

DATA621 - HW#4

Mia Chen, Wei Zhou

4/26/2020

Overview

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

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

1. DATA EXPLORATION

Describe the size and the variables in the insurance training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren’t doing your job.

Data acquisition

```
train = read.csv("https://raw.githubusercontent.com/miachen410/DATA621/master/HW%234/insurance_training
```

Data structure

There are 8161 observations and 26 variables in the training dataset.

```
dim(train)
```

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

We want to get rid of the \$ and , in the numerical data and z_ in the categorical data:

```

currencyconv = function(input) {
  out = sub("\\\\$\\", "", input)
  out = as.numeric(sub("\\\\", "", out))
  return(out)
}

# Replace spaces with underscores
underscore = function(input) {
  out = sub(" ", "_", input)
  return(out)
}

train = as.tbl(train) %>%
  mutate_at(c("INCOME", "HOME_VAL", "BLUEBOOK", "OLDCLAIM"),
            currencyconv) %>%
  mutate_at(c("EDUCATION", "JOB", "CAR_TYPE", "URBANICITY"),
            underscore) %>%
  mutate_at(c("EDUCATION", "JOB", "CAR_TYPE", "URBANICITY"),
            as.factor) %>%
  mutate(TARGET_FLAG = as.factor(TARGET_FLAG))

```

Let's look at the data structure again:

```
summary(train) %>% kable() %>% kable_styling()
```

| INDEX | TARGET_FLAG | TARGET_AMT | KIDSDRV | AGE | HOMEKIDS | YOJ |
|---------------|-------------|---------------|----------------|---------------|----------------|------------|
| Min. : 1 | 0:6008 | Min. : 0 | Min. :0.0000 | Min. :16.00 | Min. :0.0000 | Min. : 0. |
| 1st Qu.: 2559 | 1:2153 | 1st Qu.: 0 | 1st Qu.:0.0000 | 1st Qu.:39.00 | 1st Qu.:0.0000 | 1st Qu.: |
| Median : 5133 | NA | Median : 0 | Median :0.0000 | Median :45.00 | Median :0.0000 | Median : |
| Mean : 5152 | NA | Mean : 1504 | Mean :0.1711 | Mean :44.79 | Mean :0.7212 | Mean :10. |
| 3rd Qu.: 7745 | NA | 3rd Qu.: 1036 | 3rd Qu.:0.0000 | 3rd Qu.:51.00 | 3rd Qu.:1.0000 | 3rd Qu.:1. |
| Max. :10302 | NA | Max. :107586 | Max. :4.0000 | Max. :81.00 | Max. :5.0000 | Max. :23. |
| NA | NA | NA | NA | NA's :6 | NA | NA's :45 |

```
sapply(train, function(x) sum(is.na(x))) %>% kable() %>% kable_styling()
```

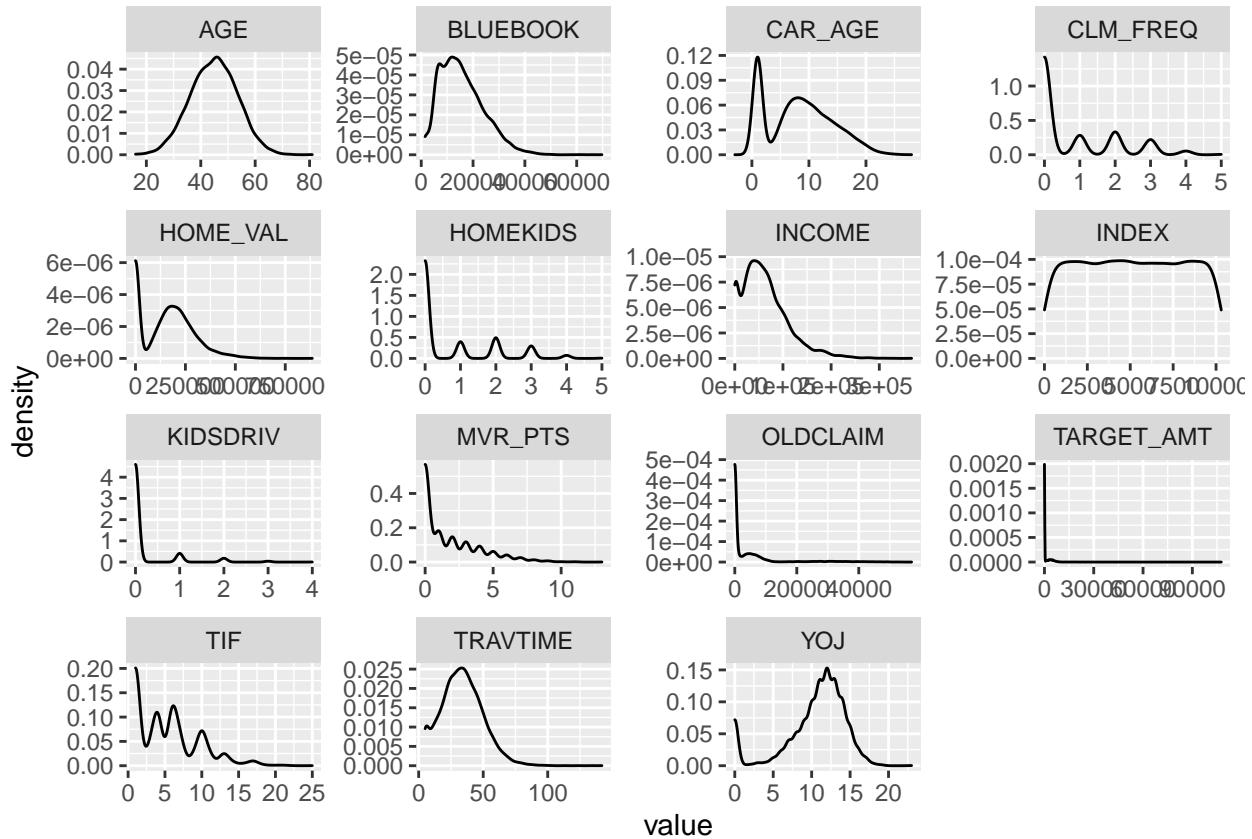
| | x |
|-------------|-----|
| INDEX | 0 |
| TARGET_FLAG | 0 |
| TARGET_AMT | 0 |
| KIDSDRV | 0 |
| AGE | 6 |
| HOMEKIDS | 0 |
| YOJ | 454 |
| INCOME | 445 |
| PARENT1 | 0 |
| HOME_VAL | 464 |
| MSTATUS | 0 |
| SEX | 0 |
| EDUCATION | 0 |
| JOB | 0 |
| TRAVTIME | 0 |
| CAR_USE | 0 |
| BLUEBOOK | 0 |
| TIF | 0 |
| CAR_TYPE | 0 |
| RED_CAR | 0 |
| OLDCLAIM | 0 |
| CLM_FREQ | 0 |
| REVOKE | 0 |
| MVR_PTS | 0 |
| CAR_AGE | 510 |
| URBANICITY | 0 |

Visulization of the data set

Let's first look at the density plots of the numerical variables to view their shapes and distributions:

```
ntrain<-select_if(train, is.numeric)
ntrain %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +   # In separate panels
    geom_density()

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



2. DATA PREPARATION

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this.

Missing values

There are 970 rows of data with NA values. We are going to replacing them with their median values.

```
# impute data for missing values
# use column mean for calculation
train$AGE[is.na(train$AGE)] <- mean(train$AGE, na.rm=TRUE)
train$YOJ[is.na(train$YOJ)] <- mean(train$YOJ, na.rm=TRUE)
train$HOME_VAL[is.na(train$HOME_VAL)] <- mean(train$HOME_VAL, na.rm=TRUE)
train$CAR_AGE[is.na(train$CAR_AGE)] <- mean(train$CAR_AGE, na.rm=TRUE)
train$INCOME[is.na(train$INCOME)] <- mean(train$INCOME, na.rm=TRUE)
#get complete cases
train <- train[complete.cases(train),]
train2<-train

train <- train[, !(colnames(train) %in% c("INDEX"))]
#
# #create variable
# train$new <- train$tax / (train$medv*10)
```

```

#
trainnum <- dplyr::select_if(train, is.numeric)
rcorr(as.matrix(trainnum))

##          TARGET_AMT KIDSDRV AGE HOMEKIDS YOJ INCOME HOME_VAL TRAVTIME
## TARGET_AMT      1.00    0.06 -0.04     0.07 -0.02 -0.06   -0.08    0.03
## KIDSDRV        0.06    1.00 -0.08     0.49  0.05 -0.05   -0.02    0.01
## AGE           -0.04   -0.08  1.00    -0.47  0.13  0.18    0.20    0.01
## HOMEKIDS       0.07    0.49 -0.47     1.00  0.08 -0.16   -0.11   -0.01
## YOJ           -0.02    0.05  0.13     0.08  1.00  0.27    0.26   -0.02
## INCOME         -0.06   -0.05  0.18    -0.16  0.27  1.00    0.54   -0.05
## HOME_VAL       -0.08   -0.02  0.20    -0.11  0.26  0.54    1.00   -0.03
## TRAVTIME       0.03    0.01  0.01    -0.01 -0.02 -0.05   -0.03    1.00
## BLUEBOOK       0.00   -0.02  0.16    -0.11  0.14  0.42    0.25   -0.02
## TIF            -0.05   -0.01  0.00     0.00  0.02 -0.01    0.00   -0.01
## OLDCLAIM        0.13    0.05 -0.04     0.05 -0.02 -0.07   -0.11   -0.01
## CLM_FREQ        0.13    0.04 -0.03     0.04 -0.02 -0.05   -0.10    0.00
## MVR_PTS         0.13    0.05 -0.07     0.06 -0.03 -0.05   -0.07    0.01
## CAR_AGE         -0.06   -0.05  0.17    -0.15  0.06  0.39    0.20   -0.04
##          BLUEBOOK TIF OLDCLAIM CLM_FREQ MVR_PTS CAR_AGE
## TARGET_AMT     0.00 -0.05    0.13    0.13    0.13   -0.06
## KIDSDRV       -0.02 -0.01    0.05    0.04    0.05   -0.05
## AGE           0.16  0.00   -0.04   -0.03   -0.07   0.17
## HOMEKIDS      -0.11  0.00    0.05    0.04    0.06   -0.15
## YOJ            0.14  0.02   -0.02   -0.02   -0.03   0.06
## INCOME         0.42 -0.01   -0.07   -0.05   -0.05   0.39
## HOME_VAL       0.25  0.00   -0.11   -0.10   -0.07   0.20
## TRAVTIME      -0.02 -0.01   -0.01    0.00    0.01   -0.04
## BLUEBOOK       1.00 -0.01   -0.04   -0.04   -0.04   0.18
## TIF            -0.01  1.00   -0.03   -0.02   -0.04   0.00
## OLDCLAIM       -0.04 -0.03    1.00    0.93    0.44   -0.02
## CLM_FREQ       -0.04 -0.02    0.93    1.00    0.41   -0.01
## MVR_PTS        -0.04 -0.04    0.44    0.41    1.00   -0.01
## CAR_AGE         0.18  0.00   -0.02   -0.01   -0.01   1.00
##
##          n= 8161
##
##
##          P
##          TARGET_AMT KIDSDRV AGE HOMEKIDS YOJ INCOME HOME_VAL TRAVTIME
## TARGET_AMT      0.0000  0.0002 0.0000  0.0585 0.0000 0.0000  0.0115
## KIDSDRV        0.0000          0.0000 0.0000  0.0000 0.0000 0.0577  0.5499
## AGE            0.0002  0.0000          0.0000  0.0000 0.0000 0.0000  0.6342
## HOMEKIDS       0.0000  0.0000 0.0000          0.0000 0.0000 0.0000  0.4230
## YOJ            0.0585  0.0000 0.0000 0.0000          0.0000 0.0000  0.1362
## INCOME          0.0000  0.0000 0.0000 0.0000 0.0000          0.0000 0.0000
## HOME_VAL        0.0000  0.0577 0.0000 0.0000 0.0000 0.0000          0.0018
## TRAVTIME        0.0115  0.5499 0.6342 0.4230 0.1362 0.0000 0.0018
## BLUEBOOK        0.6712  0.0415 0.0000 0.0000 0.0000 0.0000 0.0000  0.1246
## TIF             0.0000  0.3832 0.9404 0.6725 0.0498 0.4889 0.7280  0.2945
## OLDCLAIM         0.0000  0.0000 0.0004 0.0000 0.0987 0.0000 0.0000  0.6009
## CLM_FREQ         0.0000  0.0000 0.0054 0.0002 0.0272 0.0000 0.0000  0.7501
## MVR_PTS          0.0000  0.0000 0.0000 0.0000 0.0033 0.0000 0.0000  0.5405

