# DATA621-Homework4-SmoothOperators

*Rob Hodde, Matt Farris, Jeffrey Burmood, Bin Lin*

*4/17/2017*

**Problem Description**

The 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.

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.

Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R2, (c) F-statistic, and (d) residual plots. For the binary logistic regression model, will use a metric such as log likelihood, AIC, or ROC curve. Using the training data set, evaluate the binary logistic regression model based on (a) accuracy, (b) classification error rate, (c) precision, (d) sensitivity, (e) specificity, (f) F1 score, (g) AUC, and (h) confusion matrix. Make predictions using the evaluation data set.

Approach Steps:

1) Build a logistic regression model based on the TARGET_FLAG response variable.
2) Generate TARGET_FLAG predictions using the logistic regression model.
3) Build a linear regression model based on the non-zero values of the TARGET_AMT response variable.
4) Generate TARGET_AMT predictions using the linear regression model based on the non-zero values of the predicted TARGET_FLAG variable.

# Data Exploration

## Data Exploration

The first steps in our process was to explore our data. During this exploration, we immediately noticed the presence of inconsistent data, which is why we employed the use of the MICE package to provide weighted mean data to any missing values.

Our next step was to change the categorical data into numeric values, which was accomplished using the legend below.

After that, we repaired the offending values, standardized the data types, and add new variables to quantitatively represent the binary and categorical choices available in the data:

Below is a summary of each predictor variable's basic statistics, followed by boxplots which illustrate the spread and outliers for each variable.

| VAR | TYPE |
|---|---|
| TARGET_FLAG | double |
| TARGET_AMT | double |
| KIDSDRIV | integer |

| VAR | TYPE |
|---|---|
| AGE | integer |
| HOMEKIDS | integer |
| YOJ | integer |
| INCOME | double |
| HOME_VAL | double |
| TRAVTIME | integer |
| BLUEBOOK | double |
| TIF | integer |
| OLDCLAIM | double |
| CLM_FREQ | integer |
| MVR_PTS | integer |
| CAR_AGE | double |
| blnPARENT1 | double |
| blnMSTATUS | double |
| blnSEX | double |
| blnCAR_USE | double |
| blnNOT_RED_CAR | double |
| blnNOT_REVOKED | double |
| blnURBANICITY | double |
| intEDUCATION | double |
| intJOB | double |
| intCAR_TYPE | double |

| TARGET_FLAG | TARGET_AMT | KIDSDRIV | AGE | HOMEKIDS | YOJ |
|---|---|---|---|---|---|
| Min. :0.0000 | Min. : 0 | Min. :0.0000 | Min. :16.00 | Min. :0.0000 | Min. : 0.00 |
| 1st Qu.:0.0000 | 1st Qu.: 0 | 1st Qu.:0.0000 | 1st Qu.:39.00 | 1st Qu.:0.0000 | 1st Qu.: 9.00 |
| Median :0.0000 | Median : 0 | Median :0.0000 | Median :45.00 | Median :0.0000 | Median :11.00 |
| Mean :0.2638 | Mean : 1504 | Mean :0.1711 | Mean :44.78 | Mean :0.7212 | Mean :10.49 |
| 3rd Qu.:1.0000 | 3rd Qu.: 1036 | 3rd Qu.:0.0000 | 3rd Qu.:51.00 | 3rd Qu.:1.0000 | 3rd Qu.:13.00 |
| Max. :1.0000 | Max. :107586 | Max. :4.0000 | Max. :81.00 | Max. :5.0000 | Max. :23.00 |

| INCOME | HOME_VAL | TRAVTIME | BLUEBOOK | TIF | OLDCLAIM |
|---|---|---|---|---|---|
| Min. : 0 | Min. : 0 | Min. : 5.00 | Min. : 1500 | Min. : 1.000 | Min. : 0 |
| 1st Qu.: 28068 | 1st Qu.: 0 | 1st Qu.: 22.00 | 1st Qu.: 9280 | 1st Qu.: 1.000 | 1st Qu.: 0 |
| Median : 53628 | Median :160731 | Median : 33.00 | Median :14440 | Median : 4.000 | Median : 0 |
| Mean : 61709 | Mean :154889 | Mean : 33.49 | Mean :15710 | Mean : 5.351 | Mean : 4037 |
| 3rd Qu.: 85545 | 3rd Qu.:238850 | 3rd Qu.: 44.00 | 3rd Qu.:20850 | 3rd Qu.: 7.000 | 3rd Qu.: 4636 |
| Max. :367030 | Max. :885282 | Max. :142.00 | Max. :69740 | Max. :25.000 | Max. :57037 |

| CLM_FREQ | MVR_PTS | CAR_AGE | blnPARENT1 | blnMSTATUS | blnSEX |
|---|---|---|---|---|---|
| Min. :0.0000 | Min. : 0.000 | Min. : 0.000 | Min. :0.000 | Min. :0.0000 | Min. :0.0000 |
| 1st Qu.:0.0000 | 1st Qu.: 0.000 | 1st Qu.: 1.000 | 1st Qu.:0.000 | 1st Qu.:0.0000 | 1st Qu.:0.0000 |
| Median :0.0000 | Median : 1.000 | Median : 8.000 | Median :0.000 | Median :1.0000 | Median :1.0000 |
| Mean :0.7986 | Mean : 1.696 | Mean : 8.336 | Mean :0.132 | Mean :0.5997 | Mean :0.5361 |
| 3rd Qu.:2.0000 | 3rd Qu.: 3.000 | 3rd Qu.:12.000 | 3rd Qu.:0.000 | 3rd Qu.:1.0000 | 3rd Qu.:1.0000 |
| Max. :5.0000 | Max. :13.000 | Max. :28.000 | Max. :1.000 | Max. :1.0000 | Max. :1.0000 |

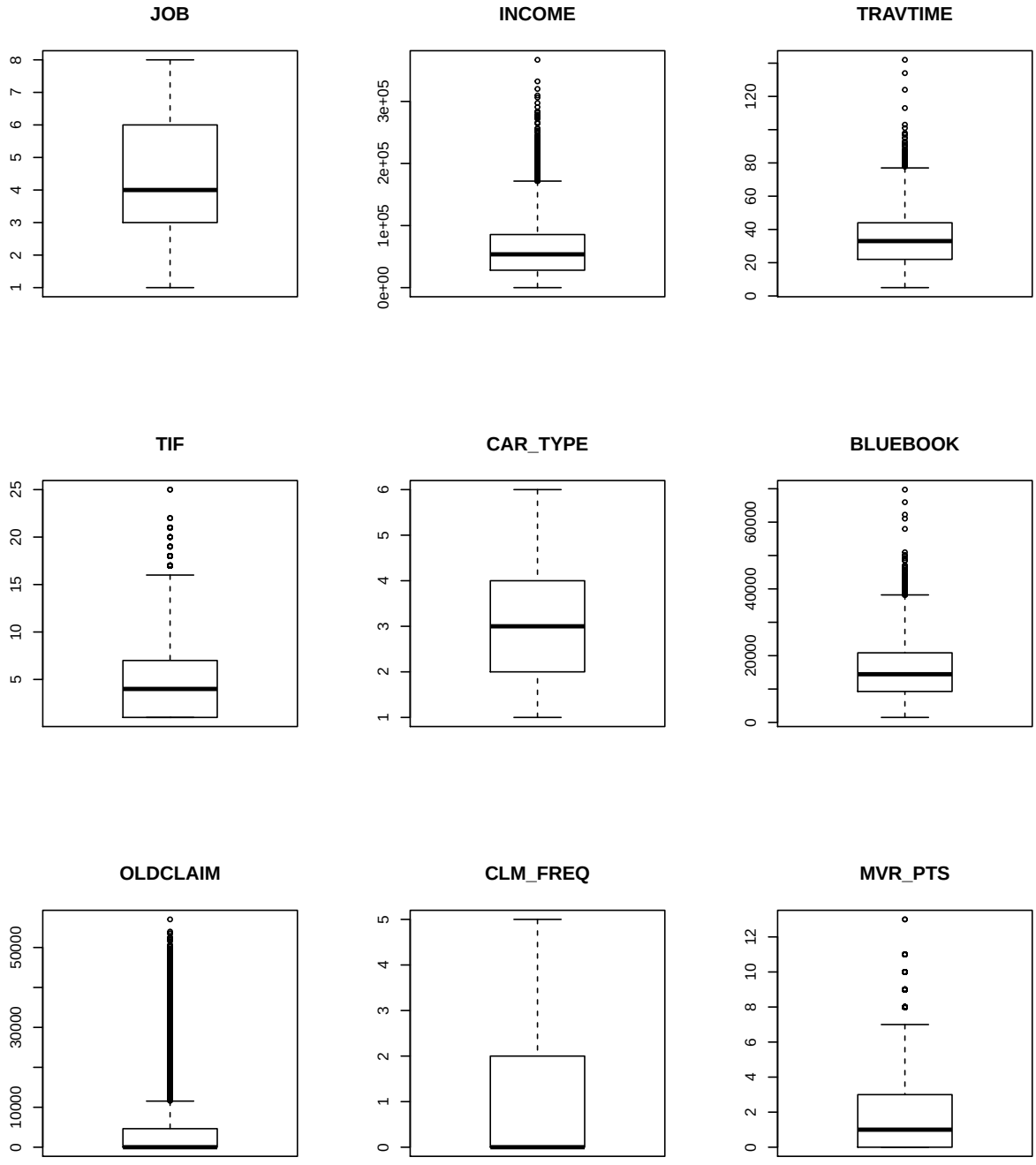
| Attribute | Binary Choice | | | Choice of Categories | | |
|---|---|---|---|---|---|---|
| | New Variable Name | GOOD = 1 | BAD = 0 | New Variable Name | Legend | Sort Order |
| INDEX | | | | | | |
| TARGET FLAG | | | | | | |
| TARGET AMT | | | | | | |
| KIDSDRIV | | | | | | |
| AGE | | | | | | |
| HOMEKIDS | | | | | | |
| YOJ | | | | | | |
| PARENT1 | blnPARENT1 | PARENT | NOT A PARENT | | | |
| HOME VAL | | | | | | |
| MSTATUS | blnMSTATUS | MARRIED | SINGLE | | | |
| SEX | blnSEX | FEMALE | MALE | | | |
| EDUCATION | | | | intEDUCATION | 1-Primary, 2-Secondary, 3-Bachelors, 4-Masters, 5-PhD | Ascending by Avg Income |
| JOB | | | | intJOB | 1-Student, 2-Home Maker, 3-Clerical, 4-Tradesperson, 5-Professional, 6-Manager, 7-Lawyer, 8-Doctor | Ascending by Avg Income |
| INCOME | | | | | | |
| TRAVTIME | | | | | | |
| CAR USE | blnCAR_USE | PRIVATE | COMMERCIAL | | | |
| TIF | | | | | | |
| CAR TYPE | | | | intCAR_TYPE | 1-Sports Car, 2-SUV, 3-Pickup, 4-Minivan, 5-Van, 6-Panel Truck | Ascending by BlueBook |
| BLUEBOOK | | | | | | |
| RED CAR | bln_NOT_RED_CAR | NOT RED CAR | RED CAR | | | |
| OLDCLAIM | | | | | | |
| CLM FREQ | | | | | | |
| REVOKED | bln_NOT_REVOKED | NOT REVOKED | LICENSE REVOKED | | | |
| MVR PTS | | | | | | |
| CAR AGE | | | | | | |
| URBANICITY | blnURBANICITY | URBAN | RURAL | | | |
| | | | | | | |
| | Everything Good is pointing the same way...to 1 | | | | The higher the number, the higher the wage / property value | |

