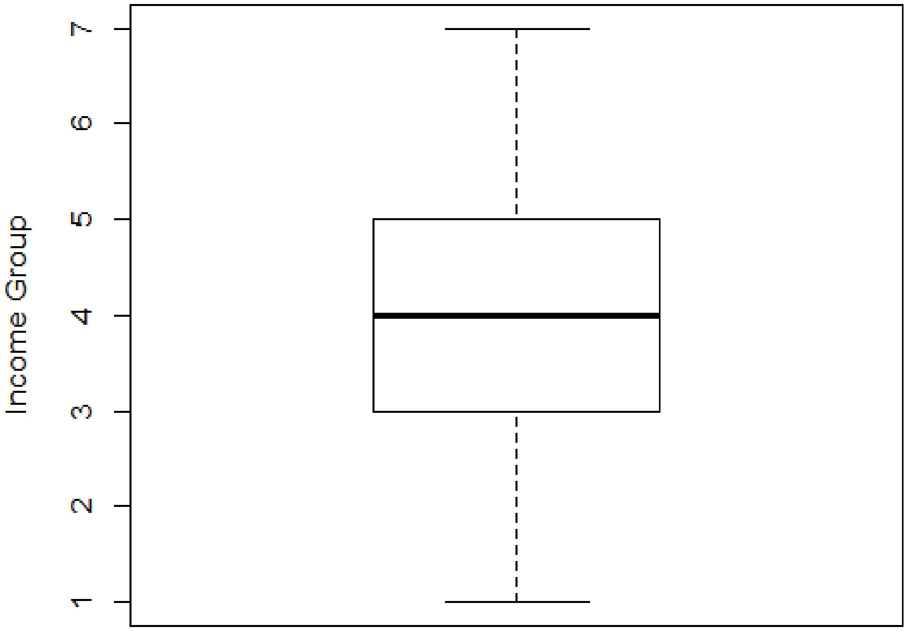


b. Comment on the skewness of the distribution.
Its highly right-skewed as seen in the histogram.
Also the calculated value of skewness comes as positive which clearly indicates its right skewed –

```
> library("moments")
> skewness(GiftData$Gift)
[1] 1.186429
```

3.

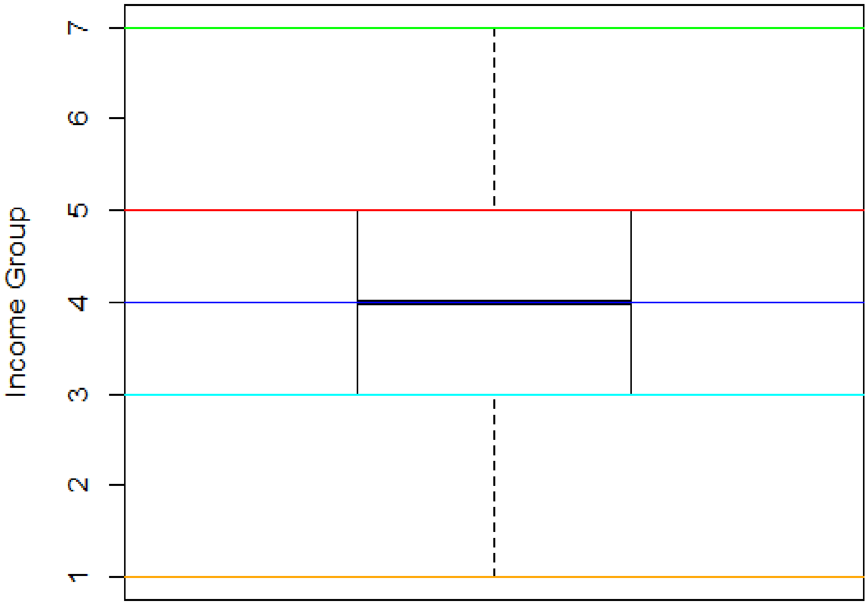
a. Produce a boxplot by Income Group and paste it here:



