hw4

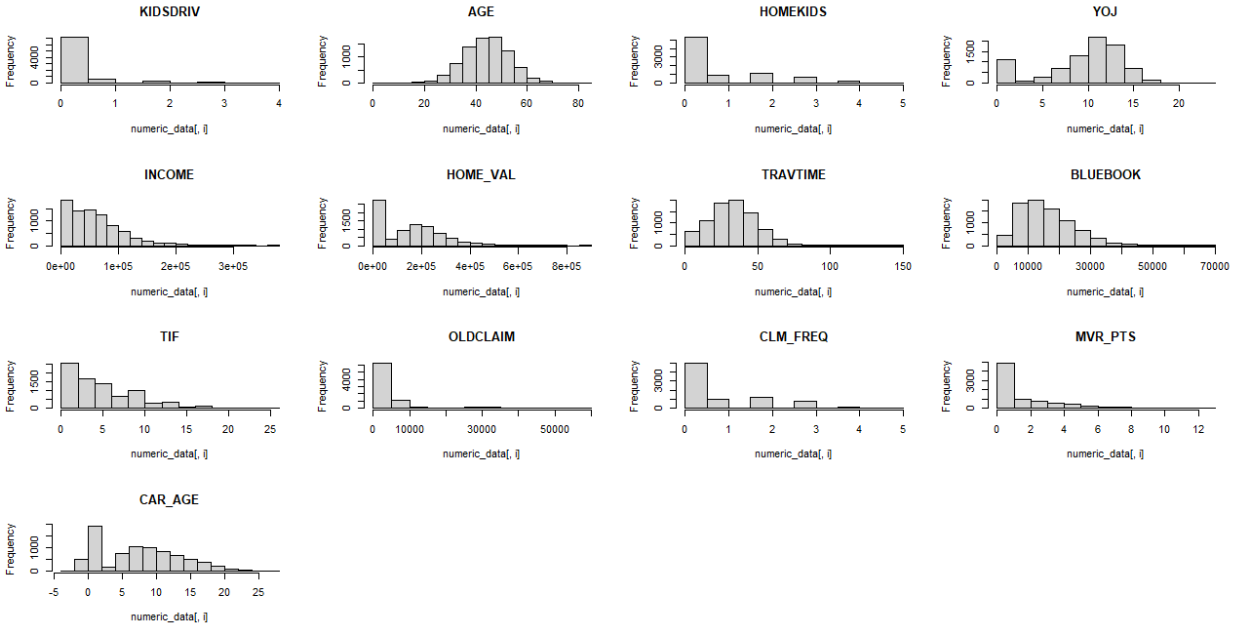
David Moste, Sadia Perveen, Vanita Thompson