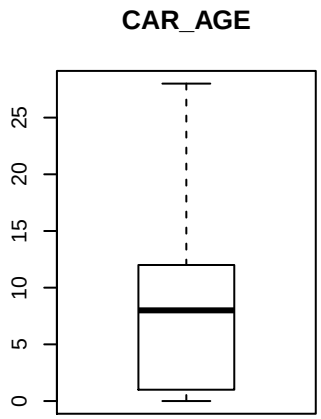
Figure 1: Legend for Categorical Variables

| blnCAR_USE | blnNOT_RED_CAR | blnNOT_REVOKED |
|---|---|---|
| Min. :0.0000 | Min. :0.0000 | Min. :0.0000 |
| 1st Qu.:0.0000 | 1st Qu.:0.0000 | 1st Qu.:1.0000 |
| Median :1.0000 | Median :1.0000 | Median :1.0000 |
| Mean :0.6288 | Mean :0.7086 | Mean :0.8775 |
| 3rd Qu.:1.0000 | 3rd Qu.:1.0000 | 3rd Qu.:1.0000 |
| Max. :1.0000 | Max. :1.0000 | Max. :1.0000 |

| blnURBANICITY | intEDUCATION | intJOB | intCAR_TYPE |
|---|---|---|---|
| Min. :0.0000 | Min. :1.000 | Min. :1.000 | Min. :1.000 |
| 1st Qu.:1.0000 | 1st Qu.:2.000 | 1st Qu.:3.000 | 1st Qu.:2.000 |
| Median :1.0000 | Median :3.000 | Median :4.000 | Median :3.000 |
| Mean :0.7955 | Mean :2.801 | Mean :4.373 | Mean :3.192 |
| 3rd Qu.:1.0000 | 3rd Qu.:4.000 | 3rd Qu.:6.000 | 3rd Qu.:4.000 |
| Max. :1.0000 | Max. :5.000 | Max. :8.000 | Max. :6.000 |





JOB
1 2 3 4 5 6 7 8
INCOME
0e+00 1e+05 2e+05 3e+05
TRAVTIME
0 20 40 60 80 120
TIF
5 10 15 20 25
CAR_TYPE
1 2 3 4 5 6
BLUEBOOK
0 20000 40000 60000
OLDCLAIM
0 10000 30000 50000
CLM_FREQ
0 1 2 3 4 5
MVR_PTS
0 2 4 6 8 10 12




Below is a correlation table illustrating the collinearity of each variable to the others.

| | KIDSDRIV | AGE | HOMEKIDS | YOJ | INCOME | HOME_VAL |
|---|---|---|---|---|---|---|
| KIDSDRIV | 1.0000000 | -0.0747945 | 0.4640152 | 0.0444575 | -0.0482671 | -0.0208227 |
| AGE | -0.0747945 | 1.0000000 | -0.4458941 | 0.1328668 | 0.1815202 | 0.2113242 |
| HOMEKIDS | 0.4640152 | -0.4458941 | 1.0000000 | 0.0846473 | -0.1597844 | -0.1100693 |
| YOJ | 0.0444575 | 0.1328668 | 0.0846473 | 1.0000000 | 0.2821729 | 0.2737353 |
| INCOME | -0.0482671 | 0.1815202 | -0.1597844 | 0.2821729 | 1.0000000 | 0.5829583 |
| HOME_VAL | -0.0208227 | 0.2113242 | -0.1100693 | 0.2737353 | 0.5829583 | 1.0000000 |
| TRAVTIME | 0.0084473 | 0.0056488 | -0.0072456 | -0.0148446 | -0.0513022 | -0.0326337 |
| BLUEBOOK | -0.0215493 | 0.1654997 | -0.1078936 | 0.1376966 | 0.4311394 | 0.2583793 |
| TIF | -0.0019887 | -0.0001610 | 0.0118133 | 0.0235317 | -0.0013089 | 0.0005486 |
| OLDCLAIM | 0.0204027 | -0.0291980 | 0.0299110 | -0.0047492 | -0.0416781 | -0.0671853 |
| CLM_FREQ | 0.0370629 | -0.0241248 | 0.0293493 | -0.0282020 | -0.0448261 | -0.0905444 |
| MVR_PTS | 0.0535664 | -0.0722487 | 0.0606013 | -0.0372666 | -0.0599870 | -0.0763485 |
| CAR_AGE | -0.0562803 | 0.1781622 | -0.1563566 | 0.0659152 | 0.4166759 | 0.2209701 |
| blnPARENT1 | 0.1966038 | -0.3145732 | 0.4492740 | -0.0513543 | -0.0723519 | -0.2581639 |
| blnMSTATUS | 0.0424609 | 0.0909813 | 0.0435259 | 0.1451118 | -0.0331084 | 0.4534094 |
| blnSEX | 0.0459344 | -0.0660283 | 0.1115114 | -0.0849984 | -0.1108873 | -0.0732736 |
| blnCAR_USE | -0.0014216 | 0.0329641 | 0.0044583 | -0.0225160 | -0.0844900 | -0.0286943 |
| blnNOT_RED_CAR | 0.0436382 | -0.0187895 | 0.0681480 | -0.0503551 | -0.0604157 | -0.0136167 |
| blnNOT_REVOKED | -0.0430620 | 0.0383623 | -0.0451156 | 0.0115272 | 0.0195288 | 0.0470423 |
| blnURBANICITY | -0.0371236 | 0.0512566 | -0.0634829 | 0.0807141 | 0.2076124 | 0.1211302 |
| intEDUCATION | -0.0714891 | 0.2446640 | -0.2036956 | 0.0854670 | 0.6033013 | 0.3488034 |
| intJOB | -0.0691631 | 0.2531767 | -0.2204296 | 0.3277284 | 0.6833423 | 0.4410682 |
| intCAR_TYPE | -0.0317930 | 0.0449765 | -0.1052492 | 0.1041217 | 0.2824104 | 0.1642502 |

| | TRAVTIME | BLUEBOOK | TIF | OLDCLAIM | CLM_FREQ | MVR_PTS |
|---|---|---|---|---|---|---|
| KIDSDRIV | 0.0084473 | -0.0215493 | -0.0019887 | 0.0204027 | 0.0370629 | 0.0535664 |
| AGE | 0.0056488 | 0.1654997 | -0.0001610 | -0.0291980 | -0.0241248 | -0.0722487 |
| HOMEKIDS | -0.0072456 | -0.1078936 | 0.0118133 | 0.0299110 | 0.0293493 | 0.0606013 |
| YOJ | -0.0148446 | 0.1376966 | 0.0235317 | -0.0047492 | -0.0282020 | -0.0372666 |
| INCOME | -0.0513022 | 0.4311394 | -0.0013089 | -0.0416781 | -0.0448261 | -0.0599870 |

| | TRAVTIME | BLUEBOOK | TIF | OLDCLAIM | CLM_FREQ | MVR_PTS |
|---|---|---|---|---|---|---|
| HOME_VAL | -0.0326337 | 0.2583793 | 0.0005486 | -0.0671853 | -0.0905444 | -0.0763485 |
| TRAVTIME | 1.0000000 | -0.0170013 | -0.0116046 | -0.0192672 | 0.0065602 | 0.0105985 |
| BLUEBOOK | -0.0170013 | 1.0000000 | -0.0054246 | -0.0295176 | -0.0363415 | -0.0391308 |
| TIF | -0.0116046 | -0.0054246 | 1.0000000 | -0.0219582 | -0.0230230 | -0.0410457 |
| OLDCLAIM | -0.0192672 | -0.0295176 | -0.0219582 | 1.0000000 | 0.4951308 | 0.2644850 |
| CLM_FREQ | 0.0065602 | -0.0363415 | -0.0230230 | 0.4951308 | 1.0000000 | 0.3966384 |
| MVR_PTS | 0.0105985 | -0.0391308 | -0.0410457 | 0.2644850 | 0.3966384 | 1.0000000 |
| CAR_AGE | -0.0397659 | 0.1911752 | 0.0095785 | -0.0137464 | -0.0078990 | -0.0161019 |
| blnPARENT1 | -0.0237406 | -0.0504582 | -0.0019519 | 0.0346893 | 0.0487424 | 0.0684526 |
| blnMSTATUS | 0.0102483 | -0.0077430 | -0.0007411 | -0.0459198 | -0.0693289 | -0.0479670 |
| blnSEX | 0.0046177 | -0.0624182 | -0.0061012 | 0.0000909 | -0.0122335 | 0.0073444 |
| blnCAR_USE | -0.0248054 | -0.2250993 | -0.0001161 | -0.0357676 | -0.0814907 | -0.0680838 |
| blnNOT_RED_CAR | -0.0039658 | -0.0218660 | 0.0008717 | -0.0138214 | -0.0260815 | -0.0060406 |
| blnNOT_REVOKED | 0.0121153 | 0.0257973 | 0.0318415 | -0.4181035 | -0.0530499 | -0.0531731 |
| blnURBANICITY | -0.1660047 | 0.0877412 | 0.0071310 | 0.1510826 | 0.2391246 | 0.1502433 |
| intEDUCATION | -0.0572046 | 0.2787819 | 0.0019466 | -0.0234187 | -0.0126025 | -0.0338609 |
| intJOB | -0.0847015 | 0.3025553 | 0.0083580 | -0.0254013 | -0.0194655 | -0.0357795 |
| intCAR_TYPE | -0.0113547 | 0.5177917 | 0.0010737 | -0.0260373 | -0.0209609 | -0.0263118 |