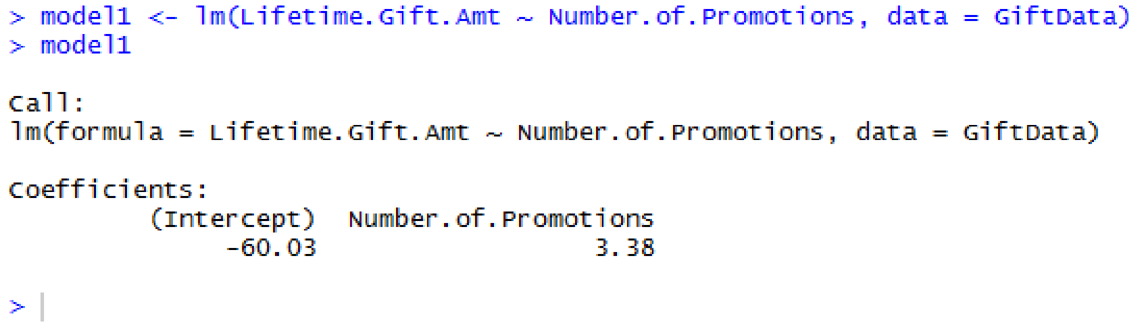
b. What do you conclude from this boxplot?

Below are key conclusions from above boxplot –

- **Maximum or Highest Income group is 7**
- **Minimum or lowest Income group is 1**
- **Median Income group is 4 (50% of Income group values are above it and 50 % values are below it)**
- **It has no outliers.**
- **First quartile is Income group 3. 25% of data is below this Income group**
- **Third quartile is Income group 5. 25% of data is greater than this Income group.**
- **The middle “box” represents the middle 50% data of Income group. The range from lower to upper quartile is called the inter-quartile range. The middle 50% of data fall within the inter-quartile range.**

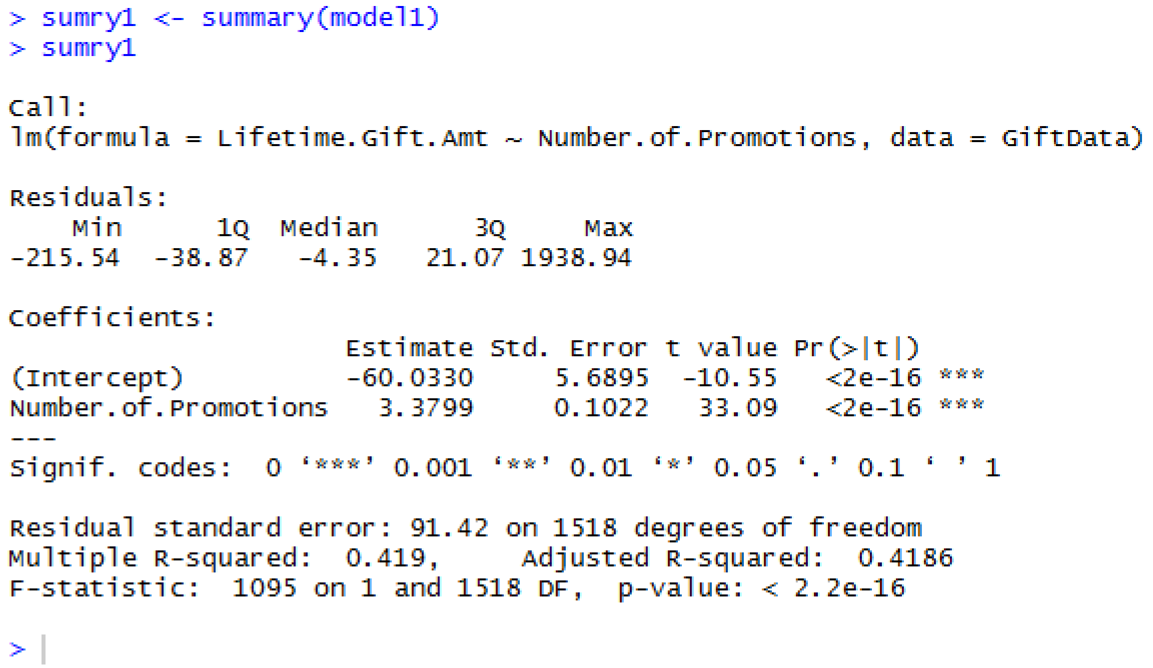


4. Run a simple regression using the entire data set with Lifetime Gift Amt as the dependent variable and Number of Promotions as the predictor variable.



```
> model1 <- lm(Lifetime.Gift.Amt ~ Number.of.Promotions, data = GiftData)
> model1

Call:
lm(formula = Lifetime.Gift.Amt ~ Number.of.Promotions, data = GiftData)

Coefficients:
         (Intercept)  Number.of.Promotions
              -60.03                  3.38

> 
```

