# Claim And Claim Cost Via Regression

Brian Weinfeld

CUNY - SPS

# Abstract

I have been tasked with developing a logistic regression and a multiple linear regression that will determine (1) the likelyhood that a policy holder will make a claim on their car and (2) given that a claim is made, how much will it cost. Using both of these models we will be able to set rates for car insurance based on a number of predictors ranging from income, distance to work or number of kids at home. Their are 8161 observations in the training set with 23 predictors. There are 2 response variables, the binary value indicating whether a claim was made and a numeric value indicating the cost of said cliam.

I will develop three logistic regression models, explore each, and ultimately select the strongest model to use on the evaluation set. I will then develop two multiple linear regressions, explore both, and select the strongest model to use on the evaluation set.

I will begin by exploring the data set as a whole and then each individual predictor.


# Data Exploration

Upon reading in the provided data, I made a small number of changes to the names of the columns to aid in my work. Categorical variables have a '.CAT' added to the end and variables with responses such as 'yes/no' have been changed to '1/0' with a more identifying value. For example 'SEX' became 'MALE.CAT'. A number of responses also had odd artifcats that needed to be addressed.

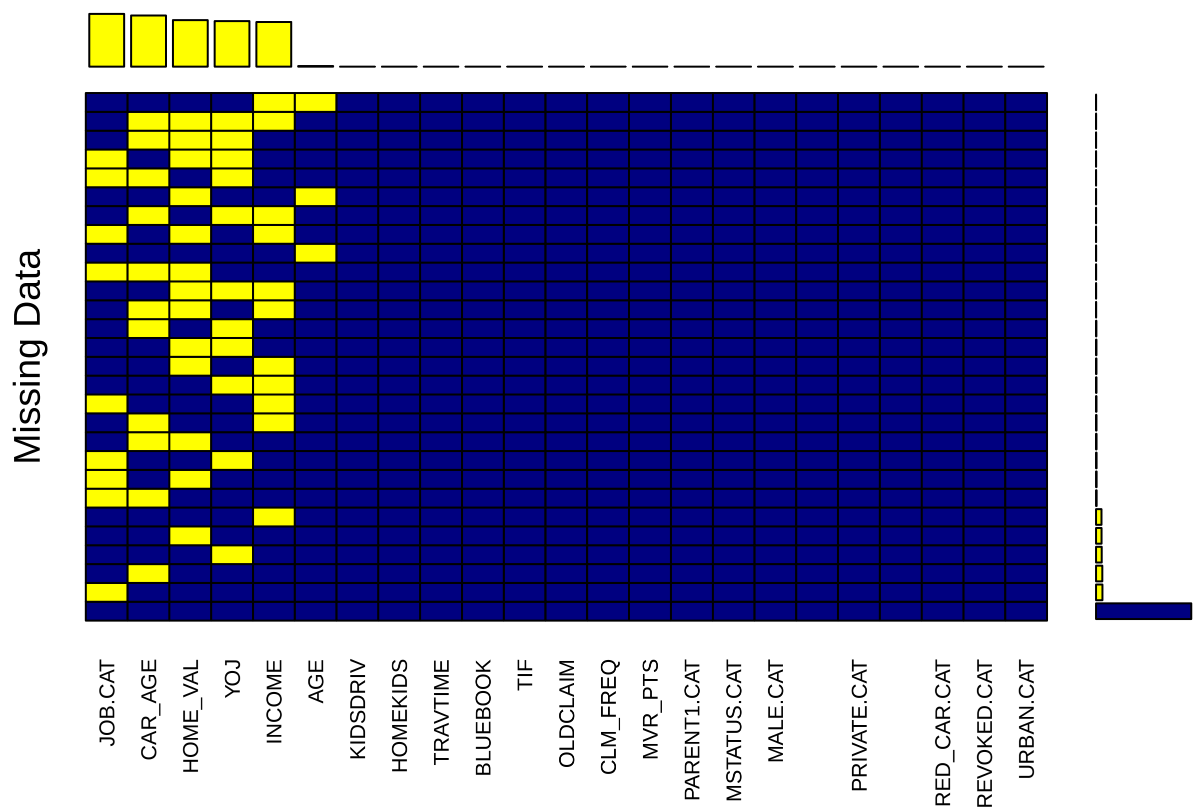
After cleaning up the data, their values were explored. Initial inspection of the data shows a small number of missing values in AGE, YOJ, INCOME, HOME_VAL, CAR_AGE and JOB.CAT. Considering the small number of missing values, it is reasonble to impute them. The below plot shows the distribution of the missing values.

```
insurance %>%
  map_dbl(~sum(is.na(.))/nrow(insurance)) %>%
  kable()
```

| | x |
|---|---|
| TARGET_FLAG | 0.0000000 |
| TARGET_AMT | 0.0000000 |
| KIDSDRIV | 0.0000000 |
| AGE | 0.0007352 |
| HOMEKIDS | 0.0000000 |
| YOJ | 0.0556304 |
| INCOME | 0.0545276 |
| HOME_VAL | 0.0568558 |
| TRAVTIME | 0.0000000 |
| BLUEBOOK | 0.0000000 |
| TIF | 0.0000000 |
| OLDCLAIM | 0.0000000 |
| CLM_FREQ | 0.0000000 |
| MVR_PTS | 0.0000000 |
| CAR_AGE | 0.0624923 |
| PARENT1.CAT | 0.0000000 |
| MSTATUS.CAT | 0.0000000 |
| MALE.CAT | 0.0000000 |
| EDUCATION.CAT | 0.0000000 |
| JOB.CAT | 0.0644529 |
| PRIVATE.CAT | 0.0000000 |
| CAR_TYPE.CAT | 0.0000000 |

| | x |
|---|---|
| RED_CAR.CAT | 0.0000000 |
| REVOKED.CAT | 0.0000000 |
| URBAN.CAT | 0.0000000 |

```
VIM::aggr(insurance[, c(-1, -2)], col=c('navyblue', 'yellow'),
          numbers=TRUE, sortVars=TRUE,
          labels=names(insurance[, c(-1, -2)]), cex.axis=.7,
          gap=3, ylab=c('Missing Data', 'Pattern'), combined=TRUE)
```




```
##
##  Variables sorted by number of missings:
##      Variable Count
##       JOB.CAT   526
##       CAR_AGE   510
##      HOME_VAL   464
##           YOJ   454
##        INCOME   445
##           AGE     6
##      KIDSDRIV     0
##      HOMEKIDS     0
##      TRAVTIME     0
##      BLUEBOOK     0
##           TIF     0
##      OLDCLAIM     0
```

```
##        CLM_FREQ     0
##         MVR_PTS     0
##     PARENT1.CAT     0
##     MSTATUS.CAT     0
##        MALE.CAT     0
##   EDUCATION.CAT     0
##     PRIVATE.CAT     0
##    CAR_TYPE.CAT     0
##     RED_CAR.CAT     0
##     REVOKED.CAT     0
##       URBAN.CAT     0
```

I will use the mice library to impute the data. Once complete I will create a new data frame that has the imputed values.

```
set.seed(123)
imputed.data <- mice::mice(insurance[, c(-1, -2)], m=5, maxit=50, method='pmm',
                           seed=500, printFlag=FALSE)
insurance.complete <- cbind(insurance[, c(1, 2)], complete(imputed.data, 1))
```

I will partition the data into a training set (80%) and a testing set (20%), seperate from the evaluation set that I will use on the selected model. After partitioning the data, I will use 10-fold cross-validation in training my models. About 25% of the customers in the full training set made a claim. This will be reflected in the partition.

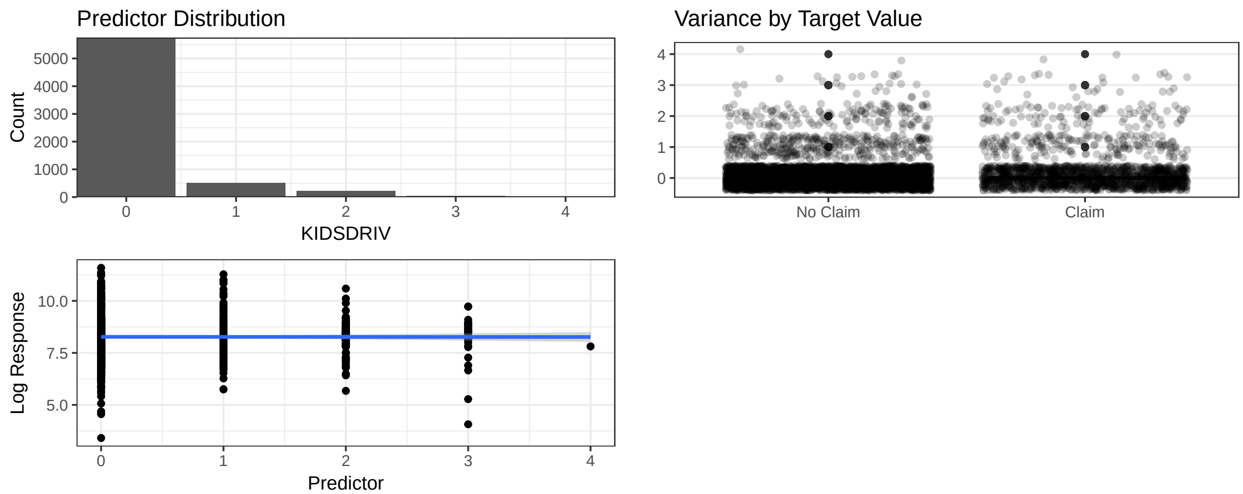
```
set.seed(1)
part <- caret::createDataPartition(insurance.complete$TARGET_FLAG, p=0.8, list=FALSE)
log.training <- insurance.complete[, -2] %>%
  filter(row_number() %in% part)
log.testing <- insurance.complete[, -2] %>%
  filter(!row_number() %in% part)

set.seed(1)
lm.insurance.complete <- insurance.complete %>%
  filter(TARGET_AMT != 0)
part <- caret::createDataPartition(lm.insurance.complete$TARGET_AMT, p=0.8, list=FALSE)
lin.training <- lm.insurance.complete[, -1] %>%
  filter(row_number() %in% part)
lin.testing <- lm.insurance.complete[, -1] %>%
  filter(!row_number() %in% part)
```

With all the values imputed, I am ready to start my initial exploration of the predictors. I created three functions to help with this analysis.

## KIDSDRIV

### The number of kids that drive the car on the policy

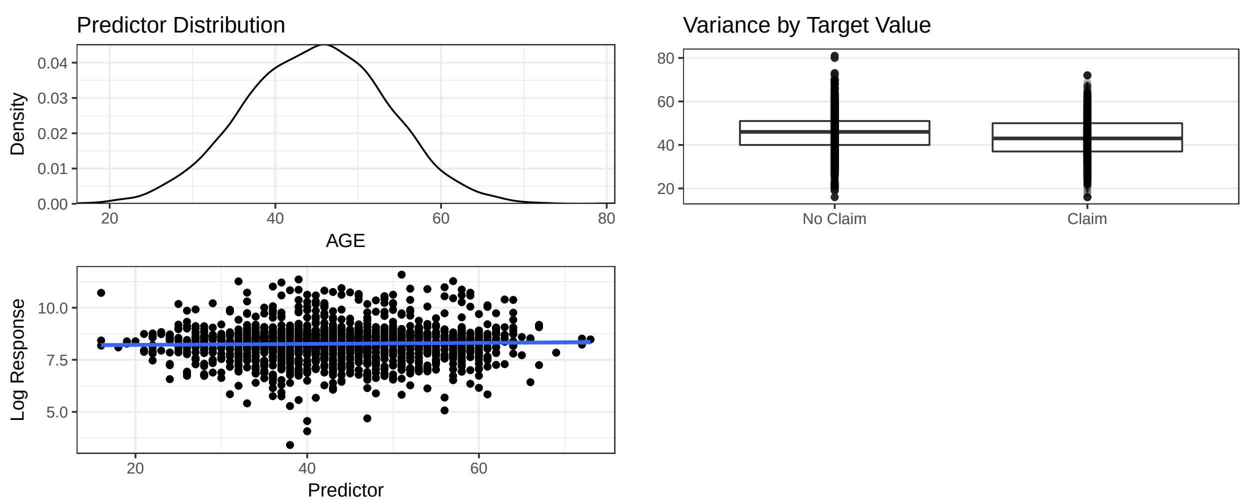
This predictor is discrete with values ranging only from 0 to 4. It is heavily skewed with most cars having 0 kid drivers. Examining the table of values, it appears that having any number of kid driver's results in a higher likelyhood of making a claim.




# AGE

### Age of the Driver

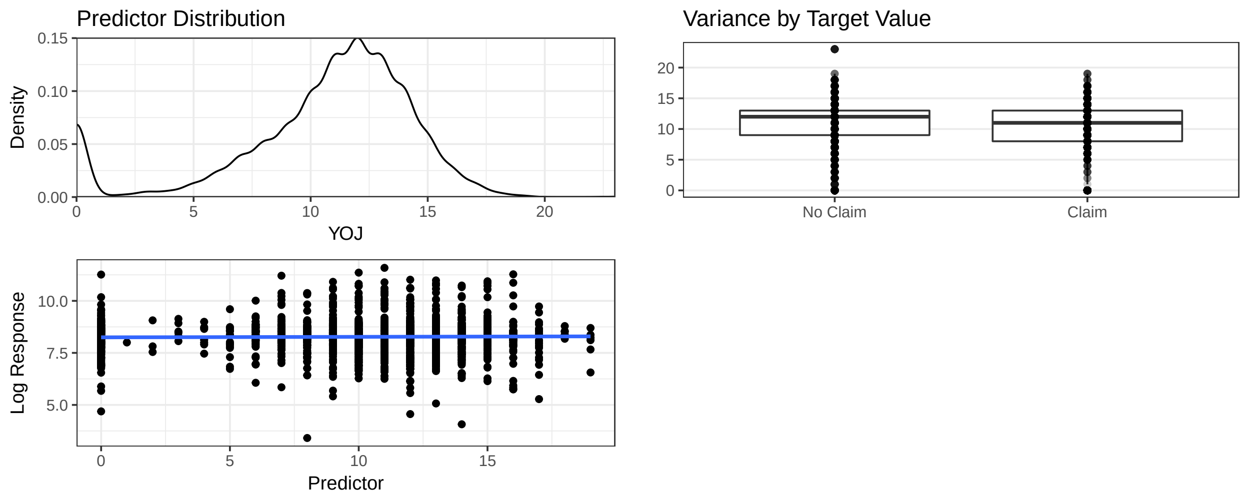
Age has a nice, normal distribution centered around 45. The distribution based on whether a claim is made or not is nearly identical. This leads me to believe that age will not be helpful in determining the likelyhook of making a claim.




