(Analytics Blog)

Multiple Linear regression and Binary Logistic Regression Model of Auto Insurance Data

Parshu Rath

February 2021 update

Abstract

Multipole Linear Rgression and Binary Logistic Models were explored in R using a set of Auto insurance data. Descriptive and predictive analytics methods were used to explore the data and the efficiency of the models. Models were compared based on several metrics and criteria for selecting an optimally performing model.

Keywords: Multiple Linear Regression, Logistic regression, Chi-square test, ROC, Homoskedasticity, Step-method, Variable selection, Confusion matrix

Primary Software Tools: R version 4.3.0 (Bunny-Wunnies Freak Out)

Contents

1. Introduction

Auto insurance data with approximately 8000 records were explored in this analysis. Data has two response variable, (1) *'TARGET_FLAG'* (whether the car was in an accident), and (2) *'TARGET_AMOUNT'* (cost of repair after the accident) . Here we will explore, analyze the data set. Linear regression method will be used for building model for the *'TARGET_AMOUNT'* response variable and binary Logistic regression model on a binary response variable *'TARGET_FLAG'* (classification model). Classifications and probabilities for a test-set of data will be evaluated by these models.

Below is a description of the some of the attributes in the dataset.

Table 1: Data Explanation

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	None
TARGET_AMT	If car was in a crash, what was the cost	None
AGE	Age of Driver	Very young people tend to be risky. Maybe very old people also.
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	# Claims (Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
HOMEKIDS	# Children at Home	Unknown effect
HOME_VAL	Home Value	In theory, home owners tend to drive more responsibly
INCOME	Income	In theory, rich people tend to get into fewer crashes
JOB	Job Category	In theory, white collar jobs tend to be safer
KIDSDRIV	# Driving Children	When teenagers drive your car, you are more likely to get into crashes
MSTATUS	Marital Status	In theory, married people drive more safely
MVR_PTS	Motor Vehicle Record Points	If you get lots of traffic tickets, you tend to get into more crashes
OLDCLAIM	Total Claims (Past 5 Years)	If your total payout over the past five years was high, this suggests future payouts will be high
PARENT1	Single Parent	Unknown effect
RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver.
SEX	Gender	Urban legend says that women have less crashes then men. Is that true?
TIF	Time in Force	People who have been customers for a long time are usually more safe.
TRAVTIME	Distance to Work	Long drives to work usually suggest greater risk
URBANICITY	Home/Work Area	Unknown
YOJ	Years on Job	People who stay at a job for a long time are usually more safe

2. Tools and Steps in Model Building

R version 4.3.0 (Bunny-Wunnies Freak Out) was used for data exploration and model building.

Below is a list of steps in data exploraation, model building and model comparison.

- Data exploration
- Data preparation
- Build Models
- Model comparison
- Model selection

3. Data Exploration

In this section, various aspects of the dataset were explored both numerically and visually. Some of the items to explore in this section are:

a. What are Mean / Standard Deviation / Median of the attributes in the dataset
b. Visualization plots: Bar Chart and Box Plot of the data
c. Correlation: Is the data correlated to the target (response) variable (or to other variables?)
d. Missing data: Are any of the variables missing and need to be imputed "fixed"?

3.1 General Information about the Dataset

The dataset contains a total of 8161 rows of observations and 25 variables in the columns. The first column is an index column and thus was excluded from data analysis. The columns *"TARGET_FLAG"* and *"TARGET_AMT"* are response variables and the remaining 23 variables are predictors.

The columns *"BLUEBOOK"*, *"INCOME"*, *"HOME_VAL"*, and *"OLDCLAIM"* are incorrectly assigned as factor variables due to punctuation marks ($ and ','). Punctuations were removed and the column data types were changed to numerical values. The data type of the variables *"TARGET_FLAG"*, *"EDUCATION"*, *"JOB"*, *"MSTATUS"*, *"PARENT1"*, *"RED_CAR"*, *"REVOKED"*, and *"SEX"* were set as factors (non numerical).

```
#Number of rows and columns
dim(ins)
```

```
## [1] 8161   25
```

```
names(ins)
```

```
##  [1] "TARGET_FLAG" "TARGET_AMT"  "KIDSDRIV"    "AGE"         "HOMEKIDS"
##  [6] "YOJ"         "INCOME"      "PARENT1"     "HOME_VAL"    "MSTATUS"
## [11] "SEX"         "EDUCATION"   "JOB"         "TRAVTIME"    "CAR_USE"
## [16] "BLUEBOOK"    "TIF"         "CAR_TYPE"    "RED_CAR"     "OLDCLAIM"
## [21] "CLM_FREQ"    "REVOKED"     "MVR_PTS"     "CAR_AGE"     "URBANICITY"
```

```
#attach(ins)
```

3.2 Variable Summary

Inspection of the data reveals that there are several missing values. The variable *"CAR_AGE"* has the most missing values numbered at 510. Summary of the variables is shown below.

Table 2: Variable summary

	Min	Q1	Median	Mean	Q3	Max	NAs	Std Dev	Skew-ness	Kurto-sis
TARGET_FLAG	No : 6008		Yes: 2153							
TARGET_AMT	0	0	0	1504	1036	107586		4704.03	8.71	115.32
KIDSDRIV	0	0	0	0.1711	0	4		0.51	3.35	14.78
AGE	16	39	45	44.79	51	81	6	8.63	-0.03	2.94
HOMEKIDS	0	0	0	0.7212	1	5		1.12	1.34	3.65
YOJ	0	9	11	10.5	13	23	454	4.09	-1.2	4.18
INCOME	0	28097	54028	61898	85986	367030	445			
PARENT1	No : 7084		Yes: 1077							
HOME_VAL	0	0	161160	154867	238724	885282	464			
MSTATUS	Yes : 4894		z_No: 3267							
SEX	M : 3786		z_F: 4375							
EDUCATION	<High School : 1203	Bachelors : 2242	Masters : 1658	PhD : 728	z_High School: 2330					
JOB	z_Blue Collar: 1825	Clerical : 1271	Professional : 1117	Manager : 988	Lawyer : 835	Student : 712	(Other) : 1413			
TRAVTIME	5	22	33	33.49	44	142		15.91	0.45	3.67
CAR_USE	Commercial: 3029		Private : 5132							
BLUEBOOK	1500	9280	14440	15710	20850	69740				
TIF	1	1	4	5.351	7	25		4.15	0.89	3.42
CAR_TYPE	Minivan :2145	Panel Truck: 676	Pickup :1389	Sports Car : 907	Van : 750	z_SUV :2294				
RED_CAR	no : 5783		yes: 2378							
OLDCLAIM	0	0	0	4037	4636	57037				
CLM_FREQ	0	0	0	0.7986	2	5		1.16	1.21	3.29
REVOKED	No : 7161	Yes: 1000								
MVR_PTS	0	0	1	1.696	3	13				
CAR_AGE	-3	1	8	8.328	12	28	510	2.15	1.35	4.38
URBANICITY	Highly Urban/ Urban :6492		z_Highly Rural/ Rural:1669							

As seen in the table above, some of the variables have median values very different from their corresponding mean values, inferring that the distribution is not normal. Similar conclusions are also reached by examining the skewness and kurtosis values for the numerical variables. In particular the response variable *"TARGET_AMT"* is strongly positively skewed with a median of zero. This is expected as most of the data (6008 observations, equals 73.6% of the total data) have *"TARGET_FLAG"* as zero indicating the driver was not involved in a crash and thus the repair amount (_"TARGET_AMT_") is zero.

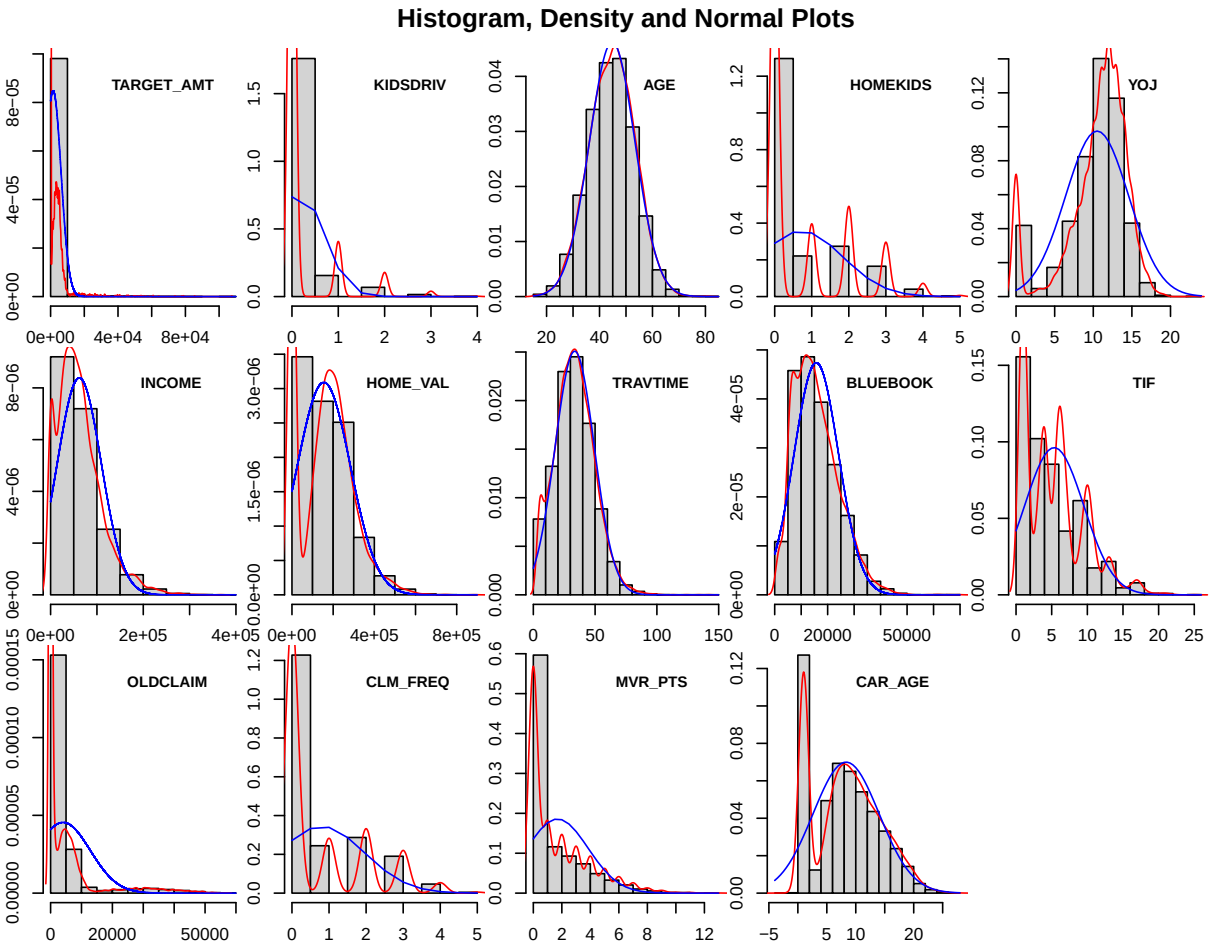
A better look at the distributions are seen in their histogram plots discussed below.

3.3 Histogram Plots

There are 14 numerical variables in the data set. They are:

```
ins_num <- ins %>% select(where(is.numeric))
names(ins_num)
```

```
##  [1] "TARGET_AMT" "KIDSDRIV"   "AGE"        "HOMEKIDS"   "YOJ"
##  [6] "INCOME"     "HOME_VAL"   "TRAVTIME"   "BLUEBOOK"   "TIF"
## [11] "OLDCLAIM"   "CLM_FREQ"   "MVR_PTS"    "CAR_AGE"
```



From the above figure it appears that *"TARGET_AMT"* is highly skewed which may be misleading due to most of the value being zero (no repair cost if there is no crash). We will take a closer look at it later in this section.

There are 11 non-numeric variables in the datasset. They are:

```
##  [1] "TARGET_FLAG" "PARENT1"     "MSTATUS"    "SEX"        "EDUCATION"
##  [6] "JOB"         "CAR_USE"     "CAR_TYPE"   "RED_CAR"    "REVOKED"
## [11] "URBANICITY"
```

3.4 Correlation

To find the interdependence of the variables, correlation values were calculated among all the 14 numeric columns. No strong correlation was observed between any two predictors. Correlation values are shown in the R script output below.