   a. Provide a summary of the regression results here:



```
> sumry1 <- summary(model1)
> sumry1

Call:
lm(formula = Lifetime.Gift.Amt ~ Number.of.Promotions, data = GiftData)

Residuals:
    Min      1Q  Median      3Q     Max 
-215.54  -38.87   -4.35   21.07 1938.94 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -60.0330     5.6895  -10.55   <2e-16 ***
Number.of.Promotions   3.3799     0.1022   33.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 91.42 on 1518 degrees of freedom
Multiple R-squared:  0.419,     Adjusted R-squared:  0.4186 
F-statistic:  1095 on 1 and 1518 DF,  p-value: < 2.2e-16

> |
```

b. Provide an interpretation of the regression coefficient on Number of Promotions.

- **Intercept estimate – is the avg value of Lifetime gift amount (Y) when Number of Promotions (X) = 0. In this case its negative, so in a way if number of promotions is 0, there is none Lifetime gift amount.**

- **Slope estimate - shows how much the dependent variable (Lifetime Gift amount) is expected to increase (since the coefficient is positive) when the independent variable (Number of Promotions) increases by one.**
In this case: 3.3799

- **We can consider a linear model to be statistically significant only when both p-values are less than the pre-determined statistical significance level, which is ideally 0.05. We can also observe this significance level from the stars at the end of the row. The more the stars beside the variable’s p value, the more significant the variable.**

  So, after observing both the p-Values in this case (as we can see the number of stars at the end is 3 which shows that the relationship is significant, and also p value is very less than 0.05), we can say coefficient is significant -
<2e-16 ***, p-value: < 2.2e-16

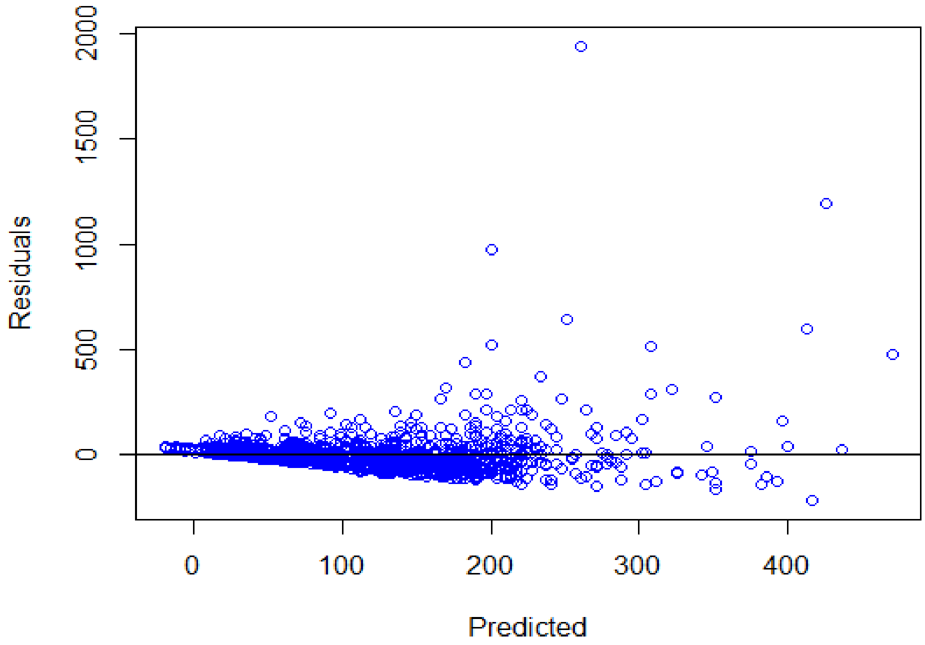
- **The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. We want it to be far away from zero as this would indicate we could reject the null**

hypothesis - that is, we could declare a relationship between Lifetime gift amount and Number of Promotions exist.
In our example, the t-statistic values (33.09) are relatively far away from zero and are large relative to the standard error, which could indicate a relationship exists.

- **The R-squared statistic provides a measure of how well the model is fitting the actual data. It takes the form of a proportion of variance. The R square is a measure of the linear relationship between our predictor variable (Number of Promotions) and our response variable (Lifetime gift amount). It always lies between 0 and 1 and the higher R square is, the better is fit.**

  In our case its 0.419, so roughly 41.9% of the variance found in the response variable (Lifetime gift amount) can be explained by the predictor variable (Number of Promotions).

c. Plot the residuals (y-axis) versus the predicted values (x-axis) from this regression; paste the plot below. What do you conclude about the distribution of the residuals.