# YOJ

### Years On Job

This predictor is nearly normal other than people who are currently unemployed. The distribution when seperated by predictor shows no meanginful difference. It is unlikely that we will use this variable.

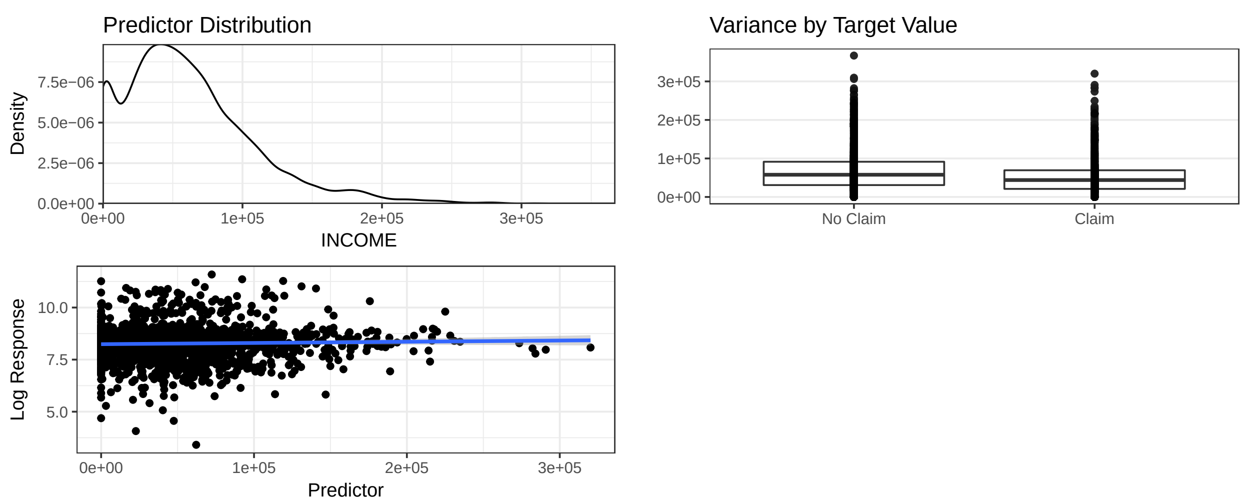



## INCOME

### Yearly Income

Income is, just like in the general population, heavily skewed. This is represented in the boxplot as well as there are numerous upper outliers in both cases. The correlation between YOJ and INCOME is not as large as one might imagine.

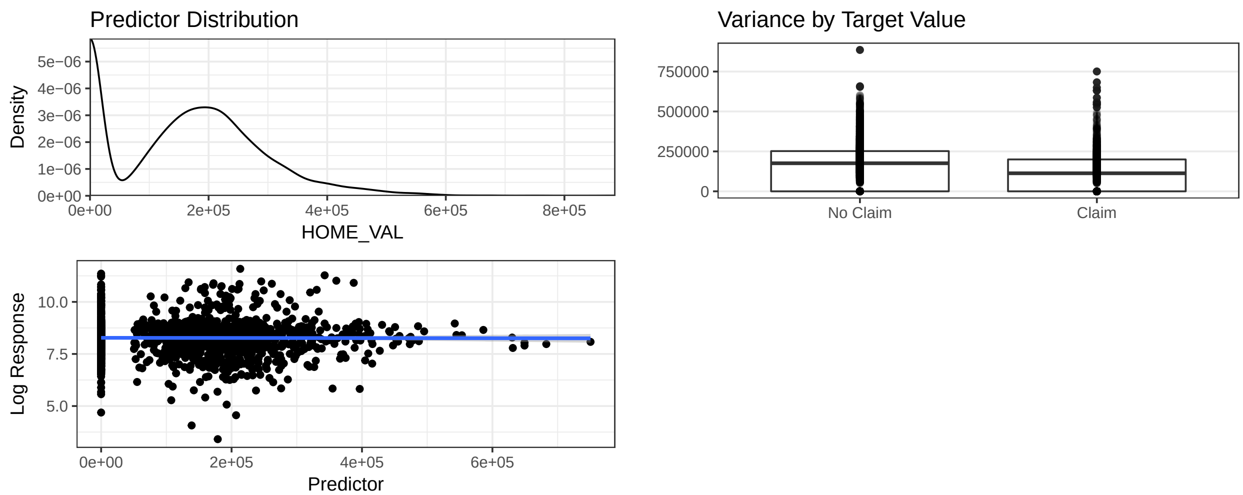
There are outliers on INCOME that may need to be addressed when refering to the TARGET_AMT.




## HOME_VAL

### Value of Home

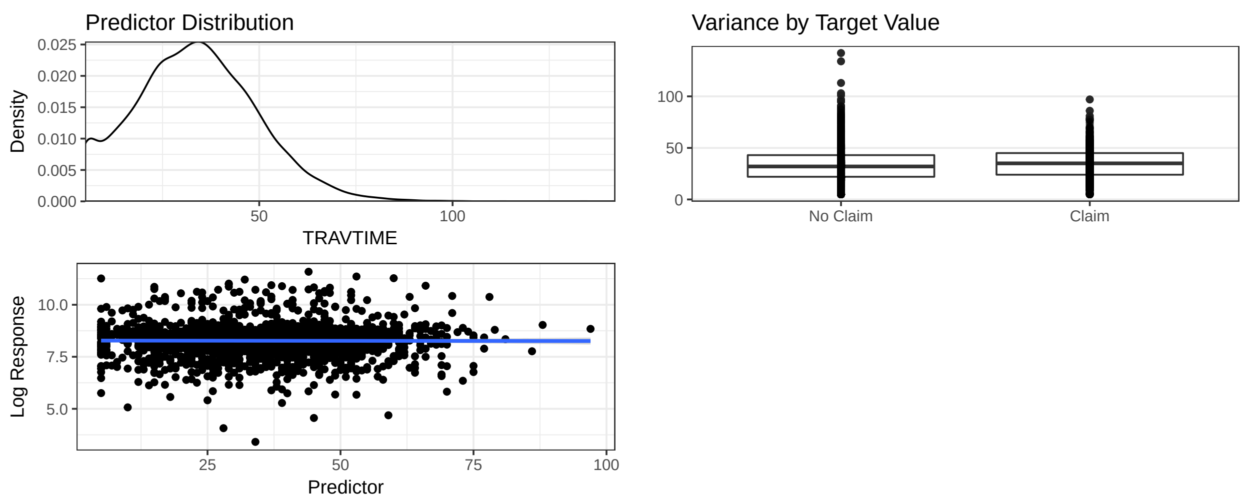
Usefullness of this predictor may be dented by the large number of people who do not own a home. It may be worth considering seperating this into a categorical variable represnting whether or not someone owns a home. The value of the home may be captured by INCOME.




## TRAVTIME

### Distance to Work

The distance travelled to work is fairly normal and the boxplots show only a subtle increase in the likelyhood of making a claim.

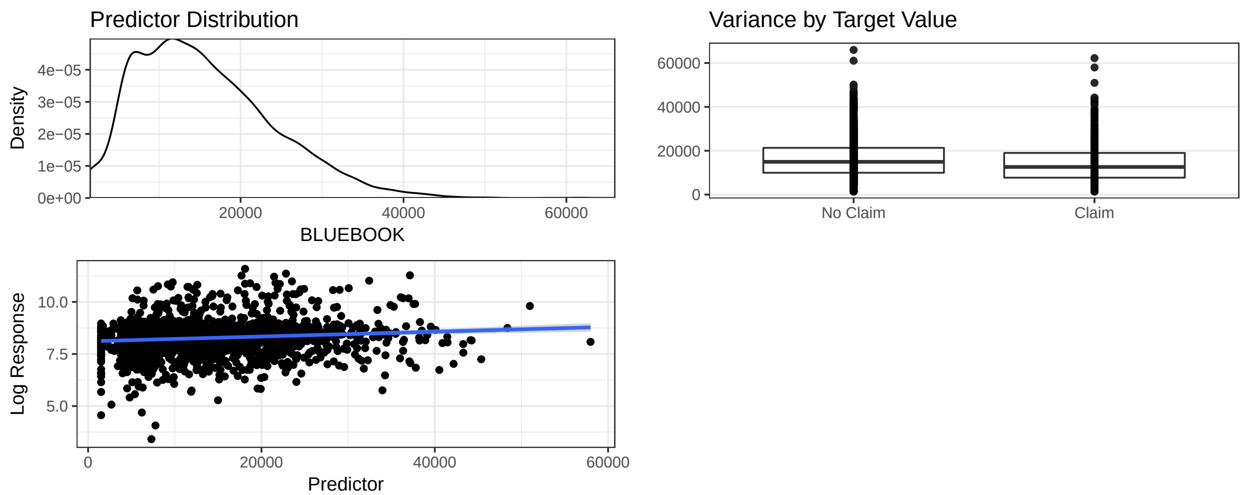



## BLUEBOOK

### Value of Vehicle

The boxplot indicates that those making a claim have a car that is lower in value. Could this be that more expensive cars are driven more carefully due to their cost or is this a confounding variable that once again measures INCOME?

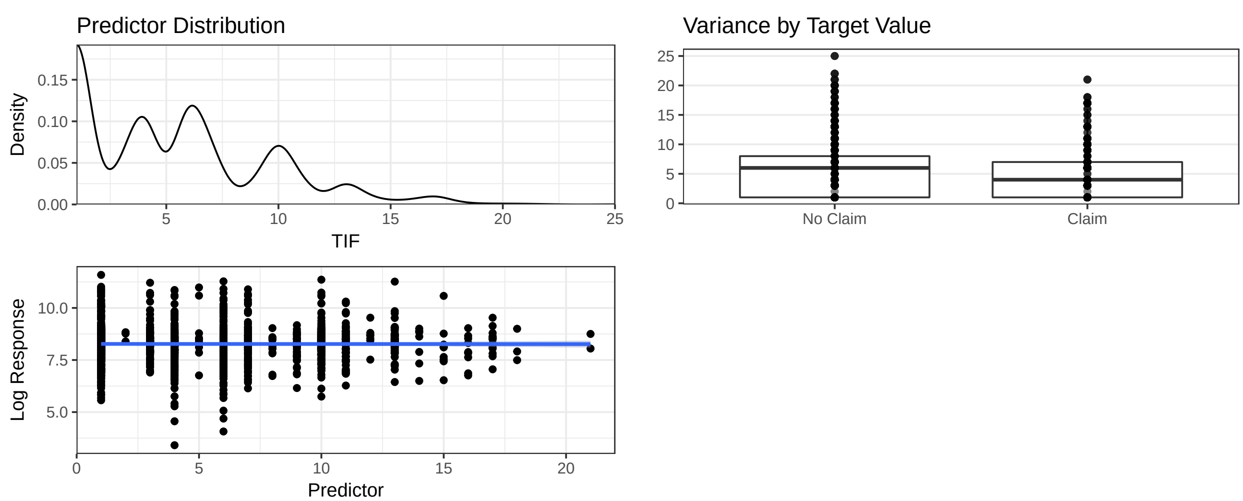
BLUEBOOK is the only variable that appears to show any ability to capture the value of the response variable TARGET_AMT.




# TIF

## Length of Stay with Company

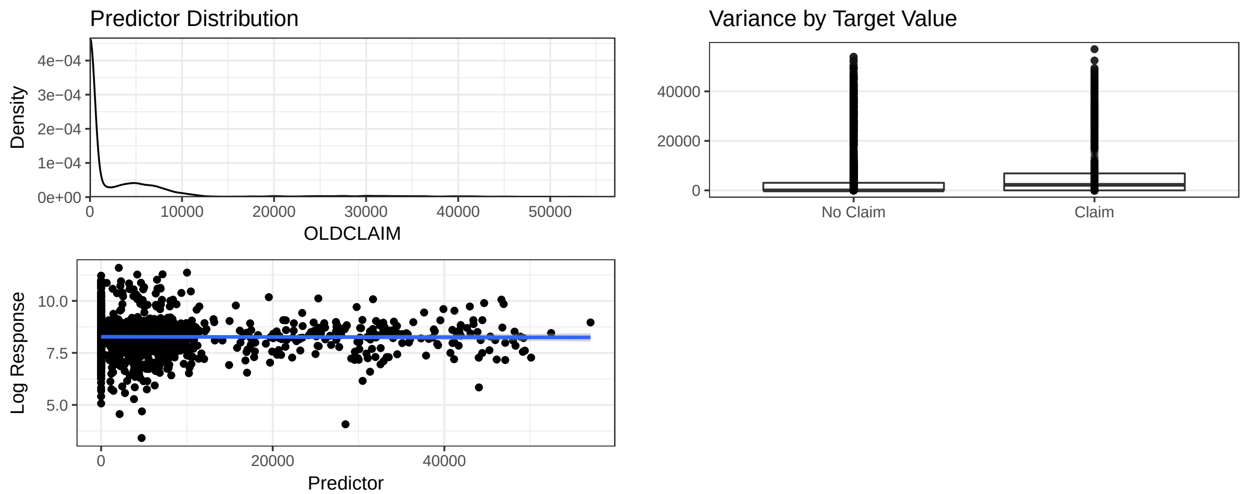
The density plot of this predictor indicates that it could be considered as discrete. There appears to be a significant decrease in the likelyhood of making a claim the longer the person has been with the company. That is, safe drivers tend to stay safe.




