DATA 621 - Homework 4

Vanita Thompson, David Moste, Sadia Perveen

April 13, 2021

In this homework assignment, you will explore, analyze and model a data set containing approximately 8000 records representing a customer at an auto insurance company. Each record has two response variables. The first response variable, TARGET_FLAG, is a 1 or a 0. A "1" means that the person was in a car crash. A zero means that the person was not in a car crash. The second response variable is TARGET_AMT. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero.

Your objective is to build multiple linear regression and binary logistic regression models on the training data to predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

1. Data Exploration:

```
names(insurance)
```

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

```
str(insurance)
```

```
## 'data.frame':    8161 obs. of  26 variables:
##  $ INDEX      : int  1 2 4 5 6 7 8 11 12 13 ...
##  $ TARGET_FLAG: int  0 0 0 0 0 1 0 1 1 0 ...
##  $ TARGET_AMT : num  0 0 0 0 0 ...
##  $ KIDSDRIV   : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ AGE        : int  60 43 35 51 50 34 54 37 34 50 ...
##  $ HOMEKIDS   : int  0 0 1 0 0 1 0 2 0 0 ...
##  $ YOJ        : int  11 11 10 14 NA 12 NA NA 10 7 ...
##  $ INCOME     : chr  "$67,349" "$91,449" "$16,039" "" ...
##  $ PARENT1    : chr  "No" "No" "No" "No" ...
##  $ HOME_VAL   : chr  "$0" "$257,252" "$124,191" "$306,251" ...
##  $ MSTATUS    : chr  "z_No" "z_No" "Yes" "Yes" ...
##  $ SEX        : chr  "M" "M" "z_F" "M" ...
##  $ EDUCATION  : chr  "PhD" "z_High School" "z_High School" "<High School"
...
```

```
##  $ JOB        : chr  "Professional" "z_Blue Collar" "Clerical" "z_Blue
Collar" ...
##  $ TRAVTIME   : int  14 22 5 32 36 46 33 44 34 48 ...
##  $ CAR_USE    : chr  "Private" "Commercial" "Private" "Private" ...
##  $ BLUEBOOK   : chr  "$14,230" "$14,940" "$4,010" "$15,440" ...
##  $ TIF        : int  11 1 4 7 1 1 1 1 1 7 ...
##  $ CAR_TYPE   : chr  "Minivan" "Minivan" "z_SUV" "Minivan" ...
##  $ RED_CAR    : chr  "yes" "yes" "no" "yes" ...
##  $ OLDCLAIM   : chr  "$4,461" "$0" "$38,690" "$0" ...
##  $ CLM_FREQ   : int  2 0 2 0 2 0 0 1 0 0 ...
##  $ REVOKED    : chr  "No" "No" "No" "No" ...
##  $ MVR_PTS    : int  3 0 3 0 3 0 0 10 0 1 ...
##  $ CAR_AGE    : int  18 1 10 6 17 7 1 7 1 17 ...
##  $ URBANICITY : chr  "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly
Urban/ Urban" "Highly Urban/ Urban" ...
```

```
dim(insurance)
```

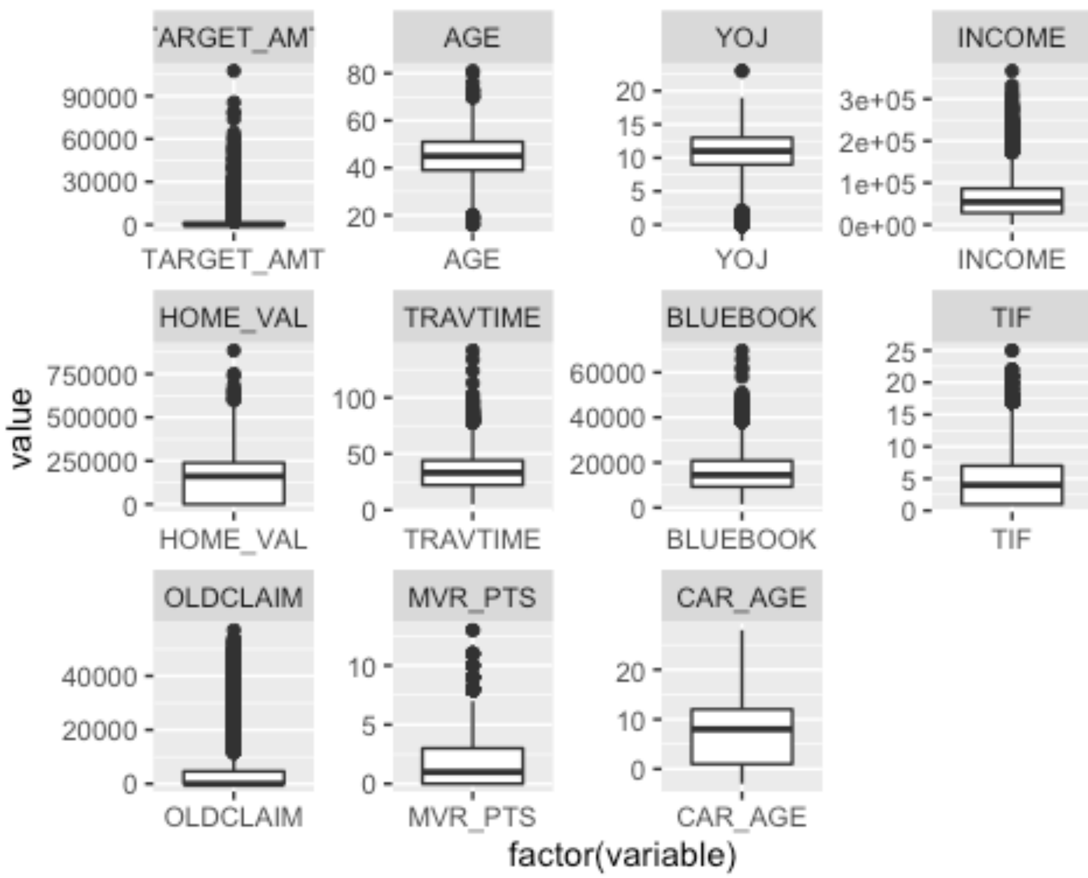
```
## [1] 8161   26
```

```
summary(insurance)
```