We can conclude from above graph that this is non-linear as the values are not randomly distributed and forming a cluster.

5. Set the seed to 1. Create training and validation samples of the data set with a split of 80% / 20% for the training and testing sets, respectively.
   a. Why are training and validation sets needed?

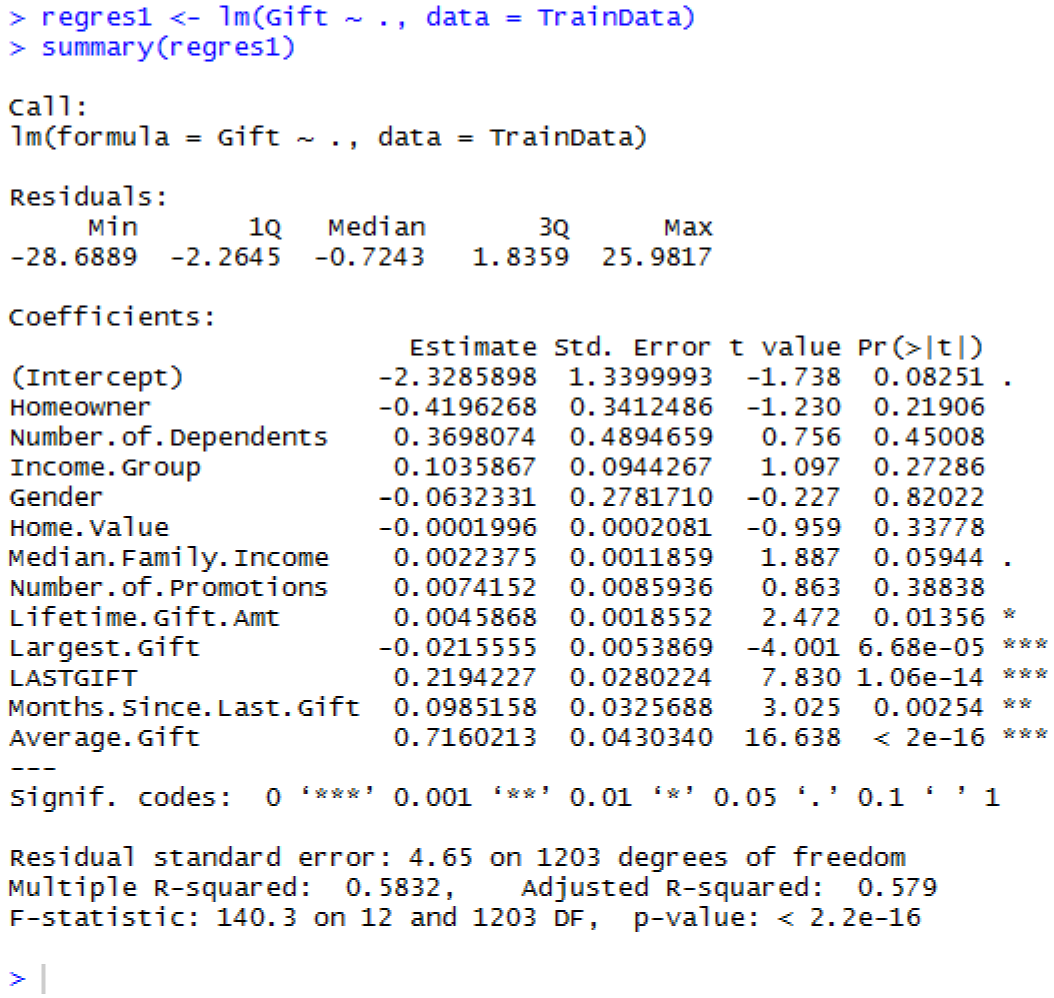
Training set is needed for learning, that is to fit the parameters [i.e., weights] of the classifier. It is used to build the model and discover potentially predictive relationships.