Table 3: Variable Correlations

Correlation	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	HOME_VAL	TRAVTIME	BLUEBOOK	TIF	OLDCLAIM	CLM_FREQ	MVR_PTS	CAR_AGE
TARGET_AMT	1	0.05	-0.1	0.06	-0.02	-0.06	-0.09	0.03	0	-0.04	0.08	0.12	0.14	-0.1
KIDSDRIV	0.05	1	-0.1	0.46	0.05	-0.04	-0.01	0	-0.02	0	0.02	0.04	0.06	-0.1
AGE	-0.05	-0.07	1	-0.45	0.14	0.19	0.22	0.01	0.17	0	-0.03	-0.03	-0.08	0.18
HOMEKIDS	0.06	0.46	-0.5	1	0.1	-0.16	-0.11	-0.02	-0.11	0	0.03	0.03	0.07	-0.2
YOJ	-0.02	0.05	0.14	0.1	1	0.28	0.27	-0.02	0.14	0.03	0	-0.03	-0.04	0.06
INCOME	-0.06	-0.04	0.19	-0.16	0.28	1	0.58	-0.04	0.43	0	-0.04	-0.05	-0.07	0.41
HOME_VAL	-0.09	-0.01	0.22	-0.11	0.27	0.58	1	-0.03	0.26	0	-0.06	-0.1	-0.09	0.22
TRAVTIME	0.03	0	0.01	-0.02	-0.02	-0.04	-0.03	1	-0.01	-0.01	-0.02	0.01	0	-0
BLUEBOOK	0	-0.02	0.17	-0.11	0.14	0.43	0.26	-0.01	1	0	-0.03	-0.05	-0.06	0.19
TIF	-0.04	0	0	0	0.03	0	0	-0.01	0	1	-0.02	-0.02	-0.04	0.01
OLDCLAIM	0.08	0.02	-0	0.03	0	-0.04	-0.06	-0.02	-0.03	-0.02	1	0.49	0.27	-0
CLM_FREQ	0.12	0.04	-0	0.03	-0.03	-0.05	-0.1	0.01	-0.05	-0.02	0.49	1	0.4	-0
MVR_PTS	0.14	0.06	-0.1	0.07	-0.04	-0.07	-0.09	0	-0.06	-0.04	0.27	0.4	1	-0
CAR_AGE	-0.06	-0.05	0.18	-0.16	0.06	0.41	0.22	-0.03	0.19	0.01	-0.01	-0.01	-0.02	1

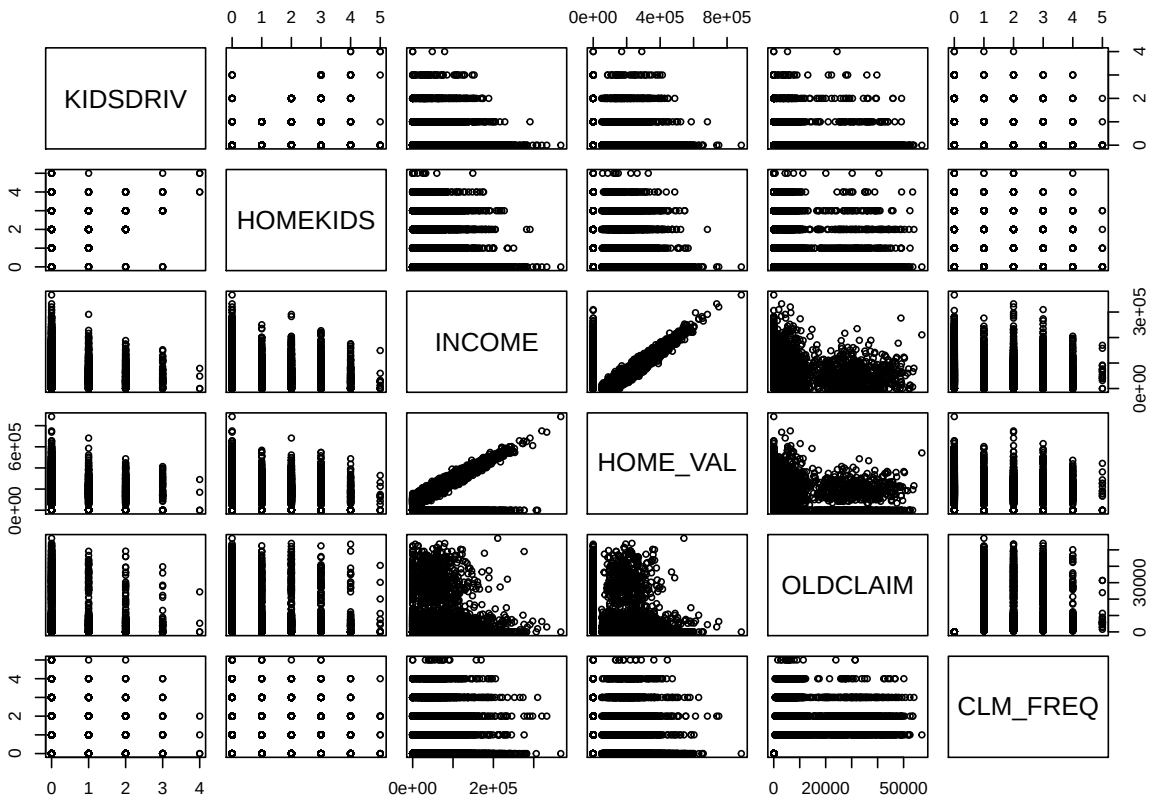
The top 10 absolute correlation values were determined and are shown below.

```
##  [1] 0.58 0.49 0.46 0.45 0.43 0.41 0.40 0.28 0.27 0.27
```

The highest correlation value observed was 0.58 between the predictors *"INCOME"* and *"HOMEVAL"*. It is intuitive to expect such a correlation. Correlation is further investigated, qualitatively, in the next section using pairs plot.

3.5 Pairs plot

Pairs plots of some of the top correlated variables are shown in the R script output below. Again the predictors *"INCOME"* and *"HOME_VAL"* look strongly correlated. The correlation value of 0.58 appears to be too low for the observed correlation in the plot due to some of the very small values for the *"INCOME"* variable.



2.6 Missing data There are several missing data in the dataset as shown in the summary table 'Table 1' and in

3.6 Missing Data

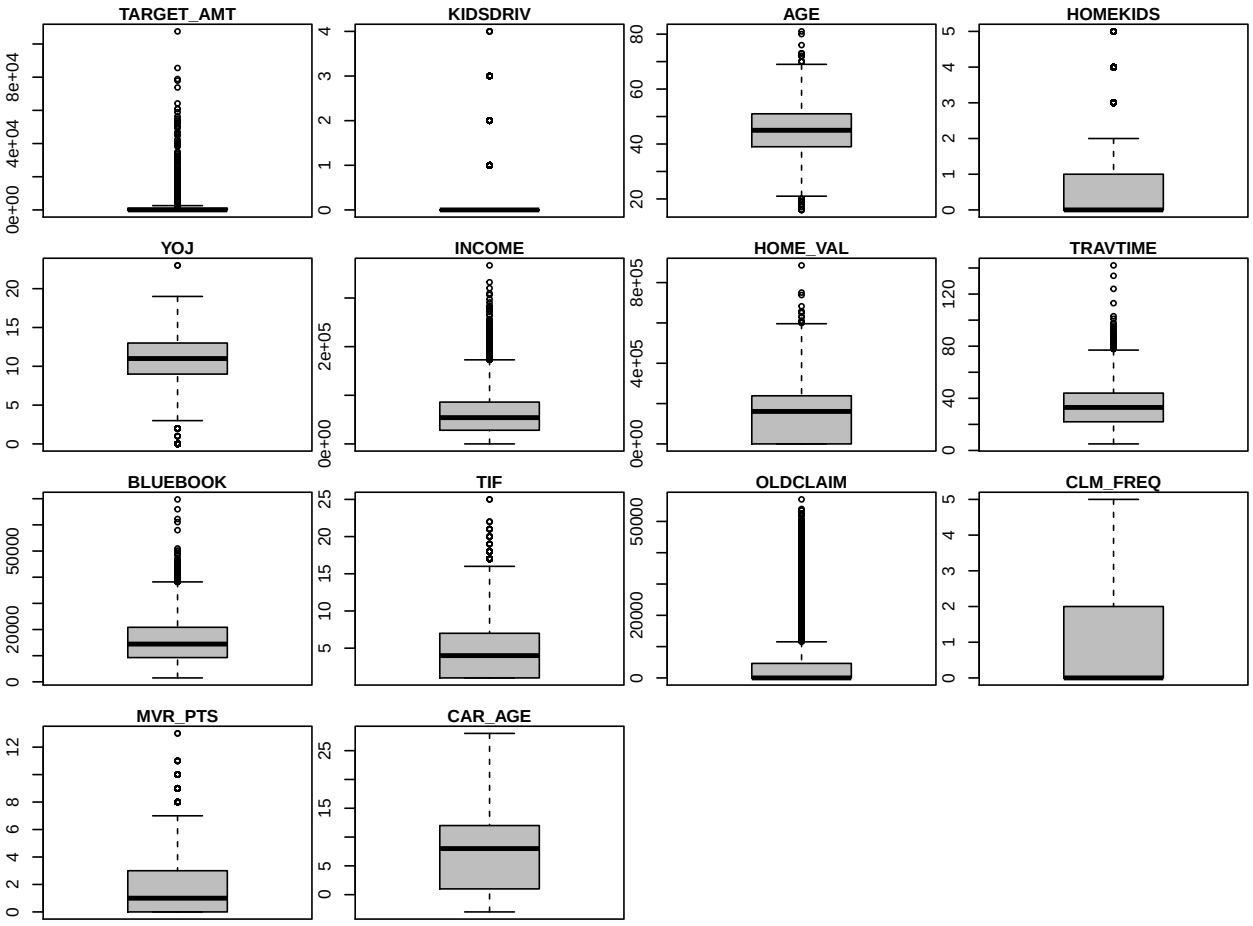
Several columns have missing data (NAs). Below is a list of columns with the numbers of missing entries in them.

```
##      AGE      YOJ   INCOME HOME_VAL  CAR_AGE
##        6      454      445      464      510

## [1] "--NA percentages are shown below--"

##      AGE      YOJ   INCOME HOME_VAL  CAR_AGE
##     0.07     5.56     5.45     5.69     6.25
```

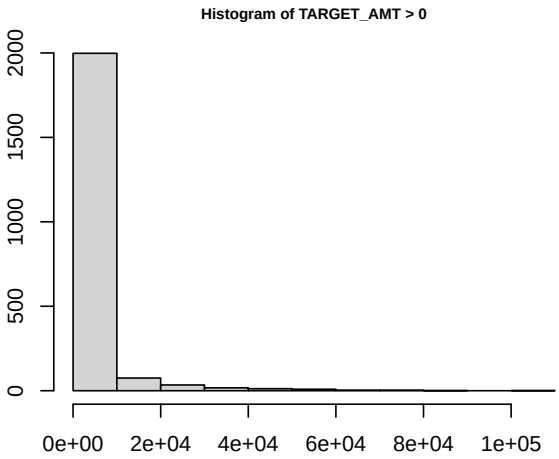
The variable *"CAR_AGE"* has the most missing data with 510 observations missing (6.25% of the total data). Typically about 5% of the missing data can be imputed without introducing bias in the dataset. Therefore, rows with NA values for selected predictor(s) will be removed and data imputed for other NA values (see the Data Preparation section).

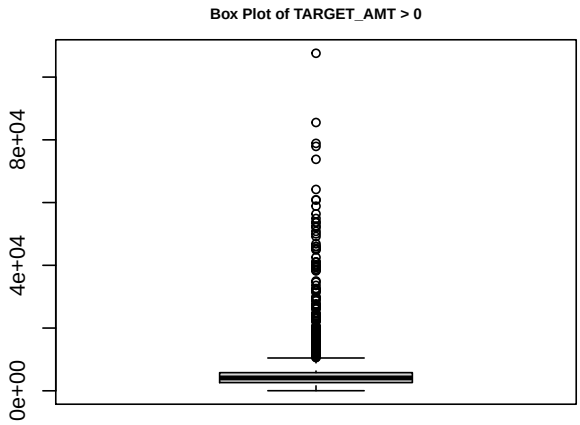
3.7 Outliers in the predictors

Outliers (values outside the whiskers $\{Q1-1.5*IQR\}$ or $\{Q3+1.5*IQR\}$) in the predictors were qualitatively investigated using boxplots. The plots are shown in the R sccript output below.