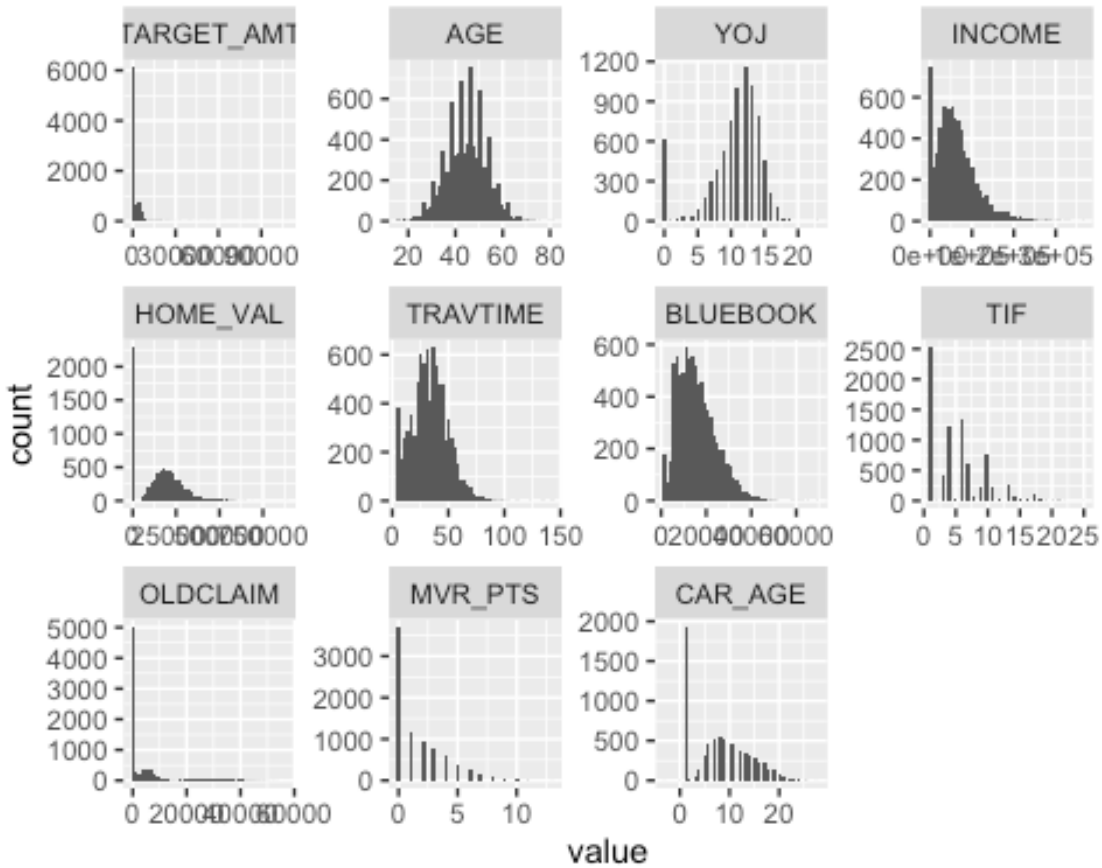
```
##      INDEX        TARGET_FLAG       TARGET_AMT        KIDSDRIV
##  Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000
##  1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000
##  Median : 5133   Median :0.0000   Median :     0   Median :0.0000
##  Mean   : 5152   Mean   :0.2638   Mean   :  1504   Mean   :0.1711
##  3rd Qu.: 7745   3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000
##  Max.   :10302   Max.   :1.0000   Max.   :107586   Max.   :4.0000
##
##       AGE           HOMEKIDS           YOJ          INCOME
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   Length:8161
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   Class :character
##  Median :45.00   Median :0.0000   Median :11.0   Mode  :character
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0
##  NA's   :6                        NA's   :454
##    PARENT1            HOME_VAL           MSTATUS              SEX
##  Length:8161        Length:8161        Length:8161        Length:8161
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   EDUCATION             JOB               TRAVTIME        CAR_USE
##  Length:8161        Length:8161        Min.   :  5.00   Length:8161
##  Class :character   Class :character   1st Qu.: 22.00   Class :character
##  Mode  :character   Mode  :character   Median : 33.00   Mode  :character
##                                        Mean   : 33.49
```

```
##                                         3rd Qu.: 44.00
##                                         Max.   :142.00
##
##    BLUEBOOK              TIF            CAR_TYPE             RED_CAR
##  Length:8161        Min.   : 1.000   Length:8161        Length:8161
##  Class :character   1st Qu.: 1.000   Class :character   Class :character
##  Mode  :character   Median : 4.000   Mode  :character   Mode  :character
##                     Mean   : 5.351
##                     3rd Qu.: 7.000
##                     Max.   :25.000
##
##    OLDCLAIM            CLM_FREQ        REVOKED              MVR_PTS
##  Length:8161        Min.   :0.0000   Length:8161        Min.   : 0.000
##  Class :character   1st Qu.:0.0000   Class :character   1st Qu.: 0.000
##  Mode  :character   Median :0.0000   Mode  :character   Median : 1.000
##                     Mean   :0.7986                      Mean   : 1.696
##                     3rd Qu.:2.0000                      3rd Qu.: 3.000
##                     Max.   :5.0000                      Max.   :13.000
##
##     CAR_AGE        URBANICITY
##  Min.   :-3.000   Length:8161
##  1st Qu.: 1.000   Class :character
##  Median : 8.000   Mode  :character
##  Mean   : 8.328
##  3rd Qu.:12.000
##  Max.   :28.000
##  NA's   :510

# The data needs to be cleaned up. We have some variables with $ and some
variables with Z_ that needs to be removed.

insurance$MSTATUS <- gsub('z_', '', insurance$MSTATUS)
insurance$SEX <- gsub('z_', '', insurance$SEX)
insurance$EDUCATION <- gsub('z_', '', insurance$EDUCATION)
insurance$JOB <- gsub('z_', '', insurance$JOB)
insurance$CAR_TYPE <- gsub('z_', '', insurance$CAR_TYPE)
insurance$URBANICITY <- gsub('z_', '', insurance$URBANICITY)
insurance$INCOME <- gsub('[\\$,]', '', insurance$INCOME)
insurance$HOME_VAL <- gsub('[\\$,]', '', insurance$HOME_VAL)
insurance$BLUEBOOK <- gsub('[\\$,]', '', insurance$BLUEBOOK)
insurance$OLDCLAIM <- gsub('[\\$,]', '', insurance$OLDCLAIM)


insurancetrain <- insurance %>%
  dplyr::select(-INDEX) %>%
  mutate(TARGET_FLAG = as.factor(TARGET_FLAG),
         KIDSDRIV = as.factor(KIDSDRIV),
         HOMEKIDS = as.factor(HOMEKIDS),
         PARENT1 = as.factor(PARENT1),
         CLM_FREQ = as.factor(CLM_FREQ),
```