5/2/2021

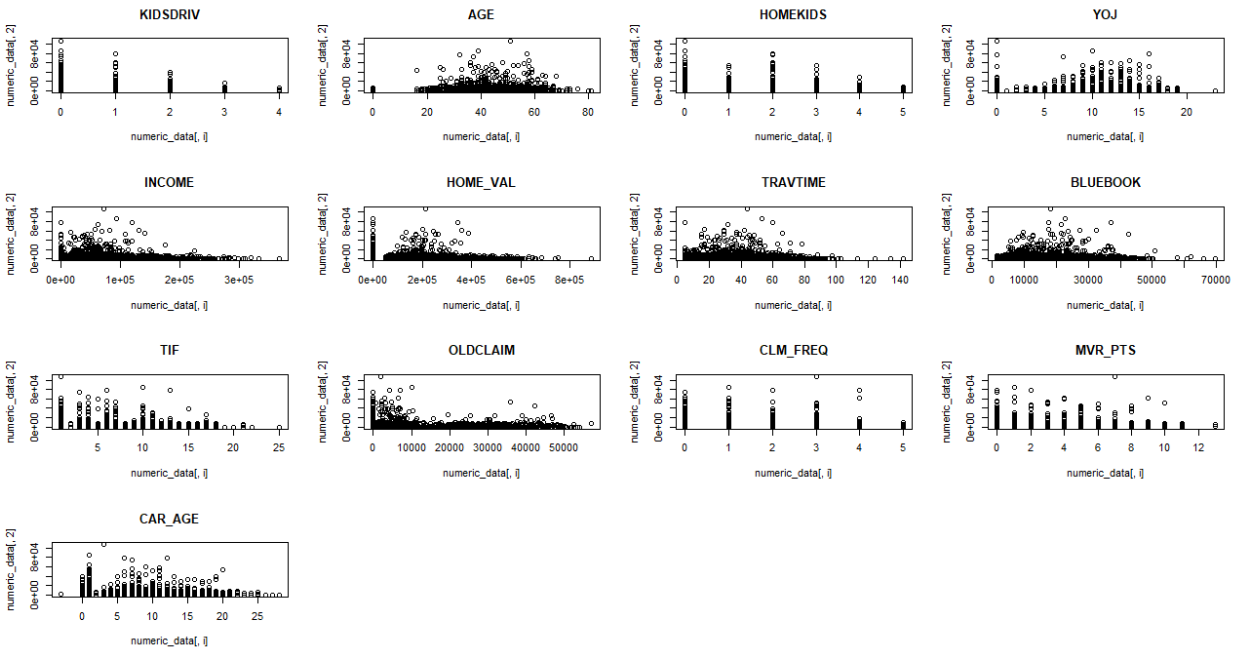
Data Exploration

After grabbing the data, we first checked out a summary of the data to see the predictor variables provided along with their summary statistics. This also allowed us to check if there was any missing data. We also noticed at this point that much of the data either needed to be converted to factors or had extraneous symbols such as a dollar sign that needed to be removed.

We then created two plots: a histogram, and a scatter plot.



The purpose of the histogram was to get a sense of the normality of each variable. Upon looking at the histogram, it was easy to see that a few of our variables were skewed and would need to be transformed.



The purpose of this scatter plot was to get a sense of any relationship between each variable and the target.

Finally, we created a correlation plot as a final check on the relationship between our variables.

Data Preparation

To start data preparation, we used the MICE package to impute any missing data. We then performed a few transformations. In order to correct some skewing, we performed a log transformation on a few predictor variables.

Build Models

The first log model we built was simply every variable in the data. The AIC for this model was 7352.6.

The second log model we built was done through backward elimination. The AIC for this model was 8568.2

The final log model we built was based on hand-picked variables. The AIC for this model was 9013.4

The first linear model was simply every variable in the data. The r-squared value was 0.29.

The second linear model was based on hand-picked variables. The r-sqaured value was 0.028.

Select Models

Based on the AIC and r-squared values, we chose to pursue both first models. When we ran our model on the training data, we recorded the following values:

Metric	Value
Accuracy	0.7919
Classification Error	0.2081
Precision	0.8177
Sensitivity	0.9233

Metric	Value
Specificity	0.4255
F1	0.8673
AUC	0.9797

”	0	1
0	5547	1237
1	461	916

When looking at the resduals for our linear model, the distribution is not normal and the qqplot shows a clear lack of normality. Additionally, theresiduals are not uniform in their variability. This indicates that this is ultimately a very poor model.

Appendix

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## Warning: package 'forcats' was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.0.4
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.0.4

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.0.3

## corrplot 0.84 loaded
```

```
library(mice)
```

```
## Warning: package 'mice' was built under R version 4.0.3

##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
# Read in data and view summary statistics
data <- read.csv("https://raw.githubusercontent.com/dmoste/DATA621/master/hw4/insurance_training_data.cs
head(data)
```

```
##   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ   INCOME PARENT1
## 1     1           0          0        0  60        0  11  $67,349      No
## 2     2           0          0        0  43        0  11  $91,449      No
## 3     4           0          0        0  35        1  10  $16,039      No
## 4     5           0          0        0  51        0  14               No
## 5     6           0          0        0  50        0  NA $114,986      No
## 6     7           1       2946        0  34        1  12 $125,301     Yes
##   HOME_VAL MSTATUS SEX     EDUCATION          JOB TRAVTIME    CAR_USE BLUEBOOK
## 1       $0    z_No   M           PhD  Professional       14    Private  $14,230
## 2 $257,252    z_No   M z_High School z_Blue Collar       22 Commercial  $14,940
## 3 $124,191     Yes z_F z_High School      Clerical        5    Private   $4,010
## 4 $306,251     Yes   M  <High School z_Blue Collar       32    Private  $15,440
## 5 $243,925     Yes z_F           PhD        Doctor       36    Private  $18,000
## 6       $0    z_No z_F     Bachelors z_Blue Collar       46 Commercial  $17,430
##   TIF   CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1  11    Minivan     yes   $4,461        2      No       3      18
## 2   1    Minivan     yes       $0        0      No       0       1
## 3   4      z_SUV      no  $38,690        2      No       3      10
## 4   7    Minivan     yes       $0        0      No       0       6
## 5   1      z_SUV      no  $19,217        2     Yes       3      17
## 6   1 Sports Car      no       $0        0      No       0       7
##            URBANICITY
## 1 Highly Urban/ Urban
## 2 Highly Urban/ Urban
## 3 Highly Urban/ Urban
## 4 Highly Urban/ Urban
## 5 Highly Urban/ Urban
## 6 Highly Urban/ Urban
```