As seen in the plot above, majority of the predictors are highly skewed. Of particular interest is the distribution of the response variable *"TARGET_AMT"*. From data summary, it is seen that the skewness for this predictor is 8.71 with a kurtosis of 115.32. One reason for the skewed distribution may be the inclusion of significant number of zero values for this predictor (i.e. observations with no repair required for the car). Taking a closer look at this variable after selecting only the non-zero values we find that the skewness is reduced to 5.64 and kurtosis is also reduced to 45.49. The histogram boxplot of the "TARGET_AMT" variables with nan-zero values are shown below.





Although, the skewness and kurtosis are reduced after removing the zero value observations (see below), the variable is still highly skewed as seen in above.

```
## [1] "Skewness of TARGET_AMT:   8.71"
## [1] "Kurtosis of TARGET_AMT:   115.32"
## [1] "Skewness of non-zero TARGET_AMT:   5.64"
## [1] "Kurtosis of non-zero TARGET_AMT:   45.49"
```

4. Data Preparation

4.1 Missing Values

As discussed in the previous section, there are several missing values in the dataset. The variable *"CAR_AGE"* has the most missing values that accounts for 6.25% of the total data. Some missing values observations will removed and the remaining missing value observations will needx to be imputed as discussed below.

4.1.1 Removing Missing Values Data

First we investigate how many observations have more than one missing variables. Below is a summary table showing the number of rows and the corresponding percentage of variables missing in them. It shows that 6448 are complete cases. One row has missing values in 4 variables (16%). Also 13 rows have missing values in 3 or more variables (12% or more). Since multiple variable values missing rows may not be very reliable, these rows (13 of them) will be excluded from analysis.

```
## Percent missing columns
##    0    4    8   12   16
## 6448 1561  139   12    1
```

After removing the 13 rows as discussed above, there were still four variable with missing values above the 5% threshold as shown below. Missing values rows of each predictors were removed in steps until the percent missing values were below the threshold.

First the rows missing values in *"CAR_AGE"* were excluded. The missing values in predictors *INCOME, HOME_VAL,* and *YOJ* are still more than 5% of the total observations (see below). This process was continued until the total missing data is below the 5% threshold. Percentage of the remaining missing values after removing rows are shown below. After removing the rows with NAs, only the *AGE* column now has 0.05% NAs which can be ignored.

```
## [1] "Percent NAs in full set of Data:"
##      AGE      YOJ   INCOME HOME_VAL  CAR_AGE
##     0.07     5.47     5.31     5.56     6.15
## [1] "Percent NAs after removing NA rows in CAR_AGE"
##      AGE      YOJ   INCOME HOME_VAL
##     0.08     5.60     5.36     5.54
## [1] "Percent NAs after removing NA rows in CAR_AGE and INCOME"
##      AGE      YOJ HOME_VAL
##     0.07     5.61     5.54
## [1] "Percent NAs after removing NA rows in CAR_AGE, INCOME and Home_VAL"
##  AGE  YOJ
## 0.04 5.63
## [1] "Percent NAs after removing NA rows in CAR_AGE, INCOME, Home_VAL and YOJ"
##  AGE
## 0.05
```

```
## [1] 6451   25
## [1] "Number of rows with NAs:  3"
```

After removing the rows with missing values there were 6451 observations left. The *"AGE"* variable still had 0.05% of missing data (3 observations). These data will be imputed as described in the next section.

```
## [1] "Remaining Dataset Dimension:  "
## [1] 6451   25
## [1] "Number of rows with NAs:  3"
```

4.1.2 Imputation of Missing values

There are only 3 rows of missing data in the variable *AGE*. The three missing values were imputed using **Predictive Mean Matching** method (Rubin (1986), Little (1988)).

4.2 Dummy Variables for Non-numeric Factors

Dummy variables can be thought of as binary (*0 0r 1*) variables in regeression analysis. It is an usefule method of using categorical predictors in regression analysis. Typicslly, one level of a variable is held fixed (assigned *0*) and the others are assigend the value *1* for building the regression model. This step is typically taken care of automatically in the regression model in R. This will be discussed more in the next section on building models.

4.3 Response Variable "TARGET_AMT" Transformation

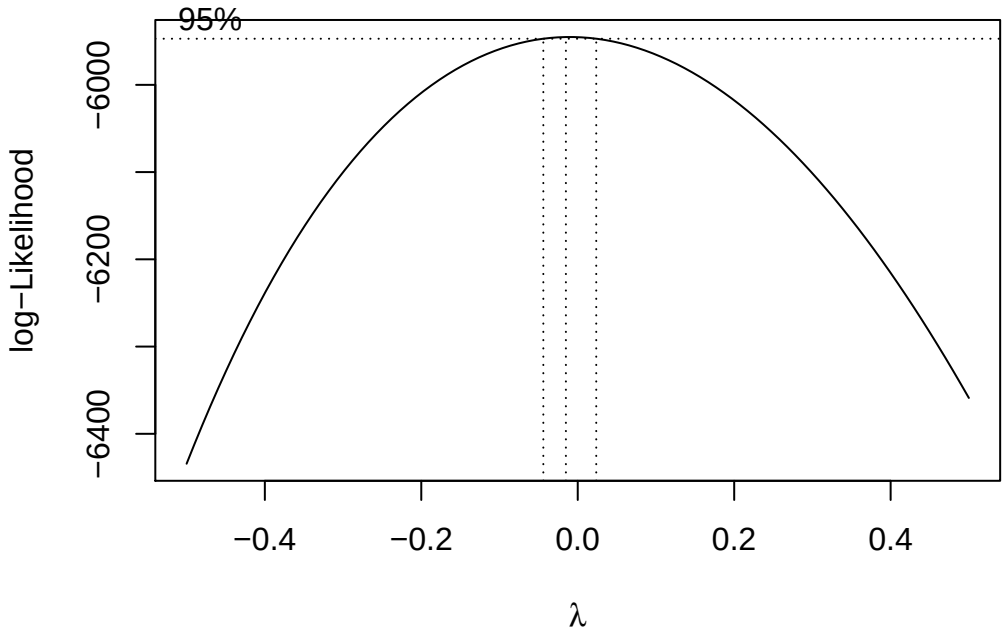
4.3.1 Box-Cox transformation and normality test of residuals

The variable "TARGET_AMT" is the response variable for the linear regression model. Normally distributed errors in linear regression model assures that the least square estimators are **BLUE (Best Linear Unbiased Estimates)**. The residuals of a linear regression of this response against all the predictors is highly skewed and is not normally distributed. This is shown by the kurtosis and the Shapiro-Wilk test for normality. The results are shown below.

```
## [1] 36.06691
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm1.1$residuals
## W = 0.52467, p-value < 2.2e-16
```

Therefore, a transformation of the response is necessary for the residuals to be normal. The **Box-Cox transformation** of the response gives a λ value for the maximum log-likelihood close to zero as shown below.




```
## [1] "lambda:  -0.015"
```

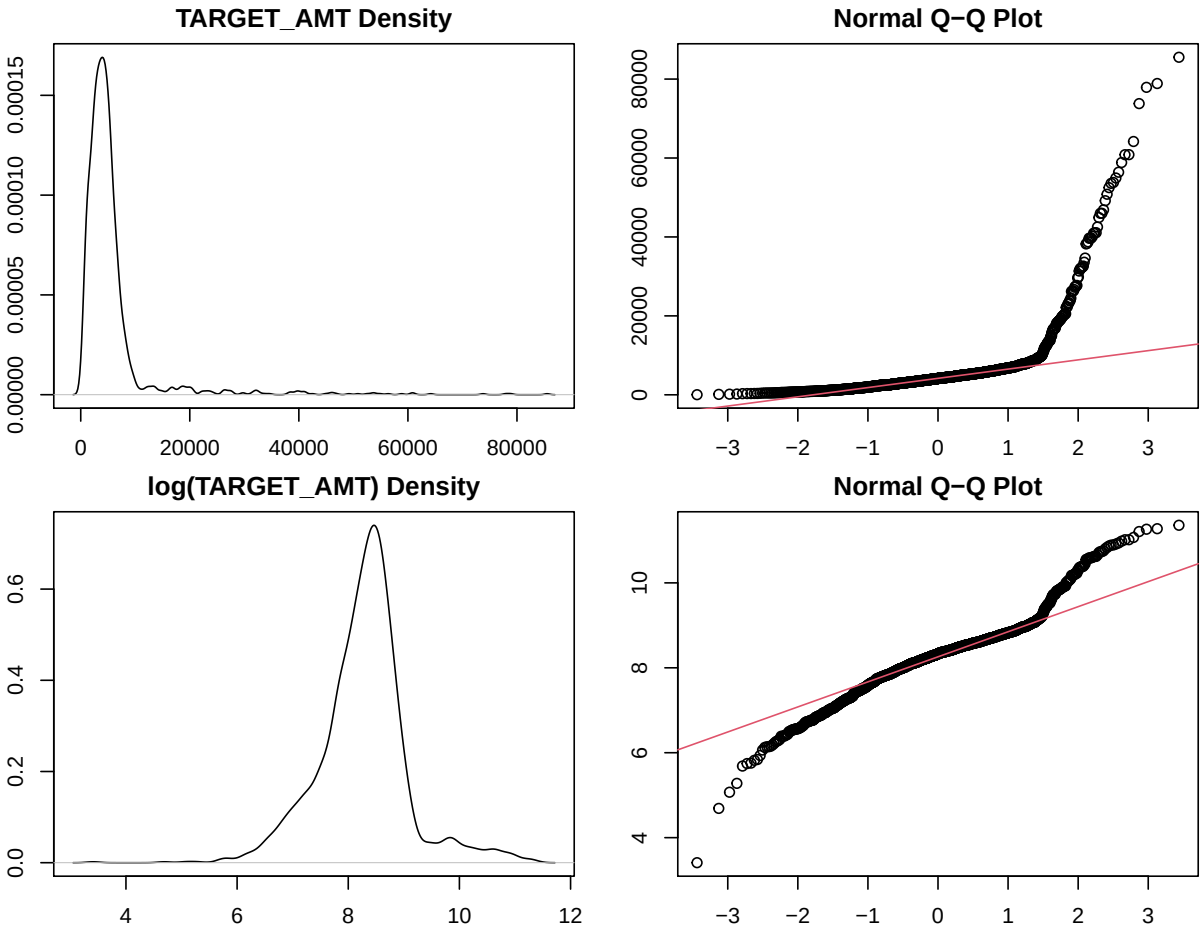
The observation that the response is highly skewed and the Box-Cox transformation indicating that the λ value for the maximum log-likelihood is near zero suggest that a log transformation of the response is appropriate. Therefore the *log(TARGET_AMT)* will be used as the response variable for the linear regression.

Log transformation of *TARGET_AMT* lowers the kurtosis significantly. However, the transformed response residuals are still not normally distributed as they fail the Shapiro-Wilk test (see below).

```
## [1] "Kurtosis:  5.389"
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm1.2$residuals
## W = 0.96297, p-value < 2.2e-16
```