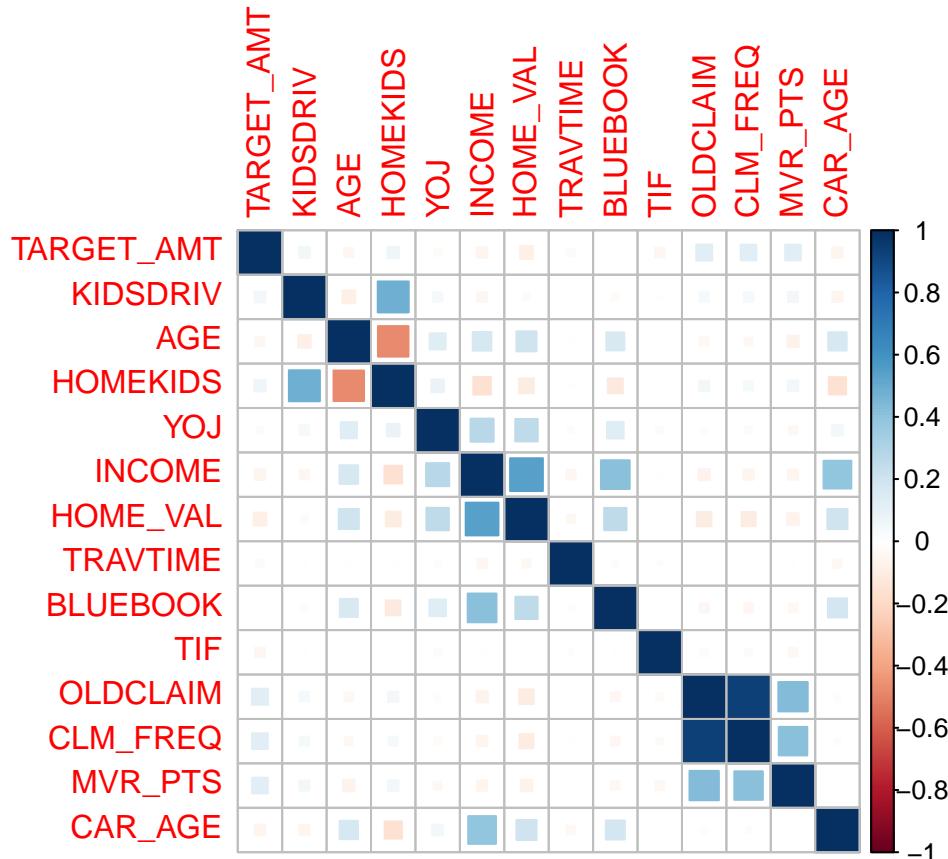
```

```

## CAR_AGE    0.0000    0.0000    0.0000  0.0000    0.0000  0.0000  0.0000    0.0009
##          BLUEBOOK   TIF    OLDCLAIM CLM_FREQ MVR PTS CAR_AGE
## TARGET_AMT 0.6712    0.0000    0.0000    0.0000    0.0000    0.0000
## KIDSDRV    0.0415    0.3832    0.0000    0.0000    0.0000    0.0000
## AGE        0.0000    0.9404    0.0004    0.0054    0.0000    0.0000
## HOMEKIDS   0.0000    0.6725    0.0000    0.0002    0.0000    0.0000
## YOJ        0.0000    0.0498    0.0987    0.0272    0.0033    0.0000
## INCOME     0.0000    0.4889    0.0000    0.0000    0.0000    0.0000
## HOME_VAL   0.0000    0.7280    0.0000    0.0000    0.0000    0.0000
## TRAVTIME   0.1246    0.2945    0.6009    0.7501    0.5405    0.0009
## BLUEBOOK   0.5420    0.0001    0.0003    0.0007    0.0000
## TIF        0.5420    0.0147    0.0408    0.0006    0.9927
## OLDCLAIM   0.0001    0.0147    0.0000    0.0000    0.0787
## CLM_FREQ   0.0003    0.0408    0.0000    0.0000    0.2247
## MVR PTS   0.0007    0.0006    0.0000    0.4250
## CAR_AGE   0.0000    0.9927    0.0787    0.2247    0.4250

```

```
corrplot(cor(trainnum), method="square")
```



```
cor.test(trainnum$HOMEKIDS,trainnum$AGE,method="pearson")
```

```

##
## Pearson's product-moment correlation
##
## data: trainnum$HOMEKIDS and trainnum$AGE

```

```

## t = -48.338, df = 8159, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.4885252 -0.4547891
## sample estimates:
##      cor
## -0.4718298

train2<-train

```

3. BUILD MODELS

Using the training data set, build at least two different multiple linear regression models and three different binary logistic regression models, using different variables (or the same variables with different transformations). You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.

Discuss the coefficients in the models, do they make sense? For example, if a person has a lot of traffic tickets, you would reasonably expect that person to have more car crashes. If the coefficient is negative (suggesting that the person is a safer driver), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

Binary Logistic Regression Models

We first create a full model by including all the variables.

Coefficients (+ or -) of variables with significant p-values:

- KIDSDRV (+): When teenagers drive your car, the car is more likely to get into crashes
- INCOME (-): Rich people are less likely to get into crashes
- PARENT1/yes (+): Single parent is more likely to get into crashes
- HOME_VAL (-): Home owners tend to drive more responsibly
- MSTATUS/yes (-): Married people tend to drive more safely
- EDUCATION/bachelor, master, phd (-): More educated people tend to drive more safely
- JOB/blue collar, clerical (+): Blue collar and clerical workers are more likely to get into crashes
- JOB/manager (-): Manegements tend to drive more safely
- TRAVTIME (+): Long drives to work suggest greater risk
- CAR_USE/private (-): Private cars are being driving less than commericial cars, thus the probability of collision is lower
- BLUEBOOK (-): Unknown effect on probability of collision, but probably effect the payout if there is a crash
- TIF (-): People who have been customers for a long time are usually more safe

- CAR_TYPE/panel truck, pickup, sports car, suv, van (+): Sports car has the highest coefficient, more likely to get into a car crash
- CLM_FREQ (+): The more claims you filed in the past 5 years, the more you are likely to file in the future
- REVOKED/yes (+): If your license was revoked in the past 7 years, you probably are a more risky driver
- MVR_PTS (+): If you get lots of traffic tickets, you tend to get into more crashes
- URBANICITY/highly urban, urban (+): If you live in the city, you are more likely to get into a crash

```
#MODEL 1
logit <- glm(formula = TARGET_FLAG ~ . - TARGET_AMT, data=train, family = "binomial" (link="logit"))
summary(logit)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - TARGET_AMT, family = binomial(link = "logit"),
##      data = train)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -2.5262   -0.7180   -0.3983    0.6545    3.1455
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -7.942e-01  3.293e-01 -2.412 0.015880 *
## KIDSDRV                   6.821e-01  1.103e-01  6.185 6.21e-10 ***
## AGE                      4.736e-05  4.078e-03  0.012 0.990734
## HOMEKIDS                  1.513e-01  8.300e-02  1.823 0.068320 .
## YOJ                      -1.353e-02  8.578e-03 -1.577 0.114756
## INCOME                   -3.457e-06  1.076e-06 -3.212 0.001317 **
## PARENT1Yes                 3.295e-01  1.144e-01  2.881 0.003970 **
## HOME_VAL                  -1.323e-06  3.419e-07 -3.871 0.000109 ***
## MSTATUSz_No                 5.146e-01  8.493e-02  6.059 1.37e-09 ***
## SEXz_F                     -8.929e-02  1.120e-01 -0.797 0.425327
## EDUCATIONBachelors          -3.720e-01  1.154e-01 -3.223 0.001267 **
## EDUCATIONMasters             -2.803e-01  1.785e-01 -1.570 0.116405
## EDUCATIONPhD                 -1.496e-01  2.135e-01 -0.701 0.483401
## EDUCATIONz_High_School          2.111e-02  9.487e-02  0.222 0.823945
## JOBclerical                  3.986e-01  1.963e-01  2.030 0.042359 *
## JOBDoctor                   -4.227e-01  2.662e-01 -1.588 0.112286
## JOBHome_Maker                  2.049e-01  2.099e-01  0.976 0.328988
## JOBLawyer                    1.172e-01  1.693e-01  0.692 0.488652
## JOBManager                  -5.616e-01  1.712e-01 -3.280 0.001038 **
## JOBPProfessional               1.673e-01  1.782e-01  0.939 0.347724
## JOBStudent                   2.038e-01  2.140e-01  0.953 0.340799
## JOBz_Blue_Collar                 3.101e-01  1.853e-01  1.674 0.094190 .
## TRAVTIME                     1.483e-02  1.880e-03  7.890 3.02e-15 ***
## CAR_USEPrivate                -7.604e-01  9.172e-02 -8.291 < 2e-16 ***
## BLUEBOOK                     -2.079e-05  5.255e-06 -3.956 7.63e-05 ***
## TIF                          -3.257e-01  4.138e-02 -7.869 3.56e-15 ***
## CAR_TYPEPanel_Truck            5.701e-01  1.613e-01  3.533 0.000410 ***
```

```

## CAR_TYPEPickup      5.578e-01  1.007e-01  5.540 3.03e-08 ***
## CAR_TYPESports_Car 1.031e+00  1.298e-01  7.942 2.00e-15 ***
## CAR_TYPEVan         6.158e-01  1.264e-01  4.872 1.10e-06 ***
## CAR_TYPEz_SUV       7.787e-01  1.111e-01  7.007 2.43e-12 ***
## RED_CARyes        -5.766e-03  8.631e-02 -0.067 0.946741
## OLDCLAIM           6.763e-03  1.697e-02  0.398 0.690300
## CLM_FREQ            3.160e-01  1.277e-01  2.474 0.013363 *
## REVOKEDYes          7.242e-01  8.184e-02  8.850 < 2e-16 ***
## MVR PTS             2.808e-01  4.202e-02  6.682 2.35e-11 ***
## CAR AGE              -1.807e-03  7.530e-03 -0.240 0.810372
## URBANICITYz_Highly_Rural/ Rural -2.371e+00  1.130e-01 -20.989 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7308.4  on 8123  degrees of freedom
## AIC: 7384.4
##
## Number of Fisher Scoring iterations: 5

```

```
exp(logit$coefficients)
```

| | | |
|----|---------------------|------------------------|
| ## | (Intercept) | KIDSDRV |
| ## | 0.45194170 | 1.97796637 |
| ## | AGE | HOMEKIDS |
| ## | 1.00004736 | 1.16334008 |
| ## | YOJ | INCOME |
| ## | 0.98656280 | 0.99999654 |
| ## | PARENT1Yes | HOME_VAL |
| ## | 1.39023185 | 0.99999868 |
| ## | MSTATUSz_No | SEXz_F |
| ## | 1.67300058 | 0.91457630 |
| ## | EDUCATIONBachelors | EDUCATIONMasters |
| ## | 0.68933598 | 0.75559308 |
| ## | EDUCATIONPhD | EDUCATIONz_High_School |
| ## | 0.86104231 | 1.02133113 |
| ## | JOBClerical | JOBDoctor |
| ## | 1.48969494 | 0.65526320 |
| ## | JOBHome_Maker | JOBLawyer |
| ## | 1.22734502 | 1.12438495 |
| ## | JOBManager | JOBProfessional |
| ## | 0.57030294 | 1.18213072 |
| ## | JOBStudent | JOBz_Blue_Collar |
| ## | 1.22609195 | 1.36361825 |
| ## | TRAVTIME | CAR_USEPrivate |
| ## | 1.01494543 | 0.46747121 |
| ## | BLUEBOOK | TIF |
| ## | 0.99997921 | 0.72204512 |
| ## | CAR_TYPEPanel_Truck | CAR_TYPEPickup |
| ## | 1.76835639 | 1.74677458 |
| ## | CAR_TYPESports_Car | CAR_TYPEVan |
| ## | 2.80272210 | 1.85105526 |