```
summary(data)
```

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

```
##
##   EDUCATION             JOB               TRAVTIME        CAR_USE
##  Length:8161        Length:8161        Min.   :  5.00   Length:8161
##  Class :character   Class :character   1st Qu.: 22.00   Class :character
##  Mode  :character   Mode  :character   Median : 33.00   Mode  :character
##                                        Mean   : 33.49
##                                        3rd Qu.: 44.00
##                                        Max.   :142.00
##
##    BLUEBOOK              TIF            CAR_TYPE           RED_CAR
##  Length:8161        Min.   : 1.000   Length:8161        Length:8161
##  Class :character   1st Qu.: 1.000   Class :character   Class :character
##  Mode  :character   Median : 4.000   Mode  :character   Mode  :character
##                     Mean   : 5.351
##                     3rd Qu.: 7.000
##                     Max.   :25.000
##
##    OLDCLAIM            CLM_FREQ        REVOKED             MVR_PTS
##  Length:8161        Min.   :0.0000   Length:8161        Min.   : 0.000
##  Class :character   1st Qu.:0.0000   Class :character   1st Qu.: 0.000
##  Mode  :character   Median :0.0000   Mode  :character   Median : 1.000
##                     Mean   :0.7986                      Mean   : 1.696
##                     3rd Qu.:2.0000                      3rd Qu.: 3.000
##                     Max.   :5.0000                      Max.   :13.000
##
##     CAR_AGE        URBANICITY
##  Min.   :-3.000   Length:8161
##  1st Qu.: 1.000   Class :character
##  Median : 8.000   Mode  :character
##  Mean   : 8.328
##  3rd Qu.:12.000
##  Max.   :28.000
##  NA's   :510
```

```
length(data$TARGET_FLAG[data$TARGET_FLAG == 0])/length(data$TARGET_FLAG)
```

```
## [1] 0.7361843
```

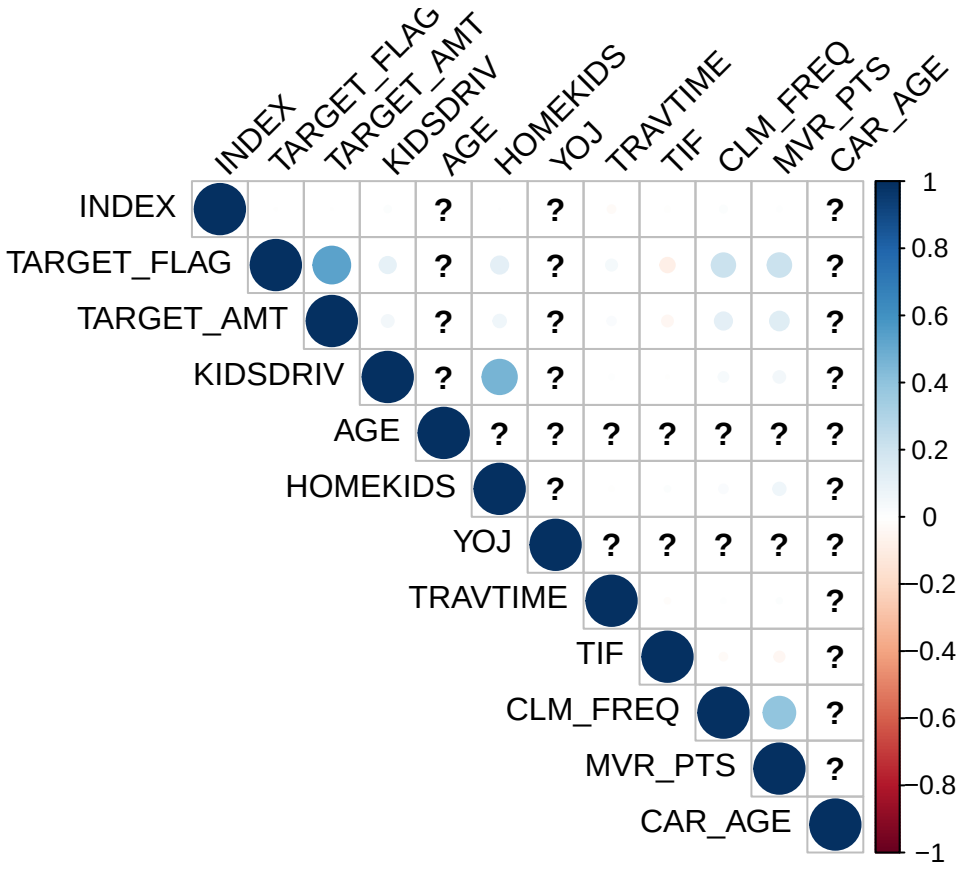
```
my_transform <- function(data){
  data <- data[-c(1)]
  data$TARGET_FLAG <- as.factor(data$TARGET_FLAG)
  data[data == ""] <- NA