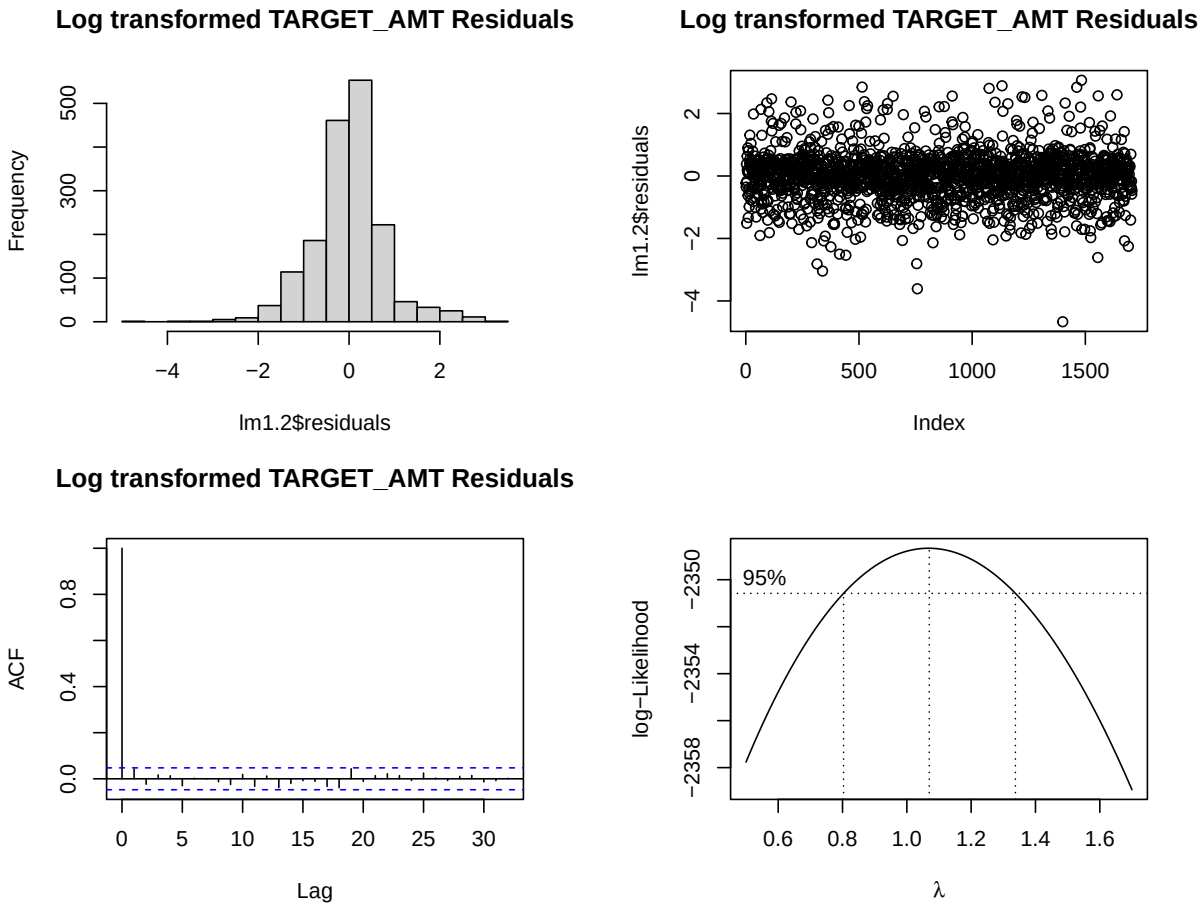
Figure below shows the effect of transformation on the distribution of the response *TARGET_AMT*.



4.3.2 Lag Autocorrelation, Breusch-Pagan heteroscedasticity tests

The residuals of the transformed response was investigated for autocorrelation, further Box-Cox transformation, and heteroscedasticity. As shown in the figurev below, the residuals are distributed in a near normal distribution. The lag plot does not show any autocorrelation in the residuals.

The Box-Cox transformation shows that the maximum *log-likelihood* occurs at $\lambda = 1$. Therefore, further transformation of the *TARGET_AMT* response is not necessary.



Below is the result of Breusch-Pagan heteroscedasticity test.

```
## [1] "Breusch-Pagan Heteroscedasticity Test for log(TARGET_AMT) Model:"
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm1.2
## BP = 43.424, df = 36, p-value = 0.1844
```

When the variance at each response is constant, the residuals are of homoscedastic (constant variance) nature. This is derived from the auxiliary regression (of the error ϵ)

$$\epsilon_i = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_p x_p$$

If the variance is constant (homescedastic) then only δ_0 is non-zero but $\delta_1, \delta_2, ...\delta_p$ are all zero. This is the null hypothesis (H_0 in the Breusch-Pagan heteroscedasticity test. Heteroscedasticity carries the risk of overestimation by the predictors. Since the *pvalue* of the Breusch-Pagan heteroscedasticity test after *log* transformation of the response *TARGET_AMT* is more than 0.05, we do not reject the null hypothesis. In other words the residuals are homescedastic.

This is an improvement, since in the residuals in the non-transformed regression model were not homesccedastic (null hypothesis was rejected) which can be seen below as the *p-value* is less than 0.05.

```
## [1] "Breusch-Pagan Heteroscedasticity Test for TARGET_AMT Model:"
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm1.1
## BP = 57.297, df = 36, p-value = 0.01347
```

4.4 Binary Logistic Response "TARGET_FLAG"

The response variable TARGET_FLAG is a factor variable and logistic regression model will be used to predict the log odd ratio for the response. Therefore. no transformation of the response is necessary.

4.5 Predictor Transformation

Predictors may need to be transformed depending on the behaviors of the predictors. For example, some correlated predictors may need to be excluded and some square *log* term of a predictor may need to be included. In logistic regression, if the predictor is normally distributed, then a square term may be added as a predictor. Similarly, a *log* term may be added when a predictor is highly skewed.

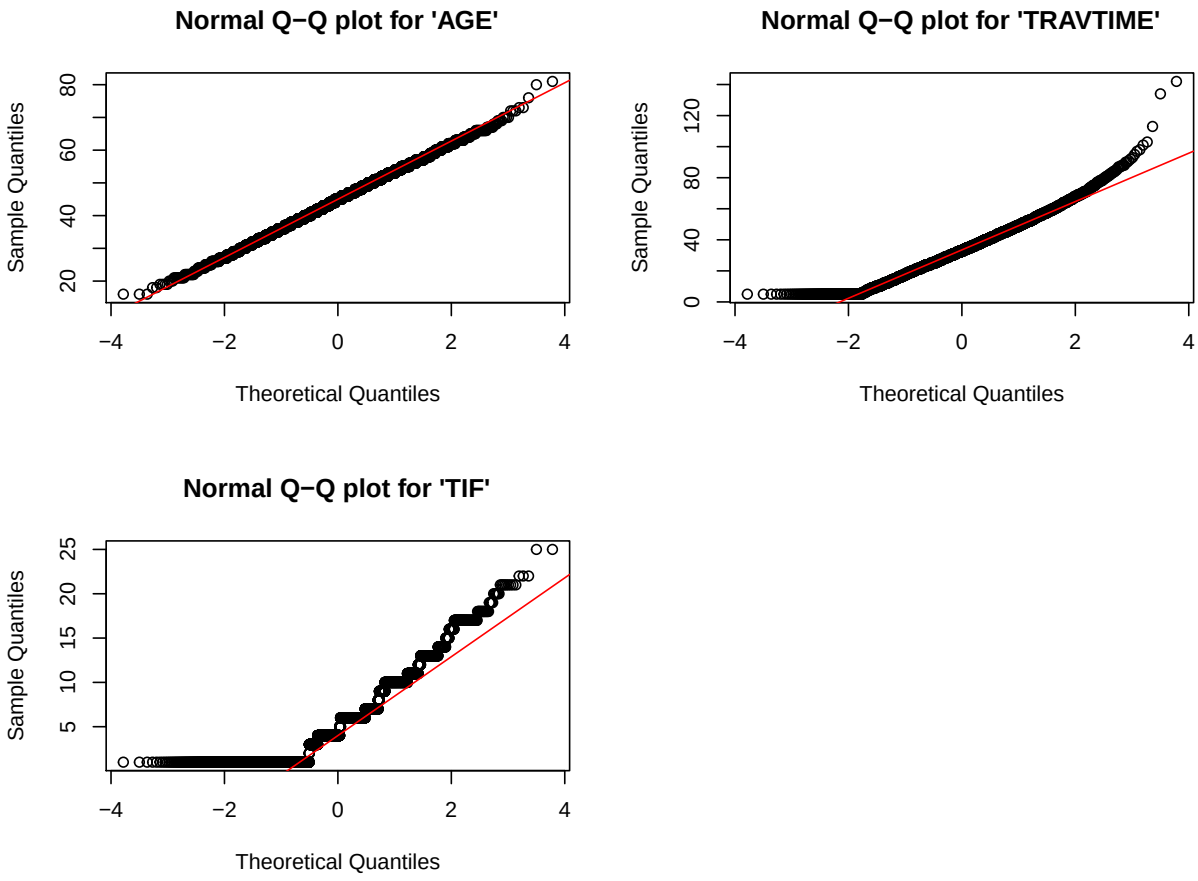
4.5.1 Normally distributed Predictors (square term for AGE)

Normally distributed predictors will have skewness near zero and kurtosis value near 3. From the variable summary information table in Table 2 we see that the predictors *AGE, TRAVTIME,* and *TIF* have skewness < 1 and the corresponding kurtosis value with 1 unit of the value 3. Therefore, these three predictors are very close to being normally distributed.

Results of the Shapiro-Wilk normality test for this predictor is shown below.

```
##
##  Shapiro-Wilk normality test
##
## data:  age
## W = 0.99837, p-value = 4.719e-05
```

Q-Q plot of these predictors are shown below. Although the normality test fails, the figure below indicates the predictor *"AGE"* may be normally distributed. Therefore a square term for *"AGE"* will be included as a predictor.



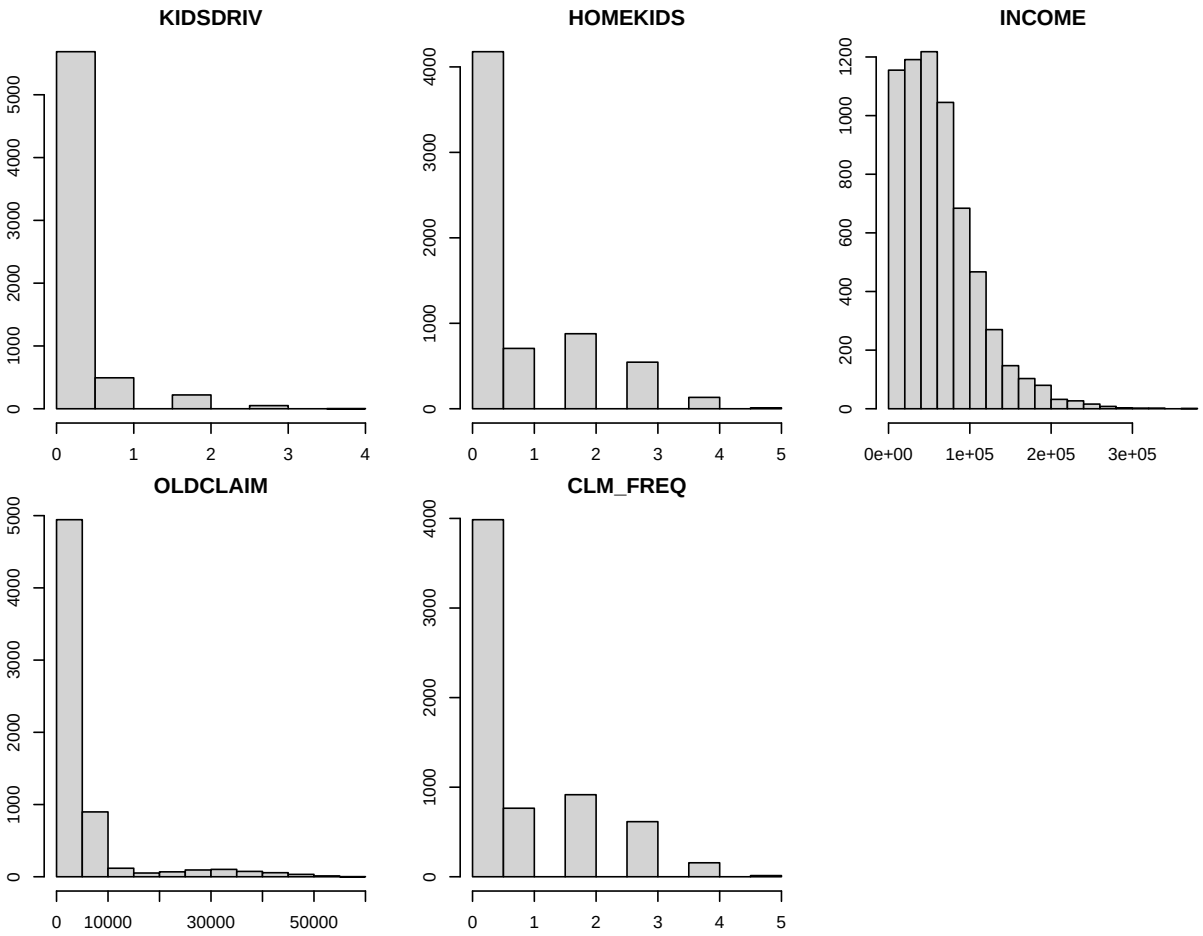
4.5.2 Highly skewed Predictors (adding square term predictors)

Skewness and kurtosis values for the variables are recalculated and shown in the table below. The predictors *KIDSDRIV, HOMEKIDS, INCOME, OLDCLAIM,* and *CLM_FREQ* have absolute skewness value > 1. Normally distributed predictors will have skewness near zero and kurtosis value near 3.