# OLDCLAIM

## Claims cost made in the Past 5 Years

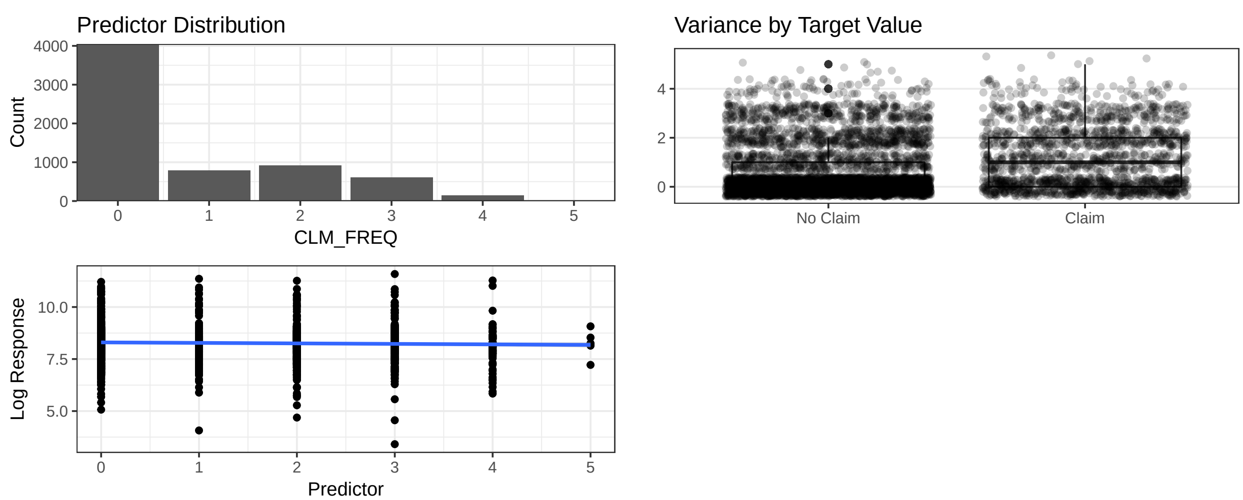
Heavily, heavily skewed predictor. Most people do not make claims.




## CLM_FREQ

### Number of claims made in the Past 5 Years

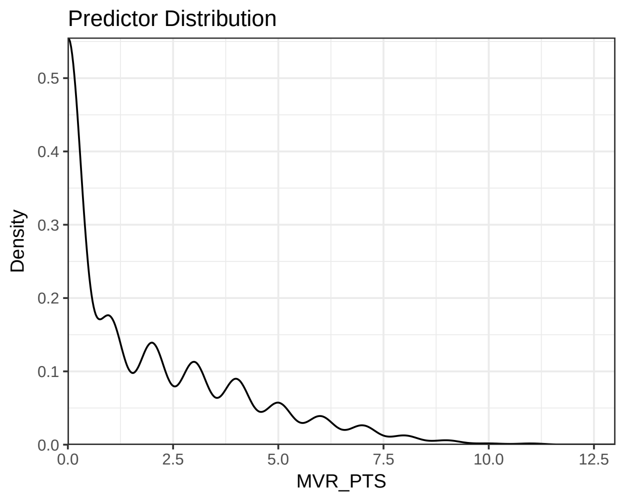
This predictor appears to be highly significant against people who have made a past claim. That is, people who have made a claim in the past 5 years are very likely to make another claim.

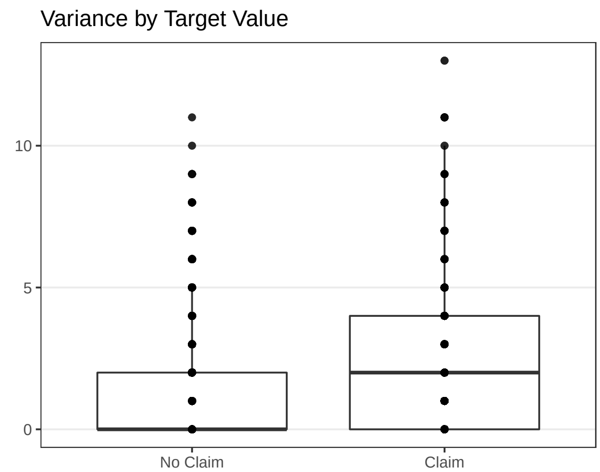



## MVR_PTS

### Motor Vehicle Record Points

This predictor can be seen as a proxy for how safe a driver someone is. Receiving points on a license indicates that the driver has likely been caught speeding, tailgating or other dangerous driving activities. The boxplot indicates that this variable is likely to be highly significant.

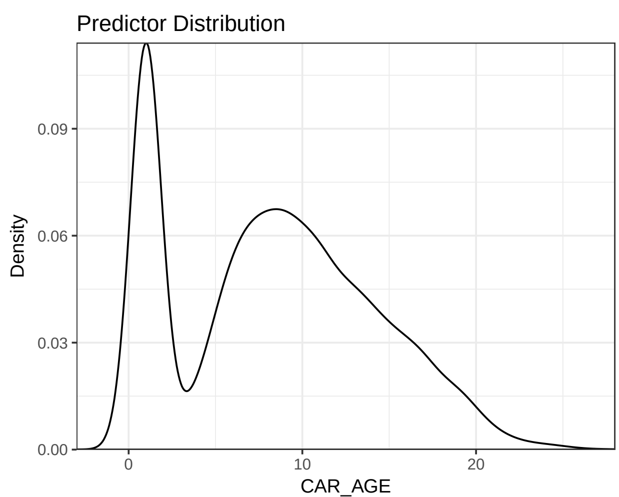


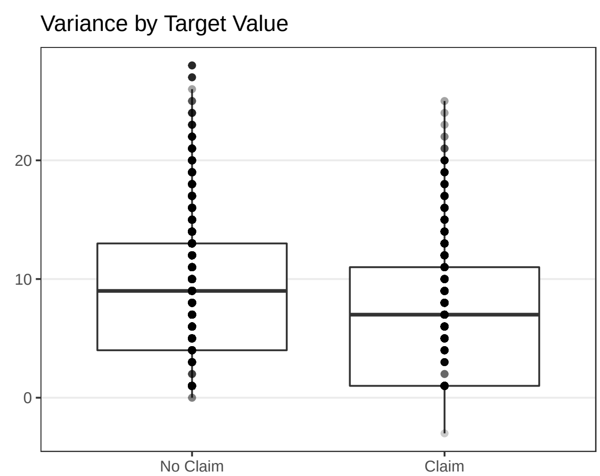




## CAR_AGE

### Age of the Vehicle

This predictor is bimodal, indicating that most cars are either brand new or quite old. There is one data point that is clearly mislabeled as it indicates the car is -3 years old. This will be corrected to 0. There is no indication whether 0, 3 or some other number is the correct choice but considering it is one value amongst many 10s of thousands it is unlikely to have any meaningful effect on the regression.




## MALE.CAT

### Categorical 0 is Female, 1 if Male

This variable was derived from SEX, just to make the variable's meaning more clear. There appears to be no meanginful difference when considering the gender of the driver.

```
##     TARGET
## MALE    0    1
##    0 2537  968
##    1 2270  755
```

## EDUCATION.CAT

### Categorical representing max education level

This variable will need to be monitored as it may be correlated with INCOME or YOJ.

```
##                TARGET
## EDUCATION         0    1
##   <High School  662  312
##   Bachelors    1390  433
##   High School  1209  625
##   Masters      1069  256
##   PhD           477   97
```

## PRIVATE.CAT

### Categorical 0 is commerical, 1 if private

This variable was derived from CAR_USE, just to make the variable's meaning more clear.

```
##        TARGET
## PRIVATE    0    1
##       0 1587  826
##       1 3220  897
```

## CAR_TYPE.CAT

### Categorical representing the car's type

Certain cars are popular with more aggressive or less safe drivers. This may assist in identifying the likelyhood of making a claim.

```
##               TARGET
## CAR_TYPE         0    1
##   Minivan     1433  272
##   Panel Truck  395  132
##   Pickup       778  360
##   Sports Car   476  227
##   SUV         1291  574
##   Van          434  158
```

## RED_CAR.CAT

### Categorical 0 if not Red, 1 if Red

Urban legend states that red cars stand out to police officers and are thus more likely to get pulled over or find themselves in perilous situations.

```
##        TARGET
## RED_CAR    0    1
##       0 3392 1232
##       1 1415  491
```

### REVOKED.CAT

#### Categorical 0 is license not revoked, 1 is revoked

The table's distribution paints a bleak picture that customers who have previously lost their license are likely to be in future accidents.

```
##         TARGET
## REVOKED    0    1
##       0 4363 1387
##       1  444  336
```

### URBAN.CAT

#### Categorical 0 is not urban home/work area, 1 is urban home/work area

This variable can be seen as a proxy for whether the driver frequently uses highways. Urban driving is more likely to result in making a claim, but highway claims are more likely to be expensive. (Collisions at 25mph are obviously less damaging than at 65mph).

```
##       TARGET
## URBAN    0    1
##     0 1261   89
##     1 3546 1634
```

## Logistic Regression

### Model 1

For the first model, I will consider only the categorical variables. This model has the advantage of being easily interpretable and the easiest to calculate for future customers.

I began by adding in all the categorical predictors and then examining which, if any, should be removed from the regression. I considered 5 different methods for model selection. [SEE APPENDIX]

**drop1** suggested keeping all the predictors

**AIC** suggested dropping RED_CAR.CAT

**BIC** suggested dropping RED_CAR.CAT

**lasso** suggested dropping RED_CAR.CAT along with specific values from JOB.CAT and CAR_TYPE.CAT, which is not recommended.

**manual selection** suggested RED_CAR.CAT only

Based on the above 5 methods, the final version of model 1 will only drop RED_CAR.CAT

```
ctrl <- trainControl(method='repeatedcv', number=10, savePredictions=TRUE)
model.1 <- train(TARGET_FLAG ~ PARENT1.CAT + MSTATUS.CAT + MALE.CAT + EDUCATION.CAT +
                   JOB.CAT + PRIVATE.CAT + CAR_TYPE.CAT + REVOKED.CAT + URBAN.CAT,
                 data=log.training, method='glm', family='binomial',
                 trControl=ctrl, tuneLength=5)
summary(model.1)
```

```
##
## Call:
```

```
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0413  -0.7562  -0.4541   0.7786   3.0499
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.80573    0.18682 -15.019  < 2e-16 ***
## PARENT1.CAT1                0.63961    0.09926   6.444 1.17e-10 ***
## MSTATUS.CAT1               -0.57838    0.07285  -7.939 2.04e-15 ***
## MALE.CAT1                   0.23577    0.09622   2.450  0.01427 *
## EDUCATION.CATBachelors     -0.62693    0.11291  -5.552 2.82e-08 ***
## `EDUCATION.CATHigh School` -0.11482    0.10149  -1.131  0.25789
## EDUCATION.CATMasters       -0.89400    0.14913  -5.995 2.04e-09 ***
## EDUCATION.CATPhD           -0.95240    0.17335  -5.494 3.93e-08 ***
## JOB.CATClerical             0.17628    0.10885   1.620  0.10532
## JOB.CATDoctor              -0.57962    0.26943  -2.151  0.03146 *
## `JOB.CATHome Maker`         0.34218    0.14056   2.434  0.01492 *
## JOB.CATLawyer               0.10660    0.17783   0.599  0.54887
## JOB.CATManager             -0.63456    0.13586  -4.671 3.00e-06 ***
## JOB.CATProfessional        -0.05117    0.12167  -0.421  0.67407
## JOB.CATStudent              0.31480    0.11943   2.636  0.00839 **
## PRIVATE.CAT1               -0.81476    0.09660  -8.435  < 2e-16 ***
## `CAR_TYPE.CATPanel Truck`   0.09198    0.15288   0.602  0.54739
## CAR_TYPE.CATPickup          0.58310    0.10804   5.397 6.77e-08 ***
## `CAR_TYPE.CATSports Car`    1.25444    0.13242   9.473  < 2e-16 ***
## CAR_TYPE.CATSUV             1.05847    0.11044   9.584  < 2e-16 ***
## CAR_TYPE.CATVan             0.42118    0.13174   3.197  0.00139 **
## REVOKED.CAT1                0.72426    0.08764   8.264  < 2e-16 ***
## URBAN.CAT1                  2.44877    0.12114  20.215  < 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: 7536.3  on 6529  degrees of freedom
## Residual deviance: 6197.0  on 6507  degrees of freedom
## AIC: 6243
##
## Number of Fisher Scoring iterations: 5
```

## Model 2

For the second model, I will begin by adding in every single predictor, running the regression and then iteratively remove terms based on my analysis. [SEE APPENDIX]