  data$TARGET_AMT <- as.numeric(data$TARGET_AMT)
  data$INCOME <- as.numeric(str_remove_all(data$INCOME, "[\\$,]"))
  data$HOME_VAL <- as.numeric(str_remove_all(data$HOME_VAL, "[\\$,]"))
  data$BLUEBOOK <- as.numeric(str_remove_all(data$BLUEBOOK, "[\\$,]"))
  data$OLDCLAIM <- as.numeric(str_remove_all(data$OLDCLAIM, "[\\$,]"))
  data$CAR_AGE <- abs(data$CAR_AGE)

  data <- data %>%
    mutate_if(is.character, as.factor)
```

```r
  data$INCOME <- log(data$INCOME + 1)
  data$HOME_VAL <- log(data$HOME_VAL + 1)
  data$BLUEBOOK <- log(data$BLUEBOOK + 1)
  data$TIF <- log(data$TIF + 1)
  data$MVR_PTS <- log(data$MVR_PTS + 1)
  data$CAR_AGE <- log(data$CAR_AGE + 1)

  return(data)
}

t_data <- my_transform(data)

temp <- mice(t_data,
             m = 5,
             maxit = 10,
             method = "pmm",
             seed = 1234)
```

```
##
##  iter imp variable
##   1   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   1   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   1   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   1   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   1   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   2   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   2   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   2   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   2   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   2   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   3   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   3   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   3   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   3   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   3   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   4   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   4   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   4   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   4   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   4   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   5   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   5   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   5   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   5   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   5   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   6   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   6   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   6   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   6   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   6   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   7   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   7   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   7   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
```

```
##   7   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   7   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   8   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   8   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   8   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   8   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   8   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   9   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   9   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   9   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   9   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   9   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##  10   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##  10   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##  10   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##  10   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##  10   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
```

```r
t_data <- complete(temp,1)

numeric_data <- data %>%
  select_if(is.numeric)

corrplot(cor(numeric_data), method = "circle",
         type = "upper",
         tl.col = "black",
         tl.srt = 45)
```




```
accident_prob_1 <- glm(TARGET_FLAG ~ . - TARGET_AMT,
                       data = t_data, family = binomial)
summary(accident_prob_1)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - TARGET_AMT, family = binomial,
##     data = t_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6461  -0.7080  -0.4015   0.6204   3.1748
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   3.057e+00  5.984e-01   5.109 3.24e-07 ***
## KIDSDRIV                      3.929e-01  6.146e-02   6.393 1.63e-10 ***
## AGE                          -4.295e-03  4.072e-03  -1.055 0.291575
## HOMEKIDS                      1.799e-02  3.777e-02   0.476 0.633919
## YOJ                           2.475e-02  1.164e-02   2.126 0.033530 *
## INCOME                       -1.044e-01  1.900e-02  -5.496 3.89e-08 ***
## PARENT1Yes                    3.753e-01  1.097e-01   3.422 0.000622 ***
## HOME_VAL                     -3.245e-02  6.994e-03  -4.639 3.49e-06 ***
## MSTATUSz_No                   4.578e-01  8.849e-02   5.173 2.30e-07 ***
## SEXz_F                       -1.204e-01  1.083e-01  -1.112 0.266141
```