```

##          CAR_TYPEz_SUV             RED_CARYes
##                2.17863803           0.99425108
##          OLDCLAIM                  CLM_FREQ
##                1.00678601           1.37166999
##          REVOKEDYes                 MVR PTS
##                2.06310705           1.32420339
##          CAR AGE URBANICITYz_Highly_Rural/
##                0.99819493           0.09334636

```

```

logitscalar <- mean(dlogis(predict(logit, type = "link")))
logitscalar * coef(logit)

```

```

##          (Intercept)            KIDSDRV
##                -1.158016e-01        9.945167e-02
##          AGE                   HOMEKIDS
##                6.904809e-06        2.206017e-02
##          YOJ                   INCOME
##                -1.972543e-03       -5.040064e-07
##          PARENT1Yes              HOME_VAL
##                4.803969e-02       -1.929523e-07
##          MSTATUSz_No             SEXz_F
##                7.503592e-02       -1.301990e-02
##          EDUCATIONBachelors     EDUCATIONMasters
##                -5.424472e-02      -4.086324e-02
##          EDUCATIONPhD            EDUCATIONz_High_School
##                -2.181469e-02       3.077558e-03
##          JOBClerical              JOBDocor
##                5.811519e-02       -6.163603e-02
##          JOBHome_Maker            JOBLawyer
##                2.986941e-02       1.709406e-02
##          JOBManager                JOBProfessional
##                -8.188439e-02       2.439650e-02
##          JOBStudent                JOBz_Blue_Collar
##                2.972047e-02       4.522137e-02
##          TRAVTIME                  CAR USEPrivate
##                2.163050e-03       -1.108755e-01
##          BLUEBOOK                  TIF
##                -3.030737e-06       -4.748519e-02
##          CAR_TYPEPanel_Truck     CAR_TYPEPickup
##                8.311836e-02       8.132789e-02
##          CAR_TYPESports_Car      CAR_TYPEVan
##                1.502692e-01       8.978260e-02
##          CAR_TYPEz_SUV             RED_CARYes
##                1.135413e-01       -8.406609e-04
##          OLDCLAIM                  CLM_FREQ
##                9.861172e-04       4.607979e-02
##          REVOKEDYes                 MVR PTS
##                1.055966e-01       4.094471e-02
##          CAR AGE URBANICITYz_Highly_Rural/
##                -2.634331e-04       -3.457765e-01

```

```

confint.default(logit)

```

```

##          2.5 %      97.5 %
## (Intercept) -1.439652e+00 -1.487526e-01
## KIDSDRV      4.659328e-01  8.982056e-01
## AGE         -7.944409e-03  8.039120e-03
## HOMEKIDS    -1.137668e-02  3.139672e-01
## YOJ          -3.034002e-02  3.283437e-03
## INCOME       -5.565742e-06 -1.347510e-06
## PARENT1Yes   1.052923e-01  5.536488e-01
## HOME_VAL     -1.993393e-06 -6.532553e-07
## MSTATUSz_No  3.481541e-01  6.810835e-01
## SEXz_F       -3.088261e-01  1.302374e-01
## EDUCATIONBachelors -5.982429e-01 -1.458101e-01
## EDUCATIONMasters -6.301049e-01  6.960026e-02
## EDUCATIONPhD   -5.680126e-01  2.687893e-01
## EDUCATIONz_High_School -1.648411e-01  2.070547e-01
## JOBClerical   1.374718e-02  7.833955e-01
## JOBDoctor    -9.444515e-01  9.901495e-02
## JOBHome_Maker -2.064601e-01  6.161667e-01
## JOBLawyer     -2.145962e-01  4.490685e-01
## JOBManager   -8.971700e-01 -2.260052e-01
## JOBProfessional -1.819185e-01  5.165555e-01
## JOBStudent   -2.155551e-01  6.232188e-01
## JOBz_Blue_Collar -5.304572e-02  6.733290e-01
## TRAVTIME      1.114967e-02  1.852002e-02
## CAR_USEPrivate -9.401775e-01 -5.806575e-01
## BLUEBOOK      -3.108426e-05 -1.048713e-05
## TIF           -4.067786e-01 -2.445566e-01
## CAR_TYPEPanel_Truck 2.538386e-01  8.862624e-01
## CAR_TYPEPickup  3.604241e-01  7.551178e-01
## CAR_TYPESports_Car 7.762439e-01  1.284938e+00
## CAR_TYPEVan    3.680626e-01  8.634492e-01
## CAR_TYPEz_SUV  5.608898e-01  9.965101e-01
## RED_CARyes    -1.749296e-01  1.633986e-01
## OLDCLAIM      -2.650452e-02  4.003069e-02
## CLM_FREQ       6.565860e-02  5.663993e-01
## REVOKEDYes    5.638194e-01  8.846068e-01
## MVR PTS       1.984459e-01  3.631762e-01
## CAR_AGE        -1.656445e-02  1.295105e-02
## URBANICITYz_Highly_Rural/ Rural -2.592883e+00 -2.149994e+00

predlogit <- predict(logit, type="response")
train2$pred1 <- predict(logit, type="response")
summary(predlogit)

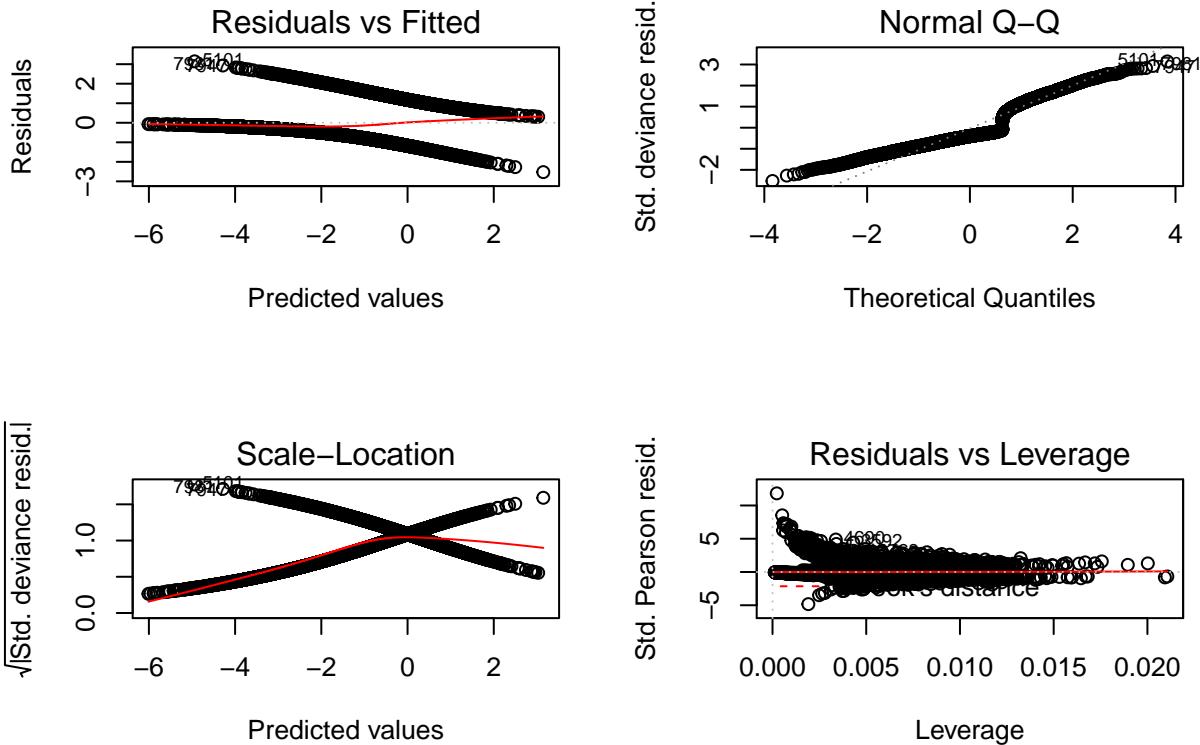
##      Min. 1st Qu. Median Mean 3rd Qu. Max.
## 0.002449 0.077438 0.201727 0.263816 0.403524 0.958860

table(true = train$TARGET_FLAG, pred = round(fitted(logit)))

##      pred
## true 0 1
## 0 5532 476
## 1 1251 902

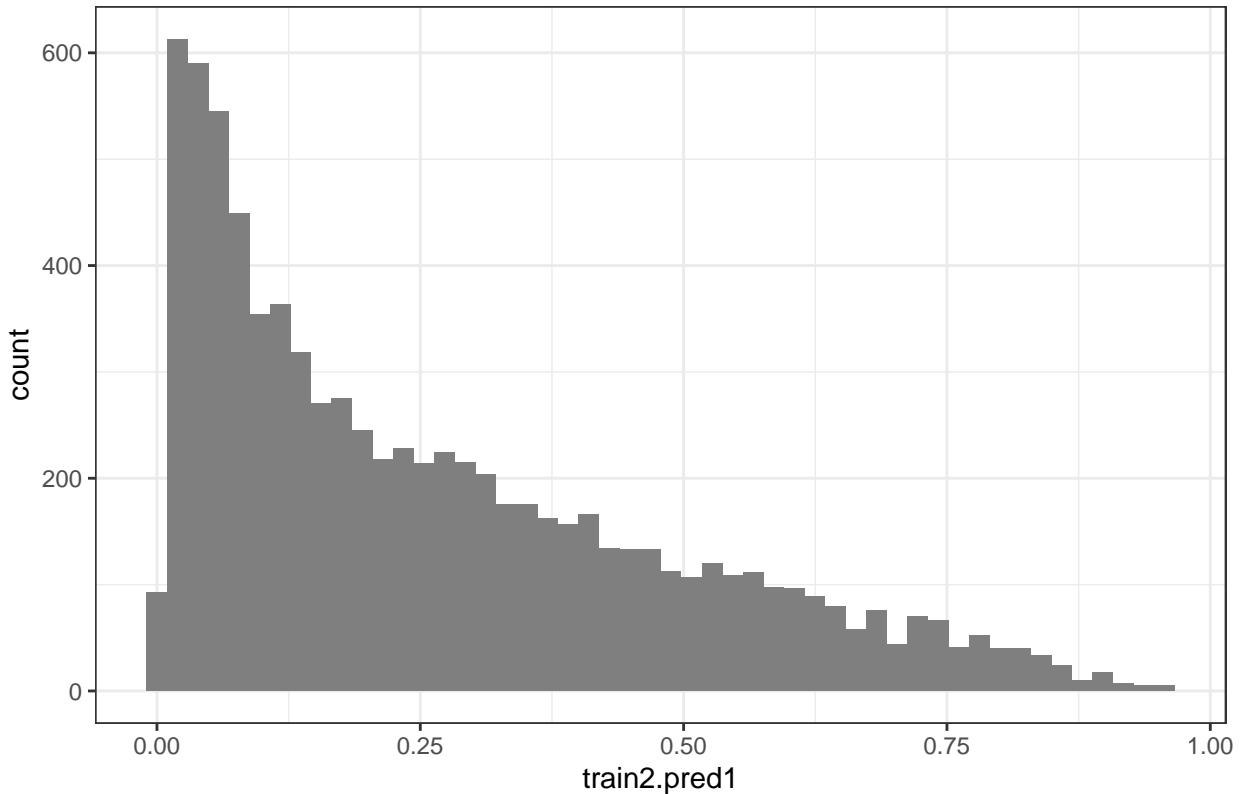
```

```
#plots for Model 1
par(mfrow=c(2,2))
plot(logit)
```



```
data.frame(train2$pred1) %>%
  ggplot(aes(x = train2.pred1)) +
  geom_histogram(bins = 50, fill = 'grey50') +
  labs(title = 'Histogram of Predictions') +
  theme_bw()
```

Histogram of Predictions