**drop1** suggested keeping all the predictors

**AIC** suggested dropping AGE, YOJ, CAR_AGE, MALE.CAT, RED_CAR.CAT

**BIC** suggested dropping AGE, YOJ, CAR_AGE, MALE.CAT, JOB.CAT, RED_CAR.CAT

**lasso** suggested dropping nothing

**manual selection** suggested dropping RED_CAR.CAT, AGE, YOJ, CAR_AGE, MALE.CAT, HOMEKIDS

Based on the above 5 methods, the final verison of model 2 will drop AGE, YOJ, CAR_AGE, MALE.CAT and RED_CAR.CAT

```
ctrl <- trainControl(method='repeatedcv', number=10, savePredictions=TRUE)
model.2 <- train(TARGET_FLAG ~ . -AGE -YOJ -CAR_AGE -MALE.CAT -RED_CAR.CAT,
                 data=log.training, method='glm', family='binomial',
                 trControl=ctrl, tuneLength=5)
summary(model.2)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6557  -0.7083  -0.3859   0.6274   3.1720
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.671e+00  2.220e-01 -12.032  < 2e-16 ***
## KIDSDRIV                    3.784e-01  6.763e-02   5.595 2.20e-08 ***
## HOMEKIDS                    4.616e-02  3.798e-02   1.215 0.224250
## INCOME                     -3.587e-06  1.222e-06  -2.936 0.003329 **
## HOME_VAL                   -1.554e-06  3.908e-07  -3.978 6.95e-05 ***
## TRAVTIME                    1.531e-02  2.121e-03   7.219 5.24e-13 ***
## BLUEBOOK                   -2.401e-05  5.335e-06  -4.500 6.80e-06 ***
## TIF                        -5.476e-02  8.177e-03  -6.697 2.13e-11 ***
## OLDCLAIM                   -1.563e-05  4.467e-06  -3.500 0.000466 ***
## CLM_FREQ                    2.053e-01  3.211e-02   6.393 1.63e-10 ***
## MVR_PTS                     1.263e-01  1.531e-02   8.250  < 2e-16 ***
## PARENT1.CAT1                3.999e-01  1.225e-01   3.265 0.001096 **
## MSTATUS.CAT1               -5.049e-01  9.555e-02  -5.284 1.26e-07 ***
## EDUCATION.CATBachelors     -4.422e-01  1.205e-01  -3.669 0.000243 ***
## `EDUCATION.CATHigh School` -4.240e-02  1.056e-01  -0.402 0.688001
## EDUCATION.CATMasters       -5.440e-01  1.624e-01  -3.350 0.000807 ***
## EDUCATION.CATPhD           -3.494e-01  2.009e-01  -1.739 0.082036 .
## JOB.CATClerical             7.080e-02  1.147e-01   0.617 0.537148
## JOB.CATDoctor              -6.667e-01  2.834e-01  -2.352 0.018653 *
## `JOB.CATHome Maker`        -3.165e-02  1.551e-01  -0.204 0.838327
## JOB.CATLawyer               4.623e-02  1.852e-01   0.250 0.802937
## JOB.CATManager             -6.036e-01  1.424e-01  -4.238 2.25e-05 ***
## JOB.CATProfessional        -5.747e-02  1.273e-01  -0.452 0.651612
## JOB.CATStudent             -1.151e-01  1.355e-01  -0.849 0.395752
## PRIVATE.CAT1               -8.021e-01  1.003e-01  -8.000 1.25e-15 ***
## `CAR_TYPE.CATPanel Truck`   5.008e-01  1.704e-01   2.939 0.003295 **
## CAR_TYPE.CATPickup          5.038e-01  1.128e-01   4.468 7.90e-06 ***
## `CAR_TYPE.CATSports Car`    9.719e-01  1.220e-01   7.968 1.61e-15 ***
## CAR_TYPE.CATSUV             7.777e-01  9.607e-02   8.095 5.73e-16 ***
## CAR_TYPE.CATVan             6.225e-01  1.374e-01   4.529 5.92e-06 ***
## REVOKED.CAT1                8.325e-01  1.028e-01   8.099 5.54e-16 ***
## URBAN.CAT1                  2.470e+00  1.279e-01  19.312  < 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: 7536.3  on 6529  degrees of freedom
## Residual deviance: 5797.9  on 6498  degrees of freedom
## AIC: 5861.9
##
## Number of Fisher Scoring iterations: 5
```

## Model 3

Examining the diagnostic plots for model 2 indicates that there are a number of predictors that may require a quadratic term. For the final model I will add in all these potential quadratic terms and select a model from there.

**drop1** suggested keeping all the predictors

**AIC** suggested dropping AGE, HOMEKIDS^2, YOJ, MVR_PTS, CAR_AGE, MALE.CAT, RED_CAR

**BIC** suggested dropping KIDSDRIV^2, AGE, HOMEKIDS, HOMEKIDS^2, YOJ, CLM_FREQ^2, MVR_PTS, CAR_AGE, MALE.CAT, JOB.CAT, RED_CAR.CAT

I will be more aggresive with this final model and select the BIC suggestion.

```
log.training <- log.training %>%
  mutate(MVR_PTS2 = MVR_PTS*MVR_PTS)
ctrl <- trainControl(method='repeatedcv', number=10, savePredictions=TRUE)
model.3 <- train(TARGET_FLAG ~ . -AGE -HOMEKIDS -YOJ -MVR_PTS -CAR_AGE -MALE.CAT
                 -JOB.CAT -RED_CAR.CAT,
                 data=log.training, method='glm', family='binomial',
                 trControl=ctrl, tuneLength=5)
summary(model.3)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6049  -0.7208  -0.4035   0.5997   3.1553
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -2.530e+00  2.120e-01 -11.935  < 2e-16 ***
## KIDSDRIV                     4.080e-01  6.191e-02   6.590 4.41e-11 ***
## INCOME                      -4.010e-06  1.126e-06  -3.563 0.000366 ***
## HOME_VAL                    -1.444e-06  3.782e-07  -3.818 0.000135 ***
## TRAVTIME                     1.554e-02  2.113e-03   7.352 1.95e-13 ***
## BLUEBOOK                    -2.434e-05  5.306e-06  -4.588 4.47e-06 ***
## TIF                         -5.386e-02  8.142e-03  -6.615 3.71e-11 ***
## OLDCLAIM                    -1.514e-05  4.433e-06  -3.415 0.000639 ***
## CLM_FREQ                     2.122e-01  3.176e-02   6.681 2.37e-11 ***
## PARENT1.CAT1                 4.780e-01  1.055e-01   4.529 5.93e-06 ***
## MSTATUS.CAT1                -4.777e-01  9.038e-02  -5.285 1.26e-07 ***
## EDUCATION.CATBachelors      -5.739e-01  1.097e-01  -5.230 1.70e-07 ***
## `EDUCATION.CATHigh School`  -9.378e-02  1.028e-01  -0.913 0.361479
## EDUCATION.CATMasters        -6.542e-01  1.235e-01  -5.299 1.17e-07 ***
## EDUCATION.CATPhD            -6.556e-01  1.696e-01  -3.867 0.000110 ***
```

```
## PRIVATE.CAT1                -8.460e-01  8.255e-02 -10.248  < 2e-16 ***
## `CAR_TYPE.CATPanel Truck`    4.577e-01  1.628e-01   2.812 0.004921 **
## CAR_TYPE.CATPickup           4.725e-01  1.104e-01   4.279 1.87e-05 ***
## `CAR_TYPE.CATSports Car`     9.672e-01  1.207e-01   8.016 1.09e-15 ***
## CAR_TYPE.CATSUV              7.820e-01  9.520e-02   8.214  < 2e-16 ***
## CAR_TYPE.CATVan              5.879e-01  1.355e-01   4.338 1.44e-05 ***
## REVOKED.CAT1                 8.376e-01  1.022e-01   8.198 2.45e-16 ***
## URBAN.CAT1                   2.437e+00  1.278e-01  19.067  < 2e-16 ***
## MVR_PTS2                     1.920e-02  2.247e-03   8.542  < 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: 7536.3  on 6529  degrees of freedom
## Residual deviance: 5832.6  on 6506  degrees of freedom
## AIC: 5880.6
##
## Number of Fisher Scoring iterations: 5
```

## Logistic Model Selection

$R^2$ does not exist for logistic regression in the traditional sense. However, there are a number of so called pseudo $R^2$ terms that can be analized. This is a good starting point for identifying the relative strength of each model.

```
model.1.diag <- glm(TARGET_FLAG ~ PARENT1.CAT + MSTATUS.CAT + MALE.CAT + EDUCATION.CAT +
                      JOB.CAT + PRIVATE.CAT + CAR_TYPE.CAT +
                   REVOKED.CAT + URBAN.CAT, data=log.training, family=binomial)
model.2.diag <- glm(TARGET_FLAG ~ . -AGE -YOJ -CAR_AGE -MALE.CAT -RED_CAR.CAT,
                    data=log.training, family=binomial)
model.3.diag <- glm(TARGET_FLAG ~ . -AGE -HOMEKIDS -YOJ - MVR_PTS -CAR_AGE -MALE.CAT
                    -JOB.CAT -RED_CAR.CAT +I(MVR_PTS^2),
                    data=log.training, family=binomial)
data_frame(name=names(pscl::pR2(model.1.diag)), value=pscl::pR2(model.1.diag)) %>%
  spread(1, 2) %>%
  kable()
```

| G2 | llh | llhNull | McFadden | r2CU | r2ML |
|---|---|---|---|---|---|
| 1339.287 | -3098.525 | -3768.168 | 0.1777106 | 0.2708375 | 0.1854321 |

```
data_frame(name=names(pscl::pR2(model.2.diag)), value=pscl::pR2(model.2.diag)) %>%
  spread(1, 2) %>%
  kable()
```

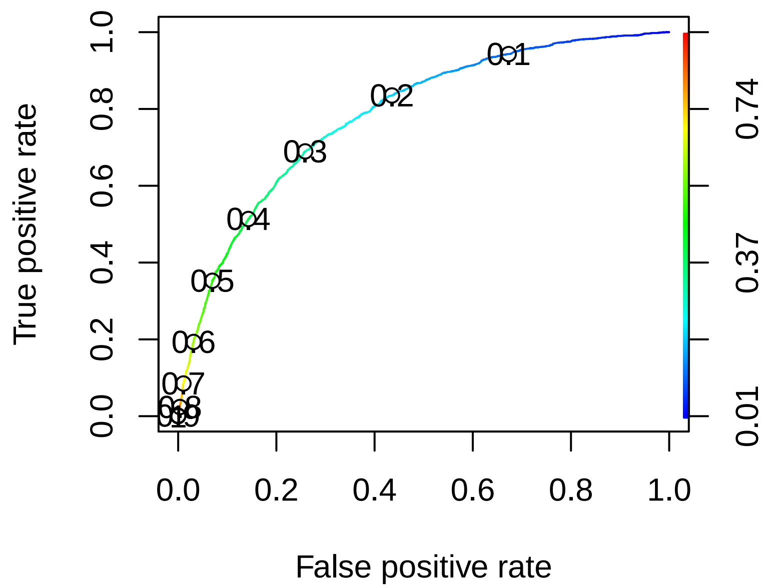
| G2 | llh | llhNull | McFadden | r2CU | r2ML |
|---|---|---|---|---|---|
| 1744.125 | -2896.106 | -3768.168 | 0.2314287 | 0.3423573 | 0.234399 |

```
data_frame(name=names(pscl::pR2(model.3.diag)), value=pscl::pR2(model.3.diag)) %>%
  spread(1, 2) %>%
  kable()
```

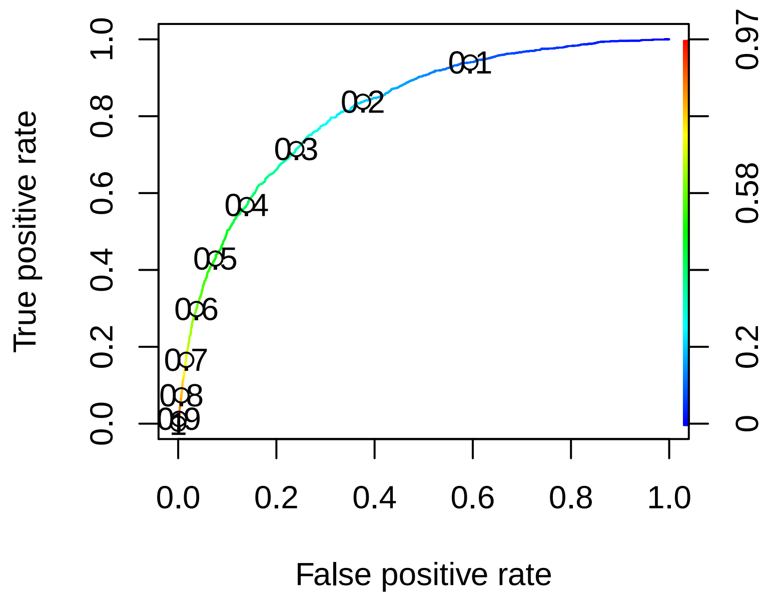
| G2 | llh | llhNull | McFadden | r2CU | r2ML |
|---|---|---|---|---|---|
| 1703.709 | -2916.314 | -3768.168 | 0.226066 | 0.335415 | 0.2296458 |

The first model is the weakest while the second and third are close in their psuedo-$R^2$. Next, we will examine the ROC curve to determine a good cutoff point for categorization against the testing data. All three models appear to have between 0.5 to 0.4 as a good compromise.