```
## EDUCATIONBachelors              -4.142e-01  1.140e-01  -3.632 0.000282 ***
## EDUCATIONMasters                -4.478e-01  1.468e-01  -3.051 0.002277 **
## EDUCATIONPhD                    -4.773e-01  1.782e-01  -2.679 0.007391 **
## EDUCATIONz_High School           1.799e-02  9.503e-02   0.189 0.849856
## JOBDoctor                       -8.177e-01  2.613e-01  -3.129 0.001752 **
## JOBHome Maker                   -4.200e-01  1.547e-01  -2.716 0.006617 **
## JOBLawyer                       -2.509e-01  1.564e-01  -1.604 0.108697
## JOBManager                      -9.788e-01  1.281e-01  -7.640 2.18e-14 ***
## JOBProfessional                 -2.940e-01  1.145e-01  -2.568 0.010234 *
## JOBStudent                      -5.855e-01  1.431e-01  -4.093 4.26e-05 ***
## JOBz_Blue Collar                -1.517e-01  1.010e-01  -1.503 0.132910
## TRAVTIME                         1.485e-02  1.888e-03   7.864 3.71e-15 ***
## CAR_USEPrivate                  -7.665e-01  9.102e-02  -8.421  < 2e-16 ***
## BLUEBOOK                        -2.941e-01  5.899e-02  -4.985 6.19e-07 ***
## TIF                             -3.229e-01  4.146e-02  -7.789 6.78e-15 ***
## CAR_TYPEPanel Truck              4.433e-01  1.487e-01   2.980 0.002882 **
## CAR_TYPEPickup                   5.645e-01  1.001e-01   5.637 1.73e-08 ***
## CAR_TYPESports Car               1.038e+00  1.281e-01   8.103 5.35e-16 ***
## CAR_TYPEVan                      6.046e-01  1.245e-01   4.857 1.19e-06 ***
## CAR_TYPEz_SUV                    8.083e-01  1.078e-01   7.499 6.45e-14 ***
## RED_CARyes                      -2.250e-02  8.638e-02  -0.260 0.794489
## OLDCLAIM                        -1.369e-05  3.922e-06  -3.491 0.000481 ***
## CLM_FREQ                         2.069e-01  2.845e-02   7.272 3.53e-13 ***
## REVOKEDYes                       8.877e-01  9.149e-02   9.703  < 2e-16 ***
## MVR_PTS                          3.111e-01  4.123e-02   7.545 4.54e-14 ***
## CAR_AGE                         -3.132e-02  4.666e-02  -0.671 0.502046
## URBANICITYz_Highly Rural/ Rural -2.437e+00  1.139e-01 -21.394  < 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: 7278.6  on 8124  degrees of freedom
## AIC: 7352.6
##
## Number of Fisher Scoring iterations: 5
```

```
accident_prob_2 <- glm(TARGET_FLAG ~ AGE + INCOME + HOME_VAL + SEX +
                         EDUCATION + TRAVTIME + CAR_TYPE + OLDCLAIM +
                         CLM_FREQ,
                       data = t_data, family = binomial)
summary(accident_prob_2)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + INCOME + HOME_VAL + SEX + EDUCATION +
##     TRAVTIME + CAR_TYPE + OLDCLAIM + CLM_FREQ, family = binomial,
##     data = t_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7865  -0.7836  -0.5782   0.9737   2.4247
##
```

```
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -5.669e-01  1.832e-01  -3.095 0.001968 **
## AGE                    -1.665e-02  3.181e-03  -5.233 1.66e-07 ***
## INCOME                 -2.725e-02  8.525e-03  -3.196 0.001391 **
## HOME_VAL               -4.899e-02  4.701e-03 -10.421  < 2e-16 ***
## SEXz_F                 -2.137e-01  8.269e-02  -2.585 0.009752 **
## EDUCATIONBachelors     -2.948e-01  8.476e-02  -3.478 0.000505 ***
## EDUCATIONMasters       -4.084e-01  9.498e-02  -4.299 1.71e-05 ***
## EDUCATIONPhD           -6.126e-01  1.257e-01  -4.875 1.09e-06 ***
## EDUCATIONz_High School  1.540e-01  7.981e-02   1.930 0.053628 .
## TRAVTIME                6.796e-03  1.664e-03   4.084 4.43e-05 ***
## CAR_TYPEPanel Truck     6.771e-01  1.133e-01   5.975 2.30e-09 ***
## CAR_TYPEPickup          8.065e-01  8.606e-02   9.371  < 2e-16 ***
## CAR_TYPESports Car      9.723e-01  1.115e-01   8.719  < 2e-16 ***
## CAR_TYPEVan             6.718e-01  1.072e-01   6.264 3.75e-10 ***
## CAR_TYPEz_SUV           8.165e-01  9.502e-02   8.593  < 2e-16 ***
## OLDCLAIM                9.417e-06  3.104e-06   3.034 0.002414 **
## CLM_FREQ                3.436e-01  2.427e-02  14.159  < 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: 8534.2  on 8144  degrees of freedom
## AIC: 8568.2
##
## Number of Fisher Scoring iterations: 4
```

```
accident_prob_3 <- glm(TARGET_FLAG ~ AGE + MVR_PTS + TRAVTIME,
                       data = t_data, family = binomial)
summary(accident_prob_3)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + MVR_PTS + TRAVTIME, family = binomial,
##     data = t_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4690  -0.8045  -0.6457   1.1948   2.1239
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.637740   0.147996  -4.309 1.64e-05 ***
## AGE         -0.024771   0.002994  -8.273  < 2e-16 ***
## MVR_PTS      0.596786   0.034539  17.279  < 2e-16 ***
## TRAVTIME     0.007114   0.001609   4.422 9.78e-06 ***
## ---
## 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: 9005.4  on 8157  degrees of freedom
## AIC: 9013.4
##
## Number of Fisher Scoring iterations: 4
```