| | CAR_AGE | blnPARENT1 | blnMSTATUS | blnSEX | blnCAR_USE | blnNOT_RED_CAI |
|---|---|---|---|---|---|---|
| KIDSDRIV | -0.0562803 | 0.1966038 | 0.0424609 | 0.0459344 | -0.0014216 | 0.043638 |
| AGE | 0.1781622 | -0.3145732 | 0.0909813 | -0.0660283 | 0.0329641 | -0.018789 |
| HOMEKIDS | -0.1563566 | 0.4492740 | 0.0435259 | 0.1115114 | 0.0044583 | 0.068148 |
| YOJ | 0.0659152 | -0.0513543 | 0.1451118 | -0.0849984 | -0.0225160 | -0.050355 |
| INCOME | 0.4166759 | -0.0723519 | -0.0331084 | -0.1108873 | -0.0844900 | -0.060415 |
| HOME_VAL | 0.2209701 | -0.2581639 | 0.4534094 | -0.0732736 | -0.0286943 | -0.013616 |
| TRAVTIME | -0.0397659 | -0.0237406 | 0.0102483 | 0.0046177 | -0.0248054 | -0.003965 |
| BLUEBOOK | 0.1911752 | -0.0504582 | -0.0077430 | -0.0624182 | -0.2250993 | -0.021866 |
| TIF | 0.0095785 | -0.0019519 | -0.0007411 | -0.0061012 | -0.0001161 | 0.000871 |
| OLDCLAIM | -0.0137464 | 0.0346893 | -0.0459198 | 0.0000909 | -0.0357676 | -0.013821 |
| CLM_FREQ | -0.0078990 | 0.0487424 | -0.0693289 | -0.0122335 | -0.0814907 | -0.026081 |
| MVR_PTS | -0.0161019 | 0.0684526 | -0.0479670 | 0.0073444 | -0.0680838 | -0.006040 |
| CAR_AGE | 1.0000000 | -0.0640196 | -0.0366952 | -0.0240469 | 0.0659686 | -0.021603 |
| blnPARENT1 | -0.0640196 | 1.0000000 | -0.4772281 | 0.0737837 | -0.0061940 | 0.042085 |
| blnMSTATUS | -0.0366952 | -0.4772281 | 1.0000000 | 0.0042094 | 0.0209315 | 0.019286 |
| blnSEX | -0.0240469 | 0.0737837 | 0.0042094 | 1.0000000 | 0.2791190 | 0.666620 |
| blnCAR_USE | 0.0659686 | -0.0061940 | 0.0209315 | 0.2791190 | 1.0000000 | 0.189997 |
| blnNOT_RED_CAR | -0.0216037 | 0.0420856 | 0.0192863 | 0.6666207 | 0.1899973 | 1.000000 |
| blnNOT_REVOKED | 0.0083538 | -0.0497187 | 0.0432305 | -0.0014340 | 0.0168969 | 0.002972 |
| blnURBANICITY | 0.1672048 | -0.0222096 | -0.0025618 | -0.0531645 | 0.0204630 | -0.046346 |
| intEDUCATION | 0.6935018 | -0.0815011 | -0.0398657 | -0.0331674 | 0.0405745 | -0.019545 |
| intJOB | 0.4990187 | -0.0901294 | -0.0298900 | -0.1177786 | 0.0933126 | -0.068912 |
| intCAR_TYPE | 0.1139984 | -0.0590901 | -0.0164472 | -0.6475502 | -0.3579857 | -0.431057 |

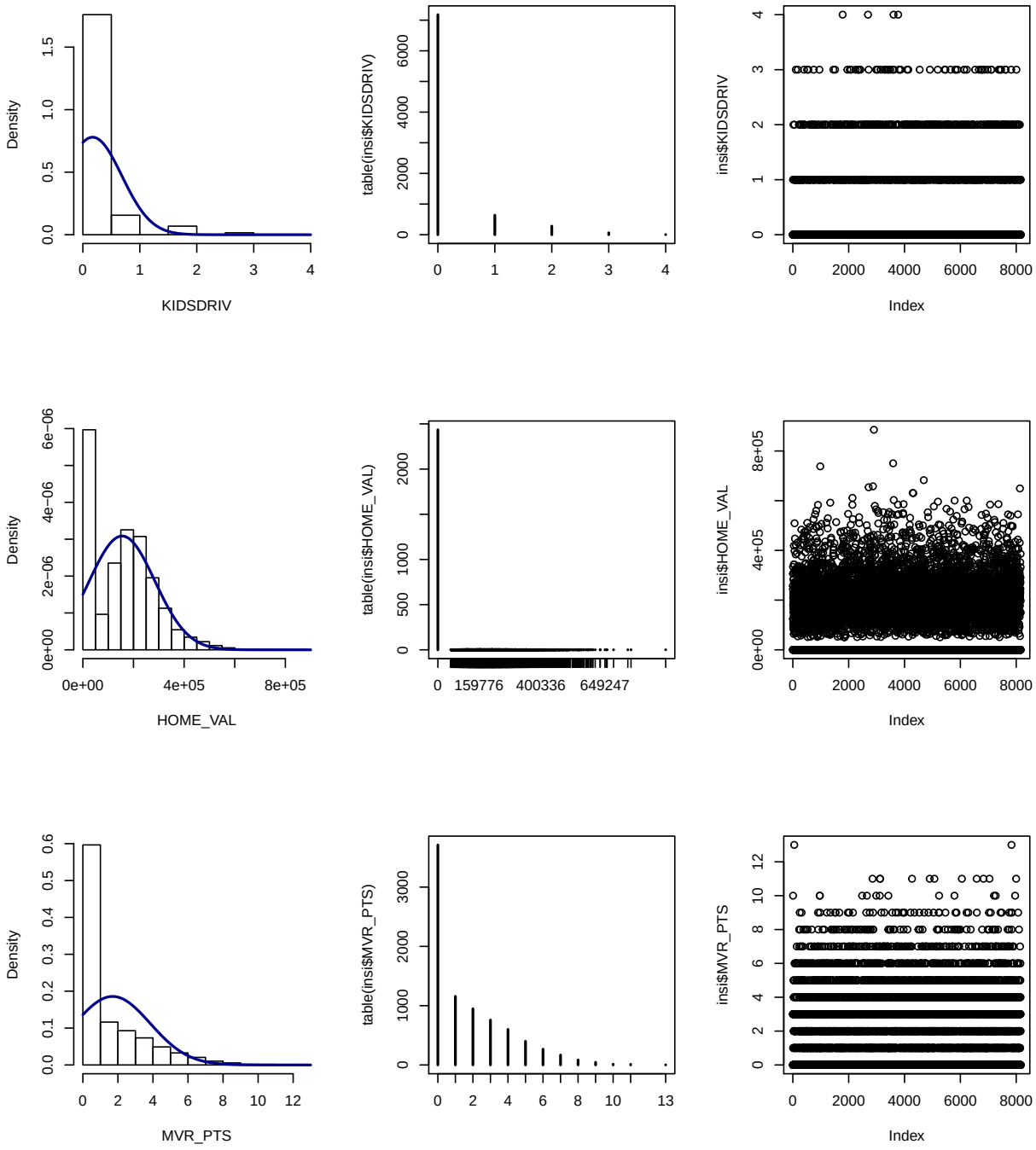
Here are the results from an analysis of the predictor variable correlations:

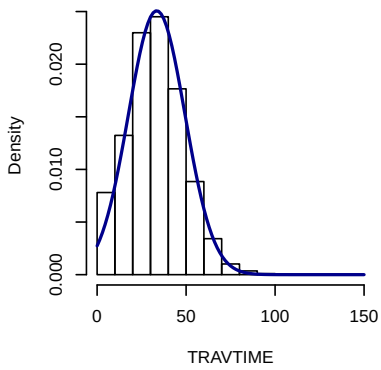
The are no strong correlations (>70%) between predictor variables, not enough to allow consideration of removing a variable from the model based on a high correlation with another variable. There is some moderate correlation (30-50%) between some variable highlighting obvious relationships such as HOMEKIDS-KIDSDRIV, HOME_VAL-INCOME, EDUCATION-INCOME, JOB-INCOME, CAR_TYPE-BLUEBOOK,

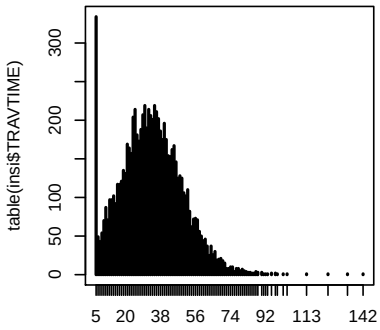
CLM_FREQ-OLDCLAIM, and MVR_PTS-CLM_FREQ.

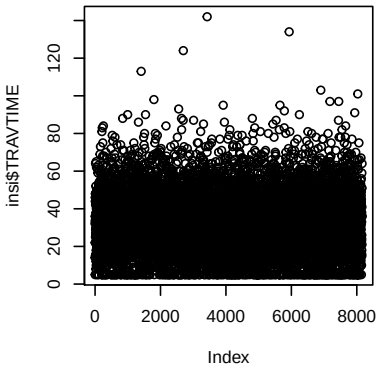
Based on an analysis of the box plots, the following variables have some outliers that may, or may not, exert influence on the regression results: - KIDSDRIV, HOME_VAL, TRAVTIME, MVR_PTS, AGE, INCOME, BLUEBOOK, OLDCLAIM

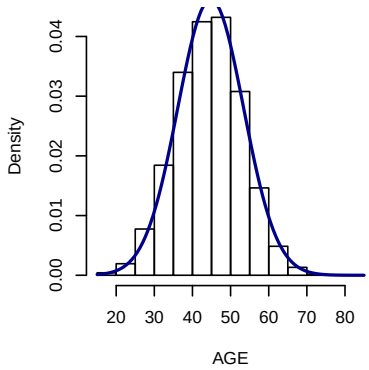
We'll next look at these variables more closely, starting with their histograms and frequency counts to better understand the nature of their distribution.