Table 5: Skewness and Kurtosis

	skewness	kurtosis			skewness	kurtosis
TARGET_AMT	8.34	100.21		TRAVTIME	0.46	3.71
KIDSDRIV	3.36	14.69		BLUEBOOK	0.82	3.93
AGE	-0.03	2.93		TIF	0.89	3.45
HOMEKIDS	1.33	3.62		OLDCLAIM	3.11	12.71
YOJ	-1.2	4.22		CLM_FREQ	1.22	3.28
INCOME	1.2	5.22		MVR_PTS	1.36	4.42
HOME_VAL	0.5	3.03		CAR_AGE	0.29	2.24

Distributions of these predictors are plotted in the figure below. It is clearly seen that the distributions are highly skewed. Therefore, *log* terms of these variables will also be included as predictors.



4.6 Adding Transformed Predictors

As discussed above, square terms and log terms of some of the predictors were added to the set of predictors. Multicollinearity was tested after adding each additional predictors and the predictor was excluded if it introduced multicollinearity.

4.6.1 Square term Predictor and Multicollinearity test

As discussed above, square term of the predictor AGE was added to the set of predictors and a linear regression model was built. Multicillinearity was checked by evaluating variance inflation factors (VIF) shown below.

```
##             GVIF Df GVIF^(1/(2*Df))
## KIDSDRIV     1.5  1             1.2
## AGE         59.9  1             7.7
## HOMEKIDS     2.2  1             1.5
## YOJ          1.8  1             1.3
## INCOME       3.4  1             1.8
## PARENT1      2.2  1             1.5
## HOME_VAL     2.2  1             1.5
## MSTATUS      2.4  1             1.5
## SEX          3.8  1             1.9
## EDUCATION   11.0  3             1.5
## JOB         32.6  8             1.2
## TRAVTIME     1.0  1             1.0
## CAR_USE      2.3  1             1.5
## BLUEBOOK     2.4  1             1.5
## TIF          1.0  1             1.0
## CAR_TYPE     7.3  5             1.2
## RED_CAR      1.8  1             1.3
## OLDCLAIM     2.0  1             1.4
## CLM_FREQ     1.4  1             1.2
## REVOKED      1.6  1             1.3
## MVR_PTS      1.2  1             1.1
## CAR_AGE      2.1  1             1.5
## URBANICITY   1.1  1             1.0
## I(AGE^2)    58.7  1             7.7
```

Both AGE and AGE^2 have VIF $>>$ 5 (arbitrary) indicating multicollinearity. Therefore the squared term predictor AGE^2 will not be used for building regression models.

4.6.2 log() term Predictors and Multicollinearity Test

The *log* terms of the predictors *KIDSDRIV, HOMEKIDS, INCOME, OLDCLAIM,* and *CLM_FREQ* were added to the set of predictors and VIF values for all the predictors were calculated. Note that a small value (0.1) was added to each to avoid *log(0)* term.

```
##                        GVIF Df GVIF^(1/(2*Df))
## KIDSDRIV               12.9  1             3.6
```

```
## AGE                      1.8  1            1.3
## HOMEKIDS                10.3  1            3.2
## YOJ                      3.9  1            2.0
## INCOME                   3.6  1            1.9
## PARENT1                  2.9  1            1.7
## HOME_VAL                 2.2  1            1.5
## MSTATUS                  2.6  1            1.6
## SEX                      3.8  1            1.9
## EDUCATION               11.2  3            1.5
## JOB                     48.2  8            1.3
## TRAVTIME                 1.0  1            1.0
## CAR_USE                  2.3  1            1.5
## BLUEBOOK                 2.4  1            1.5
## TIF                      1.0  1            1.0
## CAR_TYPE                 7.2  5            1.2
## RED_CAR                  1.8  1            1.3
## OLDCLAIM                 5.0  1            2.2
## CLM_FREQ                41.4  1            6.4
## REVOKED                  1.7  1            1.3
## MVR_PTS                  1.2  1            1.1
## CAR_AGE                  2.1  1            1.5
## URBANICITY               1.1  1            1.0
## I(log(KIDSDRIV + 0.1))  13.5  1            3.7
## I(log(HOMEKIDS + 0.1))  13.9  1            3.7
## I(log(INCOME + 0.1))     6.8  1            2.6
## log(OLDCLAIM + 0.1)    125.6  1           11.2
## I(log(CLM_FREQ + 0.1)) 245.4  1           15.7
```

As seen above, out of *KIDSDRIV, HOMEKIDS, INCOME, OLDCLAIM,* and *CLM_FREQ* only the *INCOME* predictor did not show multicollinearity *(VIF < 5)*. Therefore only *log(INCOME)* will be added as a new predictor.

5. Building Models

5.1 Linear Regression Models

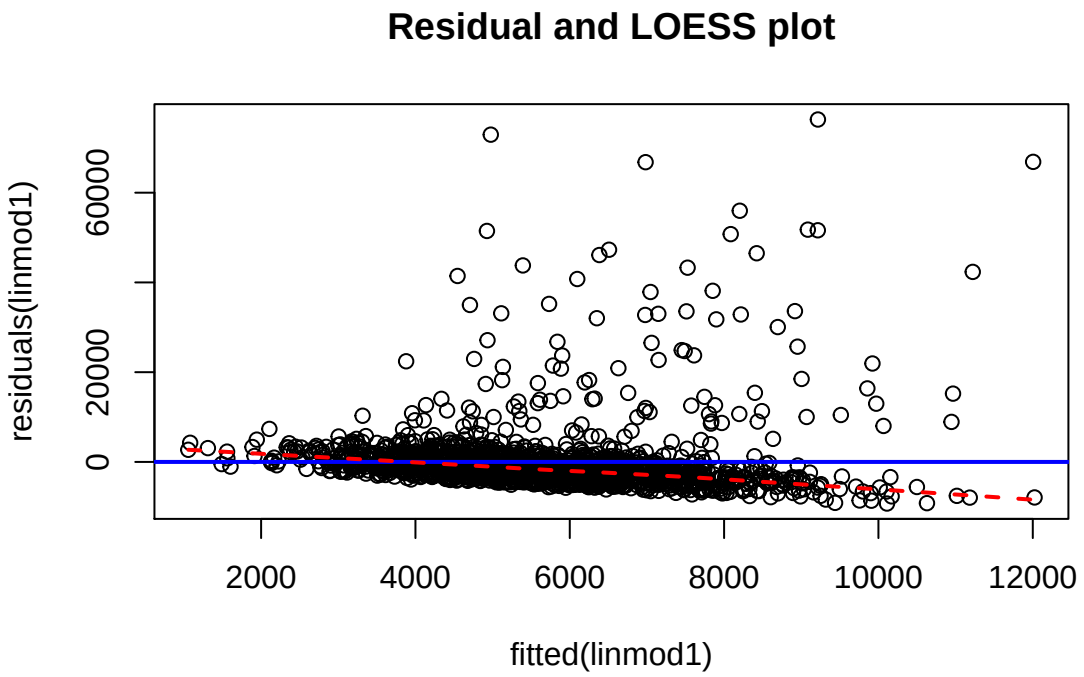
Several models were buiult by selecting sets of predictors. The predictor selection steps and model building processes for each are described below.

5.1.1 Linear Model 1

For building the first model, all the predictors including *log(INCOME)* were considered. Model summary and residual plot are shown below.

```
##
## Call:
## lm(formula = TARGET_AMT ~ . + I(log(INCOME + 0.1)), data = ins3.lin)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9272  -3164  -1440    569  76308
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.224e+03  2.277e+03   1.416   0.1570
## KIDSDRIV                     -1.798e+02  3.562e+02  -0.505   0.6137
## AGE                           3.735e+00  2.401e+01   0.156   0.8764
## HOMEKIDS                      3.075e+02  2.368e+02   1.298   0.1943
## YOJ                          -6.801e+00  8.038e+01  -0.085   0.9326
## INCOME                       -1.682e-02  8.095e-03  -2.078   0.0379 *
## PARENT1Yes                   -1.440e+02  6.453e+02  -0.223   0.8234
## HOME_VAL                      2.278e-03  2.270e-03   1.004   0.3157
## MSTATUSYes                   -1.389e+03  5.671e+02  -2.449   0.0144 *
## SEXM                          1.786e+03  7.159e+02   2.495   0.0127 *
## EDUCATIONHigh School         -6.934e+02  5.619e+02  -1.234   0.2173
## EDUCATIONMasters              6.712e+02  1.048e+03   0.640   0.5220
## EDUCATIONPhD                  3.029e+03  1.280e+03   2.367   0.0180 *
## JOBBlue Collar                5.178e+02  1.303e+03   0.397   0.6911
## JOBClerical                  -6.691e+02  1.364e+03  -0.490   0.6239
## JOBDoctor                    -3.431e+03  1.865e+03  -1.840   0.0659 .
## JOBHome Maker                -4.495e+02  1.505e+03  -0.299   0.7653
## JOBLawyer                    -2.720e+02  1.143e+03  -0.238   0.8119
## JOBManager                   -1.410e+03  1.220e+03  -1.156   0.2479
## JOBProfessional               1.121e+03  1.281e+03   0.875   0.3817
## JOBStudent                   -3.972e+02  1.527e+03  -0.260   0.7949
## TRAVTIME                      4.401e+00  1.232e+01   0.357   0.7210
## CAR_USEPrivate               -5.467e+01  5.566e+02  -0.098   0.9218
## BLUEBOOK                      1.478e-01  3.372e-02   4.384 1.24e-05 ***
## TIF                          -6.513e+00  4.687e+01  -0.139   0.8895
## CAR_TYPEPanel Truck          -1.195e+02  1.047e+03  -0.114   0.9091
## CAR_TYPEPickup                3.823e+02  6.580e+02   0.581   0.5614
## CAR_TYPESports Car            1.894e+03  8.244e+02   2.298   0.0217 *
## CAR_TYPESUV                   1.592e+03  7.348e+02   2.166   0.0304 *
## CAR_TYPEVan                  -1.395e+02  8.555e+02  -0.163   0.8705
## RED_CARyes                   -3.240e+02  5.479e+02  -0.591   0.5544
## OLDCLAIM                      5.067e-02  2.528e-02   2.004   0.0452 *
## CLM_FREQ                     -2.058e+02  1.748e+02  -1.177   0.2393
## REVOKEDYes                   -1.274e+03  5.846e+02  -2.180   0.0294 *
## MVR_PTS                       9.313e+01  7.555e+01   1.233   0.2179
## CAR_AGE                      -1.008e+02  4.856e+01  -2.076   0.0381 *
## URBANICITYHighly Urban/ Urban  7.940e+01  8.176e+02   0.097   0.9226
## I(log(INCOME + 0.1))           5.156e+01  1.120e+02   0.460   0.6454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 7583 on 1668 degrees of freedom
## Multiple R-squared:  0.04195,    Adjusted R-squared:  0.0207
## F-statistic: 1.974 on 37 and 1668 DF,  p-value: 0.0004664
```



There are 8 predictors whose coefficients are significant in this model. They are *INCOME, MSTATUS, SEX, EDUCATION (only PhD), BLUEBOOK, CAR_TYPES (SUV and Sports Car only), OLDCLAIM,* and *CAR_AGE.*

Model has a very low R^2 and *Adjusted-*R^2 values indicating a poor fit with the data. This can also be seen from the residual plot above.
The residuals are not normally distributed and there are several points with high leverages.

```
library(lmtest)
print("Homoscedasticity Test")
```

```
## [1] "Homoscedasticity Test"
```

```
bptest(linmod1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  linmod1
## BP = 57.443, df = 37, p-value = 0.01717
```

```
print("Autocorrelation Test")
```

```
## [1] "Autocorrelation Test"
```