```
accident_prob <- predict(accident_prob_1, t_data, type = "response")
accident_prob_pred <- as.factor(ifelse(accident_prob > 0.5, 1, 0))
accident_prob_data <- cbind(t_data, accident_prob, accident_prob_pred)

caret::confusionMatrix(data = accident_prob_data$accident_prob_pred,
                       reference = accident_prob_data$TARGET_FLAG,
                       mode = 'everything')
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 5547 1237
##          1  461  916
##
##                Accuracy : 0.7919
##                  95% CI : (0.783, 0.8007)
##     No Information Rate : 0.7362
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3943
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9233
##             Specificity : 0.4255
##          Pos Pred Value : 0.8177
##          Neg Pred Value : 0.6652
##               Precision : 0.8177
##                  Recall : 0.9233
##                      F1 : 0.8673
##              Prevalence : 0.7362
##          Detection Rate : 0.6797
##    Detection Prevalence : 0.8313
##       Balanced Accuracy : 0.6744
##
##        'Positive' Class : 0
##
```

```
############
cost_mod_1 <- lm(TARGET_AMT ~ .,
                     data = t_data)
summary(cost_mod_1)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = t_data)
```

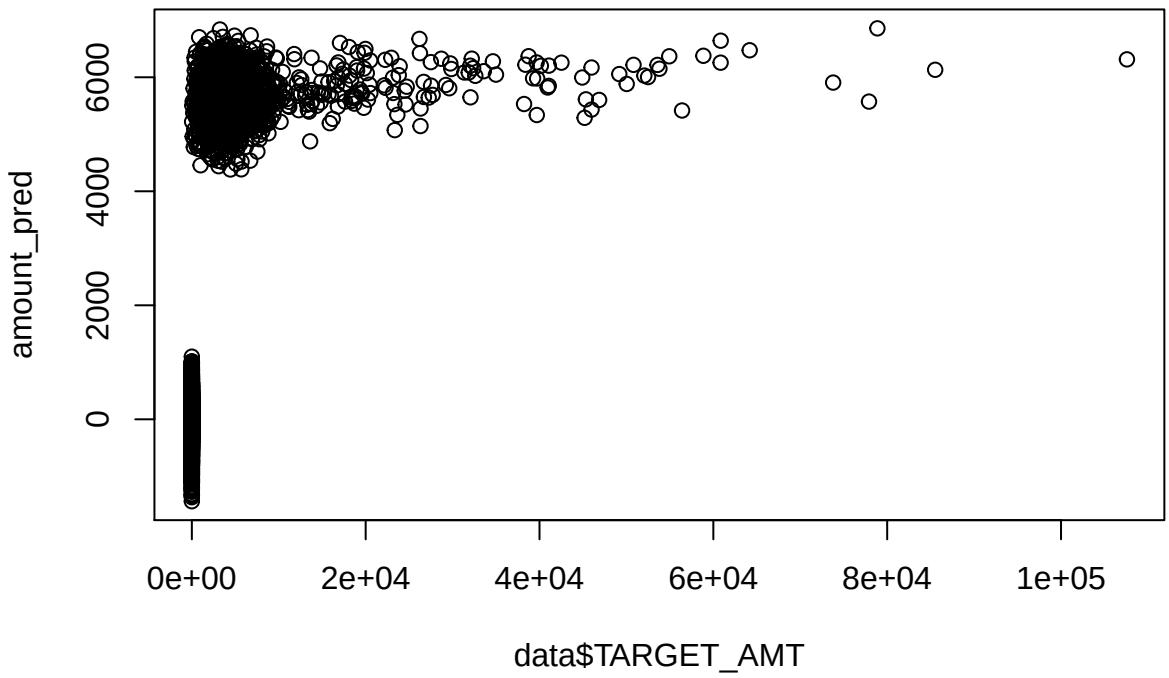
```
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6028   -442    -78    217 101272
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -4.072e+03  9.211e+02  -4.421 9.96e-06 ***
## TARGET_FLAG1                    5.736e+03  1.137e+02  50.472  < 2e-16 ***
## KIDSDRIV                       -4.242e+01  9.919e+01  -0.428   0.6689
## AGE                             6.830e+00  6.218e+00   1.098   0.2721
## HOMEKIDS                        4.841e+01  5.795e+01   0.835   0.4035
## YOJ                            -2.434e+00  1.728e+01  -0.141   0.8880
## INCOME                          9.162e+00  2.880e+01   0.318   0.7504
## PARENT1Yes                      1.554e+02  1.763e+02   0.881   0.3781
## HOME_VAL                        1.113e+01  1.110e+01   1.002   0.3163
## MSTATUSz_No                     1.785e+02  1.338e+02   1.334   0.1823
## SEXz_F                         -2.705e+02  1.564e+02  -1.729   0.0839 .
## EDUCATIONBachelors             -2.468e+01  1.767e+02  -0.140   0.8890
## EDUCATIONMasters                1.201e+01  2.254e+02   0.053   0.9575
## EDUCATIONPhD                    1.834e+02  2.681e+02   0.684   0.4939
## EDUCATIONz_High School         -1.557e+02  1.497e+02  -1.040   0.2983
## JOBDoctor                      -3.473e+02  3.490e+02  -0.995   0.3197
## JOBHome Maker                  -5.186e+01  2.311e+02  -0.224   0.8224
## JOBLawyer                       1.477e+01  2.337e+02   0.063   0.9496
## JOBManager                     -1.435e+02  1.881e+02  -0.763   0.4456
## JOBProfessional                 6.235e+01  1.758e+02   0.355   0.7229
## JOBStudent                     -4.750e+01  2.246e+02  -0.211   0.8325
## JOBz_Blue Collar                5.619e+00  1.599e+02   0.035   0.9720
## TRAVTIME                        4.781e-01  2.823e+00   0.169   0.8655
## CAR_USEPrivate                 -8.430e+01  1.425e+02  -0.591   0.5542
## BLUEBOOK                        3.991e+02  8.982e+01   4.443 9.00e-06 ***
## TIF                            -1.627e+01  6.273e+01  -0.259   0.7954
## CAR_TYPEPanel Truck             9.712e+00  2.263e+02   0.043   0.9658
## CAR_TYPEPickup                 -3.300e+01  1.484e+02  -0.222   0.8240
## CAR_TYPESports Car              2.334e+02  1.891e+02   1.234   0.2172
## CAR_TYPEVan                     7.788e+01  1.843e+02   0.423   0.6726
## CAR_TYPEz_SUV                   1.487e+02  1.525e+02   0.975   0.3294
## RED_CARyes                     -2.177e+01  1.302e+02  -0.167   0.8672
## OLDCLAIM                        3.131e-03  6.498e-03   0.482   0.6299
## CLM_FREQ                       -3.642e+01  4.800e+01  -0.759   0.4481
## REVOKEDYes                     -3.220e+02  1.525e+02  -2.111   0.0348 *
## MVR_PTS                         1.403e+02  6.570e+01   2.135   0.0328 *
## CAR_AGE                        -8.323e+01  7.119e+01  -1.169   0.2424
## URBANICITYz_Highly Rural/ Rural  5.415e+01  1.258e+02   0.430   0.6669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3969 on 8123 degrees of freedom
## Multiple R-squared:  0.2912, Adjusted R-squared:  0.288
## F-statistic:  90.2 on 37 and 8123 DF,  p-value: < 2.2e-16
```