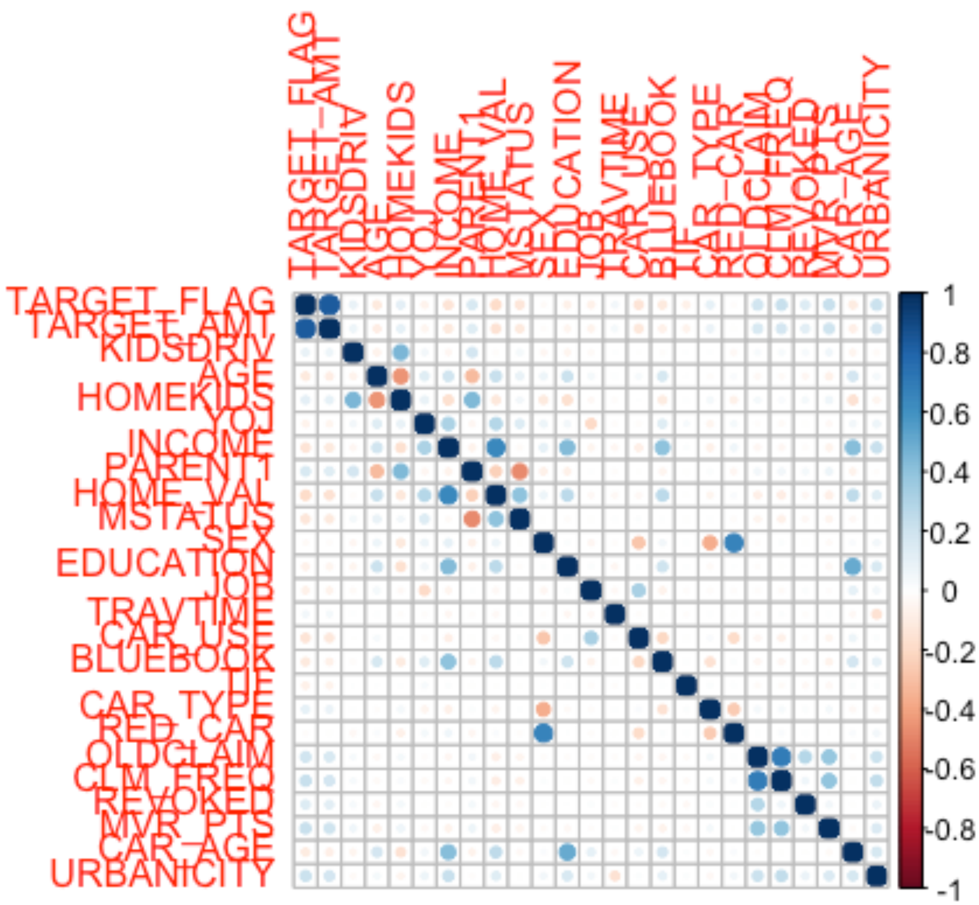
```
         OLDCLAIM = as.integer(OLDCLAIM),
         BLUEBOOK = as.integer(BLUEBOOK),
         HOME_VAL = as.integer(HOME_VAL),
         INCOME = as.integer(INCOME))

#boxplot, histogram and correlations
ggplot(melt(insurancetrain), aes(x=factor(variable), y=value)) +
facet_wrap(~variable, scale="free") + geom_boxplot()

## Using TARGET_FLAG, KIDSDRIV, HOMEKIDS, PARENT1, MSTATUS, SEX, EDUCATION,
JOB, CAR_USE, CAR_TYPE, RED_CAR, CLM_FREQ, REVOKED, URBANICITY as id
variables

## Warning: Removed 1879 rows containing non-finite values (stat_boxplot).
```




```
ggplot(melt(insurancetrain), aes(x=value)) + facet_wrap(~variable,
scale="free") + geom_histogram(bins=50)

## Using TARGET_FLAG, KIDSDRIV, HOMEKIDS, PARENT1, MSTATUS, SEX, EDUCATION,
JOB, CAR_USE, CAR_TYPE, RED_CAR, CLM_FREQ, REVOKED, URBANICITY as id
variables

## Warning: Removed 1879 rows containing non-finite values (stat_bin).
```




```
cor1 <- data.frame(lapply(insurancetrain, function(x)
as.numeric(as.factor(x))))

c <- cor(cor1, method="pearson", use="complete.obs")
corrplot(c, method="circle")
```



We observed that:

- The crime dataset contains 26 variables, with 8161 observations
- There are missing values.