```
plot.roc(train$TARGET_FLAG, train2$pred1)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

#extract variables that are significant and rerun model
sigvars <- data.frame(summary(logit)$coef[summary(logit)$coef[,4] <= .05, 4])
sigvars <- add_rownames(sigvars, "vars")

## Warning: Deprecated, use tibble::rownames_to_column() instead.

colist<-dplyr::pull(sigvars, vars)
# colist<-colist[2:11]
colist<-c("KIDSDRV", "INCOME", "PARENT1", "HOME_VAL", "MSTATUS", "EDUCATION", "JOB", "TRAVTIME", "CAR_USE", "BL")
idx <- match(colist, names(train))
trainmod2 <- cbind(train[,idx], train2['TARGET_FLAG'])
#MODEL 2
logit2 <- glm(TARGET_FLAG ~ ., data=trainmod2, family = "binomial" (link="logit"))
summary(logit2)

## 
## Call:
## glm(formula = TARGET_FLAG ~ ., family = binomial(link = "logit"),
```

```

##      data = trainmod2)
##
## Deviance Residuals:
##      Min     1Q   Median     3Q    Max
## -2.5523 -0.7190 -0.3985  0.6497  3.1365
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -8.728e-01  2.620e-01 -3.332 0.000863 ***
## KIDSDRV                    7.664e-01  9.775e-02  7.841 4.48e-15 ***
## INCOME                     -3.552e-06  1.071e-06 -3.317 0.000910 ***
## PARENT1Yes                 4.476e-01  9.451e-02  4.736 2.18e-06 ***
## HOME_VAL                   -1.367e-06  3.407e-07 -4.012 6.03e-05 ***
## MSTATUSz_No                4.766e-01  7.969e-02  5.981 2.22e-09 ***
## EDUCATIONBachelors        -3.839e-01  1.086e-01 -3.534 0.000409 ***
## EDUCATIONMasters           -3.062e-01  1.612e-01 -1.899 0.057514 .
## EDUCATIONPhD               -1.761e-01  1.997e-01 -0.882 0.377940
## EDUCATIONz_High_School    1.682e-02  9.450e-02  0.178 0.858752
## JOBClerical                4.011e-01  1.962e-01  2.044 0.040930 *
## JOBDoctor                  -4.251e-01  2.658e-01 -1.599 0.109770
## JOBHome_Maker              2.561e-01  2.038e-01  1.257 0.208790
## JOBLawyer                  1.091e-01  1.690e-01  0.646 0.518557
## JOBManager                 -5.704e-01  1.711e-01 -3.335 0.000854 ***
## JOBProfessional            1.578e-01  1.781e-01  0.886 0.375433
## JOBStudent                 2.732e-01  2.104e-01  1.299 0.194092
## JOBz_Blue_Collar           3.064e-01  1.852e-01  1.654 0.098047 .
## TRAVTIME                   1.471e-02  1.877e-03  7.837 4.61e-15 ***
## CAR_USEPrivate              -7.623e-01  9.158e-02 -8.324 < 2e-16 ***
## BLUEBOOK                   -2.321e-05  4.715e-06 -4.922 8.56e-07 ***
## TIF                         -3.257e-01  4.135e-02 -7.875 3.41e-15 ***
## CAR_TYPEPanel_Truck        6.226e-01  1.505e-01  4.137 3.53e-05 ***
## CAR_TYPEPickup              5.528e-01  1.006e-01  5.497 3.86e-08 ***
## CAR_TYPESports_Car          9.746e-01  1.074e-01  9.077 < 2e-16 ***
## CAR_TYPEVan                 6.466e-01  1.220e-01  5.301 1.15e-07 ***
## CAR_TYPEz_SUV               7.218e-01  8.585e-02  8.407 < 2e-16 ***
## CLM_FREQ                    3.624e-01  5.464e-02  6.631 3.33e-11 ***
## REVOKEDYes                 7.349e-01  8.022e-02  9.161 < 2e-16 ***
## MVR PTS                     2.863e-01  4.138e-02  6.920 4.51e-12 ***
## URBANICITYz_Highly_Rural/ Rural -2.373e+00  1.129e-01 -21.024 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 9418.0 on 8160 degrees of freedom
## Residual deviance: 7314.8 on 8130 degrees of freedom
## AIC: 7376.8
##
## Number of Fisher Scoring iterations: 5

exp(logit2$coefficients)

```

| | | |
|----|-------------|------------|
| ## | (Intercept) | KIDSDRV |
| ## | 0.41776744 | 2.15209411 |

```

##           INCOME          PARENT1Yes
##      0.99999645      1.56455092
##           HOME_VAL        MSTATUSz_No
##      0.99999863      1.61063279
## EDUCATIONBachelors EDUCATIONMasters
##          0.68119924      0.73624403
## EDUCATIONPhD       EDUCATIONz_High_School
##          0.83855457      1.01695952
##      JOBClerical        JOBDoctor
##          1.49341342      0.65370747
##     JOBHome_Maker        JOBLawyer
##          1.29192357      1.11527301
##     JOBManager        JOBProfessional
##          0.56527588      1.17097642
##     JOBStudent        JOBz_Blue_Collar
##          1.31419704      1.35848308
##      TRAVTIME          CAR_USEPrivate
##          1.01482211      0.46658415
##      BLUEBOOK            TIF
##          0.99997679      0.72205023
## CAR_TYPEPanel_Truck   CAR_TYPEPickup
##          1.86373779      1.73806537
## CAR_TYPESports_Car    CAR_TYPEVan
##          2.65019303      1.90901065
## CAR_TYPEz_SUV         CLM_FREQ
##          2.05810528      1.43671851
##     REVOKEDYes          MVR PTS
##          2.08518827      1.33154581
## URBANICITYz_Highly_Rural/ Rural
##          0.09320986

```

```

logit2scalar <- mean(dlogis(predict(logit2, type = "link")))
logit2scalar * coef(logit2)

```

```

##           (Intercept)          KIDSDRIV
##      -1.274002e-01      1.118714e-01
##           INCOME          PARENT1Yes
##      -5.185070e-07      6.533249e-02
##           HOME_VAL        MSTATUSz_No
##      -1.994968e-07      6.956953e-02
## EDUCATIONBachelors EDUCATIONMasters
##          -5.603494e-02      -4.469269e-02
## EDUCATIONPhD       EDUCATIONz_High_School
##          -2.570038e-02      2.454692e-03
##      JOBClerical        JOBDoctor
##          5.854023e-02      -6.204783e-02
##     JOBHome_Maker        JOBLawyer
##          3.738562e-02      1.592436e-02
##     JOBManager        JOBProfessional
##          -8.326286e-02      2.303837e-02
##     JOBStudent        JOBz_Blue_Collar
##          3.988064e-02      4.471824e-02
##      TRAVTIME          CAR_USEPrivate
##          2.147590e-03      -1.112694e-01

```

```

##          BLUEBOOK           TIF
## -3.387609e-06 -4.753412e-02
## CAR_TYPEPanel_Truck CAR_TYPEPickup
## 9.087371e-02 8.068389e-02
## CAR_TYPESports_Car CAR_TYPEVan
## 1.422595e-01 9.437697e-02
## CAR_TYPEz_SUV CLM_FREQ
## 1.053534e-01 5.289110e-02
## REVOKEDYes MVR PTS
## 1.072616e-01 4.179488e-02
## URBANICITYz_Highly_Rural/ Rural
## -3.463539e-01

```

```

predlogit2 <- predict(logit2, type="response")
train2$pred2 <- predict(logit2, type="response")
summary(predlogit2)

```

```

##      Min. 1st Qu. Median Mean 3rd Qu. Max.
## 0.002282 0.077191 0.202256 0.263816 0.403691 0.961502

```

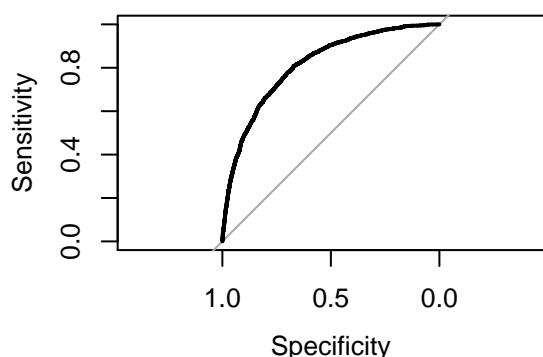
```
table(true = train$TARGET_FLAG, pred = round(fitted(logit2)))
```

```

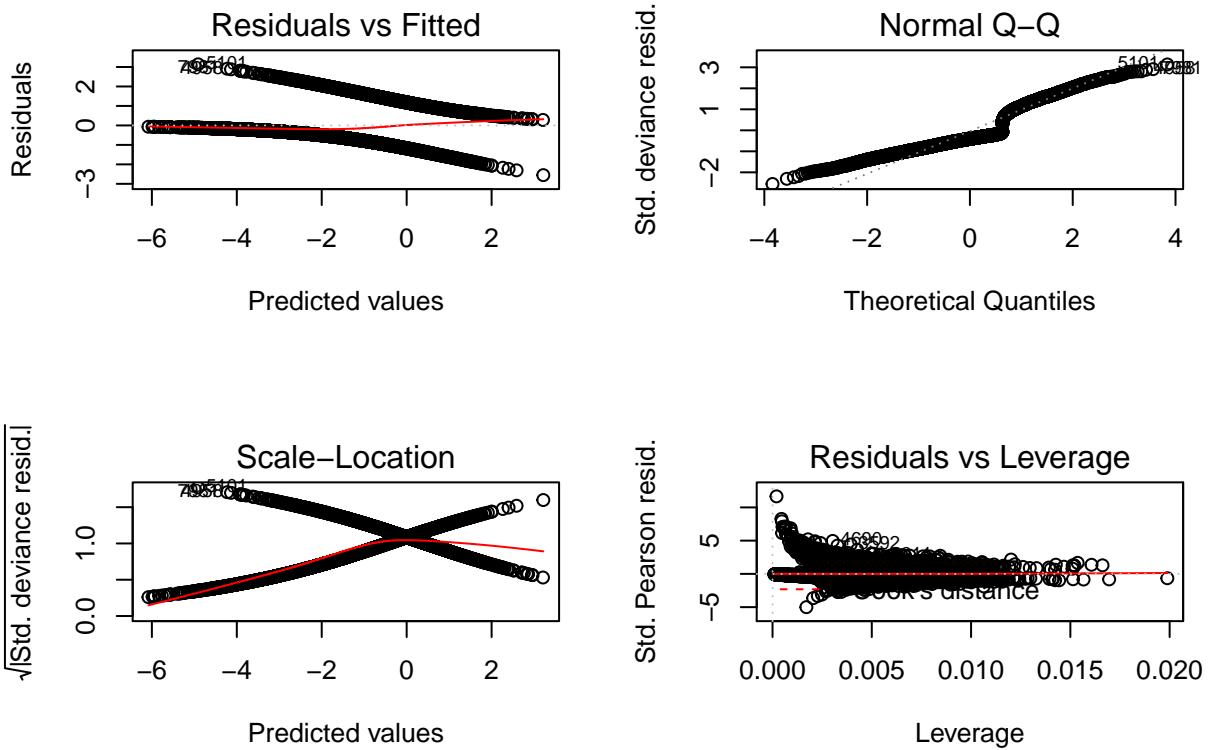
##      pred
## true 0 1
## 0 5541 467
## 1 1247 906

```