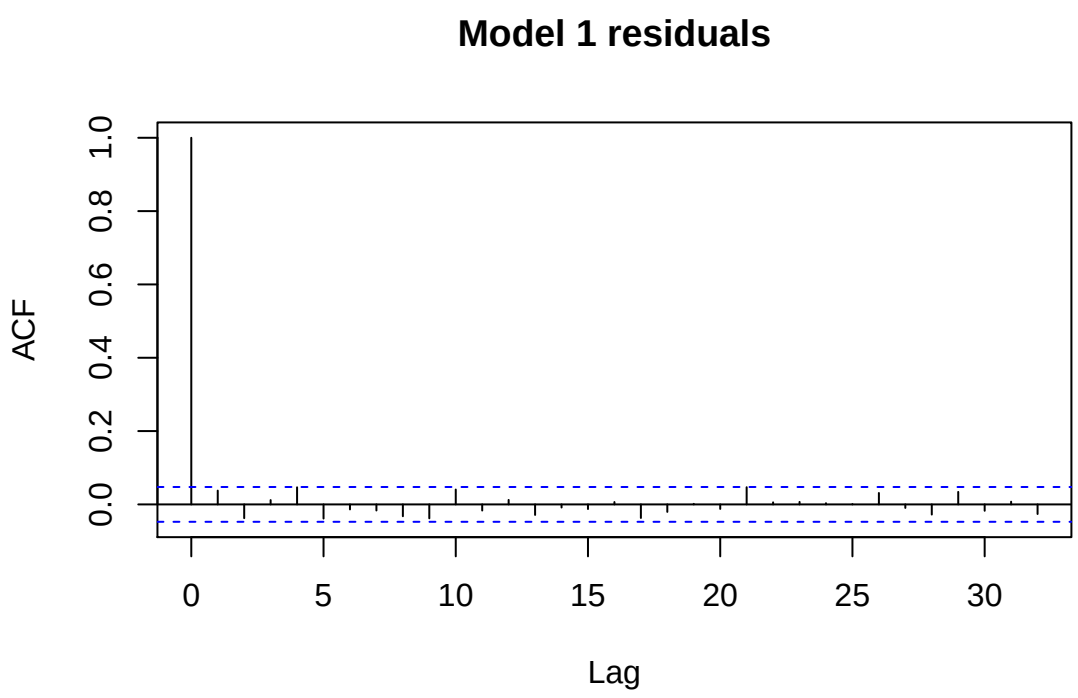
```
dwtest(linmod1)
```

```
##
##  Durbin-Watson test
##
## data:  linmod1
## DW = 1.9244, p-value = 0.05944
## alternative hypothesis: true autocorrelation is greater than 0
```

```
print("Autocorrelation Plot")
```

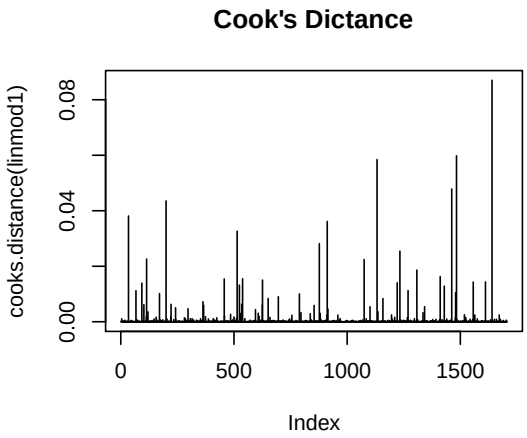
```
## [1] "Autocorrelation Plot"
```

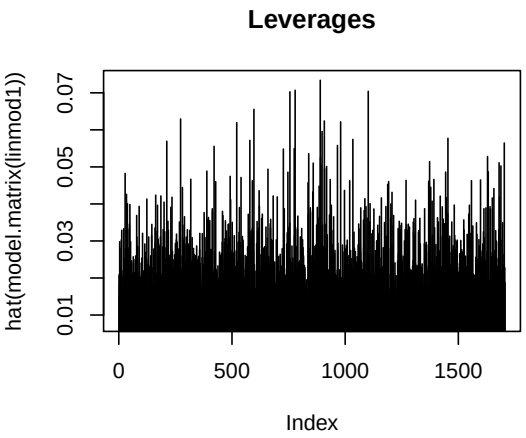
```
acf(linmod1$residuals, main='Model 1 residuals')
```

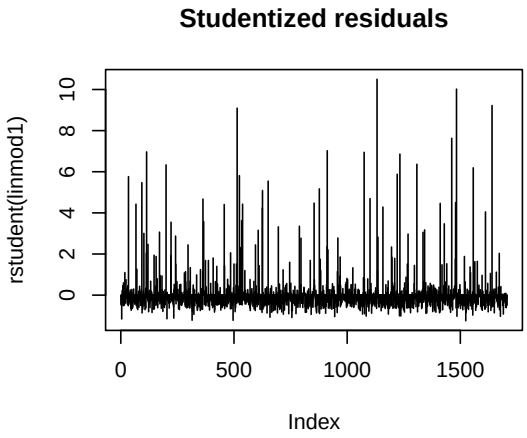


Heteroscedasticity may also be an issue with the residuals of the plot. As seen from the result of the Breusch-Pagan test, the residuals are not homoscedastic (H_0 is rejected). Also the Durbin-Watson test indicates the residulas may also be autocorrelated although autocorrelation is not very clear from the acf plot shown above (no lags are significant other than lag=0).

Reason for the poor fitting of the model is due to several influence points. Plots of Cook's distance, Leverages, and normalized (t-distribution) residuals indicate there are several influence points as shown in the plot below.

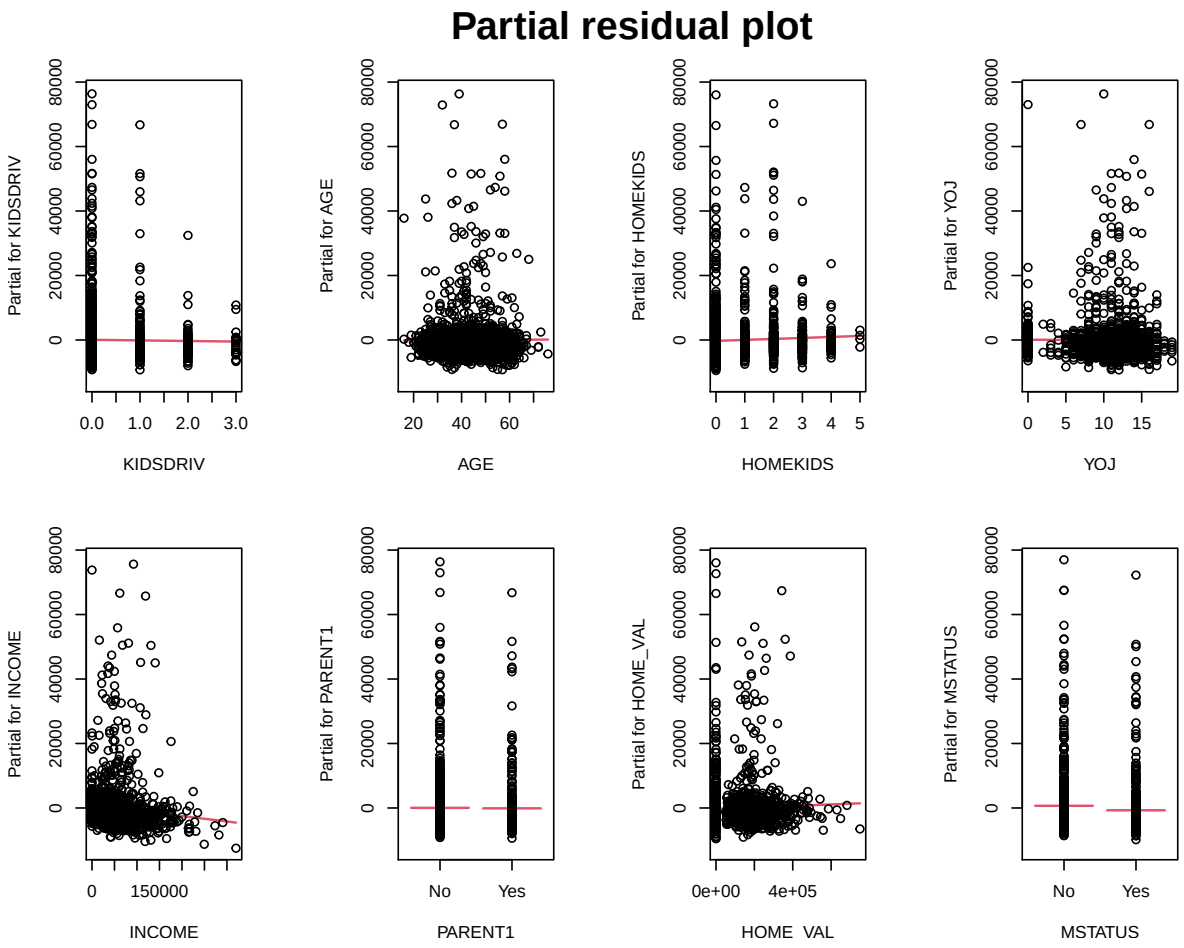






Based on the hat matrix (leverage) value of $ > (3p+1)/n$ for the influence points, there are 42 outliers in the data. Partial residual plots indicate that the residuals with high leverage points are seen for many predictors (eight of them shown in th eplot below).

```
## [1] "Number of outliers in Model1"
## [1] 42
```



5.1.1 Linear Model 2

Stepwise predictor selection method was used to select predictors for the second model. As shown below backward and both ways predictor selection methods select some of the same predictors.

Backward Stepwise Predictor Selection

```
##
## Call:
## lm(formula = TARGET_AMT ~ INCOME + MSTATUS + EDUCATION + JOB +
##     BLUEBOOK + OLDCLAIM + REVOKED + CAR_AGE, data = ins3.lin)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8981  -3161  -1494    508  76579
##
```

```
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6.054e+03  1.457e+03   4.156 3.40e-05 ***
## INCOME               -1.255e-02  7.092e-03  -1.770   0.0770 .
## MSTATUSYes           -8.866e+02  3.698e+02  -2.398   0.0166 *
## EDUCATIONHigh School -6.831e+02  5.580e+02  -1.224   0.2211
## EDUCATIONMasters      7.429e+02  1.039e+03   0.715   0.4746
## EDUCATIONPhD          3.053e+03  1.270e+03   2.404   0.0163 *
## JOBBlue Collar        5.915e+02  1.271e+03   0.465   0.6417
## JOBClerical          -6.382e+02  1.331e+03  -0.480   0.6316
## JOBDoctor            -3.535e+03  1.789e+03  -1.976   0.0483 *
## JOBHome Maker        -9.180e+02  1.360e+03  -0.675   0.4998
## JOBLawyer            -3.066e+02  1.047e+03  -0.293   0.7698
## JOBManager           -1.440e+03  1.179e+03  -1.221   0.2221
## JOBProfessional       1.088e+03  1.255e+03   0.867   0.3861
## JOBStudent           -7.334e+02  1.425e+03  -0.515   0.6068
## BLUEBOOK              1.163e-01  2.472e-02   4.706 2.73e-06 ***
## OLDCLAIM              4.098e-02  2.185e-02   1.876   0.0609 .
## REVOKEDYes           -1.151e+03  5.537e+02  -2.079   0.0378 *
## CAR_AGE              -1.059e+02  4.814e+01  -2.199   0.0280 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7569 on 1688 degrees of freedom
## Multiple R-squared:  0.03407,    Adjusted R-squared:  0.02434
## F-statistic: 3.503 on 17 and 1688 DF,  p-value: 1.671e-06
```

Bothways Stepwise Predictor Selection

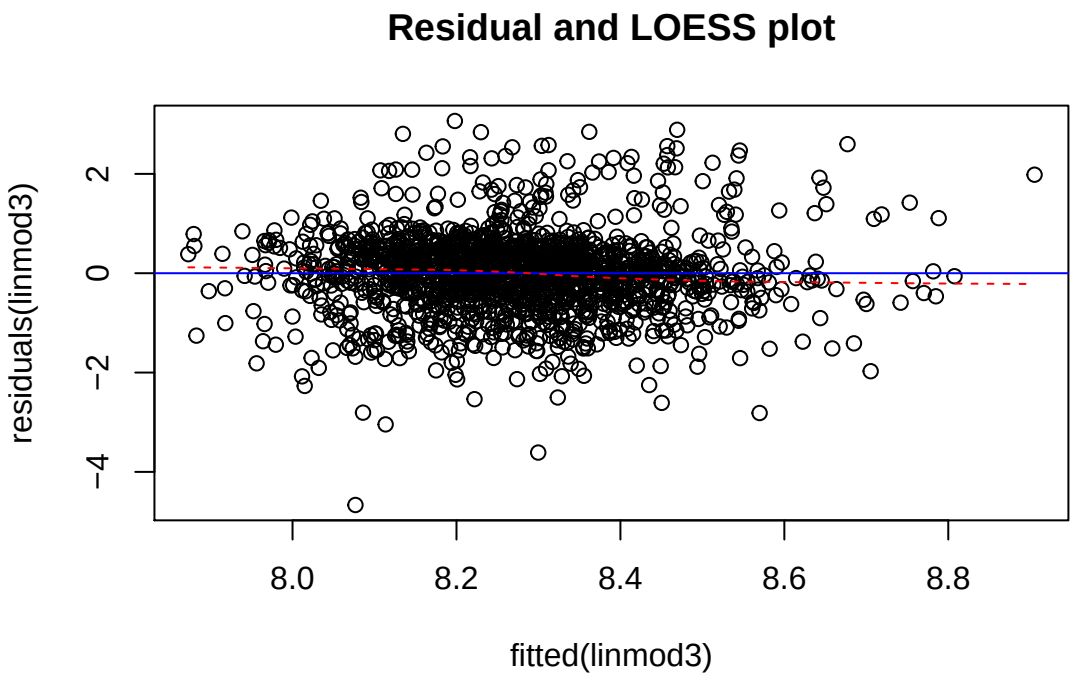
```
##
## Call:
## lm(formula = TARGET_AMT ~ BLUEBOOK + MSTATUS + CAR_AGE + MVR_PTS,
##     data = ins3.lin)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8672  -3125  -1498    352  78219
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4464.81429  496.98451   8.984  < 2e-16 ***
## BLUEBOOK       0.12211    0.02236   5.461 5.43e-08 ***
## MSTATUSYes  -846.06955  368.27165  -2.297   0.0217 *
## CAR_AGE      -55.19723   34.01948  -1.623   0.1049
## MVR_PTS      107.48158   70.74693   1.519   0.1289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7589 on 1701 degrees of freedom
## Multiple R-squared:  0.02128,    Adjusted R-squared:  0.01898
## F-statistic: 9.245 on 4 and 1701 DF,  p-value: 2.173e-07
```

The R^2 values for both models (backward and bothways predictorselecction) does not indicate a good fot for the data. The $F - statistic$ for the both ways selection is slightly better ans also this model explains the variance to a similar value using only 4 predictors (as opposes to 8 predictors for the backward selection model).

