ProjectFall23_group10

Udayveer Singh Andotra

2023-12-09

Libraries

```
# Libraries used
library(ggplot2)
library(tidyr)
library(dplyr)
library(olsrr)
library(StepReg)
library(epiDisplay)
library(caret)
library(glmnet)
library(prettyR)
library(car)
library(nortest)
```

Data cleaning

```
# Reading Data (https://www.icpsr.umich.edu/web/NACJD/studies/38604).
Household concatenated file
print(load(file ="C:\\Users\\usa7k\\OneDrive\\Regression and Time
Series\\38604-0001-Data.rda"))
```

```
## [1] "da38604.0001"
```

```
dt<-subset(da38604.0001,
select=c(YEAR,YEARQ,V2135,V2132,V2129,V2122,V2121B,V2120,V2119,V2107,V2106,V2
006,V2143,V2025,V2025B,V2026,V2031,V2033,V2034,V2036,V2038,V2043,V2046,V2073,
V2074,V2077,V2078,V2080,VFLAG))

#Factor to Numeric
lbls <- sort(levels(dt$VFLAG))
lbls <- (sub("^\\([0-9]+\\) +(.+$)", "\\1", lbls))
dt$VFLAG <- as.numeric(sub("^\\(0*([0-9]+)\\).+$", "\\1", dt$VFLAG))
dt$VFLAG <- add.value.labels(dt$VFLAG, lbls)

lbls <- sort(levels(dt$V2026))
lbls <- (sub("^\\([0-9]+\\) +(.+$)", "\\1", lbls))
dt$V2026 <- as.numeric(sub("^\\(0*([0-9]+)\\).+$", "\\1", dt$V2026))
dt$V2026 <- add.value.labels(dt$V2026, lbls)

lbls <- sort(levels(dt$V2006))
lbls <- (sub("^\\([0-9]+\\) +(.+$)", "\\1", lbls))
dt$V2006 <- as.numeric(sub("^\\(0*([0-9]+)\\).+$", "\\1", dt$V2006))
dt$V2006 <- add.value.labels(dt$V2006, lbls)
```

```
lbls <- sort(levels(dt$V2135))
lbls <- (sub("^\\([0-9]+\\) +(.+$)", "\\1", lbls))
dt$V2135 <- as.numeric(sub("^\\(0*([0-9]+)\\).+$", "\\1", dt$V2135))
dt$V2135 <- add.value.labels(dt$V2135, lbls)

## More value labels than values, only the first 13 will be used

lbls <- sort(levels(dt$V2025B))
lbls <- (sub("^\\([0-9]+\\) +(.+$)", "\\1", lbls))
dt$V2025B <- as.numeric(sub("^\\(0*([0-9]+)\\).+$", "\\1", dt$V2025B))
dt$V2025B <- add.value.labels(dt$V2025B, lbls)

lbls <- sort(levels(dt$V2078))
lbls <- (sub("^\\([0-9]+\\) +(.+$)", "\\1", lbls))
dt$V2078 <- as.numeric(sub("^\\(0*([0-9]+)\\).+$", "\\1", dt$V2078))
dt$V2078 <- add.value.labels(dt$V2078, lbls)

lbls <- sort(levels(dt$V2077))
lbls <- (sub("^\\([0-9]+\\) +(.+$)", "\\1", lbls))
dt$V2077 <- as.numeric(sub("^\\(0*([0-9]+)\\).+$", "\\1", dt$V2077))
dt$V2077 <- add.value.labels(dt$V2077, lbls)

lbls <- sort(levels(dt$V2074))
lbls <- (sub("^\\([0-9]+\\) +(.+$)", "\\1", lbls))
dt$V2074 <- as.numeric(sub("^\\(0*([0-9]+)\\).+$", "\\1", dt$V2074))
dt$V2074 <- add.value.labels(dt$V2074, lbls)

lbls <- sort(levels(dt$V2073))
lbls <- (sub("^\\([0-9]+\\) +(.+$)", "\\1", lbls))
dt$V2073 <- as.numeric(sub("^\\(0*([0-9]+)\\).+$", "\\1", dt$V2073))
dt$V2073 <- add.value.labels(dt$V2073, lbls)

lbls <- sort(levels(dt$V2038))
lbls <- (sub("^\\([0-9]+\\) +(.+$)", "\\1", lbls))
dt$V2038 <- as.numeric(sub("^\\(0*([0-9]+)\\).+$", "\\1", dt$V2038))
dt$V2038 <- add.value.labels(dt$V2038, lbls)

lbls <- sort(levels(dt$V2033))
lbls <- (sub("^\\([0-9]+\\) +(.+$)", "\\1", lbls))
dt$V2033 <- as.numeric(sub("^\\(0*([0-9]+)\\).+$", "\\1", dt$V2033))
dt$V2033 <- add.value.labels(dt$V2033, lbls)

lbls <- sort(levels(dt$V2034))
lbls <- (sub("^\\([0-9]+\\) +(.+$)", "\\1", lbls))
dt$V2034 <- as.numeric(sub("^\\(0*([0-9]+)\\).+$", "\\1", dt$V2034))
dt$V2034 <- add.value.labels(dt$V2034, lbls)

lbls <- sort(levels(dt$V2036))
```

```
lbls <- (sub("^\\([0-9]+\\) +(.+$)", "\\1", lbls))
dt$V2036 <- as.numeric(sub("^\\(0*([0-9]+)\\).+$", "\\1", dt$V2036))
dt$V2036 <- add.value.labels(dt$V2036, lbls)

## More value labels than values, only the first 2 will be used

lbls <- sort(levels(dt$V2129))
lbls <- (sub("^\\([0-9]+\\) +(.+$)", "\\1", lbls))
dt$V2129 <- as.numeric(sub("^\\(0*([0-9]+)\\).+$", "\\1", dt$V2129))
dt$V2129 <- add.value.labels(dt$V2129, lbls)

## More value labels than values, only the first 3 will be used

lbls <- sort(levels(dt$V2119))
lbls <- (sub("^\\([0-9]+\\) +(.+$)", "\\1", lbls))
dt$V2119 <- as.numeric(sub("^\\(0*([0-9]+)\\).+$", "\\1", dt$V2119))
dt$V2119 <- add.value.labels(dt$V2119, lbls)

lbls <- sort(levels(dt$V2122))
lbls <- (sub("^\\([0-9]+\\) +(.+$)", "\\1", lbls))
dt$V2122 <- as.numeric(sub("^\\(0*([0-9]+)\\).+$", "\\1", dt$V2122))
dt$V2122 <- add.value.labels(dt$V2122, lbls)

# Dropping NA responses
dt<-dt %>% drop_na(VFLAG)

# Replacing NA with 0 for the regressors
dt <- dt %>% replace(is.na(.), 0)
dt$VFlag<-ifelse(dt$VFLAG ==2,1,0) # Victim Flag (0,1)
test0<-subset(dt, YEAR>=2006 & YEAR < 2007) # Non-crisis Before
test1<-subset(dt, YEAR>=2010 & YEAR < 2012) # Non-crisis After
train<-subset(dt, YEAR>=2007 & YEAR < 2009) # Crisis (Mortgage crisis)
val<-subset(dt, YEAR>=2020 & YEAR < 2022)   # Validation (COVID-19)
```