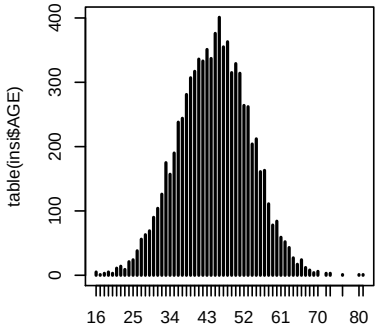


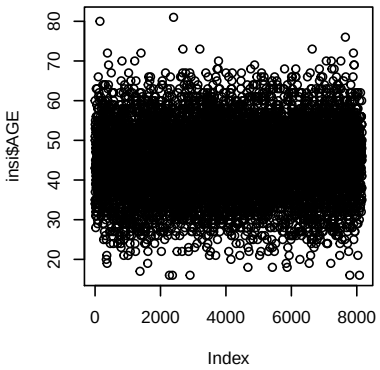

Density
0.000
0.010
0.020
0
50
100
150
TRAVTIME

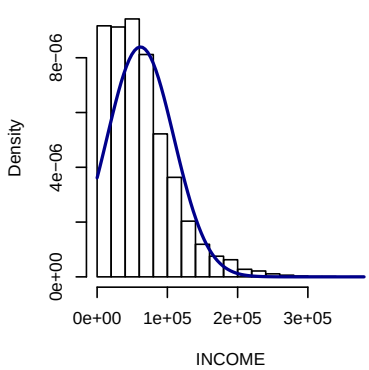

table(insi$TRAVTIME)
0
50
100
200
300
5
20
38
56
74
92
113
142

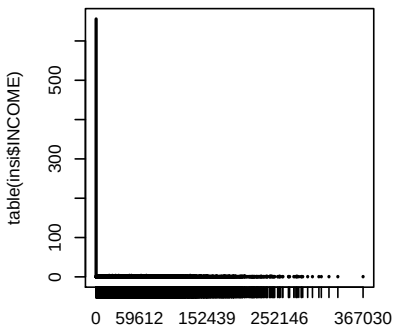

insi$TRAVTIME
0
20
40
60
80
120
0
2000
4000
6000
8000
Index

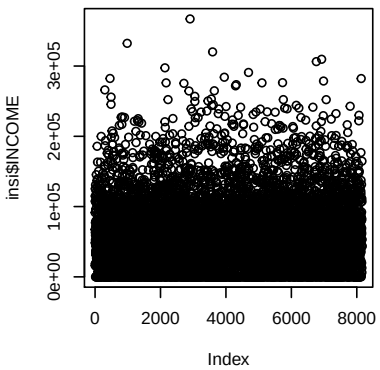

Density
0.00
0.01
0.02
0.03
0.04
20
30
40
50
60
70
80
AGE

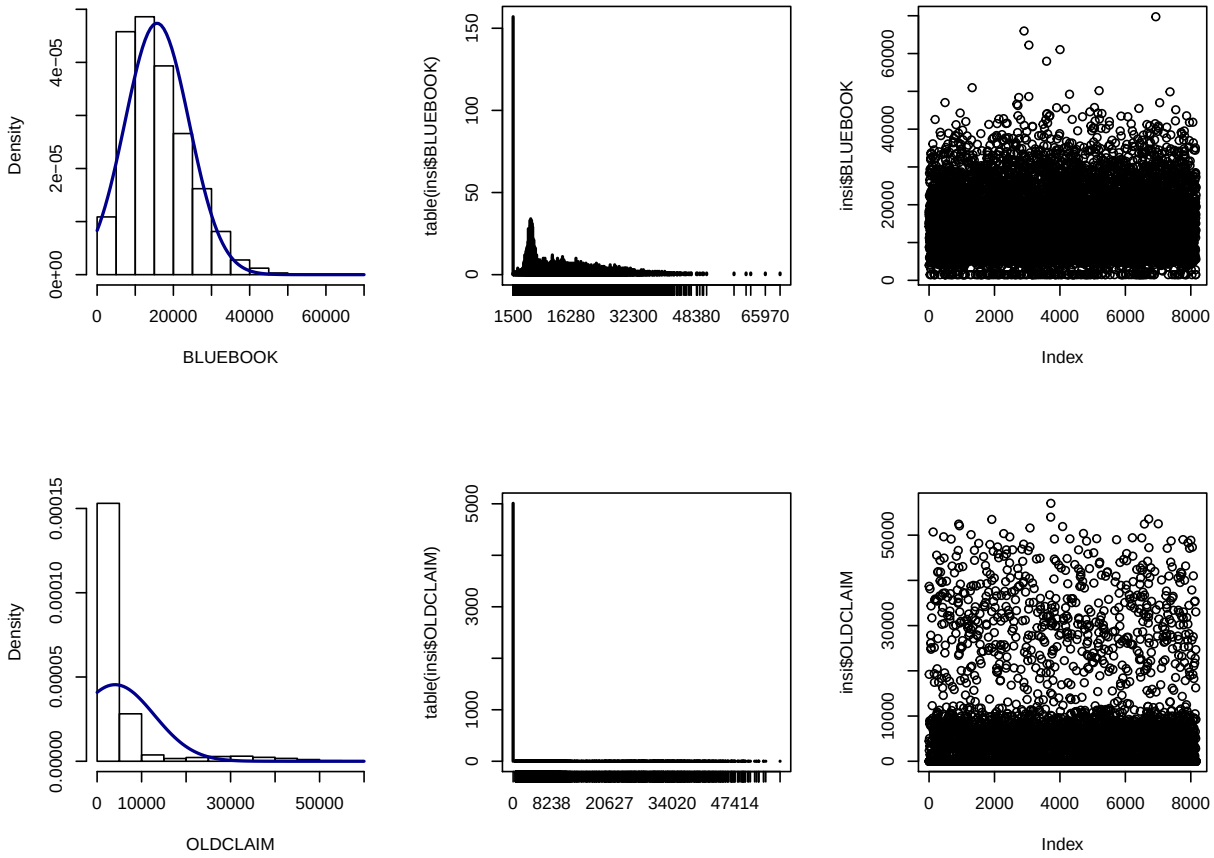

table(insi$AGE)
0
100
200
300
400
16
25
34
43
52
61
70
80


insi$AGE
20
30
40
50
60
70
80
0
2000
4000
6000
8000
Index


Density
0e+00
4e−06
8e−06
0e+00
1e+05
2e+05
3e+05
INCOME


table(insi$INCOME)
0
100
300
500
0
59612
152439
252146
367030


insi$INCOME
0e+00
1e+05
2e+05
3e+05
0
2000
4000
6000
8000
Index




The analysis of the distributions for these variables show varying degrees of skewness, except for the AGE variable, which shows a fairly normal distribution.

For the logistic regression analysis, we would like to remove as much of the skewness as possible from the candidate predictor variables. A transformation analysis was performed and a log transformation of most of the skewed variables result in a near-normal distribution consequently, for the logistic regression model, we will use the log value of the following variables for the modeling: OLDCLAIM, BLUEBOOK, TRAVTIME, HOME_VAL.

# Data Preparation

## Data Preparation

As stated earlier, most of preparation was done prior to a thorough exploration. The reason being, is that the missing value was such a hinderance to proper exploration.

Here we perform the log transformation of the variables identifed earlier with high skewness. One of the variables, OLDCLAIM, has such a distorted distribution that the log transformation will not be sufficient to make the variable a viable candidate for our model. For the OLDCLAIM variable, the difference between the median and the mean is so large we will not attempt to use the OLDCLAIM variable.

```
# perform the log transformation for the selected predictor variables and add to the data frame
# no variables used in the log transform can have a value of 0

HOME_VAL.median <- median(insi$HOME_VAL)
BLUEBOOK.median <- median(insi$BLUEBOOK)
TRAVTIME.median <- median(insi$TRAVTIME)

insi$HOME_VAL[insi$HOME_VAL<=0] <- HOME_VAL.median
insi$BLUEBOOK[insi$BLUEBOOK<=0] <- BLUEBOOK.median
insi$TRAVTIME[insi$TRAVTIME<=0] <- TRAVTIME.median

insi$BLUEBOOK <- log(insi$BLUEBOOK)
insi$TRAVTIME <- log(insi$TRAVTIME)
insi$HOME_VAL <- log(insi$HOME_VAL)
```

# Build Models

## Build Models

One method of developing multiple regression models is to take a stepwise approach. To accomplish this, we combine our knowledge from the data exploration above with logistic regression. Univariate Logistic Regression is a useful method to understand how each predictor variable interacts individually with the target (response) variable. Looking at various statistics, we determine which variable may impact our target the most.

### Logistic Regression Model

In this model-and in all the models- we set aside 20% of the training data and use 80% to train the model we then use the model to predict the outcome of the remaining 20% of the data.

In this scenario we attempt to create the simplest model possible by using only one variable - the one that provides the highest overall AUC (performance) by itself. We calculate AUC for each variable separately and then select the highest result.

| var | p_val | aic | auc |
|---|---|---|---|
| KIDSDRIV | 0.0000000 | 7496.182 | 0.5326796 |
| HOMEKIDS | 0.0000000 | 7480.467 | 0.5643226 |
| AGE | 0.0000000 | 7489.887 | 0.5677821 |
| blnPARENT1 | 0.0000000 | 7415.589 | 0.5618229 |
| BLUEBOOK | 0.0000000 | 7478.621 | 0.5664127 |
| INCOME | 0.0000000 | 7415.094 | 0.5896893 |
| CAR_AGE | 0.0000000 | 7492.059 | 0.5664136 |
| intEDUCATION | 0.0000000 | 7446.171 | 0.5818674 |
| intJOB | 0.0000000 | 7428.378 | 0.5876100 |
| blnMSTATUS | 0.0000000 | 7441.521 | 0.5704225 |
| intCAR_TYPE | 0.0000000 | 7520.419 | 0.5550436 |
| CLM_FREQ | 0.0000000 | 7272.338 | 0.6336553 |
| TIF | 0.0000000 | 7510.860 | 0.5383267 |
| MVR_PTS | 0.0000000 | 7272.810 | 0.6231339 |
| blnSEX | 0.0505324 | 7559.781 | 0.5050840 |
| blnCAR_USE | 0.0000000 | 7440.294 | 0.5881564 |
| blnNOT_RED_CAR | 0.5948267 | 7563.328 | 0.5044923 |

| var | p_val | aic | auc |
|---|---|---|---|
| blnNOT_REVOKED | 0.0000000 | 7416.317 | 0.5487418 |
| blnURBANICITY | 0.0000000 | 7163.720 | 0.6049599 |
| TRAVTIME | 0.0000004 | 7537.158 | 0.5239225 |
| HOME_VAL | 0.0000000 | 7422.825 | 0.5736370 |

The highest AUC value obtained is .63, from the variable CLM_FREQ (Claim Frequency), indicating that clients with higher past claim incidents are more likely to have claims in the future. However, .63 is not a long ways from .50, so this model is not very strong.

Next we will derive a logistic regression model by stepping backward from using all candidate variables and arriving at the variable set that maximizes the AUC value.