2. Data Preparation

```
## checking for no missing data
sapply(insurancetrain, function(x) sum(is.na(x)))

## TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS         YOJ
##           0           0           0           6           0         454
##      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX   EDUCATION
##         445           0         464           0           0           0
##         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF    CAR_TYPE
##           0           0           0           0           0           0
##     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS     CAR_AGE
##           0           0           0           0           0         510
##  URBANICITY
##           0

#Using the mice package to input the missing data.
insurancetraining2 <- mice(insurancetrain, m=5, maxit = 5, method = 'pmm')
```

```
##
##  iter imp variable
##   1   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   1   2  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   1   3  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   1   4  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   1   5  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   2   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   2   2  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   2   3  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   2   4  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   2   5  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   3   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   3   2  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   3   3  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   3   4  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   3   5  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   4   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   4   2  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   4   3  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   4   4  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   4   5  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   5   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   5   2  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   5   3  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   5   4  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   5   5  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE

## Warning: Number of logged events: 9
```

```
insurancetraining2 <- complete(insurancetraining2)
summary(insurancetraining2)
```

```
##  TARGET_FLAG   TARGET_AMT     KIDSDRIV      AGE         HOMEKIDS      YOJ
##  0:6008      Min.   :     0   0:7180   Min.   :16.00   0:5289   Min.   :
0.00
##  1:2153      1st Qu.:     0   1: 636   1st Qu.:39.00   1: 902   1st Qu.:
9.00
##              Median :     0   2: 279   Median :45.00   2:1118   Median
:11.00
##              Mean   :  1504   3:  62   Mean   :44.78   3: 674   Mean
:10.49
##              3rd Qu.:  1036   4:   4   3rd Qu.:51.00   4: 164   3rd
Qu.:13.00
##              Max.   :107586            Max.   :81.00   5:  14   Max.
:23.00
##      INCOME        PARENT1       HOME_VAL         MSTATUS
##  Min.   :     0   No :7084   Min.   :     0   Length:8161
##  1st Qu.: 27957   Yes:1077   1st Qu.:     0   Class :character
##  Median : 54009              Median :161166   Mode  :character
```

```
##  Mean   : 61751              Mean   :154983
##  3rd Qu.: 85731              3rd Qu.:238931
##  Max.   :367030              Max.   :885282
##      SEX               EDUCATION             JOB               TRAVTIME
##  Length:8161        Length:8161        Length:8161        Min.   :  5.00
##  Class :character   Class :character   Class :character   1st Qu.: 22.00
##  Mode  :character   Mode  :character   Mode  :character   Median : 33.00
##                                                           Mean   : 33.49
##                                                           3rd Qu.: 44.00
##                                                           Max.   :142.00
##    CAR_USE             BLUEBOOK          TIF            CAR_TYPE
##  Length:8161        Min.   : 1500   Min.   : 1.000   Length:8161
##  Class :character   1st Qu.: 9280   1st Qu.: 1.000   Class :character
##  Mode  :character   Median :14440   Median : 4.000   Mode  :character
##                     Mean   :15710   Mean   : 5.351
##                     3rd Qu.:20850   3rd Qu.: 7.000
##                     Max.   :69740   Max.   :25.000
##    RED_CAR             OLDCLAIM     CLM_FREQ   REVOKED
##  Length:8161        Min.   :    0   0:5009   Length:8161
##  Class :character   1st Qu.:    0   1: 997   Class :character
##  Mode  :character   Median :    0   2:1171   Mode  :character
##                     Mean   : 4037   3: 776
##                     3rd Qu.: 4636   4: 190
##                     Max.   :57037   5:  18
##     MVR_PTS          CAR_AGE        URBANICITY
##  Min.   : 0.000   Min.   :-3.00   Length:8161
##  1st Qu.: 0.000   1st Qu.: 1.00   Class :character
##  Median : 1.000   Median : 8.00   Mode  :character
##  Mean   : 1.696   Mean   : 8.33
##  3rd Qu.: 3.000   3rd Qu.:12.00
##  Max.   :13.000   Max.   :28.00
```

```
sapply(insurancetraining2, function(x) sum(is.na(x)))
```

```
## TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS         YOJ
##           0           0           0           0           0           0
##      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX   EDUCATION
##           0           0           0           0           0           0
##         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF    CAR_TYPE
##           0           0           0           0           0           0
##     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS     CAR_AGE
##           0           0           0           0           0           0
##  URBANICITY
##           0
```

```
#same for eval set
sapply(insurance_evaluation, function(x) sum(is.na(x)))
```

```
##       INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##           0        2141        2141           0           1           0
##         YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
```

```
##          94            0            0            0            0            0
##   EDUCATION          JOB     TRAVTIME      CAR_USE     BLUEBOOK          TIF
##           0            0            0            0            0            0
##    CAR_TYPE      RED_CAR     OLDCLAIM     CLM_FREQ      REVOKED      MVR_PTS
##           0            0            0            0            0            0
##     CAR_AGE   URBANICITY
##         129            0
```

```
insuranceeval2 <- mice(insurance_evaluation, m=5, maxit = 5, method = 'pmm')
```

```
## 
##  iter imp variable
##   1   1  AGE  YOJ  CAR_AGE
##   1   2  AGE  YOJ  CAR_AGE
##   1   3  AGE  YOJ  CAR_AGE
##   1   4  AGE  YOJ  CAR_AGE
##   1   5  AGE  YOJ  CAR_AGE
##   2   1  AGE  YOJ  CAR_AGE
##   2   2  AGE  YOJ  CAR_AGE
##   2   3  AGE  YOJ  CAR_AGE
##   2   4  AGE  YOJ  CAR_AGE
##   2   5  AGE  YOJ  CAR_AGE
##   3   1  AGE  YOJ  CAR_AGE
##   3   2  AGE  YOJ  CAR_AGE
##   3   3  AGE  YOJ  CAR_AGE
##   3   4  AGE  YOJ  CAR_AGE
##   3   5  AGE  YOJ  CAR_AGE
##   4   1  AGE  YOJ  CAR_AGE
##   4   2  AGE  YOJ  CAR_AGE
##   4   3  AGE  YOJ  CAR_AGE
##   4   4  AGE  YOJ  CAR_AGE
##   4   5  AGE  YOJ  CAR_AGE
##   5   1  AGE  YOJ  CAR_AGE
##   5   2  AGE  YOJ  CAR_AGE
##   5   3  AGE  YOJ  CAR_AGE
##   5   4  AGE  YOJ  CAR_AGE
##   5   5  AGE  YOJ  CAR_AGE
```

```
## Warning: Number of logged events: 16
```

```
insuranceeval2 <- complete(insuranceeval2)
insuranceeval2 <- data.frame(lapply(insuranceeval2, function(x)
as.numeric(as.factor(x))))
summary(insuranceeval2)
```

```
##      INDEX       TARGET_FLAG     TARGET_AMT      KIDSDRIV          AGE
##  Min.   :   1   Min.   : NA    Min.   : NA    Min.   :1.000   Min.   :
1.00
##  1st Qu.: 536   1st Qu.: NA    1st Qu.: NA    1st Qu.:1.000   1st
Qu.:22.00
##  Median :1071   Median : NA    Median : NA    Median :1.000   Median
```

```
:28.00
##  Mean   :1071   Mean   :NaN    Mean   :NaN    Mean   :1.163   Mean
:28.02
##  3rd Qu.:1606   3rd Qu.: NA    3rd Qu.: NA    3rd Qu.:1.000   3rd
Qu.:34.00
##  Max.   :2141   Max.   : NA    Max.   : NA    Max.   :4.000   Max.
:54.00
##                 NA's   :2141   NA's   :2141
##     HOMEKIDS          YOJ             INCOME          PARENT1
##  Min.   :1.000   Min.   : 1.00   Min.   :   1.0   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:10.00   1st Qu.: 227.0   1st Qu.:1.000
##  Median :1.000   Median :12.00   Median : 754.0   Median :1.000
##  Mean   :1.717   Mean   :11.37   Mean   : 773.1   Mean   :1.124
##  3rd Qu.:2.000   3rd Qu.:14.00   3rd Qu.:1275.0   3rd Qu.:1.000
##  Max.   :6.000   Max.   :20.00   Max.   :1804.0   Max.   :2.000
##
##     HOME_VAL         MSTATUS           SEX          EDUCATION
##  Min.   :   1.0   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:   2.0   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000
##  Median : 342.0   Median :1.000   Median :2.000   Median :3.000
##  Mean   : 463.4   Mean   :1.396   Mean   :1.546   Mean   :3.114
##  3rd Qu.: 869.0   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:5.000
##  Max.   :1398.0   Max.   :2.000   Max.   :2.000   Max.   :5.000
##
##       JOB           TRAVTIME        CAR_USE        BLUEBOOK
##  Min.   :1.000   Min.   : 1.00   Min.   :1.000   Min.   :   1.0
##  1st Qu.:4.000   1st Qu.:18.00   1st Qu.:1.000   1st Qu.: 306.0
##  Median :6.000   Median :29.00   Median :2.000   Median : 688.0
##  Mean   :5.653   Mean   :29.11   Mean   :1.645   Mean   : 702.3
##  3rd Qu.:8.000   3rd Qu.:39.00   3rd Qu.:2.000   3rd Qu.:1144.0
##  Max.   :9.000   Max.   :83.00   Max.   :2.000   Max.   :1417.0
##
##       TIF           CAR_TYPE        RED_CAR         OLDCLAIM
##  Min.   : 1.000   Min.   :1.000   Min.   :1.000   Min.   :  1.0
##  1st Qu.: 1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:  1.0
##  Median : 3.000   Median :3.000   Median :1.000   Median :  1.0
##  Mean   : 4.542   Mean   :3.517   Mean   :1.279   Mean   :169.1
##  3rd Qu.: 6.000   3rd Qu.:6.000   3rd Qu.:2.000   3rd Qu.:319.0
##  Max.   :21.000   Max.   :6.000   Max.   :2.000   Max.   :834.0
##
##     CLM_FREQ        REVOKED         MVR_PTS          CAR_AGE
##  Min.   :1.000   Min.   :1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 1.000   1st Qu.: 2.000
##  Median :1.000   Median :1.000   Median : 2.000   Median : 9.000
##  Mean   :1.809   Mean   :1.122   Mean   : 2.766   Mean   : 9.212
##  3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.: 4.000   3rd Qu.:14.000
##  Max.   :6.000   Max.   :2.000   Max.   :13.000   Max.   :27.000
##
##    URBANICITY
##  Min.   :1.000
```

```
##   1st Qu.:1.000
##   Median :1.000
##   Mean   :1.188
##   3rd Qu.:1.000
##   Max.   :2.000
##
```

```r
sapply(insurancetraining2, function(x) sum(is.na(x)))
```

```
## TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS         YOJ
##           0           0           0           0           0           0
##      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX   EDUCATION
##           0           0           0           0           0           0
##         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF    CAR_TYPE
##           0           0           0           0           0           0
##     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS     CAR_AGE
##           0           0           0           0           0           0
##  URBANICITY
##           0
```