```
#plots for Model 2
par(mfrow=c(2,2))
```

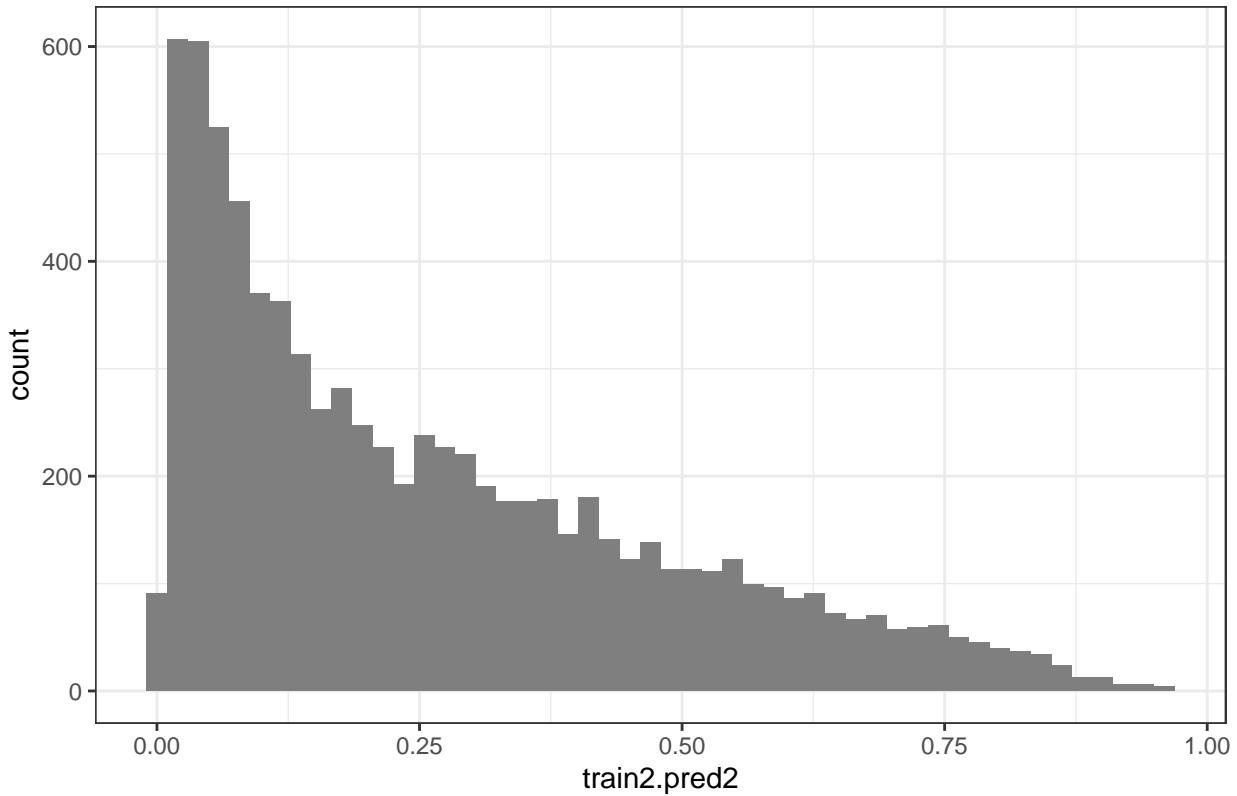


```
plot(logit2)
```



```
data.frame(train2$pred2) %>%
  ggplot(aes(x = train2.pred2)) +
  geom_histogram(bins = 50, fill = 'grey50') +
  labs(title = 'Histogram of Predictions') +
  theme_bw()
```

Histogram of Predictions



```
plot.roc(train$TARGET_FLAG, train2$pred2)
```

```
## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases
```

```
#MODEL 3  
#PC Model no racial bias  
logit3 <- glm(TARGET_FLAG ~ KIDSDRV + INCOME + HOME_VAL + TRAVTIME, data=train, family = "binomial" (1  
summary(logit3)
```

```
##  
## Call:  
## glm(formula = TARGET_FLAG ~ KIDSDRV + INCOME + HOME_VAL + TRAVTIME,  
##       family = binomial(link = "logit"), data = train)  
##  
## Deviance Residuals:  
##      Min        1Q    Median        3Q       Max  
## -1.5299  -0.8217  -0.6749   1.2315   2.8090  
##  
## Coefficients:  
##                               Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -6.876e-01  7.305e-02 -9.412 < 2e-16 ***  
## KIDSDRV     7.266e-01  8.115e-02  8.953 < 2e-16 ***  
## INCOME      -3.497e-06  6.826e-07 -5.123 3.01e-07 ***  
## HOME_VAL    -2.972e-06  2.499e-07 -11.895 < 2e-16 ***
```

```

## TRAVTIME      5.880e-03  1.598e-03   3.679  0.000234 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 9021.1  on 8156  degrees of freedom
## AIC: 9031.1
##
## Number of Fisher Scoring iterations: 4

```

```
exp(logit3$coefficients)
```

```

## (Intercept)    KIDSDRV      INCOME     HOME_VAL    TRAVTIME
## 0.5028055    2.0679778   0.9999965   0.9999970   1.0058969

```

```

predlogit3 <- predict(logit3, type="response")
train2$pred3 <- predict(logit3, type="response")
summary(predlogit3)

```

```

##      Min. 1st Qu. Median Mean 3rd Qu. Max.
## 0.01176 0.19679 0.25557 0.26382 0.32927 0.68970

```

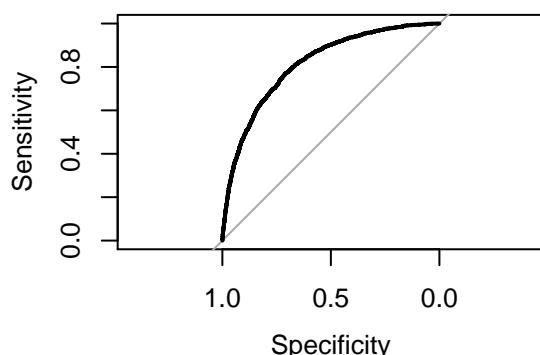
```
table(true = train$TARGET_FLAG, pred = round(fitted(logit3)))
```

```

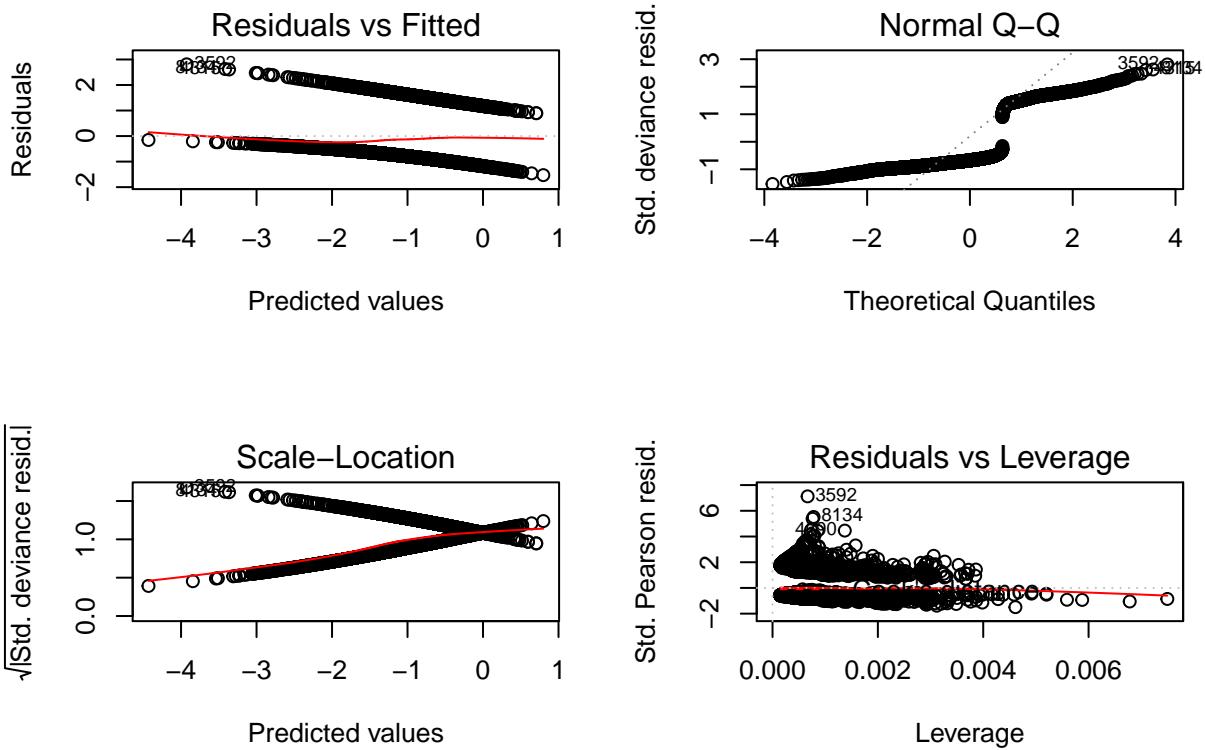
##      pred
## true 0 1
## 0 5937 71
## 1 2086 67

```

```
#plots for Model 3
par(mfrow=c(2,2))
```

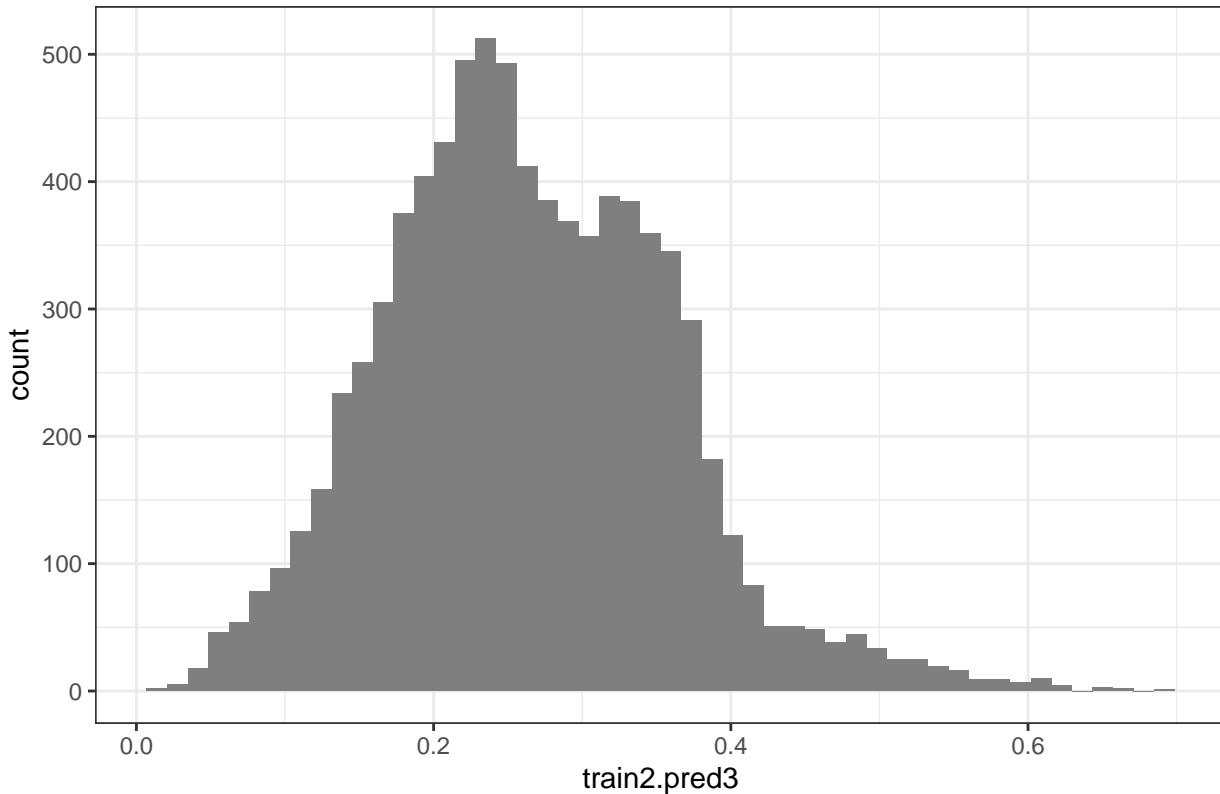