Additionally, the large negative value for REVOKED indicates lower amount of car cost if the license was revoked, which is counter intuitive. Thus we select the predictors in the both ways selection for model 2.

5.1.3 Linear Model 3

A third model was also constructed based on the Box-Cox transformation of the response variable TARGET-AMT. Below is a plot of the residuals.



Residual plot shows a better fit compared to the other 2 linear models. Overall fit (R^2) however is still very poor.

5.2 Logistic Regression Models

5.2.1 Logistic Model 1

Model 1 was built using all the predictors (excluding TARGET_AMT). Relative importance of the predictors for this model is shown below. The predictors URBANICITY, CAR_USE, REVOKED are top 3 important predictors. These predictors are intuitively important for the car to have an accident.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = binomial(link = "logit"),
##     data = ins4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5827  -0.7118  -0.4006   0.6186   3.1634
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -3.178e+00  3.765e-01  -8.440  < 2e-16 ***
## KIDSDRIV                       3.360e-01  6.906e-02   4.866 1.14e-06 ***
## AGE                           -3.658e-03  4.530e-03  -0.808 0.419286
## HOMEKIDS                       3.475e-02  4.171e-02   0.833 0.404814
## YOJ                           -1.126e-02  9.581e-03  -1.175 0.239960
## INCOME                        -3.004e-06  1.257e-06  -2.389 0.016889 *
## PARENT1Yes                     4.406e-01  1.224e-01   3.600 0.000318 ***
## HOME_VAL                      -1.296e-06  3.899e-07  -3.323 0.000889 ***
## MSTATUSYes                    -4.395e-01  9.664e-02  -4.548 5.41e-06 ***
## SEXM                           1.880e-01  1.241e-01   1.515 0.129742
## EDUCATIONHigh School           3.705e-01  9.945e-02   3.725 0.000195 ***
## EDUCATIONMasters               2.797e-02  1.605e-01   0.174 0.861671
## EDUCATIONPhD                   2.609e-01  2.051e-01   1.272 0.203313
## JOBBlue Collar                 3.046e-01  2.112e-01   1.442 0.149172
## JOBClerical                    5.121e-01  2.237e-01   2.290 0.022049 *
## JOBDoctor                     -1.935e-01  2.880e-01  -0.672 0.501707
## JOBHome Maker                  2.243e-01  2.409e-01   0.931 0.351796
## JOBLawyer                      2.880e-01  1.900e-01   1.516 0.129576
## JOBManager                    -5.767e-01  1.973e-01  -2.923 0.003468 **
## JOBProfessional                1.987e-01  2.026e-01   0.981 0.326702
## JOBStudent                     1.695e-01  2.469e-01   0.687 0.492323
## TRAVTIME                       1.554e-02  2.117e-03   7.339 2.16e-13 ***
## CAR_USEPrivate                -8.279e-01  9.883e-02  -8.377  < 2e-16 ***
## BLUEBOOK                      -2.109e-05  5.883e-06  -3.585 0.000337 ***
## TIF                           -5.308e-02  8.237e-03  -6.444 1.16e-10 ***
## CAR_TYPEPanel Truck            6.075e-01  1.791e-01   3.392 0.000694 ***
## CAR_TYPEPickup                 5.209e-01  1.126e-01   4.625 3.74e-06 ***
## CAR_TYPESports Car             1.128e+00  1.449e-01   7.786 6.89e-15 ***
## CAR_TYPESUV                    8.515e-01  1.240e-01   6.868 6.49e-12 ***
## CAR_TYPEVan                    6.310e-01  1.413e-01   4.465 7.99e-06 ***
## RED_CARyes                    -1.140e-01  9.666e-02  -1.179 0.238241
## OLDCLAIM                      -1.180e-05  4.375e-06  -2.698 0.006980 **
## CLM_FREQ                       1.943e-01  3.182e-02   6.105 1.03e-09 ***
## REVOKEDYes                     8.627e-01  1.035e-01   8.339  < 2e-16 ***
## MVR_PTS                        1.148e-01  1.527e-02   7.520 5.48e-14 ***
## CAR_AGE                       -7.181e-03  8.417e-03  -0.853 0.393568
## URBANICITYHighly Urban/ Urban  2.316e+00  1.241e-01  18.666  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7453  on 6450  degrees of freedom
## Residual deviance: 5768  on 6414  degrees of freedom
## AIC: 5842
##
## Number of Fisher Scoring iterations: 5
```

```
## [1] "Relative Variable Importance"
```

```
##                               Overall
## URBANICITYHighly Urban/ Urban    0.14
## CAR_USEPrivate                   0.06
## REVOKEDYes                       0.06
## CAR_TYPESports Car               0.06
## MVR_PTS                          0.05
## TRAVTIME                         0.05
## CAR_TYPESUV                      0.05
## TIF                              0.05
## CLM_FREQ                         0.04
## KIDSDRIV                         0.04
## CAR_TYPEPickup                   0.03
## MSTATUSYes                       0.03
## CAR_TYPEVan                      0.03
## EDUCATIONHigh School             0.03
## PARENT1Yes                       0.03
## BLUEBOOK                         0.03
## CAR_TYPEPanel Truck              0.02
## HOME_VAL                         0.02
## JOBManager                       0.02
## OLDCLAIM                         0.02
## INCOME                           0.02
## JOBClerical                      0.02
## JOBLawyer                        0.01
## SEXM                             0.01
## JOBBlue Collar                   0.01
## EDUCATIONPhD                     0.01
## RED_CARyes                       0.01
## YOJ                              0.01
## JOBProfessional                  0.01
## JOBHome Maker                    0.01
## CAR_AGE                          0.01
## HOMEKIDS                         0.01
## AGE                              0.01
## JOBStudent                       0.01
## JOBDoctor                        0.00
## EDUCATIONMasters                 0.00
```

5.2.2 Logistic Model 2

Some of the skewed predictors and their log transformed values can be added as predictors as discussed earlier. Particularly, although the variable *INCOME*, and *HOME_VAL* are are two highly skewed predictors. Log terms of these two were added to build a second logistic model and the predictors *INCOME*, and *HOME_VAL* were removed from the model to avoid multicollinearity. As seen from the summary plot, boith the log predictors are significant.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . + I(log(INCOME + 0.01)) + I(log(HOME_VAL +
##     0.01)), family = binomial(link = "logit"), data = ins4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6251  -0.7063  -0.4017   0.6000   3.1658
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -2.791e+00  3.869e-01  -7.213 5.46e-13 ***
## KIDSDRIV                      3.476e-01  6.945e-02   5.005 5.58e-07 ***
## AGE                          -7.389e-03  4.623e-03  -1.598 0.109958
## HOMEKIDS                      3.939e-03  4.271e-02   0.092 0.926531
## YOJ                           2.579e-02  1.358e-02   1.898 0.057645 .
## INCOME                       -3.299e-06  1.542e-06  -2.139 0.032455 *
## PARENT1Yes                    4.296e-01  1.230e-01   3.493 0.000478 ***
## HOME_VAL                      9.744e-08  7.676e-07   0.127 0.898994
## MSTATUSYes                   -4.164e-01  1.006e-01  -4.137 3.51e-05 ***
## SEXM                          1.896e-01  1.242e-01   1.527 0.126748
## EDUCATIONHigh School          3.765e-01  9.979e-02   3.773 0.000161 ***
## EDUCATIONMasters              3.691e-03  1.609e-01   0.023 0.981700
## EDUCATIONPhD                  2.055e-01  2.061e-01   0.997 0.318665
## JOBBlue Collar                3.025e-01  2.114e-01   1.431 0.152479
## JOBClerical                   5.163e-01  2.240e-01   2.305 0.021189 *
## JOBDoctor                    -2.042e-01  2.873e-01  -0.711 0.477098
## JOBHome Maker                -1.887e-02  2.529e-01  -0.075 0.940544
## JOBLawyer                     2.969e-01  1.899e-01   1.564 0.117858
## JOBManager                   -5.748e-01  1.971e-01  -2.917 0.003537 **
## JOBProfessional               1.952e-01  2.026e-01   0.964 0.335285
## JOBStudent                   -2.351e-01  2.651e-01  -0.887 0.375163
## TRAVTIME                      1.554e-02  2.126e-03   7.310 2.67e-13 ***
## CAR_USEPrivate               -8.351e-01  9.923e-02  -8.416  < 2e-16 ***
## BLUEBOOK                     -2.062e-05  5.889e-06  -3.502 0.000461 ***
## TIF                          -5.227e-02  8.253e-03  -6.334 2.39e-10 ***
## CAR_TYPEPanel Truck           6.030e-01  1.794e-01   3.361 0.000776 ***
## CAR_TYPEPickup                5.393e-01  1.129e-01   4.776 1.79e-06 ***
## CAR_TYPESports Car            1.134e+00  1.453e-01   7.808 5.80e-15 ***
## CAR_TYPESUV                   8.594e-01  1.242e-01   6.919 4.55e-12 ***
## CAR_TYPEVan                   6.351e-01  1.415e-01   4.489 7.17e-06 ***
## RED_CARyes                   -1.186e-01  9.674e-02  -1.226 0.220039
## OLDCLAIM                     -1.203e-05  4.389e-06  -2.741 0.006116 **
## CLM_FREQ                      1.967e-01  3.190e-02   6.168 6.94e-10 ***
## REVOKEDYes                    8.602e-01  1.037e-01   8.296  < 2e-16 ***
## MVR_PTS                       1.132e-01  1.534e-02   7.379 1.60e-13 ***
## CAR_AGE                      -6.543e-03  8.426e-03  -0.777 0.437443
## URBANICITYHighly Urban/ Urban  2.336e+00  1.247e-01  18.732  < 2e-16 ***
```

```
## I(log(INCOME + 0.01))         -6.220e-02  1.631e-02  -3.814 0.000137 ***
## I(log(HOME_VAL + 0.01))       -2.442e-02  1.134e-02  -2.153 0.031350 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7453.0  on 6450  degrees of freedom
## Residual deviance: 5747.8  on 6412  degrees of freedom
## AIC: 5825.8
##
## Number of Fisher Scoring iterations: 5
```

5.2.3 Logistic Model 3

A third model was built based on the predictor selection by stepwise method. The regression formula obtained by this method is

TARGET_FLAG ~ URBANICITY+ + JOB + MVR_PTS + MSTATUS + CAR_TYPE + REVOKED + CAR_USE + TRAVTIME + KIDSDRIV + TIF + CLM_FREQ + PARENT1 + BLUEBOOK + EDUCATION + I(log(INCOME + 1)) + HOME_VAL + OLDCLAIM + I(log(AGE)) + AGE + YOJ + RED_CAR