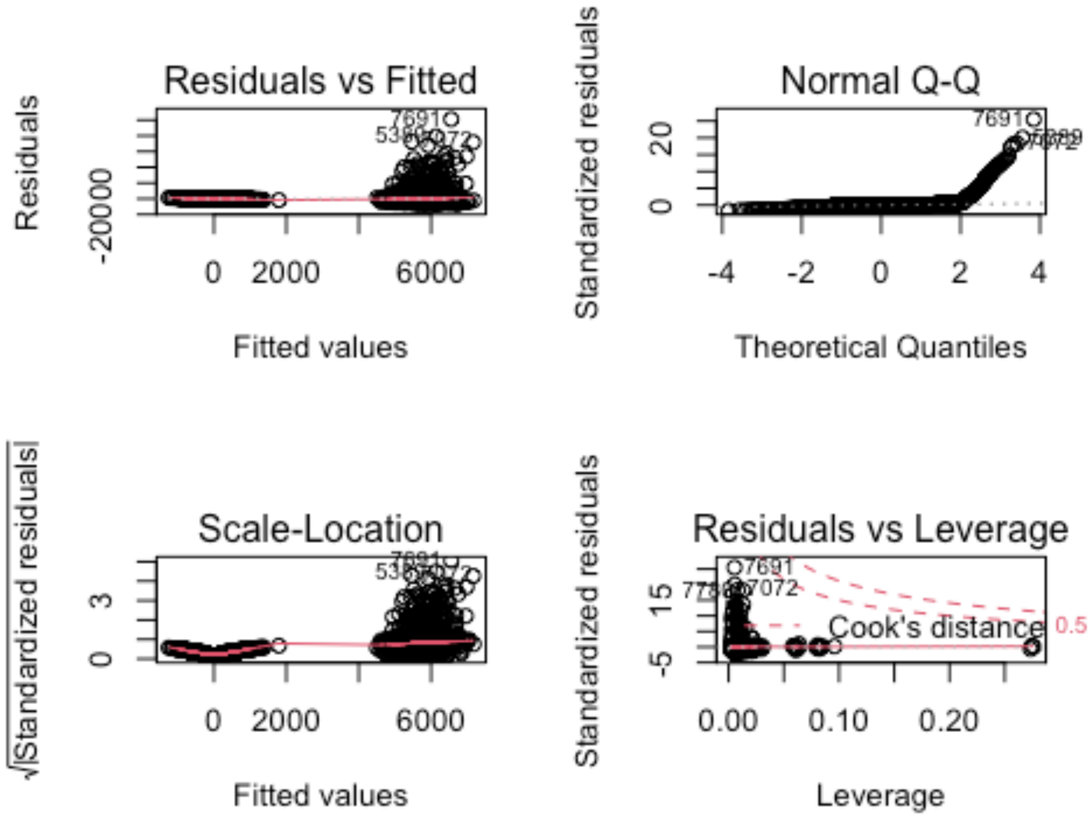
3. Build Models

```r
#We will build different multiple linear regression models and binary linear regression models.
model1 <- lm(TARGET_AMT ~ ., insurancetraining2)
summary(model1)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = insurancetraining2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6429   -476    -56    241 101031
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -7.267e+02  5.073e+02  -1.433 0.152000
## TARGET_FLAG1                   5.703e+03  1.138e+02  50.106  < 2e-16 ***
## KIDSDRIV1                      1.574e+02  1.850e+02   0.851 0.395059
## KIDSDRIV2                     -1.009e+02  2.645e+02  -0.382 0.702837
## KIDSDRIV3                     -4.164e+02  5.263e+02  -0.791 0.428835
## KIDSDRIV4                     -1.017e+03  2.078e+03  -0.489 0.624672
## AGE                            5.776e+00  6.398e+00   0.903 0.366621
## HOMEKIDS1                     -1.346e+01  1.810e+02  -0.074 0.940724
## HOMEKIDS2                      1.588e+02  1.772e+02   0.896 0.370219
## HOMEKIDS3                      2.905e+01  2.076e+02   0.140 0.888719
## HOMEKIDS4                      1.465e+02  3.436e+02   0.426 0.669865
## HOMEKIDS5                      1.329e+02  1.114e+03   0.119 0.905061
## YOJ                            1.116e+01  1.283e+01   0.871 0.384037
## INCOME                        -1.989e-03  1.599e-03  -1.244 0.213430
```

```
## PARENT1Yes                        1.150e+02  1.900e+02   0.605 0.545095    
## HOME_VAL                          3.181e-04  5.152e-04   0.617 0.536939    
## MSTATUSYes                       -1.740e+02  1.293e+02  -1.345 0.178591    
## SEXM                              2.862e+02  1.608e+02   1.780 0.075149 .  
## EDUCATIONBachelors                4.061e+01  1.788e+02   0.227 0.820282    
## EDUCATIONHigh School             -1.268e+02  1.503e+02  -0.844 0.398638    
## EDUCATIONMasters                  1.668e+02  2.610e+02   0.639 0.522843    
## EDUCATIONPhD                      3.706e+02  3.103e+02   1.194 0.232502    
## JOBBlue Collar                    5.374e+01  2.817e+02   0.191 0.848694    
## JOBClerical                      -8.764e+00  2.988e+02  -0.029 0.976598    
## JOBDoctor                        -2.865e+02  3.574e+02  -0.802 0.422801    
## JOBHome Maker                    -4.737e+01  3.187e+02  -0.149 0.881848    
## JOBLawyer                         7.333e+01  2.585e+02   0.284 0.776657    
## JOBManager                       -1.209e+02  2.523e+02  -0.479 0.631745    
## JOBProfessional                   1.764e+02  2.700e+02   0.653 0.513644    
## JOBStudent                       -1.048e+02  3.275e+02  -0.320 0.749019    
## TRAVTIME                          5.006e-01  2.826e+00   0.177 0.859398    
## CAR_USEPrivate                   -9.557e+01  1.444e+02  -0.662 0.508083    
## BLUEBOOK                          2.912e-02  7.554e-03   3.855 0.000117 ***
## TIF                              -2.898e+00  1.070e+01  -0.271 0.786499    
## CAR_TYPEPanel Truck              -3.806e+01  2.434e+02  -0.156 0.875762    
## CAR_TYPEPickup                   -2.515e+01  1.495e+02  -0.168 0.866374    
## CAR_TYPESports Car                2.056e+02  1.911e+02   1.076 0.281857    
## CAR_TYPESUV                       1.654e+02  1.573e+02   1.051 0.293169    
## CAR_TYPEVan                       9.594e+01  1.866e+02   0.514 0.607116    
## RED_CARyes                       -2.553e+01  1.303e+02  -0.196 0.844719    
## OLDCLAIM                          3.493e-03  6.986e-03   0.500 0.617085    
## CLM_FREQ1                        -2.494e+01  1.666e+02  -0.150 0.881041    
## CLM_FREQ2                        -2.090e+02  1.590e+02  -1.314 0.188760    
## CLM_FREQ3                        -2.555e+01  1.798e+02  -0.142 0.887034    
## CLM_FREQ4                        -2.105e+02  3.064e+02  -0.687 0.492056    
## CLM_FREQ5                        -5.867e+02  9.433e+02  -0.622 0.533995    
## REVOKEDYes                       -3.280e+02  1.545e+02  -2.122 0.033846 *  
## MVR_PTS                           5.521e+01  2.347e+01   2.352 0.018704 *  
## CAR_AGE                          -1.959e+01  1.073e+01  -1.826 0.067955 .  
## URBANICITYHighly Urban/ Urban -2.875e+01  1.276e+02  -0.225 0.821717    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3971 on 8111 degrees of freedom
## Multiple R-squared:  0.2916, Adjusted R-squared:  0.2873 
## F-statistic: 68.13 on 49 and 8111 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model1)
```