```
plot(logit3)
```



```
data.frame(train2$pred3) %>%
  ggplot(aes(x = train2.pred3)) +
  geom_histogram(bins = 50, fill = 'grey50') +
  labs(title = 'Histogram of Predictions') +
  theme_bw()
```

Histogram of Predictions



```
plot.roc(train$TARGET_FLAG, train2$pred3)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
logit3scalar <- mean(dlogis(predict(logit3, type = "link")))
logit3scalar * coef(logit3)
```

```
##      (Intercept)      KIDSDRV      INCOME      HOME_VAL      TRAVTIME
## -1.271908e-01  1.344090e-01 -6.468917e-07 -5.498016e-07  1.087668e-03
```

```
round(logitscalar * coef(logit), 2)
```

| | | |
|----|--------------------|------------------|
| ## | (Intercept) | KIDSDRV |
| ## | -0.12 | 0.10 |
| ## | AGE | HOMEKIDS |
| ## | 0.00 | 0.02 |
| ## | YOJ | INCOME |
| ## | 0.00 | 0.00 |
| ## | PARENT1Yes | HOME_VAL |
| ## | 0.05 | 0.00 |
| ## | MSTATUSz_No | SEXz_F |
| ## | 0.08 | -0.01 |
| ## | EDUCATIONBachelors | EDUCATIONMasters |

```

##          -0.05          -0.04
##      EDUCATIONPhD EDUCATIONz_High_School
##          -0.02          0.00
##      JOBClerical   JOBDoctor
##          0.06          -0.06
##      JOBHome_Maker  JOBLawyer
##          0.03          0.02
##      JOBManager    JOBProfessional
##          -0.08          0.02
##      JOBStudent    JOBz_Blue_Collar
##          0.03          0.05
##      TRAVTIME     CAR_USEPrivate
##          0.00          -0.11
##      BLUEBOOK      TIF
##          0.00          -0.05
##      CAR_TYPEPanel_Truck CAR_TYPEPickup
##          0.08          0.08
##      CAR_TYPESports_Car CAR_TYPEVan
##          0.15          0.09
##      CAR_TYPEz_SUV  RED_CARYes
##          0.11          0.00
##      OLDCLAIM       CLM_FREQ
##          0.00          0.05
##      REVOKEDYYes    MVR PTS
##          0.11          0.04
##      CAR AGE URBANICITYz_Highly_Rural/ Rural
##          0.00          -0.35

```

```
round(logit2scalar * coef(logit2), 2)
```

```

##          (Intercept) KIDSDRV
##          -0.13        0.11
##      INCOME        PARENT1Yes
##          0.00        0.07
##      HOME_VAL      MSTATUSz_No
##          0.00        0.07
##      EDUCATIONBachelor EDUCATIONMasters
##          -0.06        -0.04
##      EDUCATIONPhD EDUCATIONz_High_School
##          -0.03        0.00
##      JOBClerical   JOBDoctor
##          0.06        -0.06
##      JOBHome_Maker  JOBLawyer
##          0.04        0.02
##      JOBManager    JOBProfessional
##          -0.08        0.02
##      JOBStudent    JOBz_Blue_Collar
##          0.04        0.04
##      TRAVTIME     CAR_USEPrivate
##          0.00        -0.11
##      BLUEBOOK      TIF
##          0.00        -0.05
##      CAR_TYPEPanel_Truck CAR_TYPEPickup
##          0.09        0.08

```

```

##          CAR_TYPESports_Car           CAR_TYPEVan
##                      0.14                  0.09
##          CAR_TYPEz_SUV             CLM_FREQ
##                      0.11                  0.05
##          REVOKEDYes            MVR PTS
##                      0.11                  0.04
## URBANICITYz_Highly_Rural/ Rural
##                      -0.35

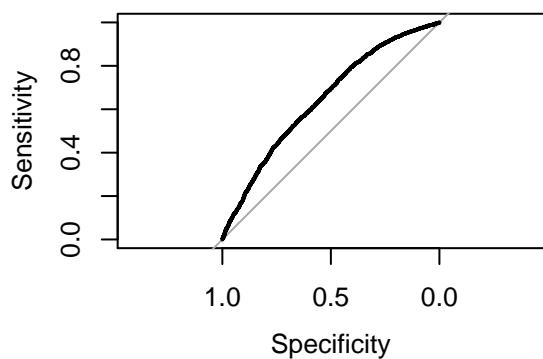
```

```
round(logit3scalar * coef(logit3),2)
```

```

## (Intercept)      KIDSDRV      INCOME     HOME_VAL     TRAVTIME
##       -0.13        0.13        0.00        0.00        0.00

```



Build Models GENERAL TARGET_AMT

```

#MODEL 1
model <- lm(TARGET_AMT ~ ., data=train)
summary(model)

##
## Call:
## lm(formula = TARGET_AMT ~ ., data = train)
##
## Residuals:
##      Min      1Q Median      3Q     Max
## -6234   -465    -58    243 101178
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.975e+02  5.010e+02 -1.193   0.2331
## TARGET_FLAG1 5.707e+03  1.134e+02 50.329  < 2e-16 ***
## KIDSDRV     -2.216e+01  1.781e+02 -0.124   0.9010
## AGE          6.145e+00  6.271e+00  0.980   0.3272
## HOMEKIDS    9.215e+01  1.256e+02  0.733   0.4633
## YOJ          7.685e+00  1.319e+01  0.583   0.5601
## INCOME      -2.258e-03  1.577e-03 -1.431   0.1524
## PARENT1Yes  1.209e+02  1.830e+02  0.661   0.5088
## HOME_VAL    3.864e-04  5.165e-04  0.748   0.4545

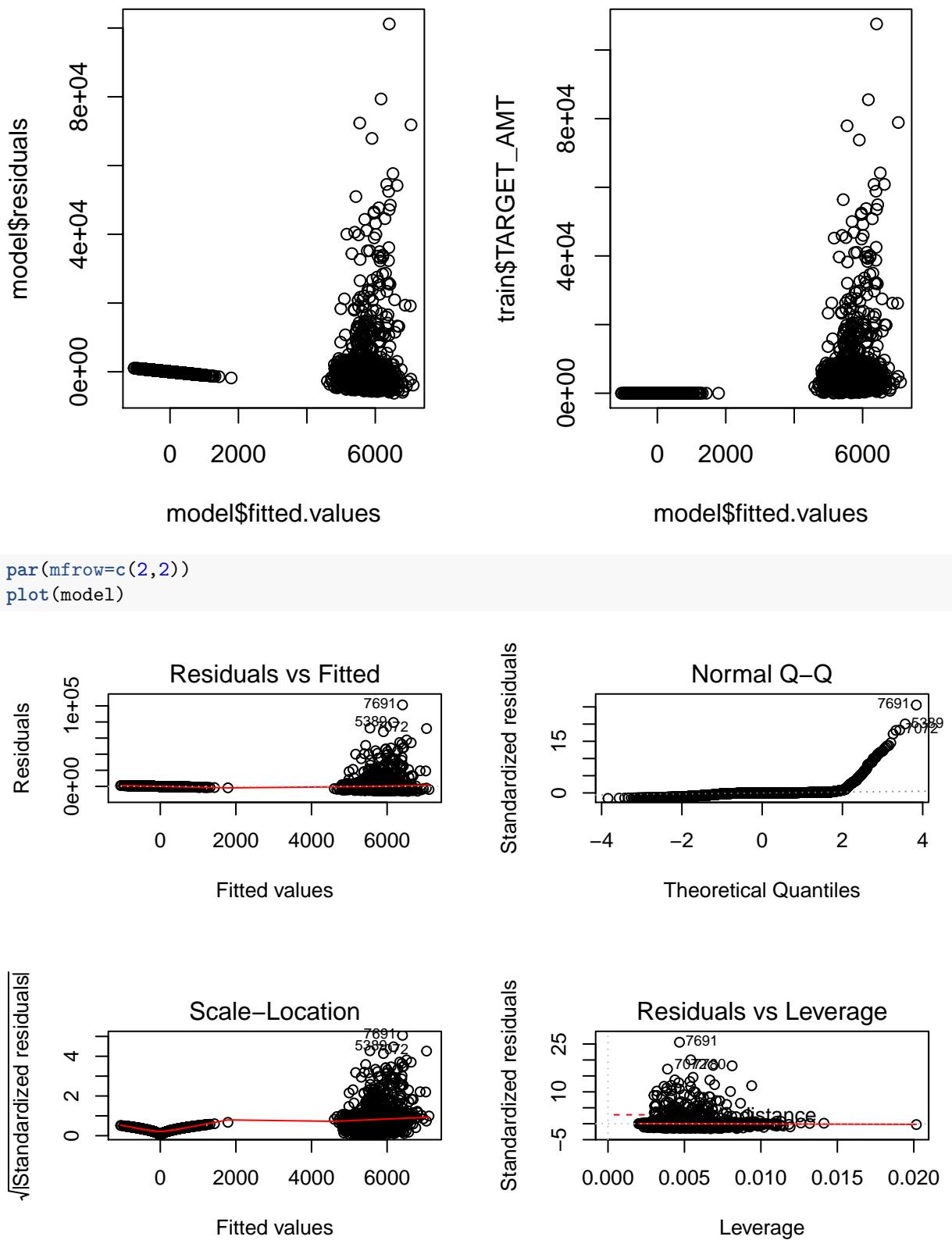
```

```

## MSTATUSz_No          1.770e+02  1.282e+02  1.381  0.1673
## SEXz_F              -2.896e+02  1.606e+02 -1.804  0.0713 .
## EDUCATIONBachelors  6.823e+01  1.790e+02  0.381  0.7031
## EDUCATIONMasters    2.235e+02  2.620e+02  0.853  0.3937
## EDUCATIONPhD         4.283e+02  3.110e+02  1.377  0.1685
## EDUCATIONz_High_School -1.243e+02  1.502e+02 -0.828  0.4077
## JOBClerical          -8.406e+00  2.984e+02 -0.028  0.9775
## JOBDoctor            -2.812e+02  3.571e+02 -0.788  0.4310
## JOBHome_Maker        -7.045e+01  3.185e+02 -0.221  0.8249
## JOBLawyer             7.660e+01  2.582e+02  0.297  0.7667
## JOBManager            -1.265e+02  2.521e+02 -0.502  0.6158
## JOBProfessional       1.733e+02  2.698e+02  0.642  0.5206
## JOBStudent            -1.306e+02  3.266e+02 -0.400  0.6892
## JOBz_Blue_Collar      5.187e+01  2.813e+02  0.184  0.8537
## TRAVTIME              5.682e-01  2.824e+00  0.201  0.8405
## CAR_USEPPrivate       -9.993e+01  1.443e+02 -0.693  0.4886
## BLUEBOOK              2.944e-02  7.536e-03  3.906  9.45e-05 ***
## TIF                   -1.653e+01  6.277e+01 -0.263  0.7922
## CAR_TYPEPanel_Truck   -5.880e+01  2.430e+02 -0.242  0.8088
## CAR_TYPEPickup        -3.318e+01  1.493e+02 -0.222  0.8241
## CAR_TYPESports_Car    2.098e+02  1.910e+02  1.099  0.2720
## CAR_TYPEVan            9.709e+01  1.865e+02  0.521  0.6026
## CAR_TYPEz_SUV          1.621e+02  1.571e+02  1.032  0.3021
## RED_CARyes             -2.696e+01  1.302e+02 -0.207  0.8360
## OLDCLAIM               4.079e+00  2.908e+01  0.140  0.8884
## CLM_FREQ                -8.551e+01  2.210e+02 -0.387  0.6989
## REVOKEDYes              -2.991e+02  1.385e+02 -2.160  0.0308 *
## MVR PTS                 1.396e+02  6.716e+01  2.079  0.0376 *
## CAR AGE                 -2.520e+01  1.118e+01 -2.254  0.0242 *
## URBANICITYz_Highly_Rural/ Rural  2.987e+01  1.272e+02  0.235  0.8143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3970 on 8122 degrees of freedom
## Multiple R-squared:  0.2912, Adjusted R-squared:  0.2879
## F-statistic:  87.8 on 38 and 8122 DF,  p-value: < 2.2e-16

par(mfrow=c(1,2))
plot(model$residuals ~ model$fitted.values)
plot(model$fitted.values,train$TARGET_AMT)

```