**MODEL 1 - Backward regression starting with all variables**

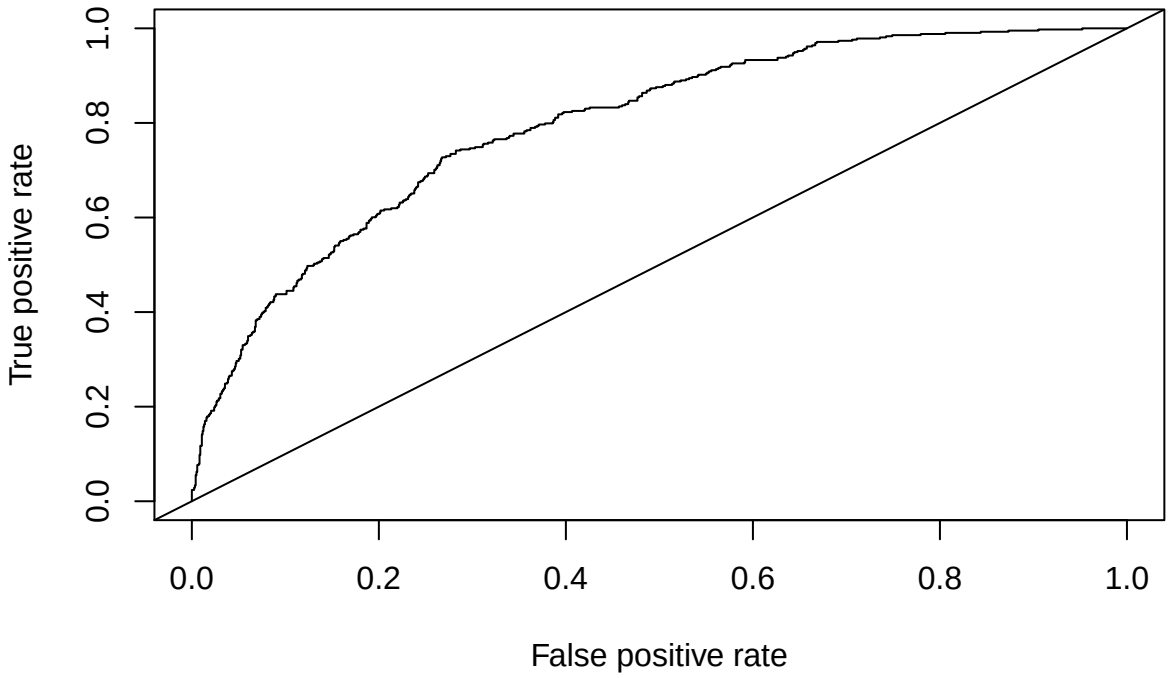
```
## Start:  AIC=5996.36
## TARGET_FLAG ~ (TARGET_AMT + KIDSDRIV + AGE + HOMEKIDS + YOJ +
##     INCOME + HOME_VAL + TRAVTIME + BLUEBOOK + TIF + OLDCLAIM +
##     CLM_FREQ + MVR_PTS + CAR_AGE + blnPARENT1 + blnMSTATUS +
##     blnSEX + blnCAR_USE + blnNOT_RED_CAR + blnNOT_REVOKED + blnURBANICITY +
##     intEDUCATION + intJOB + intCAR_TYPE) - TARGET_AMT - OLDCLAIM
##
##                   Df Deviance    AIC
## - AGE              1   5950.4 5994.4
## - blnSEX           1   5950.6 5994.6
## - blnNOT_RED_CAR   1   5950.6 5994.6
## - CAR_AGE          1   5950.9 5994.9
## - YOJ              1   5951.0 5995.0
## - HOMEKIDS         1   5951.9 5995.9
## <none>                 5950.4 5996.4
## - intJOB           1   5954.2 5998.2
## - HOME_VAL         1   5954.6 5998.6
## - intEDUCATION     1   5957.3 6001.3
## - blnPARENT1       1   5957.7 6001.7
## - INCOME           1   5958.9 6002.9
## - intCAR_TYPE      1   5965.8 6009.8
## - BLUEBOOK         1   5968.8 6012.8
## - CLM_FREQ         1   5981.0 6025.0
## - KIDSDRIV         1   5981.8 6025.8
## - TIF              1   6003.0 6047.0
## - blnMSTATUS       1   6007.4 6051.4
## - MVR_PTS          1   6009.4 6053.4
## - TRAVTIME         1   6013.9 6057.9
## - blnNOT_REVOKED   1   6023.3 6067.3
## - blnCAR_USE       1   6124.6 6168.6
## - blnURBANICITY    1   6440.3 6484.3
##
## Step:  AIC=5994.41
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + HOME_VAL +
##     TRAVTIME + BLUEBOOK + TIF + CLM_FREQ + MVR_PTS + CAR_AGE +
##     blnPARENT1 + blnMSTATUS + blnSEX + blnCAR_USE + blnNOT_RED_CAR +
##     blnNOT_REVOKED + blnURBANICITY + intEDUCATION + intJOB +
##     intCAR_TYPE
```

```
##
##                   Df Deviance    AIC
## - blnNOT_RED_CAR   1   5950.7 5992.7
## - blnSEX           1   5950.7 5992.7
## - CAR_AGE          1   5951.0 5993.0
## - YOJ              1   5951.1 5993.1
## - HOMEKIDS         1   5952.4 5994.4
## <none>                 5950.4 5994.4
## - intJOB           1   5954.3 5996.3
## - HOME_VAL         1   5954.8 5996.8
## - intEDUCATION     1   5957.5 5999.5
## - blnPARENT1       1   5958.0 6000.0
## - INCOME           1   5958.9 6000.9
## - intCAR_TYPE      1   5965.8 6007.8
## - BLUEBOOK         1   5969.4 6011.4
## - CLM_FREQ         1   5981.1 6023.1
## - KIDSDRIV         1   5982.5 6024.5
## - TIF              1   6003.0 6045.0
## - blnMSTATUS       1   6007.6 6049.6
## - MVR_PTS          1   6009.7 6051.7
## - TRAVTIME         1   6013.9 6055.9
## - blnNOT_REVOKED   1   6023.4 6065.4
## - blnCAR_USE       1   6124.9 6166.9
## - blnURBANICITY    1   6441.0 6483.0
##
## Step:  AIC=5992.66
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + HOME_VAL +
##     TRAVTIME + BLUEBOOK + TIF + CLM_FREQ + MVR_PTS + CAR_AGE +
##     blnPARENT1 + blnMSTATUS + blnSEX + blnCAR_USE + blnNOT_REVOKED +
##     blnURBANICITY + intEDUCATION + intJOB + intCAR_TYPE
##
##                   Df Deviance    AIC
## - blnSEX           1   5950.7 5990.7
## - CAR_AGE          1   5951.2 5991.2
## - YOJ              1   5951.4 5991.4
## <none>                 5950.7 5992.7
## - HOMEKIDS         1   5952.7 5992.7
## - intJOB           1   5954.6 5994.6
## - HOME_VAL         1   5955.1 5995.1
## - intEDUCATION     1   5957.8 5997.8
## - blnPARENT1       1   5958.2 5998.2
## - INCOME           1   5959.2 5999.2
## - intCAR_TYPE      1   5966.0 6006.0
## - BLUEBOOK         1   5969.8 6009.8
## - CLM_FREQ         1   5981.4 6021.4
## - KIDSDRIV         1   5982.6 6022.6
## - TIF              1   6003.2 6043.2
## - blnMSTATUS       1   6007.9 6047.9
## - MVR_PTS          1   6010.0 6050.0
## - TRAVTIME         1   6014.1 6054.1
## - blnNOT_REVOKED   1   6023.6 6063.6
## - blnCAR_USE       1   6125.0 6165.0
## - blnURBANICITY    1   6441.1 6481.1
##
```

```
## Step:  AIC=5990.74
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + HOME_VAL +
##     TRAVTIME + BLUEBOOK + TIF + CLM_FREQ + MVR_PTS + CAR_AGE +
##     blnPARENT1 + blnMSTATUS + blnCAR_USE + blnNOT_REVOKED + blnURBANICITY +
##     intEDUCATION + intJOB + intCAR_TYPE
##
##                   Df Deviance    AIC
## - CAR_AGE          1   5951.3 5989.3
## - YOJ              1   5951.5 5989.5
## <none>                 5950.7 5990.7
## - HOMEKIDS         1   5952.8 5990.8
## - intJOB           1   5954.8 5992.8
## - HOME_VAL         1   5955.2 5993.2
## - intEDUCATION     1   5957.8 5995.8
## - blnPARENT1       1   5958.3 5996.3
## - INCOME           1   5959.2 5997.2
## - BLUEBOOK         1   5971.7 6009.7
## - CLM_FREQ         1   5981.5 6019.5
## - intCAR_TYPE      1   5981.9 6019.9
## - KIDSDRIV         1   5982.7 6020.7
## - TIF              1   6003.3 6041.3
## - blnMSTATUS       1   6008.0 6046.0
## - MVR_PTS          1   6010.1 6048.1
## - TRAVTIME         1   6014.2 6052.2
## - blnNOT_REVOKED   1   6023.6 6061.6
## - blnCAR_USE       1   6126.6 6164.6
## - blnURBANICITY    1   6441.2 6479.2
##
## Step:  AIC=5989.28
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + HOME_VAL +
##     TRAVTIME + BLUEBOOK + TIF + CLM_FREQ + MVR_PTS + blnPARENT1 +
##     blnMSTATUS + blnCAR_USE + blnNOT_REVOKED + blnURBANICITY +
##     intEDUCATION + intJOB + intCAR_TYPE
##
##                   Df Deviance    AIC
## - YOJ              1   5952.0 5988.0
## <none>                 5951.3 5989.3
## - HOMEKIDS         1   5953.4 5989.4
## - intJOB           1   5955.6 5991.6
## - HOME_VAL         1   5955.7 5991.7
## - blnPARENT1       1   5958.9 5994.9
## - INCOME           1   5959.8 5995.8
## - intEDUCATION     1   5964.8 6000.8
## - BLUEBOOK         1   5972.2 6008.2
## - CLM_FREQ         1   5981.9 6017.9
## - intCAR_TYPE      1   5982.5 6018.5
## - KIDSDRIV         1   5983.3 6019.3
## - TIF              1   6004.1 6040.1
## - blnMSTATUS       1   6008.4 6044.4
## - MVR_PTS          1   6010.6 6046.6
## - TRAVTIME         1   6014.6 6050.6
## - blnNOT_REVOKED   1   6024.3 6060.3
## - blnCAR_USE       1   6128.4 6164.4
## - blnURBANICITY    1   6441.8 6477.8
```

```
##
## Step:  AIC=5988.01
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + INCOME + HOME_VAL + TRAVTIME +
##     BLUEBOOK + TIF + CLM_FREQ + MVR_PTS + blnPARENT1 + blnMSTATUS +
##     blnCAR_USE + blnNOT_REVOKED + blnURBANICITY + intEDUCATION +
##     intJOB + intCAR_TYPE
##
##                  Df Deviance    AIC
## - HOMEKIDS        1   5953.8 5987.8
## <none>                5952.0 5988.0
## - HOME_VAL        1   5956.6 5990.6
## - intJOB          1   5957.8 5991.8
## - blnPARENT1      1   5959.7 5993.7
## - INCOME          1   5961.0 5995.0
## - intEDUCATION    1   5964.8 5998.8
## - BLUEBOOK        1   5973.3 6007.3
## - CLM_FREQ        1   5982.6 6016.6
## - intCAR_TYPE     1   5983.2 6017.2
## - KIDSDRIV        1   5984.4 6018.4
## - TIF             1   6005.2 6039.2
## - blnMSTATUS      1   6011.1 6045.1
## - MVR_PTS         1   6011.8 6045.8
## - TRAVTIME        1   6015.2 6049.2
## - blnNOT_REVOKED  1   6025.0 6059.0
## - blnCAR_USE      1   6128.6 6162.6
## - blnURBANICITY   1   6441.9 6475.9
##
## Step:  AIC=5987.76
## TARGET_FLAG ~ KIDSDRIV + INCOME + HOME_VAL + TRAVTIME + BLUEBOOK +
##     TIF + CLM_FREQ + MVR_PTS + blnPARENT1 + blnMSTATUS + blnCAR_USE +
##     blnNOT_REVOKED + blnURBANICITY + intEDUCATION + intJOB +
##     intCAR_TYPE
##
##                  Df Deviance    AIC
## <none>                5953.8 5987.8
## - HOME_VAL        1   5958.9 5990.9
## - intJOB          1   5960.3 5992.3
## - INCOME          1   5962.2 5994.2
## - intEDUCATION    1   5967.1 5999.1
## - blnPARENT1      1   5969.6 6001.6
## - BLUEBOOK        1   5975.5 6007.5
## - CLM_FREQ        1   5984.5 6016.5
## - intCAR_TYPE     1   5985.2 6017.2
## - KIDSDRIV        1   5999.2 6031.2
## - TIF             1   6006.8 6038.8
## - blnMSTATUS      1   6012.5 6044.5
## - MVR_PTS         1   6013.8 6045.8
## - TRAVTIME        1   6016.4 6048.4
## - blnNOT_REVOKED  1   6027.5 6059.5
## - blnCAR_USE      1   6130.2 6162.2
## - blnURBANICITY   1   6443.5 6475.5
##
## Call:
```

```
## glm(formula = TARGET_FLAG ~ KIDSDRIV + INCOME + HOME_VAL + TRAVTIME +
##     BLUEBOOK + TIF + CLM_FREQ + MVR_PTS + blnPARENT1 + blnMSTATUS +
##     blnCAR_USE + blnNOT_REVOKED + blnURBANICITY + intEDUCATION +
##     intJOB + intCAR_TYPE, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4270  -0.7250  -0.4193   0.6678   2.9515
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     4.301e+00  1.559e+00   2.758 0.005822 **
## KIDSDRIV        4.134e-01  6.110e-02   6.765 1.33e-11 ***
## INCOME         -3.635e-06  1.266e-06  -2.871 0.004093 **
## HOME_VAL       -2.831e-01  1.243e-01  -2.277 0.022768 *
## TRAVTIME        4.345e-01  5.641e-02   7.704 1.32e-14 ***
## BLUEBOOK       -2.779e-01  5.932e-02  -4.686 2.79e-06 ***
## TIF            -5.823e-02  8.155e-03  -7.140 9.30e-13 ***
## CLM_FREQ        1.575e-01  2.820e-02   5.585 2.33e-08 ***
## MVR_PTS         1.165e-01  1.508e-02   7.724 1.12e-14 ***
## blnPARENT1      4.133e-01  1.039e-01   3.977 6.97e-05 ***
## blnMSTATUS     -5.981e-01  7.778e-02  -7.690 1.48e-14 ***
## blnCAR_USE     -9.525e-01  7.253e-02 -13.133  < 2e-16 ***
## blnNOT_REVOKED -7.674e-01  8.862e-02  -8.660  < 2e-16 ***
## blnURBANICITY   2.289e+00  1.238e-01  18.496  < 2e-16 ***
## intEDUCATION   -1.389e-01  3.814e-02  -3.641 0.000271 ***
## intJOB         -6.922e-02  2.706e-02  -2.558 0.010531 *
## intCAR_TYPE    -1.492e-01  2.667e-02  -5.594 2.22e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7559.6  on 6527  degrees of freedom
## Residual deviance: 5953.8  on 6511  degrees of freedom
## AIC: 5987.8
##
## Number of Fisher Scoring iterations: 5
```




```
## [1] 0.7948707
```

```
##            Reference
## Prediction    0    1
##          0 1115  245
##          1  100  173
```