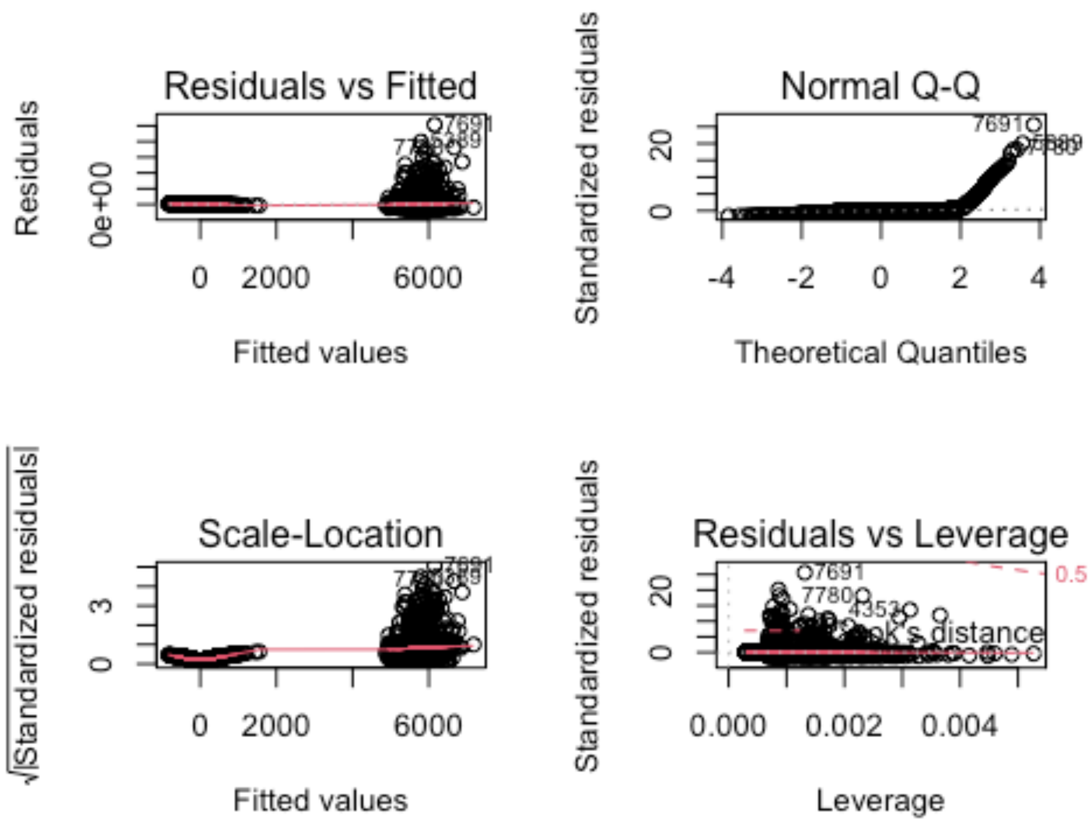



```
model2 <- stepAIC(model1, direction = "both", trace = FALSE)
summary(model2)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ TARGET_FLAG + PARENT1 + SEX + BLUEBOOK +
##     REVOKED + MVR_PTS, data = insurancetraining2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6092   -405    -37    202 101433
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.143e+02  1.123e+02  -5.470 4.63e-08 ***
## TARGET_FLAG1  5.716e+03  1.047e+02  54.618  < 2e-16 ***
## PARENT1Yes    2.237e+02  1.319e+02   1.696   0.0899 .
## SEXM          1.935e+02  8.844e+01   2.187   0.0287 *
## BLUEBOOK      2.824e-02  5.256e-03   5.373 7.95e-08 ***
## REVOKEDYes   -2.963e+02  1.356e+02  -2.186   0.0289 *
## MVR_PTS       4.951e+01  2.098e+01   2.360   0.0183 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3966 on 8154 degrees of freedom
## Multiple R-squared:  0.2895, Adjusted R-squared:  0.289
## F-statistic: 553.8 on 6 and 8154 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model2)
```




```
#box-cox
insurancebc <- preProcess(insurancetraining2, c("BoxCox"))
insurancebc_transformed <- predict(insurancebc, insurancetraining2)
model4 <- lm(TARGET_AMT ~ ., insurancebc_transformed)
summary(model4)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = insurancebc_transformed)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6416   -478    -70    239 101019
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                      -1.191e+03  5.402e+02  -2.206   0.0274 *
## TARGET_FLAG1                      5.709e+03  1.139e+02  50.130  < 2e-16 ***
## KIDSDRIV1                         1.547e+02  1.850e+02   0.836   0.4029
## KIDSDRIV2                        -1.078e+02  2.644e+02  -0.408   0.6834
## KIDSDRIV3                        -4.168e+02  5.262e+02  -0.792   0.4283
## KIDSDRIV4                        -1.028e+03  2.077e+03  -0.495   0.6207
## AGE                               5.472e+00  6.394e+00   0.856   0.3921
## HOMEKIDS1                        -1.283e+01  1.809e+02  -0.071   0.9435
## HOMEKIDS2                         1.615e+02  1.771e+02   0.912   0.3618
## HOMEKIDS3                         3.186e+01  2.076e+02   0.153   0.8780
## HOMEKIDS4                         1.465e+02  3.435e+02   0.426   0.6698
## HOMEKIDS5                         1.057e+02  1.114e+03   0.095   0.9244
## YOJ                               1.068e+01  1.282e+01   0.833   0.4048
## INCOME                           -1.922e-03  1.592e-03  -1.207   0.2274
## PARENT1Yes                        1.146e+02  1.899e+02   0.603   0.5463
## HOME_VAL                          3.217e-04  5.150e-04   0.625   0.5323
## MSTATUSYes                       -1.725e+02  1.293e+02  -1.334   0.1822
## SEXM                              2.969e+02  1.595e+02   1.862   0.0627 .
## EDUCATIONBachelors                3.140e+01  1.788e+02   0.176   0.8606
## EDUCATIONHigh School             -1.319e+02  1.502e+02  -0.878   0.3801
## EDUCATIONMasters                  1.552e+02  2.610e+02   0.595   0.5521
## EDUCATIONPhD                      3.687e+02  3.103e+02   1.188   0.2348
## JOBBlue Collar                    4.416e+01  2.816e+02   0.157   0.8754
## JOBClerical                      -1.148e+01  2.987e+02  -0.038   0.9693
## JOBDoctor                        -2.965e+02  3.573e+02  -0.830   0.4067
## JOBHome Maker                    -3.282e+01  3.187e+02  -0.103   0.9180
## JOBLawyer                         6.616e+01  2.584e+02   0.256   0.7980
## JOBManager                       -1.296e+02  2.522e+02  -0.514   0.6074
## JOBProfessional                   1.685e+02  2.700e+02   0.624   0.5326
## JOBStudent                       -8.598e+01  3.275e+02  -0.263   0.7929
## TRAVTIME                          9.893e-01  7.906e+00   0.125   0.9004
## CAR_USEPrivate                   -9.315e+01  1.444e+02  -0.645   0.5188
## BLUEBOOK                          3.895e+00  8.983e-01   4.336 1.47e-05 ***
## TIF                              -9.568e+00  3.655e+01  -0.262   0.7935
## CAR_TYPEPanel Truck              -2.509e+01  2.375e+02  -0.106   0.9159
## CAR_TYPEPickup                   -1.367e+01  1.495e+02  -0.091   0.9271
## CAR_TYPESports Car                2.385e+02  1.912e+02   1.248   0.2122
## CAR_TYPESUV                       1.790e+02  1.560e+02   1.148   0.2512
## CAR_TYPEVan                       7.430e+01  1.866e+02   0.398   0.6905
## RED_CARyes                       -2.382e+01  1.303e+02  -0.183   0.8549
## OLDCLAIM                          3.473e-03  6.984e-03   0.497   0.6190
## CLM_FREQ1                        -2.798e+01  1.666e+02  -0.168   0.8666
## CLM_FREQ2                        -2.101e+02  1.590e+02  -1.321   0.1864
## CLM_FREQ3                        -2.578e+01  1.798e+02  -0.143   0.8860
## CLM_FREQ4                        -2.065e+02  3.063e+02  -0.674   0.5003
## CLM_FREQ5                        -5.948e+02  9.430e+02  -0.631   0.5283
## REVOKEDYes                       -3.269e+02  1.545e+02  -2.116   0.0343 *
## MVR_PTS                           5.532e+01  2.347e+01   2.357   0.0184 *
## CAR_AGE                          -1.962e+01  1.073e+01  -1.830   0.0674 .
## URBANICITYHighly Urban/ Urban -3.160e+01  1.274e+02  -0.248   0.8042
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3970 on 8111 degrees of freedom
## Multiple R-squared:  0.2919, Adjusted R-squared:  0.2877
## F-statistic: 68.25 on 49 and 8111 DF,  p-value: < 2.2e-16
```

```r
#For the first three models we see similar results. Where the Q1 and Q3 are not evenly distributed.
#The r-squared is .29, .28 and .29. Lets look at more models.

glm_data <- data.frame(lapply(insurancetraining2, function(x)
as.numeric(as.factor(x)))) %>%
  mutate(TARGET_FLAG = as.factor(TARGET_FLAG))
glm_data1 <- glm_data %>%
  dplyr::select(-"TARGET_AMT")

model5 <- glm(TARGET_FLAG ~ ., family = "binomial", glm_data1)
summary(model5)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = glm_data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5412  -0.7266  -0.4142   0.6511   3.1414
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.650e+00  4.007e-01 -11.606  < 2e-16 ***
## KIDSDRIV     3.694e-01  6.044e-02   6.112 9.85e-10 ***
## AGE         -1.066e-03  3.940e-03  -0.271 0.786656
## HOMEKIDS     6.467e-02  3.659e-02   1.767 0.077151 .
## YOJ         -1.416e-02  7.717e-03  -1.835 0.066523 .
## INCOME      -1.449e-04  2.255e-05  -6.426 1.31e-10 ***
## PARENT1      3.690e-01  1.084e-01   3.403 0.000666 ***
## HOME_VAL    -9.056e-05  2.579e-05  -3.511 0.000446 ***
## MSTATUS     -4.979e-01  8.059e-02  -6.179 6.45e-10 ***
## SEX         -7.906e-02  8.400e-02  -0.941 0.346581
## EDUCATION   -4.925e-03  2.945e-02  -0.167 0.867163
## JOB         -4.668e-02  1.160e-02  -4.022 5.77e-05 ***
## TRAVTIME     1.512e-02  1.878e-03   8.050 8.25e-16 ***
## CAR_USE     -8.536e-01  6.646e-02 -12.843  < 2e-16 ***
## BLUEBOOK    -2.975e-04  4.586e-05  -6.488 8.71e-11 ***
## TIF         -5.458e-02  7.290e-03  -7.487 7.04e-14 ***
## CAR_TYPE     1.259e-01  1.834e-02   6.869 6.49e-12 ***
## RED_CAR     -2.044e-02  8.568e-02  -0.239 0.811412
## OLDCLAIM    -4.986e-05  4.510e-05  -1.106 0.268896
## CLM_FREQ     1.725e-01  3.217e-02   5.363 8.19e-08 ***
```