Validation set is needed to tune the parameters [i.e., architecture, not weights] of a classifier. It is needed to estimate how good our model has been trained (that is dependent upon the size of data, the value we would like to predict, input etc) and to estimate model properties (mean error for numeric predictors, classification errors for classifiers etc.)

b. Fill in the following table:

	Number of cases	Average Gift Amount Target	Median Gift Amount Target
Training set	1216	11.89744	10
Validation set	304	11.98849	10

6. Run a least squares regression using the training data with Gift as the dependent variable and all of the other variables as predictors.
   a. Provide a summary of the regression results here.



```
> regres1 <- lm(Gift ~ ., data = TrainData)
> summary(regres1)

Call:
lm(formula = Gift ~ ., data = TrainData)

Residuals:
     Min       1Q   Median       3Q      Max 
-28.6889  -2.2645  -0.7243   1.8359  25.9817 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)            -2.3285898  1.3399993  -1.738  0.08251 .  
Homeowner              -0.4196268  0.3412486  -1.230  0.21906    
Number.of.Dependents    0.3698074  0.4894659   0.756  0.45008    
Income.Group            0.1035867  0.0944267   1.097  0.27286    
Gender                 -0.0632331  0.2781710  -0.227  0.82022    
Home.Value             -0.0001996  0.0002081  -0.959  0.33778    
Median.Family.Income    0.0022375  0.0011859   1.887  0.05944 .  
Number.of.Promotions    0.0074152  0.0085936   0.863  0.38838    
Lifetime.Gift.Amt       0.0045868  0.0018552   2.472  0.01356 *  
Largest.Gift           -0.0215555  0.0053869  -4.001 6.68e-05 ***
LASTGIFT                0.2194227  0.0280224   7.830 1.06e-14 ***
Months.Since.Last.Gift  0.0985158  0.0325688   3.025  0.00254 ** 
Average.Gift            0.7160213  0.0430340  16.638  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.65 on 1203 degrees of freedom
Multiple R-squared:  0.5832,    Adjusted R-squared:  0.579 
F-statistic: 140.3 on 12 and 1203 DF,  p-value: < 2.2e-16

> 
```

   b. Which predictors are significant? Are the signs on the significant variable coefficients sensible? Discuss.

We can consider a linear model to be statistically significant only when both p-values are less than the pre-determined statistical significance level, which is ideally 0.05. We can also observe this significance level from the stars at the end of the row. The more the stars beside the variable’s p value, the more significant the variable.

Based on above point, the below predictors are significant –

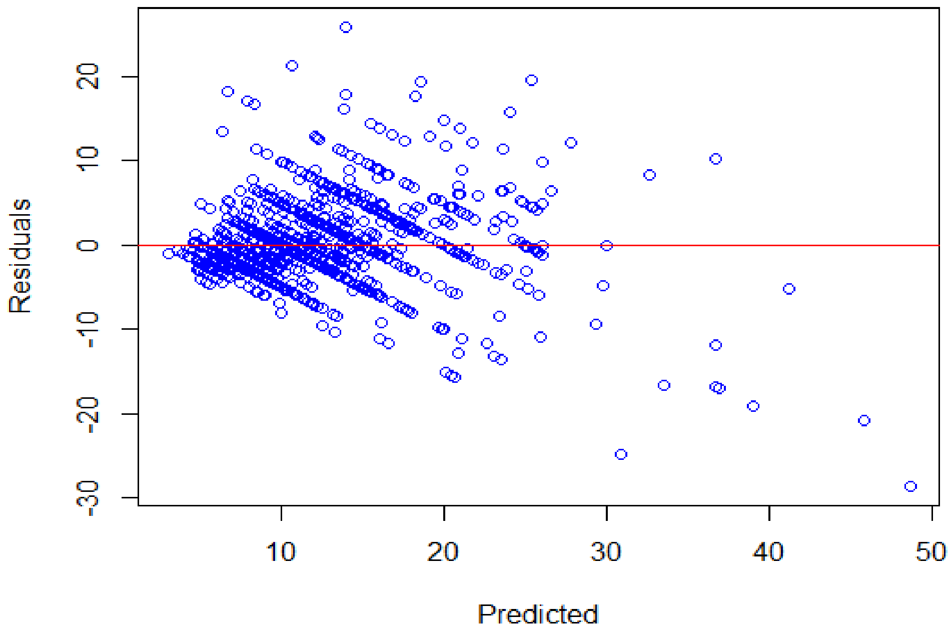
Highly significant - Average.Gift, LASTGIFT, Largest.Gift