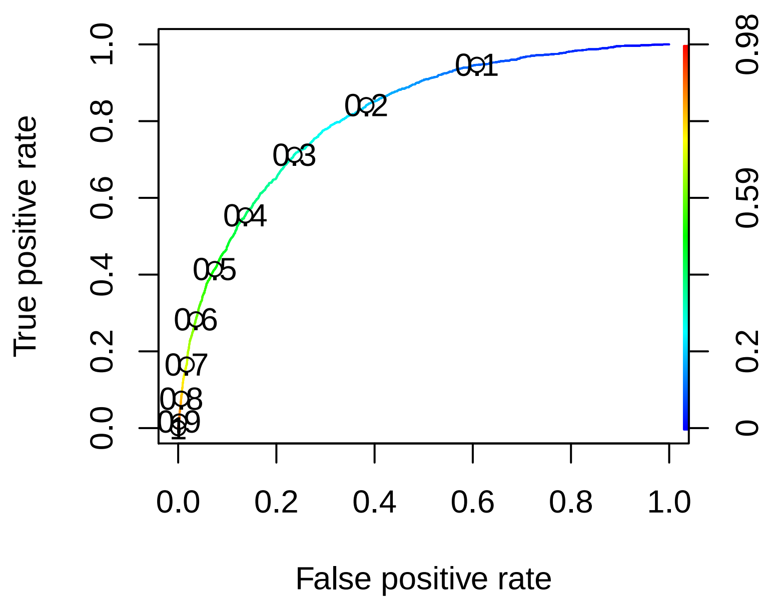
```
ROCRPred <- prediction(predict(model.1.diag, type='response'), log.training$TARGET_FLAG)
ROCRPref <- performance(ROCRPred, 'tpr', 'fpr')
plot(ROCRPref, colorize=TRUE, print.cutoffs.at = seq(0.1, by=0.1))
```




```
ROCRPred <- prediction(predict(model.2.diag, type='response'), log.training$TARGET_FLAG)
ROCRPref <- performance(ROCRPred, 'tpr', 'fpr')
plot(ROCRPref, colorize=TRUE, print.cutoffs.at = seq(0.1, by=0.1))
```




```r
ROCRPred <- prediction(predict(model.3.diag, type='response'), log.training$TARGET_FLAG)
ROCRPref <- performance(ROCRPred, 'tpr', 'fpr')
plot(ROCRPref, colorize=TRUE, print.cutoffs.at = seq(0.1, by=0.1))
```




Finally, I will create a confusion matrix for each model against the testing data.

```r
predictions <- ifelse(predict(model.1, newdata=log.testing, type='prob')[2] < 0.4, 0, 1)
caret::confusionMatrix(table(predicted=predictions, actual = log.testing$TARGET_FLAG))
```

```
## Confusion Matrix and Statistics
##
##          actual
## predicted    0    1
##         0 1010  203
##         1  191  227
##
##                Accuracy : 0.7584
##                  95% CI : (0.7369, 0.779)
##     No Information Rate : 0.7364
##     P-Value [Acc > NIR] : 0.02227
##
##                   Kappa : 0.3722
##  Mcnemar's Test P-Value : 0.57946
##
##             Sensitivity : 0.8410
##             Specificity : 0.5279
##          Pos Pred Value : 0.8326
##          Neg Pred Value : 0.5431
##              Prevalence : 0.7364
##          Detection Rate : 0.6193
##    Detection Prevalence : 0.7437
##       Balanced Accuracy : 0.6844
##
##        'Positive' Class : 0
##
```

```r
predictions <- ifelse(predict(model.2, newdata=log.testing, type='prob')[2] < 0.4, 0, 1)
caret::confusionMatrix(table(predicted=predictions, actual = log.testing$TARGET_FLAG))
```

```
## Confusion Matrix and Statistics
##
##          actual
## predicted    0    1
##         0 1012  180
##         1  189  250
##
##                Accuracy : 0.7738
##                  95% CI : (0.7527, 0.7939)
##     No Information Rate : 0.7364
##     P-Value [Acc > NIR] : 0.0002795
##
##                   Kappa : 0.4212
##  Mcnemar's Test P-Value : 0.6770710
##
##             Sensitivity : 0.8426
##             Specificity : 0.5814
##          Pos Pred Value : 0.8490
##          Neg Pred Value : 0.5695
##              Prevalence : 0.7364
##          Detection Rate : 0.6205
```

```
##    Detection Prevalence : 0.7308
##       Balanced Accuracy : 0.7120
##
##        'Positive' Class : 0
##
```

```
log.testing <- log.testing %>%
  mutate(MVR_PTS2 = MVR_PTS*MVR_PTS)
predictions <- ifelse(predict(model.3, newdata=log.testing, type='prob')[2] < 0.4, 0, 1)
caret::confusionMatrix(table(predicted=predictions, actual = log.testing$TARGET_FLAG))
```

```
## Confusion Matrix and Statistics
##
##          actual
## predicted    0    1
##         0 1010  189
##         1  191  241
##
##                Accuracy : 0.767
##                  95% CI : (0.7457, 0.7873)
##     No Information Rate : 0.7364
##     P-Value [Acc > NIR] : 0.002449
##
##                   Kappa : 0.4008
##  Mcnemar's Test P-Value : 0.959087
##
##             Sensitivity : 0.8410
##             Specificity : 0.5605
##          Pos Pred Value : 0.8424
##          Neg Pred Value : 0.5579
##              Prevalence : 0.7364
##          Detection Rate : 0.6193
##    Detection Prevalence : 0.7351
##       Balanced Accuracy : 0.7007
##
##        'Positive' Class : 0
##
```

Based on all the available diagnostics, I will select **model 2**. It has the highest pseudo-$R^2$, it performed the strongest on the testing set and has the lowest AIC.

# LINEAR REGRESSION

## Model 4

The initial predictor exploration does not bode well for the linear regression. None of the predictors appeared to show any real ability to predict the response variable. I will begin by include all the predictors compared against the log transformed TARGET_AMT. The log transformation was deemed necessary by viewing the distribution of TARGET_AMT and it's relationship to each of the predictors.

```
ln.model.1 <- lm(log(TARGET_AMT) ~ ., data=lin.training)
summary(ln.model.1)
```

```
##
```

```
## Call:
## lm(formula = log(TARGET_AMT) ~ ., data = lin.training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6040 -0.4127  0.0473  0.4096  3.3172
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               8.012e+00  1.922e-01  41.693  < 2e-16 ***
## KIDSDRIV                 -5.082e-03  3.807e-02  -0.133  0.89383
## AGE                       1.845e-03  2.519e-03   0.732  0.46400
## HOMEKIDS                  4.052e-03  2.518e-02   0.161  0.87219
## YOJ                       4.272e-04  5.848e-03   0.073  0.94178
## INCOME                   -1.406e-06  8.175e-07  -1.719  0.08576 .
## HOME_VAL                  2.953e-08  2.430e-07   0.122  0.90331
## TRAVTIME                 -3.742e-04  1.327e-03  -0.282  0.77798
## BLUEBOOK                  1.333e-05  3.653e-06   3.650  0.00027 ***
## TIF                      -1.114e-03  5.120e-03  -0.218  0.82781
## OLDCLAIM                  3.452e-06  2.686e-06   1.285  0.19890
## CLM_FREQ                 -4.358e-02  1.869e-02  -2.332  0.01982 *
## MVR_PTS                   1.402e-02  8.222e-03   1.705  0.08842 .
## CAR_AGE                  -2.502e-03  5.261e-03  -0.476  0.63446
## PARENT1.CAT1              2.182e-02  7.081e-02   0.308  0.75805
## MSTATUS.CAT1             -8.348e-02  6.068e-02  -1.376  0.16905
## MALE.CAT1                 3.141e-02  7.844e-02   0.400  0.68887
## EDUCATION.CATBachelors   -4.942e-02  7.720e-02  -0.640  0.52214
## EDUCATION.CATHigh School -4.099e-03  6.149e-02  -0.067  0.94685
## EDUCATION.CATMasters      1.058e-01  1.194e-01   0.886  0.37562
## EDUCATION.CATPhD          3.108e-01  1.513e-01   2.054  0.04009 *
## JOB.CATClerical          -1.816e-02  6.839e-02  -0.265  0.79067
## JOB.CATDoctor            -2.537e-01  1.991e-01  -1.274  0.20270
## JOB.CATHome Maker        -9.452e-02  1.032e-01  -0.915  0.36010
## JOB.CATLawyer            -7.305e-02  1.199e-01  -0.609  0.54259
## JOB.CATManager           -7.515e-02  1.006e-01  -0.747  0.45529
## JOB.CATProfessional       1.875e-02  7.867e-02   0.238  0.81165
## JOB.CATStudent           -4.833e-02  8.426e-02  -0.574  0.56632
## PRIVATE.CAT1             -4.940e-03  6.167e-02  -0.080  0.93617
## CAR_TYPE.CATPanel Truck  -1.036e-02  1.147e-01  -0.090  0.92808
## CAR_TYPE.CATPickup        2.738e-02  7.179e-02   0.381  0.70298
## CAR_TYPE.CATSports Car    2.294e-02  8.897e-02   0.258  0.79653
## CAR_TYPE.CATSUV           6.544e-02  7.922e-02   0.826  0.40891
## CAR_TYPE.CATVan          -3.698e-02  9.228e-02  -0.401  0.68868
## RED_CAR.CAT1              8.830e-02  5.965e-02   1.480  0.13898
## REVOKED.CAT1             -9.921e-02  6.120e-02  -1.621  0.10519
## URBAN.CAT1                9.446e-02  9.175e-02   1.030  0.30337
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8252 on 1688 degrees of freedom
## Multiple R-squared:  0.02992,    Adjusted R-squared:  0.009227
## F-statistic: 1.446 on 36 and 1688 DF,  p-value: 0.04291
```

The summary shows that only BLUEBOOK appears to have any predictive ability. This was the initial conclusion of the exploratory analysis as well. [SEE APPENDIX]

## Model 5

This model was selected by using model shrinkage. [SEE APPENDIX]

```
ln.model.2 <- lm(log(TARGET_AMT) ~ BLUEBOOK + CLM_FREQ + MVR_PTS + MSTATUS.CAT +
                   RED_CAR.CAT, data=lin.training)
summary(ln.model.2)
```

```
##
## Call:
## lm(formula = log(TARGET_AMT) ~ BLUEBOOK + CLM_FREQ + MVR_PTS +
##     MSTATUS.CAT + RED_CAR.CAT, data = lin.training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6245 -0.3970  0.0399  0.4045  3.2369
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.118e+00  5.147e-02 157.726  < 2e-16 ***
## BLUEBOOK      1.134e-05  2.369e-06   4.789 1.82e-06 ***
## CLM_FREQ     -3.224e-02  1.650e-02  -1.954   0.0509 .
## MVR_PTS       1.588e-02  8.029e-03   1.977   0.0481 *
## MSTATUS.CAT1 -6.964e-02  3.959e-02  -1.759   0.0787 .
## RED_CAR.CAT1  8.049e-02  4.334e-02   1.857   0.0635 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8215 on 1719 degrees of freedom
## Multiple R-squared:  0.02077,    Adjusted R-squared:  0.01792
## F-statistic: 7.291 on 5 and 1719 DF,  p-value: 9.048e-07
```

Despite including only the most important predictors, 3 of them are still not statistically significant. The below anova test demonstrates that there was no apparent loss of predicive ability by dropping the other terms.

```
anova(ln.model.1, ln.model.2)
```

| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|
| 1688 | 1149.328 | NA | NA | NA | NA |
| 1719 | 1160.168 | -31 | -10.83953 | 0.5135434 | 0.9881165 |

## Model 6

I noticed during the exploratory analysis that there were a number of upper and lower outliers that may be skewing the results of the regression. I will perform a robusted regression for this model and further explore BLUEBOOK, the most significant of the predictors in more depth. [SEE APPENDIX]

```
ln.model.3 <- MASS::rlm(log(TARGET_AMT) ~ poly(BLUEBOOK, 2), data=lin.training)
summary(ln.model.3)
```

```
##
```

```
## Call: rlm(formula = log(TARGET_AMT) ~ poly(BLUEBOOK, 2), data = lin.training)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.81229 -0.39109  0.03545  0.39412  3.25143
##
## Coefficients:
##                    Value    Std. Error t value
## (Intercept)          8.2784   0.0159   519.3645
## poly(BLUEBOOK, 2)1   2.1194   0.6620     3.2014
## poly(BLUEBOOK, 2)2  -1.6185   0.6620    -2.4447
##
## Residual standard error: 0.5826 on 1722 degrees of freedom
```

```
anova(ln.model.2, ln.model.3)
```

| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|
| 1719 | 1160.168 | NA | NA | NA | NA |
| NA | 1167.063 | NA | -6.895664 | NA | NA |

There does not appear to be any significant difference in the predictive ability of all three models.

### MODEL SELECTION

None of the three models have demonstrated a strong predictive ability and all have a small, similar $R^2$. It appears that we do not have the necessary predictors to conclusively determine the response. To help determine which model to select, I will test their predictions against the testing set.

```
pred.1 <- exp(predict(ln.model.1, newdata=lin.testing, interval='confidence'))
sum(ifelse(pred.1[, 2] < lin.testing[, 1] & pred.1[, 3] > lin.testing[, 1], 1, 0))
```

```
## [1] 123
```

```
pred.2 <- exp(predict(ln.model.2, newdata=lin.testing, interval='confidence'))
sum(ifelse(pred.2[, 2] < lin.testing[, 1] & pred.2[, 3] > lin.testing[, 1], 1, 0))
```

```
## [1] 45
```

```
pred.3 <- exp(predict(ln.model.3, newdata=lin.testing, interval='confidence'))
sum(ifelse(pred.3[, 2] < lin.testing[, 1] & pred.3[, 3] > lin.testing[, 1], 1, 0))
```

```
## [1] 22
```

While none of the models did a particularly strong job of predicting TARGET_AMT, the first model did by far the best. This is the model that included over a dozen non-significant predictors. This has me concerned about overfitting. This is especially due to the fact that an anova test determined no significant difference in the first and second models. For that reason, I have decied to select **model 5**.

## CONCLUSION

As required, the predictions for the logistic regression and linear regression have been written out to csv files. However, in the end neither model turned out to be particularly powerful. The logistic regression was able to get near an 80% accurate prediction rate, however considering that a 75% prediction rate could be obtained by simply guessing that a claim would NOT be made, 80% is not as strong as it appears. In the same veign, it appears that essentially none of our predictors is useful for identifying the cost of a claim other than the current value of the car. The $R^2$ on the linear regression was particularly poor at about 0.02. No modeling techniques were able to overcome this barrier.