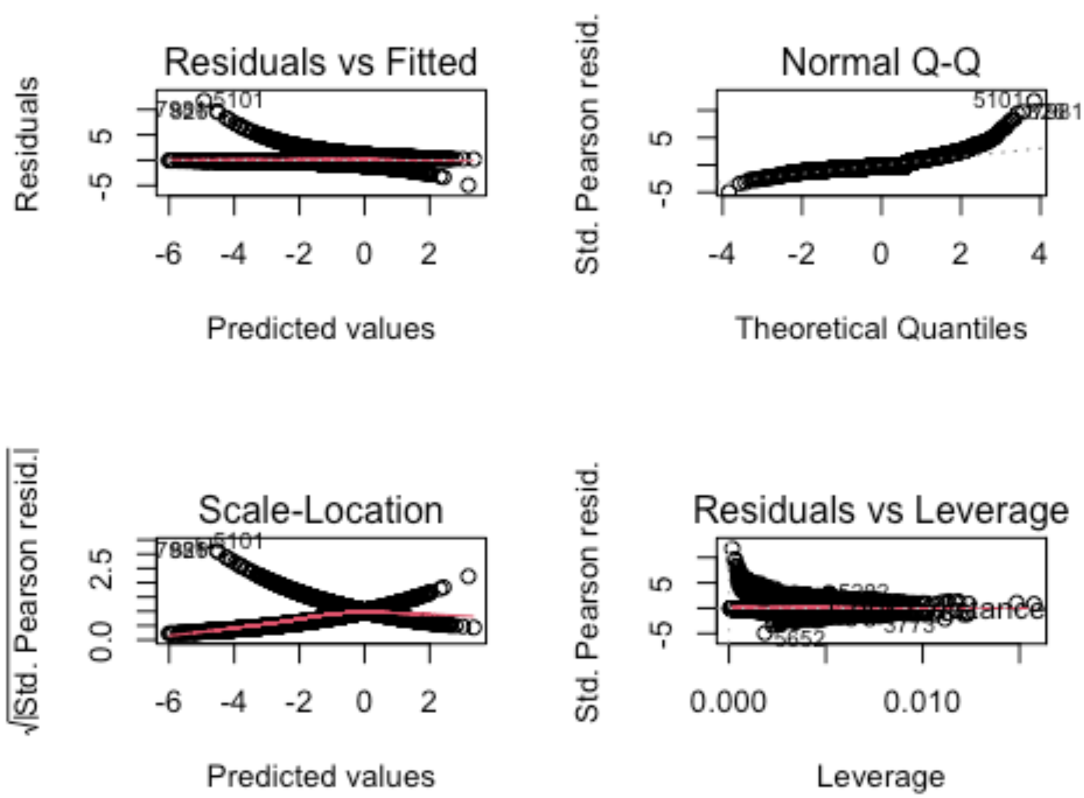
```
## REVOKED      7.759e-01  8.471e-02   9.159  < 2e-16 ***
## MVR_PTS      1.162e-01  1.362e-02   8.528  < 2e-16 ***
## CAR_AGE     -2.121e-02  6.113e-03  -3.470 0.000520 ***
## URBANICITY   2.317e+00  1.121e-01  20.676  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9418  on 8160  degrees of freedom
## Residual deviance: 7398  on 8137  degrees of freedom
## AIC: 7446
##
## Number of Fisher Scoring iterations: 5
```

```
par(mfrow=c(2,2))
plot(model5)
# Model5 has evenly distributed deviance. many of our variables are
significant.


model6 <- stepAIC(model5, direction = "both", trace = FALSE)
summary(model6)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME +
##     PARENT1 + HOME_VAL + MSTATUS + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS +
##     CAR_AGE + URBANICITY, family = "binomial", data = glm_data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5489  -0.7280  -0.4122   0.6503   3.1252
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.842e+00  3.463e-01 -13.980  < 2e-16 ***
## KIDSDRIV     3.681e-01  5.942e-02   6.196 5.80e-10 ***
## HOMEKIDS     7.173e-02  3.378e-02   2.123 0.033742 *
## YOJ         -1.514e-02  7.586e-03  -1.995 0.046029 *
## INCOME      -1.462e-04  2.207e-05  -6.625 3.48e-11 ***
## PARENT1      3.763e-01  1.077e-01   3.493 0.000477 ***
## HOME_VAL    -9.127e-05  2.567e-05  -3.555 0.000378 ***
## MSTATUS     -4.958e-01  8.053e-02  -6.157 7.41e-10 ***
## JOB         -4.741e-02  1.156e-02  -4.100 4.14e-05 ***
## TRAVTIME     1.518e-02  1.876e-03   8.088 6.08e-16 ***
## CAR_USE     -8.281e-01  6.343e-02 -13.056  < 2e-16 ***
## BLUEBOOK    -2.953e-04  4.556e-05  -6.482 9.03e-11 ***
## TIF         -5.448e-02  7.284e-03  -7.479 7.46e-14 ***
```

```
## CAR_TYPE      1.348e-01  1.706e-02   7.901 2.77e-15 ***
## CLM_FREQ      1.501e-01  2.526e-02   5.943 2.81e-09 ***
## REVOKED       7.428e-01  7.951e-02   9.342  < 2e-16 ***
## MVR_PTS       1.141e-01  1.345e-02   8.483  < 2e-16 ***
## CAR_AGE      -2.163e-02  5.703e-03  -3.793 0.000149 ***
## URBANICITY    2.310e+00  1.119e-01  20.639  < 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: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7401.3  on 8142  degrees of freedom
## AIC: 7439.3
##
## Number of Fisher Scoring iterations: 5
```

```r
par(mfrow=c(2,2))
plot(model5)
```




```r
#This model has similar distribution however the AIC has not improved.

#box-cox
```

```r
glm_data12 <- preProcess(glm_data1, c("BoxCox"))
glmbc_transformed <- predict(glm_data12, glm_data1)
model7 <- glm(TARGET_FLAG ~ ., family = "binomial", glmbc_transformed)
summary(model7)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data =
glmbc_transformed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3434  -0.7272  -0.4114   0.6684   3.1748
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.5356112  0.3954075  -8.942  < 2e-16 ***
## KIDSDRIV     1.3998112  0.2444793   5.726 1.03e-08 ***
## AGE          0.0002871  0.0040905   0.070 0.944052
## HOMEKIDS     0.4735114  0.2137900   2.215 0.026771 *
## YOJ         -0.0001842  0.0022015  -0.084 0.933326
## INCOME      -0.0069581  0.0008723  -7.977 1.50e-15 ***
## PARENT1      0.2689187  0.1183525   2.272 0.023075 *
## HOME_VAL    -0.0168165  0.0043407  -3.874 0.000107 ***
## MSTATUS     -0.5132528  0.0868412  -5.910 3.42e-09 ***
## SEX         -0.0542853  0.0843341  -0.644 0.519774
## EDUCATION   -0.0247696  0.0379749  -0.652 0.514232
## JOB         -0.1237916  0.0247478  -5.002 5.67e-07 ***
## TRAVTIME     0.0412499  0.0049980   8.253  < 2e-16 ***
## CAR_USE     -0.8160522  0.0674303 -12.102  < 2e-16 ***
## BLUEBOOK    -0.0049333  0.0006987  -7.060 1.66e-12 ***
## TIF         -0.1832012  0.0238805  -7.672 1.70e-14 ***
## CAR_TYPE     0.1857806  0.0255250   7.278 3.38e-13 ***
## RED_CAR     -0.0205521  0.0856448  -0.240 0.810354
## OLDCLAIM    -0.0206122  0.0318136  -0.648 0.517046
## CLM_FREQ     1.1420151  0.4247696   2.689 0.007176 **
## REVOKED      0.7540602  0.0813805   9.266  < 2e-16 ***
## MVR_PTS      0.4148092  0.0623269   6.655 2.83e-11 ***
## CAR_AGE     -0.0725739  0.0177255  -4.094 4.23e-05 ***
## URBANICITY   2.2990506  0.1124285  20.449  < 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: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7380.8  on 8137  degrees of freedom
## AIC: 7428.8
##
## Number of Fisher Scoring iterations: 5
```