Less significant – Months.Since.Last.Gift, Lifetime.Gift.Amt

Very less significant – Median.Family.Income

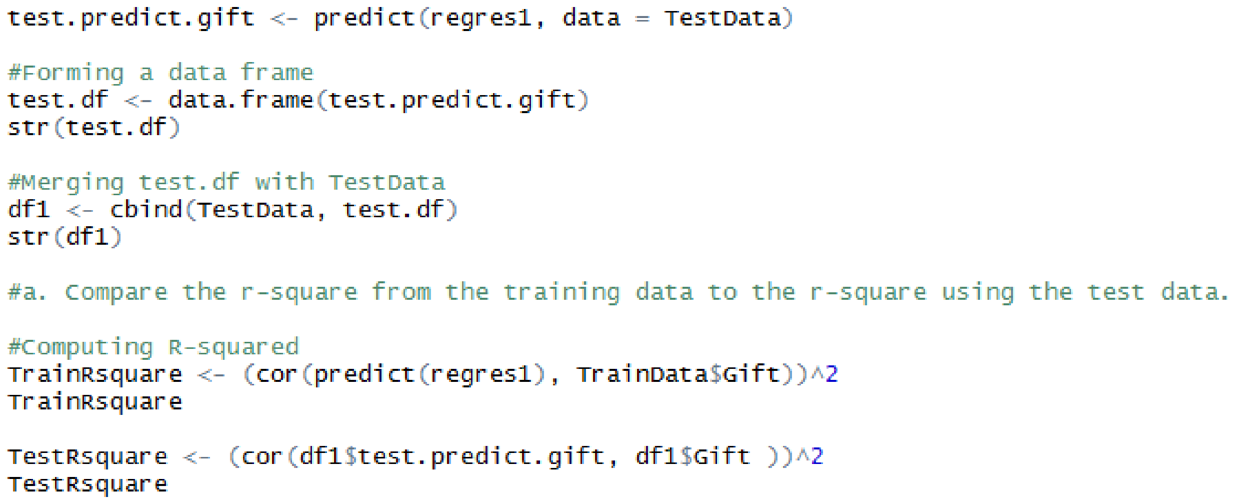
The sign on significant variable - Largest.Gift is negative and not sensible as it should also increase as per the increase in predictor variable.

c. Plot the residuals (y-axis) versus the predicted values (x-axis) from this regression; paste the plot below. What do you conclude about the distribution of the residuals.



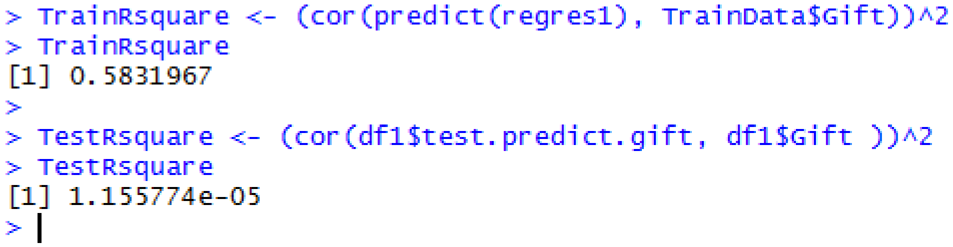
We can conclude from above graph that this is non-linear as the values are not randomly distributed and forming a cluster.

7. Using the regression you just ran using all predictors, whether significant or not, determine the r-square (the square of the correlation) between the actual value of Gift in the <u>test data</u> with the <u>prediction</u> of Gift.



```
test.predict.gift <- predict(regres1, data = TestData)

#Forming a data frame
test.df <- data.frame(test.predict.gift)
str(test.df)

#Merging test.df with TestData
df1 <- cbind(TestData, test.df)
str(df1)

#a. Compare the r-square from the training data to the r-square using the test data.

#Computing R-squared
TrainRsquare <- (cor(predict(regres1), TrainData$Gift))^2
TrainRsquare

TestRsquare <- (cor(df1$test.predict.gift, df1$Gift ))^2
TestRsquare
```

a. Compare the r-square from the training data to the r-square using the test data.



```
> TrainRsquare <- (cor(predict(regres1), TrainData$Gift))^2
> TrainRsquare
[1] 0.5831967
>
> TestRsquare <- (cor(df1$test.predict.gift, df1$Gift ))^2
> TestRsquare
[1] 1.155774e-05
> 
```

b. Is there evidence of overfitting? Discuss.

Yes here is overfitting as we can see in the above screenshot that R square from test data (1.155774e-05) is much lower than from training data (0.5831967).