Our derived logistic regression model has a maximize AUC value of .79 with great p-values on all of the selected variables.

The table below illustrates the various fitness parameters that describe the effectiveness of the logistic regression model.

| Parameters | Model1 |
|---|---|
| Accuracy | 0.7887324 |
| Classification Error Rate | 0.2112676 |
| Precision | 0.8198529 |
| Sensitivity | 0.9176955 |
| Specificity | 0.4138756 |
| F1 Score | 0.8660194 |

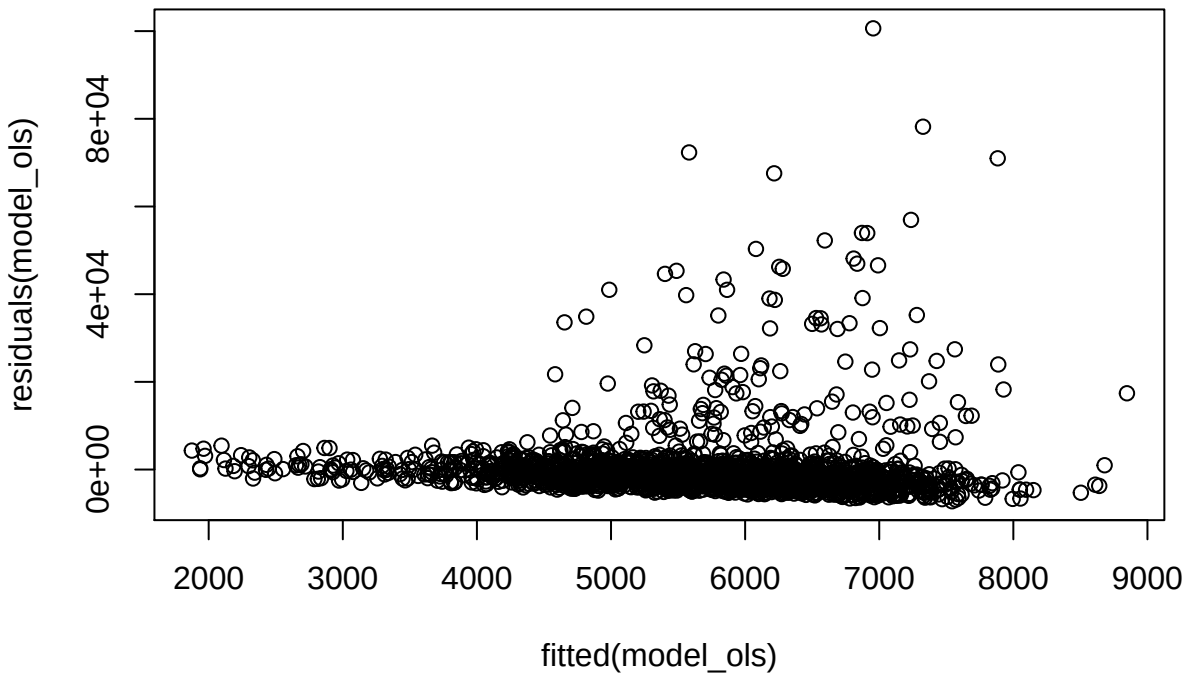
**Linear Regression Models**

For the linear regression model, we went ahead and attempted to produce several models to try and find the cost of repair based on several factors. Unlike in the Logistic Model, where we utilized all the variables and work backwards, for the linear approach we decided to limit to 8 variables, which most "intuitively"

made sense to include when trying to predict cost. Certain variables, even when made into numeric values, just would not make any sense to predict cost. For instance, all binary data was eliminated from this model, it just didn't seem reasonable to predict a continous value using these independent variables. So, we started modeling from these Variables: "TARGET_FLAG","INCOME","TRAVTIME","BLUEBOOK", "YOJ","intEDUCATION","intJOB","CAR_AGE"

Our first attempt at modelling will be to take a generic linear model with our two tranformed variables that we kept:

```
##
## Call:
## lm(formula = TARGET_AMT ~ INCOME + TRAVTIME + BLUEBOOK + YOJ +
##     intEDUCATION + intJOB + CAR_AGE, data = lin_model)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7227  -3095  -1564    303 100631
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -8.259e+03  2.655e+03  -3.110  0.00189 **
## INCOME       -4.255e-03  5.940e-03  -0.716  0.47387
## TRAVTIME     -3.684e+01  2.954e+02  -0.125  0.90075
## BLUEBOOK      1.436e+03  2.717e+02   5.285 1.38e-07 ***
## YOJ           3.570e+01  4.251e+01   0.840  0.40111
## intEDUCATION  4.156e+02  2.434e+02   1.707  0.08791 .
## intJOB        4.143e+01  1.491e+02   0.278  0.78111
## CAR_AGE      -1.009e+02  4.212e+01  -2.395  0.01671 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7684 on 2145 degrees of freedom
## Multiple R-squared:  0.01831,    Adjusted R-squared:  0.01511
## F-statistic: 5.715 on 7 and 2145 DF,  p-value: 1.448e-06
```

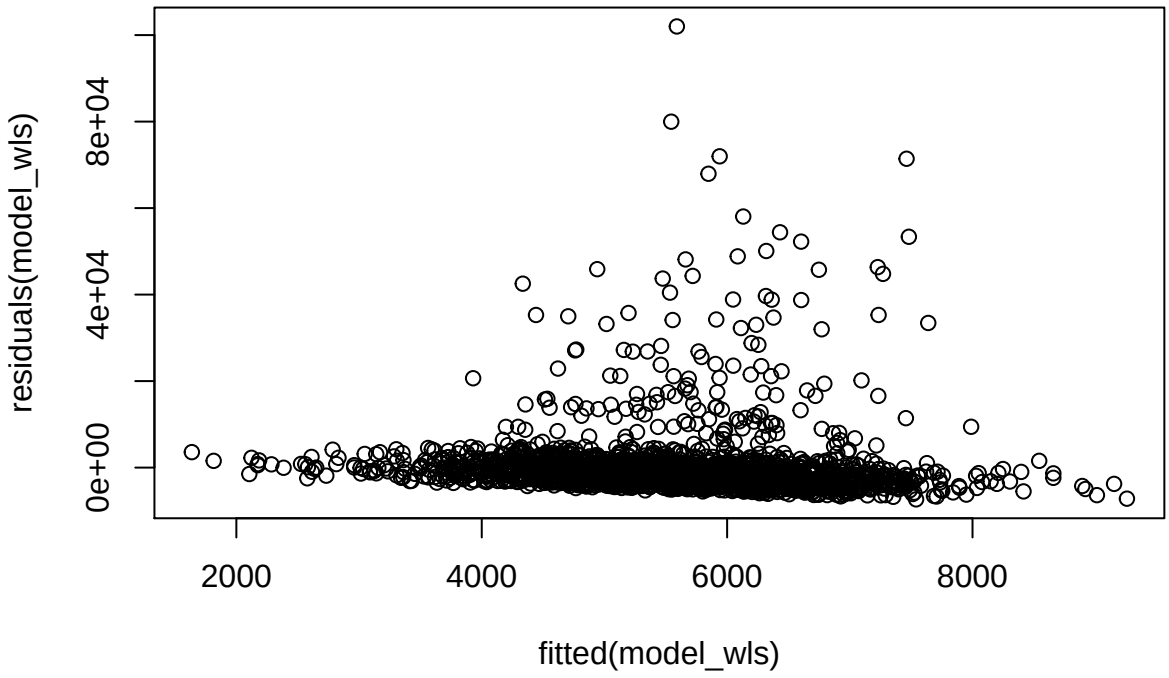



From the above, we can see that a straight linear model was not very effective at producing any results.