Some of the othetr model parameters are

`_Null deviance: 7453.0  on 6450  degrees of freedom_`

Residual deviance: 5701.9 on 6416 degrees of freedom AIC: 5771.9

Number of Fisher Scoring iterations: 5

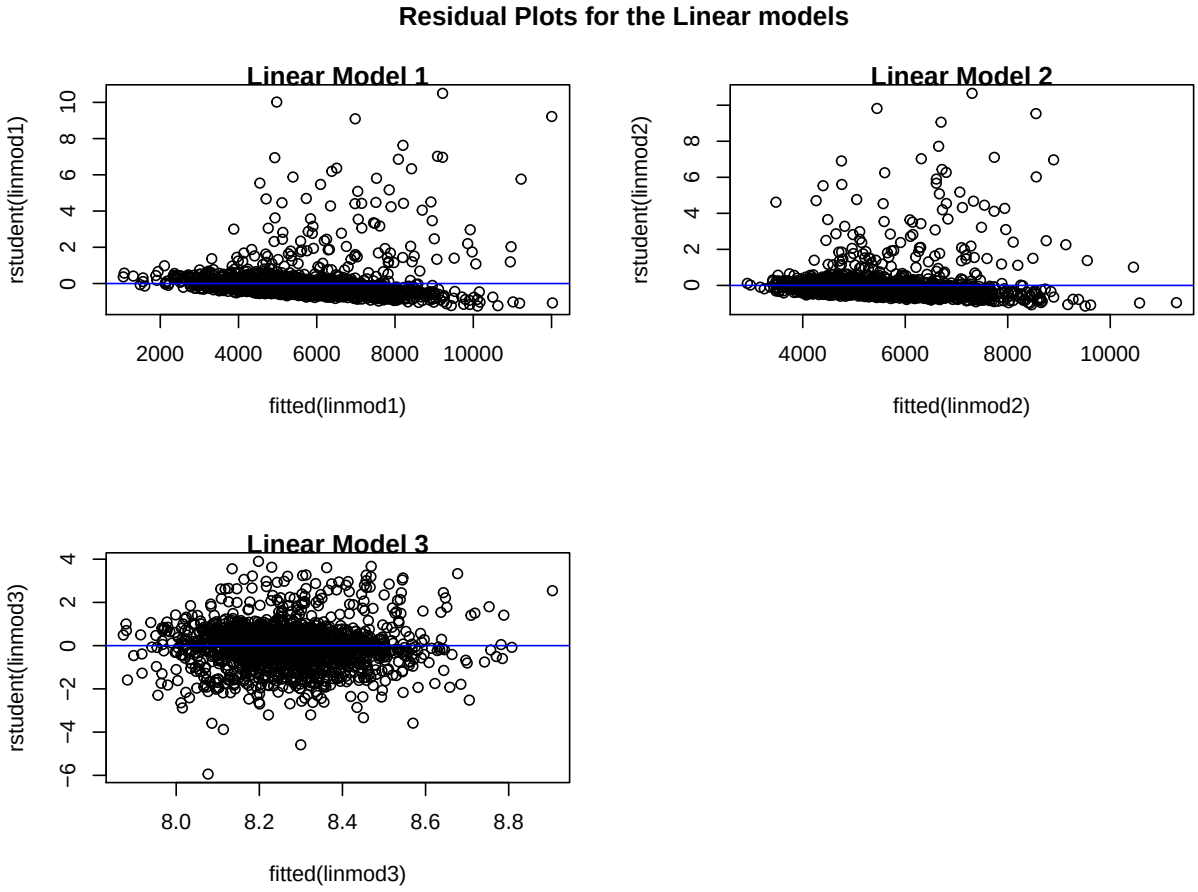
6. Model Selection

Linear models and logistic models were evaluated separately based on criteria described below.

6.1 Linear Model Evaluation

6.1.1 Residual plots

Residual plots of the three models were evaluated to investigate the goodness of fit of the data. The plots are shown below.



It is clear from the above plot that the residuals are not symmetrically distributed in models 1 and 2. They do appear to have a symmetric distribution in model 3. We look at the homoscedasticity of the models next.

6.1.2 Homoscedasticity

Breusch-Pagan homoscedasticity tests were performed on the three models. Results indicate that model 1 and model 2 have heteroscedastic residuals. Model 3 has homoscedastic residuals. Thus OLS can be applied to model 3.

##

```
##  Linear Model 1 Homoscedasticity Test:
##
##  studentized Breusch-Pagan test
##
## data:  linmod1
## BP = 57.443, df = 37, p-value = 0.01717
##
##
##  Linear Model 2 Homoscedasticity Test:
##
##  studentized Breusch-Pagan test
##
## data:  linmod2
## BP = 26.605, df = 4, p-value = 2.389e-05
##
##
##  Linear Model 3 Homoscedasticity Test:
##
##  studentized Breusch-Pagan test
##
## data:  linmod3
## BP = 43.424, df = 36, p-value = 0.1844
```

6.1.3 Influence Points

Total number of influence points for each of the models were calculated based on the criteria that outliers will have [*leverage values* $> 3p/n$] where $p = k+1$, k being the number of independent predictors (Pituch and Stevens (2016), page 108). Model with a least number of influence points is preferred.

6.1.4 Overall Comparison (Adjusted R-square, RSE, and F-statistic, etc)

Adjusted R-square, residual standard error, AIC, F-statistic, number of outliers, and homoscedasticity were evaluated for the best model selection. Results are shown in the table below.

```
##
##
## | Adj.R.square|    RSE|      AIC| F.statistic| Number.of.outliers| Homoscedasticity.Test|
## |------------:|------:|--------:|-----------:|------------------:|---------------------:|
## |        0.021| 7582.7| 35362.53|        1.97|                 60|             0.0171723|
## |        0.019| 7589.4| 35332.95|        9.24|                 13|             0.0000239|
## |        0.011|    0.8|  4112.19|        1.51|                 60|             0.1844407|
```

Based on lowest Adj. R^2, RSE, AIC and also being homoscedastic, Linear Model 3 seem to perform the best and will be selected for evaluation.

6.2 Logistic Model Evaluation

6.2.1 Deviance Chi-square Test

Deviances of the models have *chi-square* distributions. The models were compared based on their degrees of freedom and the deviance in the model. The *p-values* for the model care shown below.

```
##
##
## |H0.Model |H1.Model |  p.value|
## |:--------|:--------|--------:|
## |Model 1  |Model 2  | 4.12e-05|
## |Model 3  |Model 2  | 1.00e+00|
```

Based on the p-values, model 3 seems to be a better model.

6.2.2 Likelihood ratio test and R-square values

Likelihood ratio has a *Chi-square* distribution and can be used to test the probability of the *chi-square* value being larger than the critical value. Out of the three models, the first two models (logmod1, logmod2) have the same number of predictors and thus *log likelihood* ratio test cannot be used ($df = 0$). Log likelihood ratio, *R-square* values, and *pseudo-R-square* values were also calculated and are shown below.

```
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
```

```
##
##
## |Model            | R.square | McFadden.Pseudo.R.square |   Log.likelihood.DF    | p.value..Log.likel
## |:----------------|:--------:|:------------------------:|:----------------------:|:------------------
## |Logistic Model 1 |  0.336   |          0.226           | (Model 2/Model 1):  -2 |           4.12e-05
## |Logistic Model 2 |  0.339   |          0.229           | (Model 3/Model 1):  2  |           0.00e+00
## |Logistic Model 3 |  0.347   |          0.235           | (Model 3/Model 2):  4  |           0.00e+00
```

All three models have similar but low R-square, and pseudo R-square values. Log likelihood test indicate model 2 (model with the largest number of predictors) is better than model 1 and 3.

6.2.3 Receiver Operating Characteristic (ROC) Curves and AUC

ROC (receiver operating characteristic) curve is a good qualitative metric to evaluate goodness of fit of a model. The *AUC* (area under the curve) values of the ROC curve gives a quantitative assessment of how good the mode is. Figure elow shows the ROC curves for the 3 models. All the models appear to perform to the same extent in this data set.

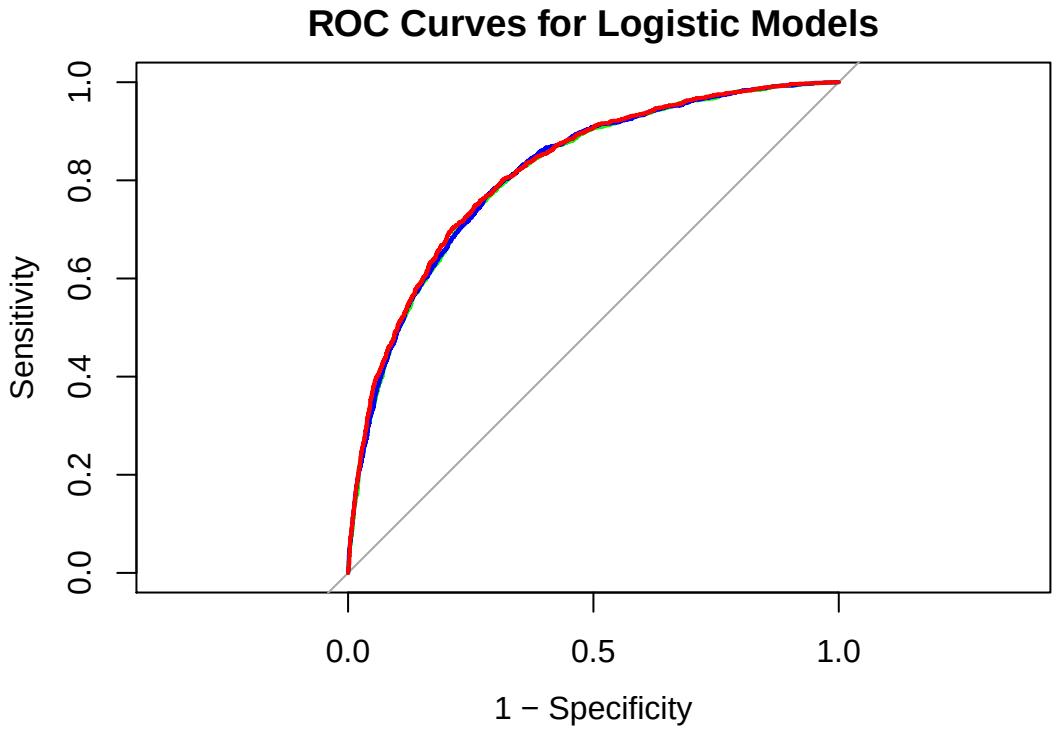


Table below shows the AUC of the ROC curves and their corresponding 95% confidence intervals. All three models have similar AUC values.

```
##
##
## |Model   | AUC.of.ROC | X95.CI..DeLong. |
## |:-------|:----------:|:---------------:|
## |Model 1 |   0.814    | 0.803  -  0.826 |
## |Model 2 |   0.816    | 0.804  -  0.827 |
## |Model 3 |   0.820    | 0.808  -  0.831 |
```

6.2.4 Confusion matrix

Confusion matrix for the three models were evaluated using a threshold of 0.5. The results are shown below.

```
## --Logistic Model 1--
##           Reference
## Prediction    0    1
##          0 4382  975
##          1  363  731
##
##
##  --Logistic Model 2--
##           Reference
## Prediction    0    1
##          0 4379  972
##          1  366  734
##
##
##  --Logistic Model 3--
##           Reference
## Prediction    0    1
##          0 4380  952
##          1  365  754
```

Model 3 performs the best with lowest number of false negative values. Numerical parameters for these predictions will be evaluated which will help select the best model for the data.

6.2.5 Regression Metrics

The numerical metrics such as accuracy, classification error, precision, sensitivity, specificity, and F1-score. The results are shown in the table below.

```
##
##
## |Metrics                   | Model.1 | Model.2 |Model.3 |
## |:-------------------------|:-------:|:-------:|:-------|
## |Accuracy                  |  0.79   |  0.79   |0.80    |
## |Classification Error rate |  0.21   |  0.21   |0.20    |
## |Precission                |  0.82   |  0.82   |0.82    |
## |Sensitivity               |  0.92   |  0.92   |0.92    |
```

```
## |Specificity               |  0.43   |  0.43   |0.44    |
## |F-1 score                 |  0.87   |  0.87   |0.87    |
```

The metric values for the three models are very similar. However, based on confusion matrix, Model 3 can be used for TARGET_FLAG predictions.

7. References

Little, Roderick J. A. 1988. "Missing-Data Adjustments in Large Surveys." Journal Article. *Journal of Business & Economic Statistics* 6: 287–96.

Pituch, K., and J. P. Stevens. 2016. *Applied Multivariate Statistics for the Social Sciences: Analyses with Sas and Ibm's Spss*. Book. 6th Edison. Routledge Publishing.

Rubin, Donald B. 1986. "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations." Journal Article. *Journal of Business & Economic Statistics* 4: 87–94.