In a way, this exploration has only reaffirmed the need for insurance. Car collisions are somewhat rare and can happen to anyone at any time. The likelyhood and cost of the incident can vary widely. As a result of not being able to predict accurately when such incidents will occur, we require that all drivers protect themselves with insurance.

```
evaluation.data <- read_csv('C:\\Users\\Brian\\Desktop\\GradClasses\\Summer18\\621\\621week4\\insurance-
  select(-INDEX) %>% #don't need index
  mutate(TARGET_FLAG = factor(TARGET_FLAG),
         INCOME = as.numeric(str_replace_all(INCOME, '\\D', '')),
         PARENT1.CAT = factor(ifelse(PARENT1 == 'Yes', 1, 0)),
         HOME_VAL = as.numeric(str_replace_all(HOME_VAL, '\\D', '')),
         MSTATUS.CAT = factor(ifelse(MSTATUS == 'Yes', 1, 0)),
         MALE.CAT = factor(ifelse(SEX == 'M', 1, 0)),
         EDUCATION.CAT = factor(ifelse(EDUCATION == 'z_High School', 'High School', EDUCATION)),
         JOB.CAT = factor(ifelse(JOB == 'z_Blue Collar', 'Blue Collar', JOB)),
         PRIVATE.CAT = factor(ifelse(CAR_USE == 'Private', 1, 0)),
         BLUEBOOK = as.numeric(str_replace_all(BLUEBOOK, '\\D', '')),
         CAR_TYPE.CAT = factor(ifelse(CAR_TYPE == 'z_SUV', 'SUV', CAR_TYPE)),
         RED_CAR.CAT = factor(ifelse(RED_CAR == 'yes', 1, 0)),
         OLDCLAIM = as.numeric(str_replace_all(OLDCLAIM, '\\D', '')),
         REVOKED.CAT = factor(ifelse(REVOKED == 'Yes', 1, 0)),
         URBAN.CAT = factor(ifelse(URBANICITY == 'Highly Urban/ Urban', 1, 0)),
         MVR_PTS2 = MVR_PTS*MVR_PTS) %>%
  select(-SEX, -URBANICITY, -CAR_USE, -MSTATUS, -PARENT1, -JOB, -CAR_TYPE, -REVOKED, -RED_CAR, -EDUCATI

write_csv(as_data_frame(predict(model.2.diag, newdata=evaluation.data, type='response')), 'logistic_reg
write_csv(as_data_frame(predict(ln.model.2, newdata=evaluation.data, interval='confidence')), 'linear_r
```

# APPENDIX

## LOGISTIC REGRESSION

### Model 1 Selection

```
model.1.full <- glm(TARGET_FLAG ~ PARENT1.CAT + MSTATUS.CAT + MALE.CAT + EDUCATION.CAT +
                      JOB.CAT + PRIVATE.CAT + CAR_TYPE.CAT +
                      RED_CAR.CAT + REVOKED.CAT + URBAN.CAT, family=binomial, data=log.training)
drop1(model.1.full)
```

| | Df | Deviance | AIC |
|---|---|---|---|
| | NA | 6196.277 | 6244.277 |
| PARENT1.CAT | 1 | 6238.009 | 6284.009 |
| MSTATUS.CAT | 1 | 6258.423 | 6304.423 |
| MALE.CAT | 1 | 6199.063 | 6245.063 |
| EDUCATION.CAT | 4 | 6266.443 | 6306.443 |
| JOB.CAT | 7 | 6265.210 | 6299.210 |
| PRIVATE.CAT | 1 | 6268.689 | 6314.689 |
| CAR_TYPE.CAT | 5 | 6323.457 | 6361.457 |

| | Df | Deviance | AIC |
|---|---|---|---|
| RED_CAR.CAT | 1 | 6197.050 | 6243.050 |
| REVOKED.CAT | 1 | 6263.184 | 6309.184 |
| URBAN.CAT | 1 | 6836.627 | 6882.627 |

```
MASS::stepAIC(model.1.full, trace=0)
```

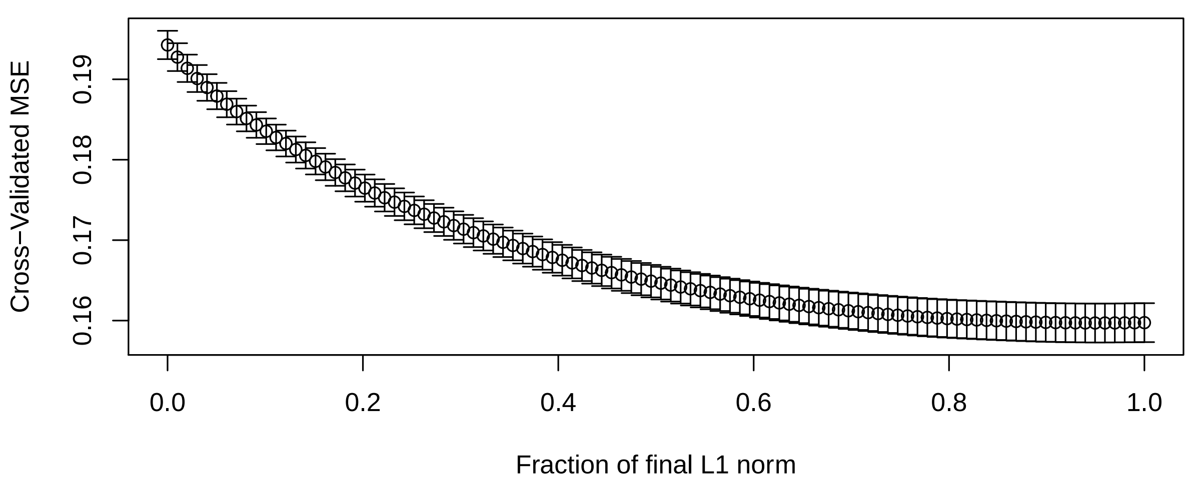
```
##
## Call:  glm(formula = TARGET_FLAG ~ PARENT1.CAT + MSTATUS.CAT + MALE.CAT +
##     EDUCATION.CAT + JOB.CAT + PRIVATE.CAT + CAR_TYPE.CAT + REVOKED.CAT +
##     URBAN.CAT, family = binomial, data = log.training)
##
## Coefficients:
##              (Intercept)               PARENT1.CAT1
##                 -2.80573                    0.63961
##             MSTATUS.CAT1                  MALE.CAT1
##                 -0.57838                    0.23577
##   EDUCATION.CATBachelors  EDUCATION.CATHigh School
##                 -0.62693                   -0.11482
##     EDUCATION.CATMasters           EDUCATION.CATPhD
##                 -0.89400                   -0.95240
##         JOB.CATClerical             JOB.CATDoctor
##                  0.17628                   -0.57962
##        JOB.CATHome Maker             JOB.CATLawyer
##                  0.34218                    0.10660
##           JOB.CATManager       JOB.CATProfessional
##                 -0.63456                   -0.05117
##           JOB.CATStudent               PRIVATE.CAT1
##                  0.31480                   -0.81476
##  CAR_TYPE.CATPanel Truck        CAR_TYPE.CATPickup
##                  0.09198                    0.58310
##    CAR_TYPE.CATSports Car           CAR_TYPE.CATSUV
##                  1.25444                    1.05847
##          CAR_TYPE.CATVan               REVOKED.CAT1
##                  0.42118                    0.72426
##               URBAN.CAT1
##                  2.44877
##
## Degrees of Freedom: 6529 Total (i.e. Null);  6507 Residual
## Null Deviance:       7536
## Residual Deviance: 6197  AIC: 6243
```

```
MASS::stepAIC(model.1.full, k=log(nrow(log.training)), trace=0)
```

```
##
## Call:  glm(formula = TARGET_FLAG ~ PARENT1.CAT + MSTATUS.CAT + EDUCATION.CAT +
##     JOB.CAT + PRIVATE.CAT + CAR_TYPE.CAT + REVOKED.CAT + URBAN.CAT,
##     family = binomial, data = log.training)
##
## Coefficients:
##              (Intercept)               PARENT1.CAT1
##                 -2.64083                    0.62287
##             MSTATUS.CAT1    EDUCATION.CATBachelors
##                 -0.58112                   -0.63460
```

```
## EDUCATION.CATHigh School      EDUCATION.CATMasters
##                 -0.11741                 -0.90163
##         EDUCATION.CATPhD          JOB.CATClerical
##                 -0.98037                  0.17777
##            JOB.CATDoctor         JOB.CATHome Maker
##                 -0.56865                  0.31235
##            JOB.CATLawyer           JOB.CATManager
##                  0.10457                 -0.63257
##      JOB.CATProfessional           JOB.CATStudent
##                 -0.05546                  0.32244
##             PRIVATE.CAT1   CAR_TYPE.CATPanel Truck
##                 -0.81017                  0.16522
##        CAR_TYPE.CATPickup    CAR_TYPE.CATSports Car
##                  0.59304                  1.10501
##           CAR_TYPE.CATSUV          CAR_TYPE.CATVan
##                  0.91069                  0.47154
##             REVOKED.CAT1               URBAN.CAT1
##                  0.72789                  2.45024
##
## Degrees of Freedom: 6529 Total (i.e. Null);  6508 Residual
## Null Deviance:       7536
## Residual Deviance: 6203  AIC: 6247
```

```
set.seed(123)
model.1.lasso <- lars(model.matrix(~ PARENT1.CAT + MSTATUS.CAT + MALE.CAT + EDUCATION.CAT
                                   + JOB.CAT + PRIVATE.CAT + CAR_TYPE.CAT +
                                     RED_CAR.CAT + REVOKED.CAT + URBAN.CAT, log.training),
                      as.numeric(log.training$TARGET_FLAG))
cvlmod <- cv.lars(model.matrix(~ PARENT1.CAT + MSTATUS.CAT + MALE.CAT + EDUCATION.CAT +
                                 JOB.CAT + PRIVATE.CAT + CAR_TYPE.CAT +
                                 RED_CAR.CAT + REVOKED.CAT + URBAN.CAT, log.training),
                  as.numeric(log.training$TARGET_FLAG))
```



```r
predict(model.1.lasso, s=0.9494949, type='coef', mode='fraction')$coef
```

```
##              (Intercept)              PARENT1.CAT1              MSTATUS.CAT1
##             0.0000000000              0.1213647128             -0.0874451342
##                 MALE.CAT1   EDUCATION.CATBachelors EDUCATION.CATHigh School
##             0.0220785941             -0.0879004191             -0.0010461384
##      EDUCATION.CATMasters         EDUCATION.CATPhD          JOB.CATClerical
##            -0.1263733918             -0.1422308833              0.0236443618
##            JOB.CATDoctor         JOB.CATHome Maker           JOB.CATLawyer
##            -0.0914008251              0.0505833464             -0.0004754156
##           JOB.CATManager       JOB.CATProfessional          JOB.CATStudent
##            -0.1175211570             -0.0163205022              0.0488848878
##             PRIVATE.CAT1   CAR_TYPE.CATPanel Truck       CAR_TYPE.CATPickup
##            -0.1310238177             -0.0051469131              0.0767786518
##     CAR_TYPE.CATSports Car           CAR_TYPE.CATSUV          CAR_TYPE.CATVan
##             0.1742879086              0.1463183928              0.0477593605
##             RED_CAR.CAT1              REVOKED.CAT1               URBAN.CAT1
##             0.0095991727              0.1333193452              0.3180415792
```

## Model 2 Selection

```r
model.2.full <- glm(TARGET_FLAG ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME +
                      HOME_VAL + TRAVTIME + BLUEBOOK + TIF + OLDCLAIM + CLM_FREQ +
                      MVR_PTS + CAR_AGE + PARENT1.CAT + MSTATUS.CAT + MALE.CAT +
                      EDUCATION.CAT + JOB.CAT + PRIVATE.CAT + CAR_TYPE.CAT +
                      RED_CAR.CAT + REVOKED.CAT + URBAN.CAT, data=log.training, family=binomial)
drop1(model.2.full)
```

| | Df | Deviance | AIC |
|---|---|---|---|
| | NA | 5794.729 | 5868.729 |
| KIDSDRIV | 1 | 5825.885 | 5897.885 |
| AGE | 1 | 5795.377 | 5867.377 |
| HOMEKIDS | 1 | 5795.842 | 5867.842 |
| YOJ | 1 | 5796.558 | 5868.558 |
| INCOME | 1 | 5802.724 | 5874.724 |
| HOME_VAL | 1 | 5809.702 | 5881.702 |
| TRAVTIME | 1 | 5847.488 | 5919.488 |
| BLUEBOOK | 1 | 5809.055 | 5881.055 |
| TIF | 1 | 5840.861 | 5912.861 |
| OLDCLAIM | 1 | 5807.171 | 5879.171 |
| CLM_FREQ | 1 | 5835.207 | 5907.207 |
| MVR_PTS | 1 | 5862.213 | 5934.213 |
| CAR_AGE | 1 | 5794.819 | 5866.819 |
| PARENT1.CAT | 1 | 5804.437 | 5876.437 |
| MSTATUS.CAT | 1 | 5820.708 | 5892.708 |
| MALE.CAT | 1 | 5794.750 | 5866.750 |
| EDUCATION.CAT | 4 | 5813.004 | 5879.004 |
| JOB.CAT | 7 | 5831.347 | 5891.347 |
| PRIVATE.CAT | 1 | 5859.021 | 5931.021 |
| CAR_TYPE.CAT | 5 | 5865.533 | 5929.533 |
| RED_CAR.CAT | 1 | 5794.850 | 5866.850 |
| REVOKED.CAT | 1 | 5858.895 | 5930.895 |

| | Df | Deviance | AIC |
|---|---|---|---|
| URBAN.CAT | 1 | 6345.545 | 6417.545 |