Aside (Matt's point of view)- At this point, I had tried several different ways to mitigate the problem shown above, and that was that the only variable I saw of any signficance was the BLUEBOOK value. I tried different a differnt transformation type mostly of the log(TARGET_AMT) ~ log(Dependent Variables). I then tried several different weighting types like 1/SD and 1/fitted (shown below), but I still got the same results. Through entirely happenstance, as we were working off the same dataset, one of my compatriates overwrote the variables to transform only two variables with the log function, the TRAVTIME and BLUEBOOK variable. Since I left the code sitting there, lo and behold, the transformations of just those two skewed values changed the entire code.

```
##
## Call:
## lm(formula = TARGET_AMT ~ INCOME + TRAVTIME + BLUEBOOK + YOJ +
##     intEDUCATION + intJOB + CAR_AGE, data = lin_model, weights = wts)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -15.0707  -0.8423  -0.3900   0.1229  23.1725
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.632e+03  6.942e+02  -8.113 8.21e-16 ***
## INCOME        3.926e-03  2.881e-03   1.363  0.17306
## TRAVTIME     -2.548e+01  8.420e+01  -0.303  0.76221
## BLUEBOOK      1.184e+03  8.404e+01  14.095  < 2e-16 ***
## YOJ           1.235e+02  1.492e+01   8.274 2.25e-16 ***
```

```
## intEDUCATION  1.057e+02  1.173e+02   0.901  0.36745
## intJOB       -5.194e+02  5.350e+01  -9.708  < 2e-16 ***
## CAR_AGE       7.337e+01  2.392e+01   3.067  0.00219 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.126 on 2145 degrees of freedom
## Multiple R-squared:  0.2312, Adjusted R-squared:  0.2287
## F-statistic: 92.15 on 7 and 2145 DF,  p-value: < 2.2e-16
```

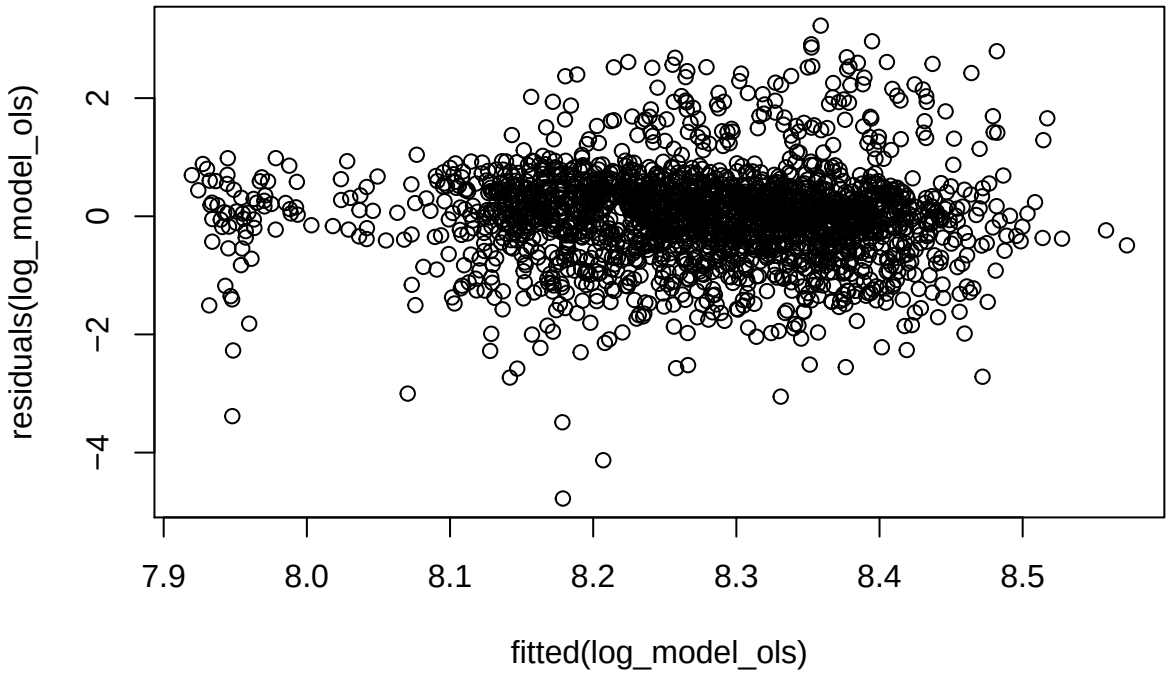



Using this weighing value, which weights each variable according to $1/(fitted)^2$. As you can see from the p values and the $R^2$-values our model become much more approproate. Originally we were has a "goodness of fit" measure of 1% barely describing the variance in our model. With the weighting, we now have a much higher $R^2$.

Originally, our model we were going to choose relied on the below, which used log transformations on most of the independent variables AND our independent variable.

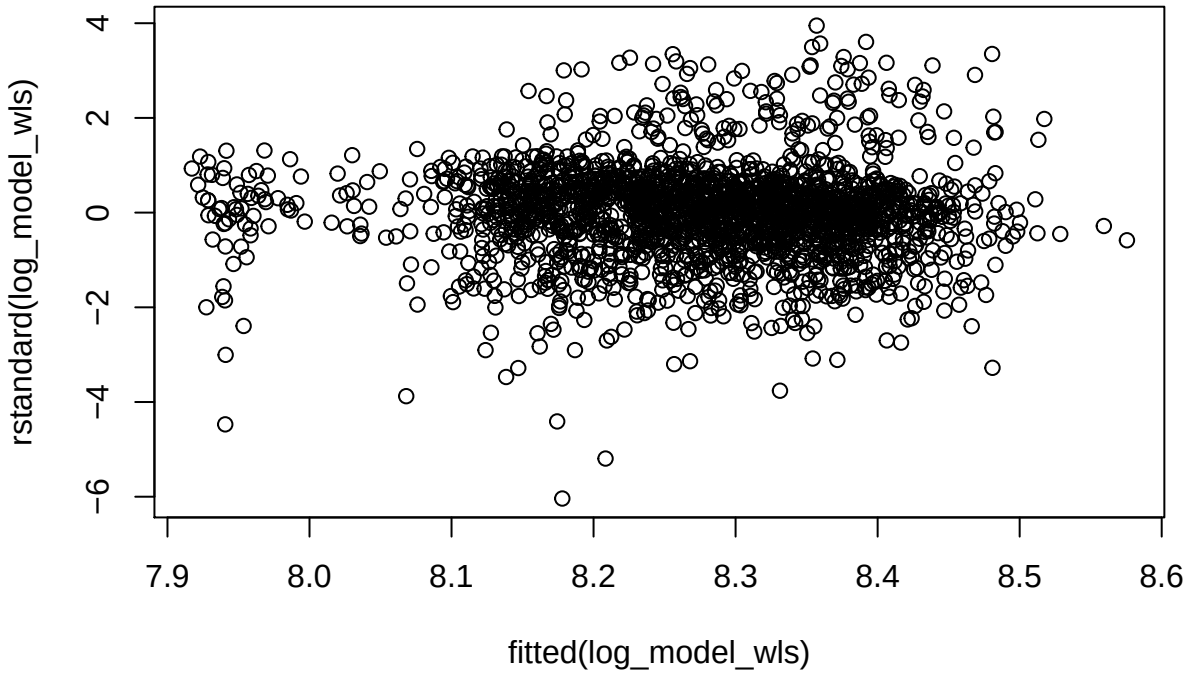
```
##
## Call:
## lm(formula = TARGET_AMT ~ INCOME + TRAVTIME + BLUEBOOK + YOJ +
##     intEDUCATION + intJOB + CAR_AGE, data = log_lin_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7777 -0.3866  0.0375  0.3975  3.2271
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.823919   0.268879  25.379  < 2e-16 ***
## INCOME       -0.002772   0.008844  -0.313    0.754
## TRAVTIME     -0.016039   0.031011  -0.517    0.605
## BLUEBOOK      0.158843   0.028144   5.644 1.88e-08 ***
## YOJ           0.001646   0.005977   0.275    0.783
## intEDUCATION  0.017260   0.024830   0.695    0.487
## intJOB        0.001379   0.015956   0.086    0.931
## CAR_AGE      -0.003015   0.004424  -0.681    0.496
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8068 on 2145 degrees of freedom
## Multiple R-squared:  0.01739,    Adjusted R-squared:  0.01418
## F-statistic: 5.422 on 7 and 2145 DF,  p-value: 3.503e-06
```




```
##
## Call:
## lm(formula = TARGET_AMT ~ INCOME + TRAVTIME + BLUEBOOK + YOJ +
##     intEDUCATION + intJOB + CAR_AGE, data = log_lin_model, weights = log_wts)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4762 -0.6683  0.0654  0.6946  5.5440
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.808658   0.264493  25.742  < 2e-16 ***
## INCOME       -0.003191   0.008818  -0.362    0.717
## TRAVTIME     -0.017669   0.030974  -0.570    0.568
## BLUEBOOK      0.161411   0.027699   5.827 6.48e-09 ***
## YOJ           0.001927   0.005974   0.322    0.747
## intEDUCATION  0.014480   0.024818   0.583    0.560
## intJOB        0.001973   0.015936   0.124    0.901
## CAR_AGE      -0.002657   0.004433  -0.599    0.549
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.405 on 2145 degrees of freedom
## Multiple R-squared:  0.01831,    Adjusted R-squared:  0.01511
## F-statistic: 5.715 on 7 and 2145 DF,  p-value: 1.446e-06
```