```
#extract variables that are significant and rerun model
sigvars <- data.frame(summary(model)$coef[summary(model)$coef[,4] <= .05, 4])
sigvars <- add_rownames(sigvars, "vars")
```

```

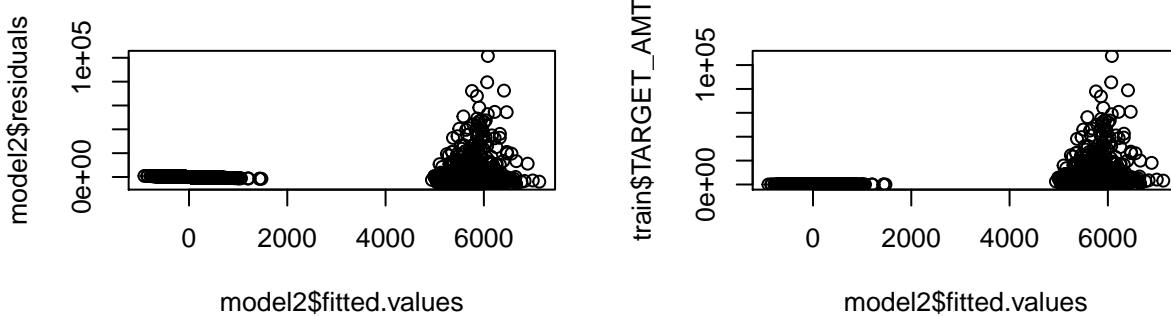
## Warning: Deprecated, use tibble::rownames_to_column() instead.

colist<-dplyr::pull(sigvars, vars)
colist<-c("TARGET_FLAG", "BLUEBOOK", "REVOKED", "MVR PTS", "CAR AGE")
idx <- match(colist, names(train))
trainmod2 <- cbind(train[,idx], train['TARGET_AMT'])
#MODEL 2
model2<-lm(TARGET_AMT ~ ., data=trainmod2)
summary(model2)

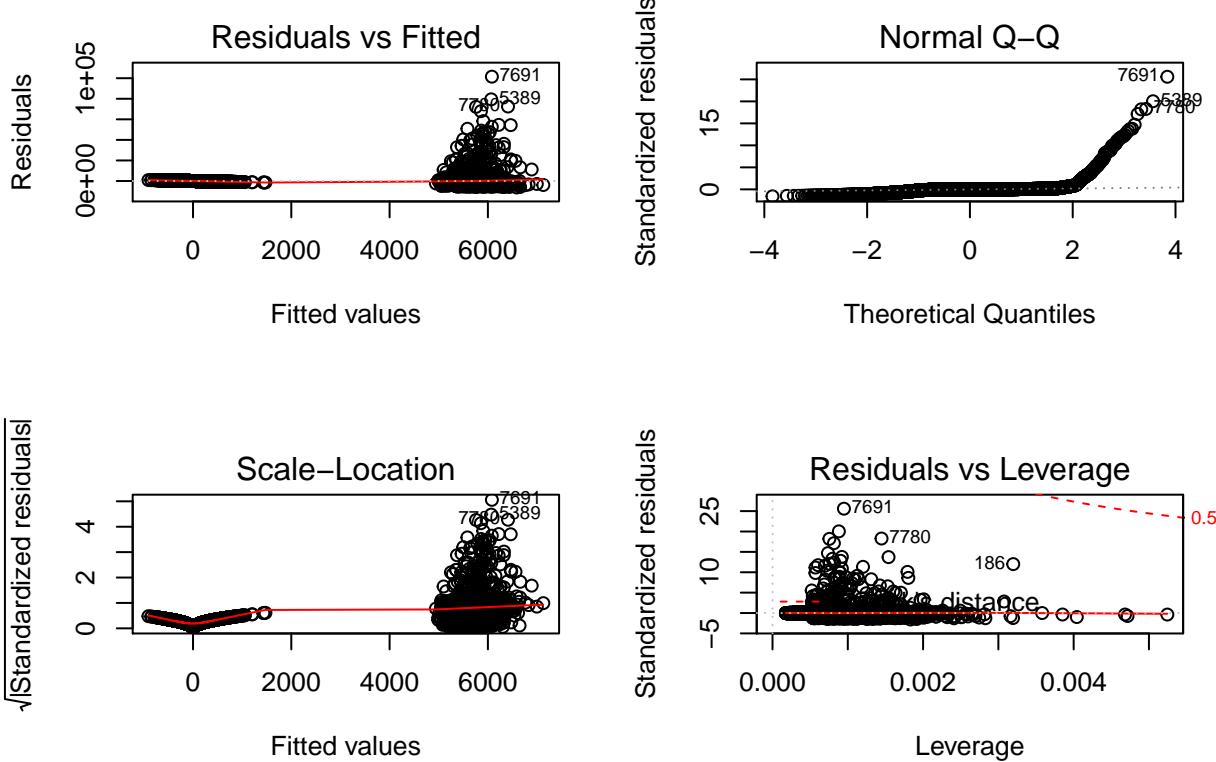
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = trainmod2)
##
## Residuals:
##     Min      1Q Median      3Q     Max
## -6269   -378    -34    192 101505
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.315e+02  1.206e+02 -3.579 0.000347 ***
## TARGET_FLAG1 5.735e+03  1.036e+02  55.334 < 2e-16 ***
## BLUEBOOK     3.010e-02  5.328e-03   5.649 1.67e-08 ***
## REVOKEDYes  -2.874e+02  1.356e+02 -2.120 0.034021 *
## MVR PTS      1.309e+02  6.101e+01   2.145 0.031986 *
## CAR AGE     -1.291e+01  8.122e+00  -1.590 0.111894
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3968 on 8155 degrees of freedom
## Multiple R-squared:  0.289, Adjusted R-squared:  0.2886
## F-statistic: 662.9 on 5 and 8155 DF, p-value: < 2.2e-16

par(mfrow=c(2,2))
plot(model2$residuals ~ model2$fitted.values)
plot(model2$fitted.values, train$TARGET_AMT)
par(mfrow=c(2,2))

```

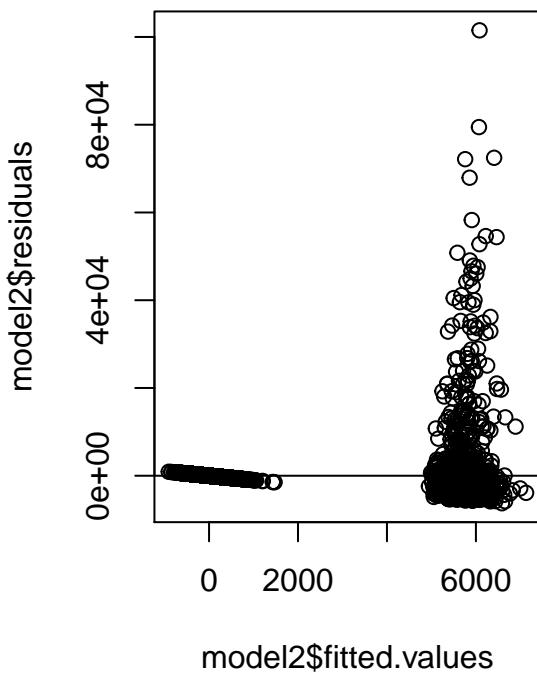