```
cost_mod_2 <- lm(TARGET_AMT ~ AGE + INCOME + HOME_VAL +
                      SEX + EDUCATION + TRAVTIME + CAR_TYPE +
```

```
                      OLDCLAIM + CLM_FREQ,
                    data = t_data)
summary(cost_mod_2)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ AGE + INCOME + HOME_VAL + SEX + EDUCATION +
##     TRAVTIME + CAR_TYPE + OLDCLAIM + CLM_FREQ, data = t_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4164  -1633  -1005    -90 105092
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.760e+03  3.660e+02   4.809 1.54e-06 ***
## AGE                     -1.215e+01  6.223e+00  -1.953 0.050879 .
## INCOME                  -6.241e+00  1.802e+01  -0.346 0.729178
## HOME_VAL                -5.223e+01  9.540e+00  -5.475 4.50e-08 ***
## SEXz_F                  -2.157e+02  1.480e+02  -1.458 0.144913
## EDUCATIONBachelors      -3.262e+02  1.693e+02  -1.927 0.054001 .
## EDUCATIONMasters        -4.499e+02  1.835e+02  -2.451 0.014247 *
## EDUCATIONPhD            -5.536e+02  2.277e+02  -2.432 0.015054 *
## EDUCATIONz_High School   3.961e+01  1.653e+02   0.240 0.810690
## TRAVTIME                 7.044e+00  3.241e+00   2.173 0.029774 *
## CAR_TYPEPanel Truck      1.012e+03  2.114e+02   4.788 1.71e-06 ***
## CAR_TYPEPickup           6.694e+02  1.607e+02   4.165 3.14e-05 ***
## CAR_TYPESports Car       8.528e+02  2.084e+02   4.092 4.31e-05 ***
## CAR_TYPEVan              8.894e+02  2.002e+02   4.443 8.98e-06 ***
## CAR_TYPEz_SUV            6.294e+02  1.681e+02   3.745 0.000182 ***
## OLDCLAIM                 8.302e-03  6.750e-03   1.230 0.218765
## CLM_FREQ                 3.914e+02  5.132e+01   7.626 2.69e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4642 on 8144 degrees of freedom
## Multiple R-squared:  0.02799,    Adjusted R-squared:  0.02608
## F-statistic: 14.66 on 16 and 8144 DF,  p-value: < 2.2e-16
```