```r
# Getting the confusion matix, roc curve for each model
confusionMatrix1 <- confusionMatrix(as.factor(as.integer(fitted(model5) >
.5)), as.factor(model5$y), positive = "1")
rocmodel1 <- roc(glm_data$TARGET_FLAG,  predict(model5, glm_data))

## Setting levels: control = 1, case = 2

## Setting direction: controls < cases

confusionMatrix1

## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 5553 1293
##          1  455  860
##
##                Accuracy : 0.7858
##                  95% CI : (0.7767, 0.7947)
##     No Information Rate : 0.7362
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3699
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.3994
##             Specificity : 0.9243
##          Pos Pred Value : 0.6540
##          Neg Pred Value : 0.8111
##              Prevalence : 0.2638
##          Detection Rate : 0.1054
##    Detection Prevalence : 0.1611
##       Balanced Accuracy : 0.6619
##
##        'Positive' Class : 1
##

rocmodel1

##
## Call:
## roc.default(response = glm_data$TARGET_FLAG, predictor = predict(model5,
glm_data))
##
## Data: predict(model5, glm_data) in 6008 controls (glm_data$TARGET_FLAG 1)
< 2153 cases (glm_data$TARGET_FLAG 2).
## Area under the curve: 0.8067
```

```r
confusionMatrix2  <- confusionMatrix(as.factor(as.integer(fitted(model6) >
.5)), as.factor(model6$y), positive = "1")
rocmodel2 <- roc(glm_data$TARGET_FLAG,  predict(model6, glm_data))
```

```
## Setting levels: control = 1, case = 2
## Setting direction: controls < cases
```

```r
confusionMatrix2
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 5547 1295
##          1  461  858
##
##                Accuracy : 0.7848
##                  95% CI : (0.7758, 0.7937)
##     No Information Rate : 0.7362
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3674
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.3985
##             Specificity : 0.9233
##          Pos Pred Value : 0.6505
##          Neg Pred Value : 0.8107
##              Prevalence : 0.2638
##          Detection Rate : 0.1051
##    Detection Prevalence : 0.1616
##       Balanced Accuracy : 0.6609
##
##        'Positive' Class : 1
##
```

```r
rocmodel2
```

```
##
## Call:
## roc.default(response = glm_data$TARGET_FLAG, predictor = predict(model6,
glm_data))
##
## Data: predict(model6, glm_data) in 6008 controls (glm_data$TARGET_FLAG 1)
< 2153 cases (glm_data$TARGET_FLAG 2).
## Area under the curve: 0.8064
```

```r
confusionMatrix3  <- confusionMatrix(as.factor(as.integer(fitted(model7) >
.5)), as.factor(model7$y), positive = "1")
rocmodel3 <- roc(glm_data$TARGET_FLAG,  predict(model7, glm_data))
```

```
## Setting levels: control = 1, case = 2
## Setting direction: controls < cases
```

```
confusionMatrix3
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction    0    1
##           0 5553 1293
##           1  455  860
##
##                Accuracy : 0.7858
##                  95% CI : (0.7767, 0.7947)
##     No Information Rate : 0.7362
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3699
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.3994
##             Specificity : 0.9243
##          Pos Pred Value : 0.6540
##          Neg Pred Value : 0.8111
##              Prevalence : 0.2638
##          Detection Rate : 0.1054
##    Detection Prevalence : 0.1611
##       Balanced Accuracy : 0.6619
##
##        'Positive' Class : 1
##
```

```
rocmodel3
```

```
##
## Call:
## roc.default(response = glm_data$TARGET_FLAG, predictor = predict(model7,
glm_data))
##
## Data: predict(model7, glm_data) in 6008 controls (glm_data$TARGET_FLAG 1)
< 2153 cases (glm_data$TARGET_FLAG 2).
## Area under the curve: 0.5841
```

```
#Model5 has the highest AUC
# predict

predict <- predict(model5, insuranceeval2, interval = "prediction")
eval <- table(as.integer(predict > .5))
eval
```

```
##
##    0    1
## 2074   67
```