As you can see from the statistical output, with a log transformed dependent variable, our model does not hold up. However, or F-statistic is much lower, and the reason for this, is that the F-statistic is not corrected for the weights.

```
backward <- step(model_wls)
```

```
## Start:  AIC=3255.59
## TARGET_AMT ~ INCOME + TRAVTIME + BLUEBOOK + YOJ + intEDUCATION +
##     intJOB + CAR_AGE
##
```

```
##                Df Sum of Sq     RSS    AIC
## - TRAVTIME      1      0.41  9694.8 3253.7
## - intEDUCATION  1      3.67  9698.1 3254.4
## - INCOME        1      8.39  9702.8 3255.4
## <none>                       9694.4 3255.6
## - CAR_AGE       1     42.52  9736.9 3263.0
## - YOJ           1    309.39 10003.8 3321.2
## - intJOB        1    425.91 10120.3 3346.2
## - BLUEBOOK      1    897.83 10592.2 3444.3
##
## Step:  AIC=3253.68
## TARGET_AMT ~ INCOME + BLUEBOOK + YOJ + intEDUCATION + intJOB +
##     CAR_AGE
##
##                Df Sum of Sq     RSS    AIC
## - intEDUCATION  1      4.21  9699.0 3252.6
## - INCOME        1      8.74  9703.5 3253.6
## <none>                       9694.8 3253.7
## - CAR_AGE       1     43.27  9738.1 3261.3
## - YOJ           1    309.04 10003.8 3319.2
## - intJOB        1    439.95 10134.7 3347.2
## - BLUEBOOK      1    897.74 10592.5 3442.3
##
## Step:  AIC=3252.61
## TARGET_AMT ~ INCOME + BLUEBOOK + YOJ + intJOB + CAR_AGE
##
##            Df Sum of Sq     RSS    AIC
## <none>                   9699.0 3252.6
## - INCOME    1     13.34  9712.3 3253.6
## - CAR_AGE   1    179.99  9879.0 3290.2
## - YOJ       1    333.87 10032.9 3323.5
## - intJOB    1    439.73 10138.7 3346.1
## - BLUEBOOK  1    900.79 10599.8 3441.8
```

```
summary(backward)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ INCOME + BLUEBOOK + YOJ + intJOB +
##     CAR_AGE, data = lin_model, weights = wts)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2970  -0.8414  -0.3878   0.1339  23.2058
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.621e+03  6.348e+02  -8.855  < 2e-16 ***
## INCOME       4.748e-03  2.763e-03   1.718   0.0859 .
## BLUEBOOK     1.186e+03  8.400e+01  14.121  < 2e-16 ***
## YOJ          1.176e+02  1.368e+01   8.597  < 2e-16 ***
## intJOB      -5.121e+02  5.190e+01  -9.866  < 2e-16 ***
## CAR_AGE      8.915e+01  1.412e+01   6.312 3.33e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.125 on 2147 degrees of freedom
## Multiple R-squared:  0.2308, Adjusted R-squared:  0.229
## F-statistic: 128.9 on 5 and 2147 DF,  p-value: < 2.2e-16
```

# Choose Model

---

## Choose Model

We chose the Backward logistic regression model to make our predictions for the evaluation dataset (non-zero value). This model has accuracy rate as high as 80%. In the meantime, the precision and sensitivity are at the level of 82% and 92%, which indicate this model is very good at eliminating false negative and false positive situations. Both AUC and F1 Score are around 80%, which also indicates that it has high accuracy in terms of predicting the final response variables.

For linear regression model, we are showing the outputs of each of the 4 tests we had. M1 = straight LM with 2 transformations, m2 = weight lm, m3 = straight lm with log independent, m4 = same as previous but weighted. According to the following summary statistics, the weighted lm out performs all the others. As mentioned previously, the F-statistic is high, which might indicate our model lacks validity. Or it could also indicate the claim amount of motor vehicle accident tends to be unpredictable, and wit

| Parameters | m1 | m2 | m3 | m4 |
|---|---|---|---|---|
| p-value | 1.000000e-07 | 0.0000000 | 0.0000000 | 0.0000000 |
| Mean Squared Error | 5.883165e+07 | 4.5027319 | 0.6484358 | 1.9675620 |
| R^2 | 1.830890e-02 | 0.2311872 | 0.0173879 | 0.0183103 |
| F-Statistics | 5.715013e+00 | 92.1451254 | 5.4224278 | 5.7154583 |

```
##
## Call:
## lm(formula = TARGET_AMT ~ INCOME + TRAVTIME + BLUEBOOK + YOJ +
##     intEDUCATION + intJOB + CAR_AGE, data = lin_model, weights = wts)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -15.0707  -0.8423  -0.3900   0.1229  23.1725
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.632e+03  6.942e+02  -8.113 8.21e-16 ***
## INCOME        3.926e-03  2.881e-03   1.363  0.17306
## TRAVTIME     -2.548e+01  8.420e+01  -0.303  0.76221
## BLUEBOOK      1.184e+03  8.404e+01  14.095  < 2e-16 ***
## YOJ           1.235e+02  1.492e+01   8.274 2.25e-16 ***
## intEDUCATION  1.057e+02  1.173e+02   0.901  0.36745
## intJOB       -5.194e+02  5.350e+01  -9.708  < 2e-16 ***
## CAR_AGE       7.337e+01  2.392e+01   3.067  0.00219 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.126 on 2145 degrees of freedom
## Multiple R-squared:  0.2312, Adjusted R-squared:  0.2287
```

```
## F-statistic: 92.15 on 7 and 2145 DF,  p-value: < 2.2e-16
```

Below is the evaluation of the first 20 values on the evaluation test dataset.

| TARGET_FLAG | TARGET_AMT | KIDSDRIV | AGE | HOMEKIDS | YOJ | INCOME | HOME_VAL |
|---|---|---|---|---|---|---|---|
| 0 | 5626.848 | 0 | 48 | 0 | 11 | 52881 | 11.97756 |
| 0 | 4681.689 | 1 | 40 | 1 | 11 | 50815 | 11.97756 |
| 0 | 5085.579 | 0 | 44 | 2 | 12 | 43486 | 11.97756 |
| 0 | 4102.642 | 0 | 35 | 2 | 0 | 21204 | 11.97756 |
| 0 | 4686.697 | 0 | 59 | 0 | 12 | 87460 | 11.97756 |
| 0 | 6914.618 | 0 | 46 | 0 | 14 | 61509 | 12.24298 |
| 0 | 5188.159 | 0 | 60 | 0 | 12 | 37940 | 12.11581 |
| 0 | 6530.771 | 0 | 54 | 0 | 12 | 33212 | 11.97308 |
| 0 | 7315.587 | 2 | 36 | 2 | 12 | 130540 | 12.74896 |
| 0 | 4745.353 | 0 | 50 | 0 | 8 | 167469 | 11.97756 |
| 0 | 6959.906 | 0 | 42 | 0 | 13 | 52988 | 12.07980 |
| 1 | 5184.814 | 0 | 41 | 2 | 7 | 17755 | 11.91046 |
| 1 | 5506.092 | 1 | 37 | 2 | 13 | 59379 | 11.97756 |
| 0 | 4951.185 | 0 | 36 | 3 | 12 | 56048 | 11.83382 |
| 0 | 5946.418 | 0 | 34 | 3 | 12 | 22510 | 11.72713 |
| 1 | 5105.635 | 0 | 35 | 2 | 12 | 39066 | 11.97756 |
| 1 | 3175.312 | 2 | 44 | 2 | 14 | 45576 | 11.96160 |
| 0 | 6607.592 | 0 | 48 | 0 | 9 | 61509 | 12.18300 |
| 1 | 6250.149 | 0 | 62 | 0 | 15 | 40656 | 12.24646 |
| 0 | 6419.953 | 0 | 39 | 0 | 11 | 33727 | 11.97756 |

| TRAVTIME | BLUEBOOK | TIF | OLDCLAIM | CLM_FREQ | MVR_PTS | CAR_AGE | blnPARENT1 |
|---|---|---|---|---|---|---|---|
| 3.258097 | 9.997433 | 1 | 0 | 0 | 2 | 10 | 0 |
| 3.044522 | 9.848503 | 6 | 3295 | 1 | 2 | 1 | 1 |
| 3.401197 | 8.682708 | 10 | 0 | 0 | 0 | 10 | 1 |
| 4.304065 | 9.130214 | 6 | 0 | 0 | 0 | 4 | 1 |
| 3.806662 | 9.643421 | 1 | 44857 | 2 | 4 | 1 | 0 |
| 1.945910 | 10.152689 | 1 | 2119 | 1 | 2 | 12 | 0 |
| 2.772589 | 9.331673 | 1 | 0 | 0 | 0 | 1 | 0 |
| 3.295837 | 10.085809 | 4 | 0 | 0 | 5 | 9 | 0 |
| 1.609438 | 10.210972 | 4 | 0 | 0 | 0 | 9 | 1 |
| 3.091042 | 10.438518 | 4 | 0 | 0 | 3 | 1 | 0 |
| 3.178054 | 10.062626 | 1 | 0 | 0 | 2 | 11 | 0 |
| 3.367296 | 9.580524 | 1 | 0 | 0 | 2 | 1 | 1 |
| 4.127134 | 9.144201 | 4 | 0 | 0 | 0 | 6 | 1 |
| 2.708050 | 9.277999 | 6 | 2045 | 2 | 2 | 16 | 0 |
| 3.258097 | 9.594922 | 4 | 0 | 0 | 0 | 1 | 0 |
| 3.761200 | 9.182969 | 5 | 0 | 0 | 4 | 4 | 1 |
| 3.295837 | 7.313220 | 4 | 25276 | 1 | 8 | 4 | 0 |
| 3.688880 | 10.053200 | 6 | 0 | 0 | 1 | 17 | 0 |
| 4.043051 | 9.123693 | 4 | 42342 | 2 | 2 | 7 | 0 |
| 3.295837 | 9.291920 | 1 | 8350 | 1 | 2 | 12 | 0 |

| blnMSTATUS | blnSEX | blnCAR_USE | blnNOT_RED_CAR | blnNOT_REVOKED | blnURBANICITY | intEDUC |
|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 1 | 1 | |
| 0 | 0 | 1 | 1 | 1 | 1 | |

| blnMSTATUS | blnSEX | blnCAR_USE | blnNOT_RED_CAR | blnNOT_REVOKED | blnURBANICITY | intEDUC |
|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 1 | 1 | 0 | |
| 0 | 0 | 1 | 1 | 0 | 0 | |
| 0 | 0 | 1 | 0 | 1 | 1 | |
| 1 | 0 | 0 | 1 | 1 | 1 | |
| 1 | 1 | 0 | 1 | 1 | 1 | |
| 1 | 0 | 0 | 1 | 1 | 1 | |
| 0 | 1 | 0 | 1 | 1 | 0 | |
| 0 | 1 | 1 | 1 | 1 | 1 | |
| 1 | 0 | 0 | 1 | 1 | 0 | |
| 0 | 1 | 1 | 0 | 0 | 1 | |
| 0 | 1 | 0 | 1 | 1 | 1 | |
| 1 | 0 | 1 | 1 | 1 | 1 | |
| 1 | 0 | 1 | 0 | 0 | 0 | |
| 0 | 1 | 1 | 1 | 1 | 1 | |
| 1 | 1 | 1 | 1 | 1 | 1 | |
| 1 | 0 | 0 | 0 | 1 | 1 | |
| 1 | 1 | 1 | 1 | 0 | 1 | |
| 0 | 1 | 1 | 1 | 1 | 1 | |

The Smooth Operators of R Fusion Have Struck Again.