```
amount_pred <- predict(cost_mod_1,t_data)
plot(data$TARGET_AMT, amount_pred)
```

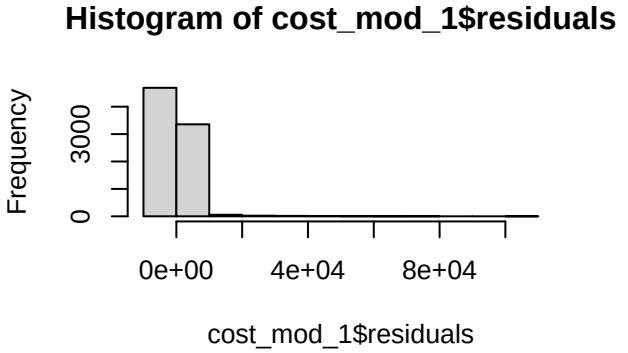


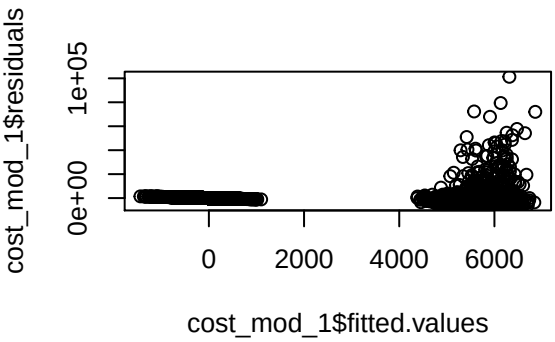

```
par(mfrow=c(2,2))
hist(cost_mod_1$residuals)
plot(cost_mod_1$residuals ~ cost_mod_1$fitted.values)
qqnorm(cost_mod_1$residuals)
qqline(cost_mod_1$residuals)

####
eval <- read.csv("https://raw.githubusercontent.com/dmoste/DATA621/master/hw4/insurance-evaluation-data

t_eval <- my_transform(eval)

temp <- mice(t_eval,
             m = 5,
             maxit = 10,
             method = "pmm",
             seed = 1234)
```

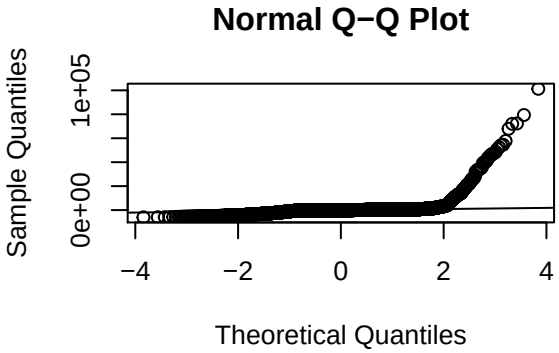
```
##
##  iter imp variable
##   1   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   1   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   1   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   1   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   1   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   2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

```
##    2   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    2   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    2   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    3   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    3   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    3   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    3   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    3   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    4   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    4   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    4   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    4   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    4   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    5   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    5   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    5   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    5   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    5   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    6   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    6   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    6   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    6   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    6   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    7   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    7   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    7   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    7   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    7   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    8   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    8   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    8   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    8   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    8   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    9   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    9   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    9   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    9   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##    9   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   10   1  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   10   2  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   10   3  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   10   4  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE
##   10   5  AGE  YOJ  INCOME  HOME_VAL  JOB  CAR_AGE

## Warning: Number of logged events: 2
```

```r
t_eval <- complete(temp,1)

eval_prob_pred <- predict(accident_prob_1, t_eval, type = "response")
eval_prob_pred <- ifelse(eval_prob_pred > 0.5, 1, 0)
eval_amount_pred <- predict(cost_mod_1,t_eval)
```


Histogram of cost_mod_1$residuals
Frequency
0
3000
0e+00
4e+04
8e+04
cost_mod_1$residuals


cost_mod_1$residuals
0e+00
1e+05
0
2000
4000
6000
cost_mod_1$fitted.values


Normal Q−Q Plot
Sample Quantiles
0e+00
1e+05
−4
−2
0
2
4
Theoretical Quantiles