```
plot(model2)
```

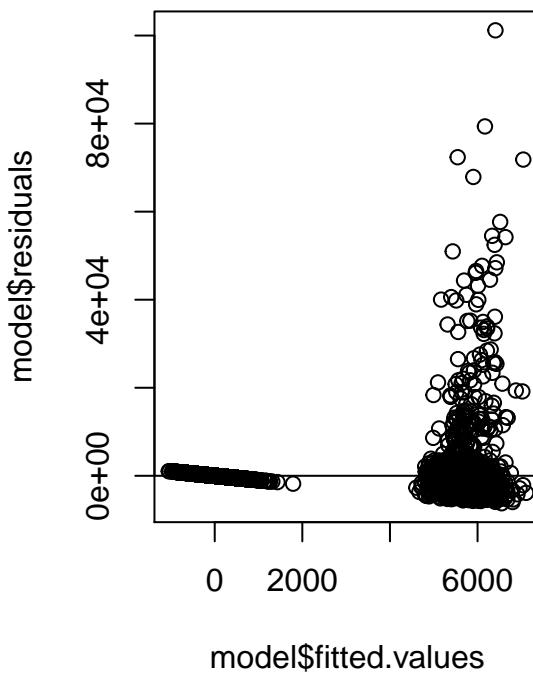


```
par(mfrow=c(1,2))
plot(model2$residuals ~ model2$fitted.values, main="New Reduced Var Model")
abline(h = 0)
plot(model$residuals ~ model$fitted.values, main="Original Model All Vars")
abline(h = 0)
```

New Reduced Var Model



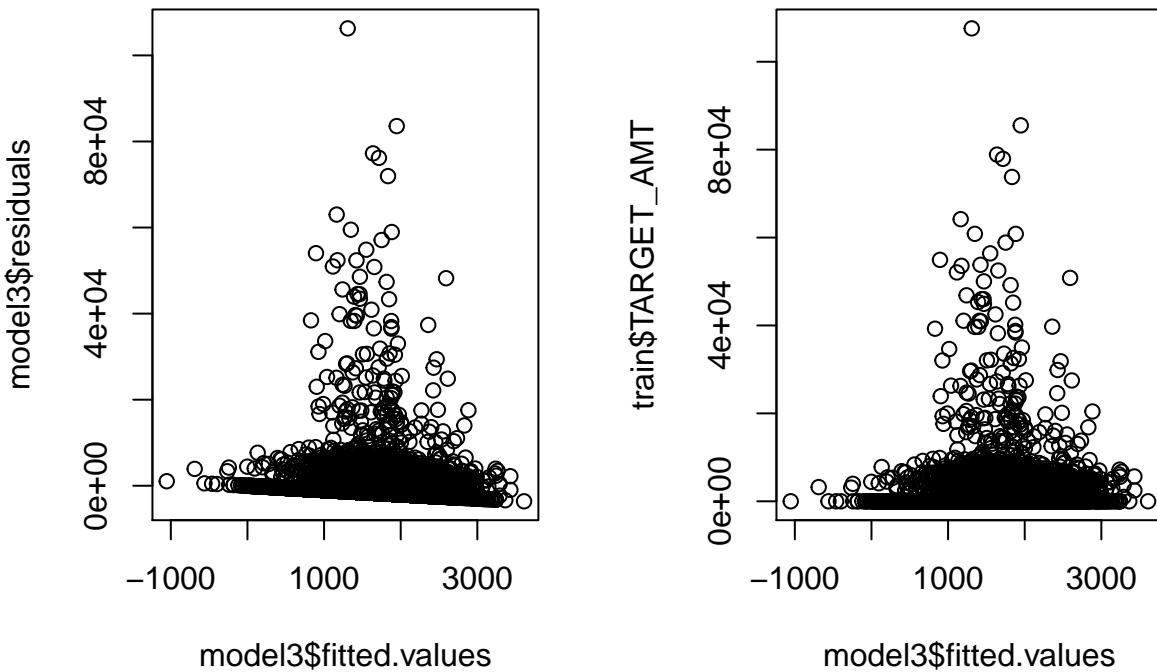
Original Model All Vars



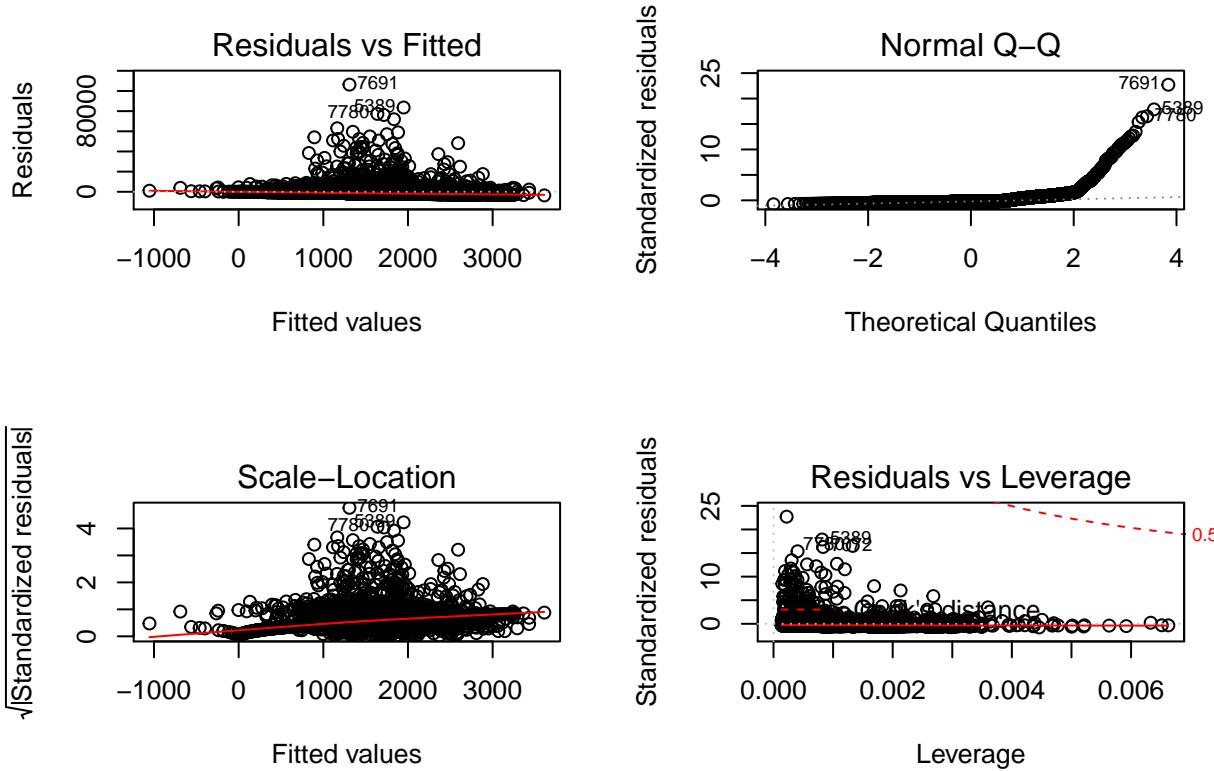
```
#MODEL 3
#remove variables with opposite coefficients
model3<-lm(TARGET_AMT ~ KIDSDRV + INCOME + HOME_VAL + TRAVTIME, data=train)
summary(model3)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ KIDSDRV + INCOME + HOME_VAL + TRAVTIME,
##      data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3610  -1652  -1239   -318 106277
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.680e+03  1.470e+02 11.426 < 2e-16 ***
## KIDSDRV     9.172e+02  1.789e+02  5.126 3.03e-07 ***
## INCOME      -1.242e-03  1.336e-03 -0.930  0.3522
## HOME_VAL    -2.809e-03  4.920e-04 -5.710 1.17e-08 ***
## TRAVTIME     7.234e+00  3.260e+00  2.219  0.0265 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4679 on 8156 degrees of freedom
## Multiple R-squared:  0.01096,    Adjusted R-squared:  0.01047
## F-statistic: 22.59 on 4 and 8156 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(1,2))
plot(model3$residuals ~ model3$fitted.values)
plot(model3$fitted.values,train$TARGET_AMT)
```



```
par(mfrow=c(2,2))
plot(model3)
```



4. SELECT MODELS

Decide on the criteria for selecting the best multiple linear regression model and the best binary logistic regression model. Will you select models with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your models.

For the multiple linear regression model, will you use a metric such as Adjusted R2, RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. 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 you use a metric such as log likelihood, AIC, ROC curve, etc.? 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.

```
test = read.csv("https://raw.githubusercontent.com/miachen410/DATA621/master/HW%234/insurance-evaluation.csv")
test2<- test
dim(test)

## [1] 2141    26

test$TARGET_AMT <- 0
test$TARGET_FLAG <- 0
test = as.tbl(test) %>%
  mutate_at(c("INCOME", "HOME_VAL", "BLUEBOOK", "OLDCLAIM"),
            currencyconv) %>%
  mutate_at(c("EDUCATION", "JOB", "CAR_TYPE", "URBANICITY"),
            underscore) %>%
  mutate_at(c("EDUCATION", "JOB", "CAR_TYPE", "URBANICITY"),
            as.factor) %>%
  mutate(TARGET_FLAG = as.factor(TARGET_FLAG))
# impute data for missing values
# use column mean for calculation
test$HOMEKIDS <- log(test$HOMEKIDS+1)
test$MVR_PTS <- log(test$MVR_PTS+1)
test$OLDCLAIM <- log(test$OLDCLAIM+1)
test$TIF <- log(test$TIF+1)
test$KIDSDRIV <- log(test$KIDSDRIV+1)
test$CLM_FREQ <- log(test$CLM_FREQ+1)
# use column mean for calculation
test$AGE[is.na(test$AGE)] <- mean(test$AGE, na.rm=TRUE)
test$YOJ[is.na(test$YOJ)] <- mean(test$YOJ, na.rm=TRUE)
test$HOME_VAL[is.na(test$HOME_VAL)] <- mean(test$HOME_VAL, na.rm=TRUE)
test$CAR_AGE[is.na(test$CAR_AGE)] <- mean(test$CAR_AGE, na.rm=TRUE)
test$INCOME[is.na(test$INCOME)] <- mean(test$INCOME, na.rm=TRUE)
#get complete cases
#remove rad per correlation in prior section
test <- test[, !(colnames(test) %in% c("INDEX"))]
TARGET_FLAG <- predict(logit, newdata = test, type="response")
y_pred_num <- ifelse(TARGET_FLAG > 0.5, 1, 0)
y_pred <- factor(y_pred_num, levels=c(0, 1))
summary(y_pred)

##      0      1
```

```

## 1776 365

rbind(round(summary(predlogit),4), round(summary(TARGET_FLAG),4)) %>% kable()



|        | Min.   | 1st Qu. | Median | Mean   | 3rd Qu. | Max. |
|--------|--------|---------|--------|--------|---------|------|
| 0.0024 | 0.0774 | 0.2017  | 0.2638 | 0.4035 | 0.9589  |      |
| 0.0031 | 0.0777 | 0.2183  | 0.2708 | 0.4102 | 0.9464  |      |



test$TARGET_FLAG <- as.factor(test$TARGET_FLAG)
test2 <- test[, !(colnames(test) %in% c("TARGET_FLAG"))]
TARGET_AMT<- predict(model, newdata = test, interval='confidence') #data from scaling originally to get
summary(TARGET_AMT)

##      fit                  lwr                  upr
## Min. :-1206.170  Min. :-1870.4  Min. :-542.0
## 1st Qu.: -255.615 1st Qu.: -782.6  1st Qu.: 256.4
## Median : -22.708  Median : -538.1  Median : 478.1
## Mean   : -8.173   Mean   : -540.5  Mean   : 524.1
## 3rd Qu.: 223.762  3rd Qu.: -303.8  3rd Qu.: 774.3
## Max.   : 1251.287  Max.   : 521.4   Max.   :1998.7

summary(model)

## 
## Call:
## lm(formula = TARGET_AMT ~ ., data = train)
## 
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -6234   -465   -58    243 101178 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -5.975e+02 5.010e+02 -1.193 0.2331    
## TARGET_FLAG1 5.707e+03 1.134e+02 50.329 < 2e-16 ***
## KIDSDRV     -2.216e+01 1.781e+02 -0.124 0.9010    
## AGE          6.145e+00 6.271e+00 0.980 0.3272    
## HOMEKIDS    9.215e+01 1.256e+02 0.733 0.4633    
## YOJ          7.685e+00 1.319e+01 0.583 0.5601    
## INCOME       -2.258e-03 1.577e-03 -1.431 0.1524    
## PARENT1Yes   1.209e+02 1.830e+02 0.661 0.5088    
## HOME_VAL     3.864e-04 5.165e-04 0.748 0.4545    
## MSTATUSz_No  1.770e+02 1.282e+02 1.381 0.1673    
## SEXz_F       -2.896e+02 1.606e+02 -1.804 0.0713 .  
## EDUCATIONBachelors 6.823e+01 1.790e+02 0.381 0.7031    
## EDUCATIONMasters 2.235e+02 2.620e+02 0.853 0.3937    
## EDUCATIONPhD   4.283e+02 3.110e+02 1.377 0.1685    
## EDUCATIONz_High_School -1.243e+02 1.502e+02 -0.828 0.4077    
## JOBClerical   -8.406e+00 2.984e+02 -0.028 0.9775    
## JOBDoctor     -2.812e+02 3.571e+02 -0.788 0.4310    
## JOBHome_Maker -7.045e+01 3.185e+02 -0.221 0.8249    
## JOBLawyer     7.660e+01 2.582e+02 0.297 0.7667

```

```

## JOBManager          -1.265e+02  2.521e+02  -0.502   0.6158
## JOBProfessional    1.733e+02  2.698e+02   0.642   0.5206
## JOBStudent          -1.306e+02  3.266e+02  -0.400   0.6892
## JOBz_Blue_Collar    5.187e+01  2.813e+02   0.184   0.8537
## TRAVTIME            5.682e-01  2.824e+00   0.201   0.8405
## CAR_USEPrivate      -9.993e+01  1.443e+02  -0.693   0.4886
## BLUEBOOK            2.944e-02  7.536e-03   3.906  9.45e-05 ***
## TIF                 -1.653e+01  6.277e+01  -0.263   0.7922
## CAR_TYPEPanel_Truck -5.880e+01  2.430e+02  -0.242   0.8088
## CAR_TYPEPickup      -3.318e+01  1.493e+02  -0.222   0.8241
## CAR_TYPESports_Car  2.098e+02  1.910e+02   1.099   0.2720
## CAR_TYPEVan          9.709e+01  1.865e+02   0.521   0.6026
## CAR_TYPEz_SUV        1.621e+02  1.571e+02   1.032   0.3021
## RED_CARyes          -2.696e+01  1.302e+02  -0.207   0.8360
## OLDCLAIM             4.079e+00  2.908e+01   0.140   0.8884
## CLM_FREQ              -8.551e+01  2.210e+02  -0.387   0.6989
## REVOKEDYes           -2.991e+02  1.385e+02  -2.160   0.0308 *
## MVR PTS              1.396e+02  6.716e+01   2.079   0.0376 *
## CAR AGE              -2.520e+01  1.118e+01  -2.254   0.0242 *
## URBANICITYz_Highly_Rural/ Rural  2.987e+01  1.272e+02   0.235   0.8143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3970 on 8122 degrees of freedom
## Multiple R-squared:  0.2912, Adjusted R-squared:  0.2879
## F-statistic:  87.8 on 38 and 8122 DF,  p-value: < 2.2e-16

```