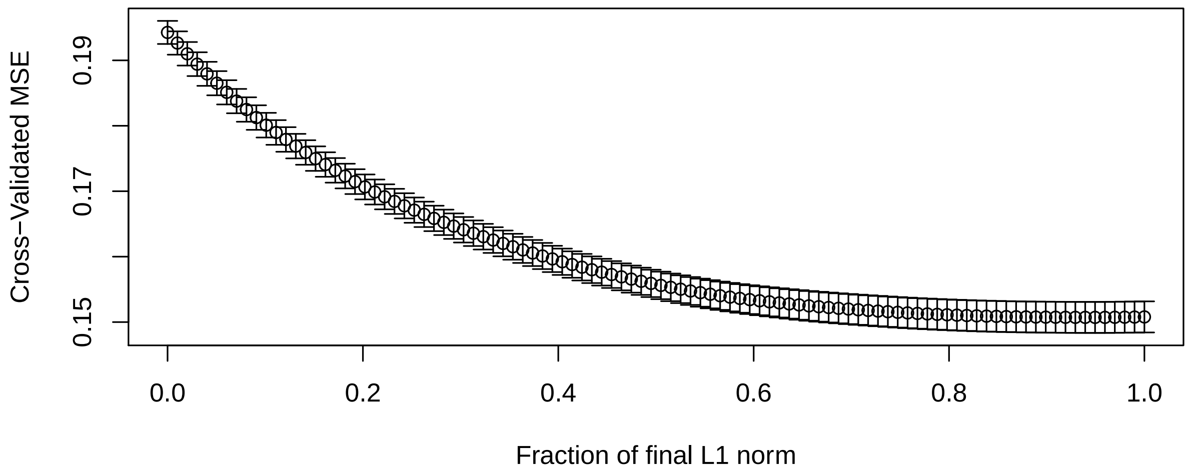
```
MASS::stepAIC(model.2.full, trace=0)
```

```
##
## Call:  glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME +
##     HOME_VAL + TRAVTIME + BLUEBOOK + TIF + OLDCLAIM + CLM_FREQ +
##     MVR_PTS + PARENT1.CAT + MSTATUS.CAT + EDUCATION.CAT + JOB.CAT +
##     PRIVATE.CAT + CAR_TYPE.CAT + REVOKED.CAT + URBAN.CAT, family = binomial,
##     data = log.training)
##
## Coefficients:
##             (Intercept)                 KIDSDRIV
##              -2.550e+00                3.739e-01
##                HOMEKIDS                      YOJ
##               5.603e-02               -1.403e-02
##                  INCOME                 HOME_VAL
##              -3.404e-06               -1.539e-06
##                TRAVTIME                 BLUEBOOK
##               1.536e-02               -2.364e-05
##                     TIF                 OLDCLAIM
##              -5.456e-02               -1.553e-05
##                CLM_FREQ                  MVR_PTS
##               2.054e-01                1.255e-01
##             PARENT1.CAT1             MSTATUS.CAT1
##               3.923e-01               -4.920e-01
##   EDUCATION.CATBachelors  EDUCATION.CATHigh School
##              -4.421e-01               -4.212e-02
##     EDUCATION.CATMasters         EDUCATION.CATPhD
##              -5.408e-01               -3.541e-01
##         JOB.CATClerical           JOB.CATDoctor
##               7.109e-02               -6.796e-01
##        JOB.CATHome Maker           JOB.CATLawyer
##              -1.060e-01                3.385e-02
##          JOB.CATManager      JOB.CATProfessional
##              -6.089e-01               -6.452e-02
##          JOB.CATStudent             PRIVATE.CAT1
##              -1.775e-01               -7.964e-01
##  CAR_TYPE.CATPanel Truck      CAR_TYPE.CATPickup
##               4.998e-01                5.065e-01
##   CAR_TYPE.CATSports Car         CAR_TYPE.CATSUV
##               9.704e-01                7.782e-01
##         CAR_TYPE.CATVan             REVOKED.CAT1
##               6.174e-01                8.315e-01
##              URBAN.CAT1
##               2.474e+00
##
## Degrees of Freedom: 6529 Total (i.e. Null);  6497 Residual
## Null Deviance:       7536
## Residual Deviance: 5796  AIC: 5862
```

```r
MASS::stepAIC(model.2.full, k=log(nrow(log.training)), trace=0)
```

```
##
## Call:  glm(formula = TARGET_FLAG ~ KIDSDRIV + INCOME + HOME_VAL + TRAVTIME +
##     BLUEBOOK + TIF + OLDCLAIM + CLM_FREQ + MVR_PTS + PARENT1.CAT +
##     MSTATUS.CAT + EDUCATION.CAT + PRIVATE.CAT + CAR_TYPE.CAT +
##     REVOKED.CAT + URBAN.CAT, family = binomial, data = log.training)
##
## Coefficients:
##             (Intercept)                 KIDSDRIV
##              -2.593e+00                4.041e-01
##                  INCOME                 HOME_VAL
##              -4.002e-06               -1.472e-06
##                TRAVTIME                 BLUEBOOK
##               1.557e-02               -2.435e-05
##                     TIF                 OLDCLAIM
##              -5.334e-02               -1.559e-05
##                CLM_FREQ                  MVR_PTS
##               2.036e-01                1.310e-01
##            PARENT1.CAT1             MSTATUS.CAT1
##               4.801e-01               -4.755e-01
##   EDUCATION.CATBachelors  EDUCATION.CATHigh School
##              -5.773e-01               -1.005e-01
##     EDUCATION.CATMasters         EDUCATION.CATPhD
##              -6.551e-01               -6.588e-01
##            PRIVATE.CAT1   CAR_TYPE.CATPanel Truck
##              -8.501e-01                4.477e-01
##      CAR_TYPE.CATPickup    CAR_TYPE.CATSports Car
##               4.673e-01                9.633e-01
##         CAR_TYPE.CATSUV          CAR_TYPE.CATVan
##               7.785e-01                5.822e-01
##            REVOKED.CAT1               URBAN.CAT1
##               8.430e-01                2.432e+00
##
## Degrees of Freedom: 6529 Total (i.e. Null);  6506 Residual
## Null Deviance:       7536
## Residual Deviance: 5837  AIC: 5885
```

```r
set.seed(123)
model.2.lasso <- lars(model.matrix(~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + HOME_VAL +
                                     TRAVTIME + BLUEBOOK + TIF + OLDCLAIM + CLM_FREQ +
                                     MVR_PTS + CAR_AGE + PARENT1.CAT + MSTATUS.CAT +
                                     MALE.CAT + EDUCATION.CAT + JOB.CAT +
                                     PRIVATE.CAT + CAR_TYPE.CAT + RED_CAR.CAT + REVOKED.CAT +
                                     URBAN.CAT, log.training), as.numeric(log.training$TARGET_FLAG))
cvlmod <- cv.lars(model.matrix(~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + HOME_VAL +
                                 TRAVTIME + BLUEBOOK + TIF + OLDCLAIM + CLM_FREQ +
                                 MVR_PTS + CAR_AGE + PARENT1.CAT + MSTATUS.CAT + MALE.CAT +
                                 EDUCATION.CAT + JOB.CAT +
                                 PRIVATE.CAT + CAR_TYPE.CAT + RED_CAR.CAT + REVOKED.CAT +
                                 URBAN.CAT, log.training), as.numeric(log.training$TARGET_FLAG))
```




```
cvlmod$index[which.min(cvlmod$cv)]
```

```
## [1] 0.9494949
```

```
predict(model.2.lasso, s=0.9292929, type='coef', mode='fraction')$coef
```

```
##              (Intercept)                 KIDSDRIV                      AGE
##             0.000000e+00             5.723853e-02            -4.911684e-04
##                 HOMEKIDS                      YOJ                   INCOME
##             5.099281e-03            -2.040275e-03            -4.039648e-07
##                 HOME_VAL                 TRAVTIME                 BLUEBOOK
##            -1.930173e-07             1.992438e-03            -2.696929e-06
##                      TIF                 OLDCLAIM                 CLM_FREQ
##            -7.500450e-03            -2.215477e-06             3.313635e-02
##                  MVR_PTS                  CAR_AGE             PARENT1.CAT1
##             2.287662e-02            -7.243079e-04             7.881198e-02
##             MSTATUS.CAT1                MALE.CAT1   EDUCATION.CATBachelors
##            -6.689047e-02             0.000000e+00            -5.902253e-02
## EDUCATION.CATHigh School     EDUCATION.CATMasters         EDUCATION.CATPhD
##             2.609209e-03            -6.785259e-02            -4.929497e-02
##         JOB.CATClerical            JOB.CATDoctor       JOB.CATHome Maker
##             1.263407e-02            -7.684859e-02             0.000000e+00
##           JOB.CATLawyer           JOB.CATManager      JOB.CATProfessional
##             0.000000e+00            -9.334725e-02            -4.739336e-03
##          JOB.CATStudent             PRIVATE.CAT1  CAR_TYPE.CATPanel Truck
##            -8.742197e-03            -1.262808e-01             2.559194e-02
##      CAR_TYPE.CATPickup   CAR_TYPE.CATSports Car           CAR_TYPE.CATSUV
##             5.385015e-02             1.229269e-01             9.723174e-02
##         CAR_TYPE.CATVan             RED_CAR.CAT1             REVOKED.CAT1
##             5.790253e-02             1.359614e-04             1.352238e-01
##               URBAN.CAT1
##             2.942721e-01
```

## Model 3 Selection

```
model.3.full <- glm(TARGET_FLAG ~ KIDSDRIV + I(KIDSDRIV ^ 2) + AGE + HOMEKIDS +
                      I(HOMEKIDS ^ 2) + YOJ + INCOME + HOME_VAL + TRAVTIME + BLUEBOOK +
                      TIF + OLDCLAIM + CLM_FREQ + I(CLM_FREQ ^ 2) + MVR_PTS + I(MVR_PTS ^ 2) +
                      CAR_AGE + PARENT1.CAT + MSTATUS.CAT + MALE.CAT + EDUCATION.CAT + JOB.CAT +
                      PRIVATE.CAT + CAR_TYPE.CAT + RED_CAR.CAT + REVOKED.CAT +
                      URBAN.CAT, data=log.training, family=binomial)
drop1(model.3.full)
```

| | Df | Deviance | AIC |
|---|---|---|---|
| | NA | 5781.167 | 5863.167 |
| KIDSDRIV | 1 | 5788.018 | 5868.018 |
| I(KIDSDRIV^2) | 1 | 5781.500 | 5861.500 |
| AGE | 1 | 5781.530 | 5861.530 |
| HOMEKIDS | 1 | 5783.051 | 5863.051 |
| I(HOMEKIDS^2) | 1 | 5782.358 | 5862.358 |
| YOJ | 1 | 5783.111 | 5863.111 |
| INCOME | 1 | 5789.711 | 5869.711 |
| HOME_VAL | 1 | 5794.815 | 5874.815 |
| TRAVTIME | 1 | 5833.679 | 5913.679 |
| BLUEBOOK | 1 | 5795.568 | 5875.568 |
| TIF | 1 | 5827.674 | 5907.674 |
| OLDCLAIM | 1 | 5797.943 | 5877.943 |
| CLM_FREQ | 1 | 5801.231 | 5881.231 |
| I(CLM_FREQ^2) | 1 | 5786.990 | 5866.990 |
| MVR_PTS | 1 | 5781.876 | 5861.876 |
| I(MVR_PTS^2) | 1 | 5786.662 | 5866.662 |
| CAR_AGE | 1 | 5781.205 | 5861.205 |
| PARENT1.CAT | 1 | 5786.976 | 5866.976 |
| MSTATUS.CAT | 1 | 5809.044 | 5889.044 |
| MALE.CAT | 1 | 5781.173 | 5861.173 |
| EDUCATION.CAT | 4 | 5799.959 | 5873.959 |
| JOB.CAT | 7 | 5817.394 | 5885.394 |
| PRIVATE.CAT | 1 | 5843.400 | 5923.400 |
| CAR_TYPE.CAT | 5 | 5851.549 | 5923.549 |
| RED_CAR.CAT | 1 | 5781.345 | 5861.345 |
| REVOKED.CAT | 1 | 5849.739 | 5929.739 |
| URBAN.CAT | 1 | 6312.472 | 6392.472 |

```
MASS::stepAIC(model.3.full, trace=0)
```

```
##
## Call:  glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + I(HOMEKIDS^2) +
##     YOJ + INCOME + HOME_VAL + TRAVTIME + BLUEBOOK + TIF + OLDCLAIM +
##     CLM_FREQ + I(CLM_FREQ^2) + I(MVR_PTS^2) + PARENT1.CAT + MSTATUS.CAT +
##     EDUCATION.CAT + JOB.CAT + PRIVATE.CAT + CAR_TYPE.CAT + REVOKED.CAT +
##     URBAN.CAT, family = binomial, data = log.training)
##
## Coefficients:
##              (Intercept)                     KIDSDRIV
##               -2.501e+00                    3.641e-01
```

```
##                HOMEKIDS            I(HOMEKIDS^2)
##               1.945e-01                -4.214e-02
##                     YOJ                    INCOME
##              -1.416e-02                -3.544e-06
##                HOME_VAL                  TRAVTIME
##              -1.467e-06                 1.536e-02
##                BLUEBOOK                       TIF
##              -2.352e-05                -5.497e-02
##                OLDCLAIM                  CLM_FREQ
##              -1.891e-05                 4.330e-01
##           I(CLM_FREQ^2)              I(MVR_PTS^2)
##              -6.565e-02                 1.734e-02
##             PARENT1.CAT1              MSTATUS.CAT1
##               3.197e-01                -5.224e-01
##   EDUCATION.CATBachelors  EDUCATION.CATHigh School
##              -4.328e-01                -3.060e-02
##     EDUCATION.CATMasters          EDUCATION.CATPhD
##              -5.246e-01                -3.163e-01
##         JOB.CATClerical             JOB.CATDoctor
##               5.827e-02                -6.947e-01
##       JOB.CATHome Maker             JOB.CATLawyer
##              -1.216e-01                 2.454e-02
##          JOB.CATManager       JOB.CATProfessional
##              -6.187e-01                -7.278e-02
##          JOB.CATStudent              PRIVATE.CAT1
##              -1.827e-01                -7.849e-01
##  CAR_TYPE.CATPanel Truck        CAR_TYPE.CATPickup
##               5.142e-01                 5.095e-01
##   CAR_TYPE.CATSports Car           CAR_TYPE.CATSUV
##               9.679e-01                 7.743e-01
##          CAR_TYPE.CATVan              REVOKED.CAT1
##               6.268e-01                 8.684e-01
##              URBAN.CAT1
##               2.453e+00
##
## Degrees of Freedom: 6529 Total (i.e. Null);  6495 Residual
## Null Deviance:       7536
## Residual Deviance: 5783  AIC: 5853
```

```
MASS::stepAIC(model.3.full, k=log(nrow(log.training)), trace=0)
```

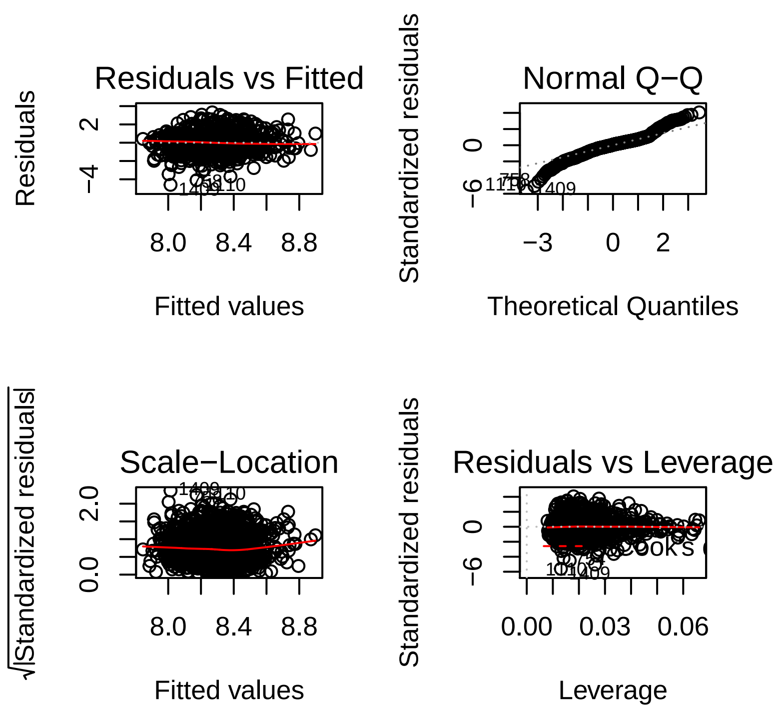
```
##
## Call:  glm(formula = TARGET_FLAG ~ KIDSDRIV + INCOME + HOME_VAL + TRAVTIME +
##     BLUEBOOK + TIF + OLDCLAIM + CLM_FREQ + I(MVR_PTS^2) + PARENT1.CAT +
##     MSTATUS.CAT + EDUCATION.CAT + PRIVATE.CAT + CAR_TYPE.CAT +
##     REVOKED.CAT + URBAN.CAT, family = binomial, data = log.training)
##
## Coefficients:
##             (Intercept)                  KIDSDRIV
##              -2.530e+00                 4.080e-01
##                  INCOME                  HOME_VAL
##              -4.010e-06                -1.444e-06
##                TRAVTIME                  BLUEBOOK
##               1.554e-02                -2.434e-05
##                     TIF                  OLDCLAIM
```

```
##                 -5.386e-02                    -1.514e-05
##                   CLM_FREQ                 I(MVR_PTS^2)
##                  2.122e-01                     1.920e-02
##               PARENT1.CAT1                  MSTATUS.CAT1
##                  4.780e-01                    -4.777e-01
##     EDUCATION.CATBachelors  EDUCATION.CATHigh School
##                 -5.739e-01                    -9.378e-02
##       EDUCATION.CATMasters             EDUCATION.CATPhD
##                 -6.542e-01                    -6.556e-01
##               PRIVATE.CAT1   CAR_TYPE.CATPanel Truck
##                 -8.460e-01                     4.577e-01
##         CAR_TYPE.CATPickup    CAR_TYPE.CATSports Car
##                  4.725e-01                     9.672e-01
##            CAR_TYPE.CATSUV              CAR_TYPE.CATVan
##                  7.820e-01                     5.879e-01
##               REVOKED.CAT1                   URBAN.CAT1
##                  8.376e-01                     2.437e+00
##
## Degrees of Freedom: 6529 Total (i.e. Null);  6506 Residual
## Null Deviance:       7536
## Residual Deviance: 5833  AIC: 5881
```

## Model 4 Selection

Model 4 shows no apparent outliers or bad leverage points. This is a valid regression.

```r
par(mfrow=c(2,2))
plot(ln.model.1)
```



## Model 5 Selection

Model 5 was simplified using stepAIC

```
MASS::stepAIC(ln.model.1, trace=0)
```

```
##
## Call:
## lm(formula = log(TARGET_AMT) ~ BLUEBOOK + CLM_FREQ + MVR_PTS +
##     MSTATUS.CAT + RED_CAR.CAT, data = lin.training)
##
## Coefficients:
##  (Intercept)      BLUEBOOK      CLM_FREQ       MVR_PTS  MSTATUS.CAT1
##    8.118e+00     1.134e-05    -3.224e-02     1.588e-02    -6.964e-02
## RED_CAR.CAT1
##    8.049e-02
```

## DIAGNOSTICS

## LOGISTIC REGRESSION

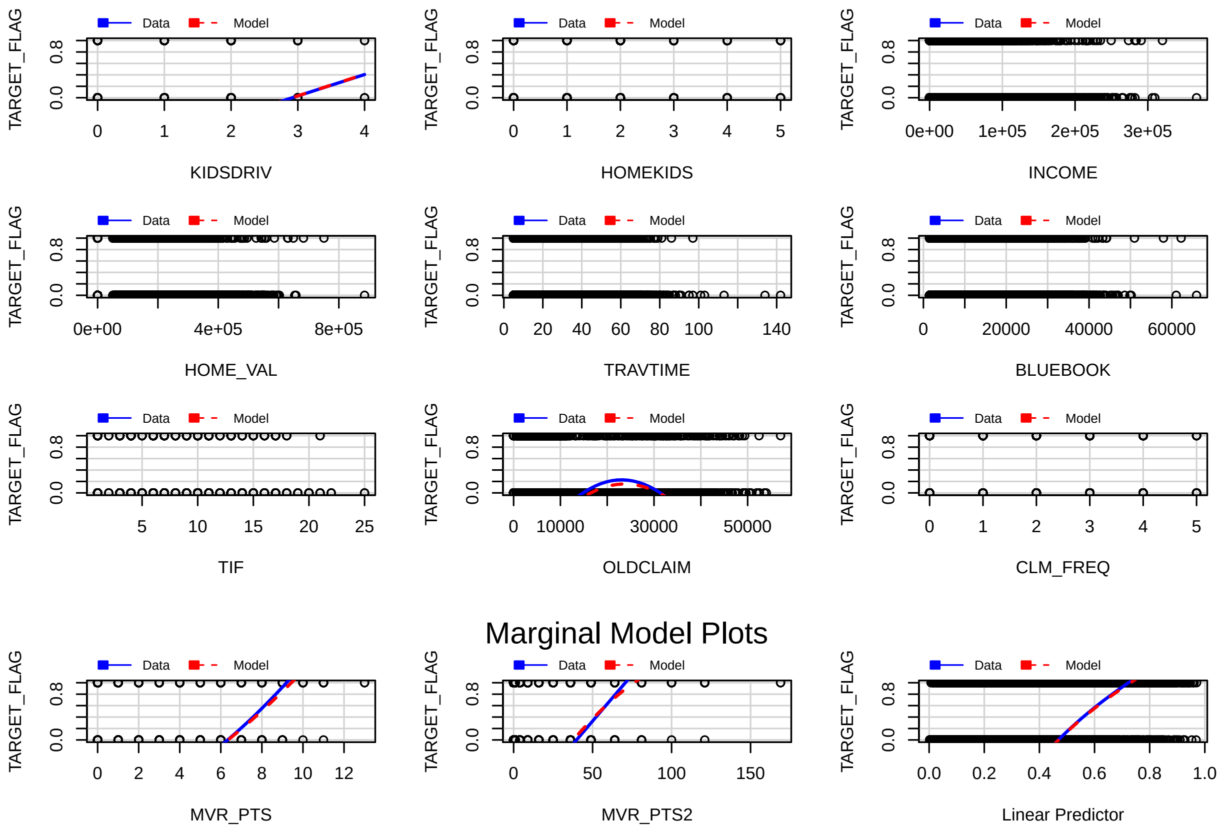
Multi-collinearity is not an issue for any of the three models.

```
car::vif(model.2.diag)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## KIDSDRIV        1.309855  1        1.144489
## HOMEKIDS        1.828832  1        1.352343
## INCOME          2.616859  1        1.617671
## HOME_VAL        2.032597  1        1.425692
## TRAVTIME        1.039361  1        1.019491
## BLUEBOOK        1.740927  1        1.319442
## TIF             1.010203  1        1.005088
## OLDCLAIM        1.655198  1        1.286545
## CLM_FREQ        1.473428  1        1.213849
## MVR_PTS         7.481396  1        2.735214
## PARENT1.CAT     1.897780  1        1.377599
## MSTATUS.CAT     2.135983  1        1.461500
## EDUCATION.CAT 5.091979  4        1.225634
## JOB.CAT         7.906833  7        1.159159
## PRIVATE.CAT     2.319325  1        1.522933
## CAR_TYPE.CAT    2.558079  5        1.098478
## REVOKED.CAT     1.293346  1        1.137254
## URBAN.CAT       1.138633  1        1.067067
## MVR_PTS2        7.217232  1        2.686491
```

Finally, I have the mmps plots to demonstrate a good fit.

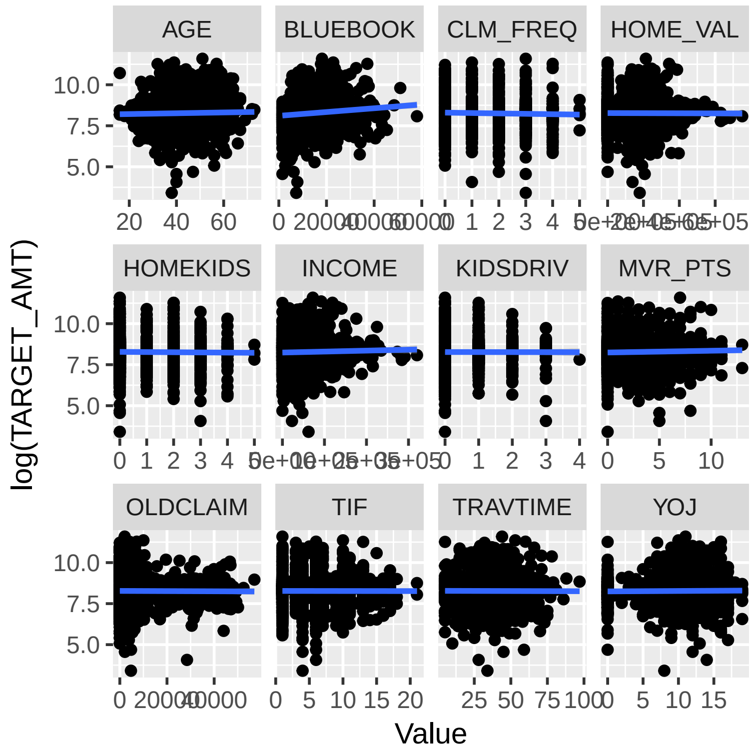
```
car::marginalModelPlots(model.2.diag)
```




# LINEAR REGRESSION

Examination of the predictors against the response variable TARGET_AMT reveal that TARGET_AMT should be log transformed.

```
lin.training %>%
  select_if(is.numeric) %>%
  gather(key='Predictor', value='Value', 2:13) %>%
  ggplot(aes(Value, log(TARGET_AMT))) +
  geom_point() +
  geom_smooth(method='lm') +
  facet_wrap(~Predictor, scale='free_x')
```



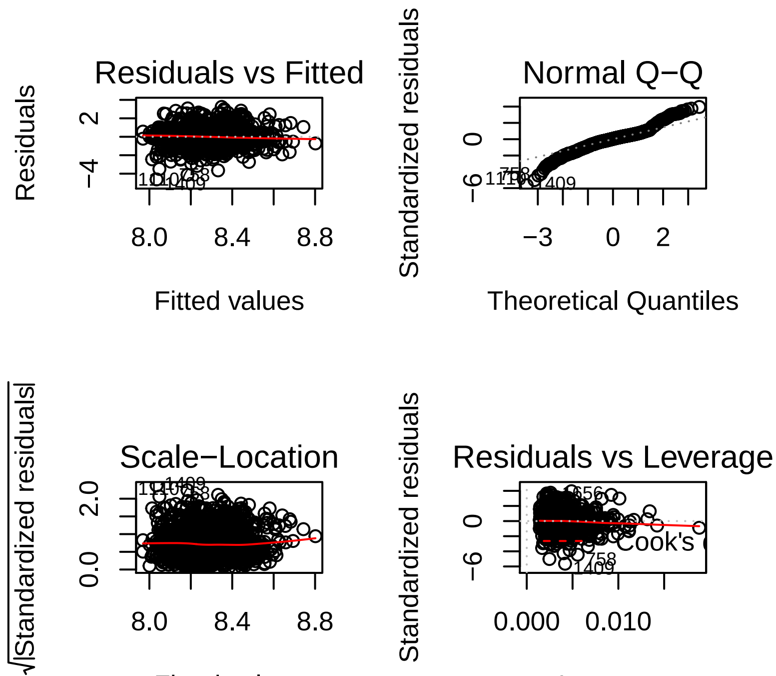

Multi-collinearity is not an issue for any of the three models.

```
car::vif(ln.model.2)
```

```
##     BLUEBOOK     CLM_FREQ      MVR_PTS MSTATUS.CAT RED_CAR.CAT
##     1.002772     1.103982     1.104214    1.000183    1.002189
```

A plot of diagnostics and mmps demonstrate that this is a valid regression. There are a number of lower outliers present due to the extremely small size of the claim.

```
par(mfrow=c(2,2))
plot(ln.model.2)
```




```
car::mmps(ln.model.2)
```