SLR:Normality and Heteroscadicity

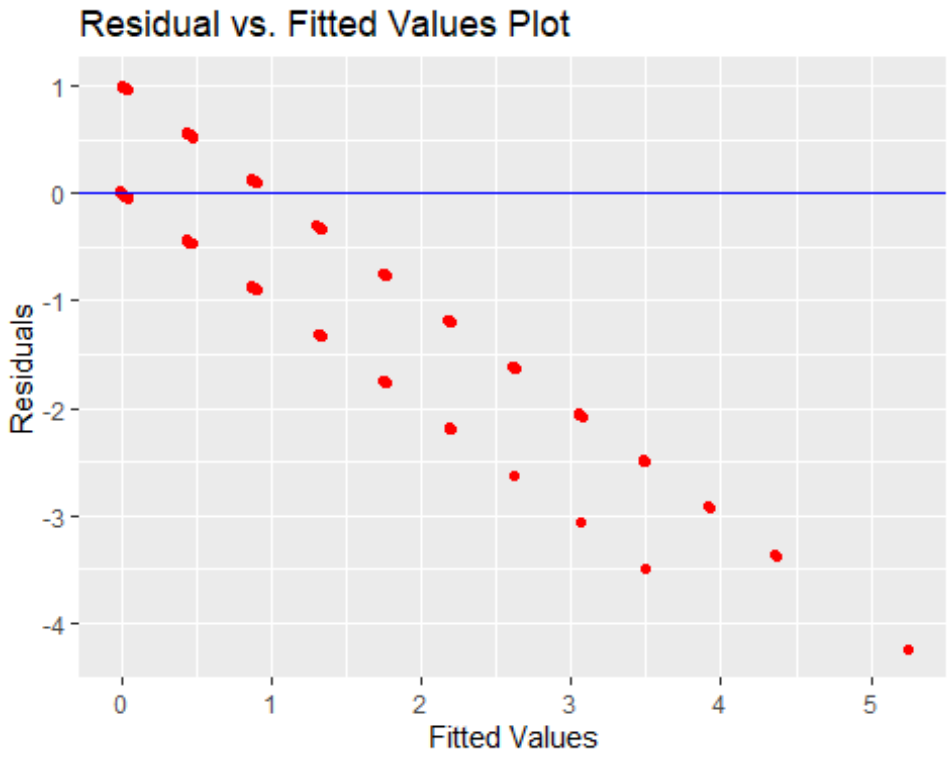
```
summary(train)
```

```
##       YEAR          YEARQ          V2135
##  Min.   :2007   Min.   :2007   Min.   :11.0
##  1st Qu.:2007   1st Qu.:2007   1st Qu.:98.0
##  Median :2007   Median :2007   Median :98.0
##  Mean   :2007   Mean   :2008   Mean   :94.4
##  3rd Qu.:2008   3rd Qu.:2008   3rd Qu.:98.0
##  Max.   :2008   Max.   :2008   Max.   :98.0
##
##                            V2132           V2129          V2122
##  (0) Regular school           :  1026   Min.   :1.000   Min.   : 1.00
##  (1) College/university       :  9989   1st Qu.:1.000   1st Qu.: 8.00
##  (2) Trade school             :   273   Median :2.000   Median :16.00
##  (3) Vocational school        :   658   Mean   :1.861   Mean   :18.07
##  (4) None of the above schools:146378   3rd Qu.:2.000   3rd Qu.:28.00
```

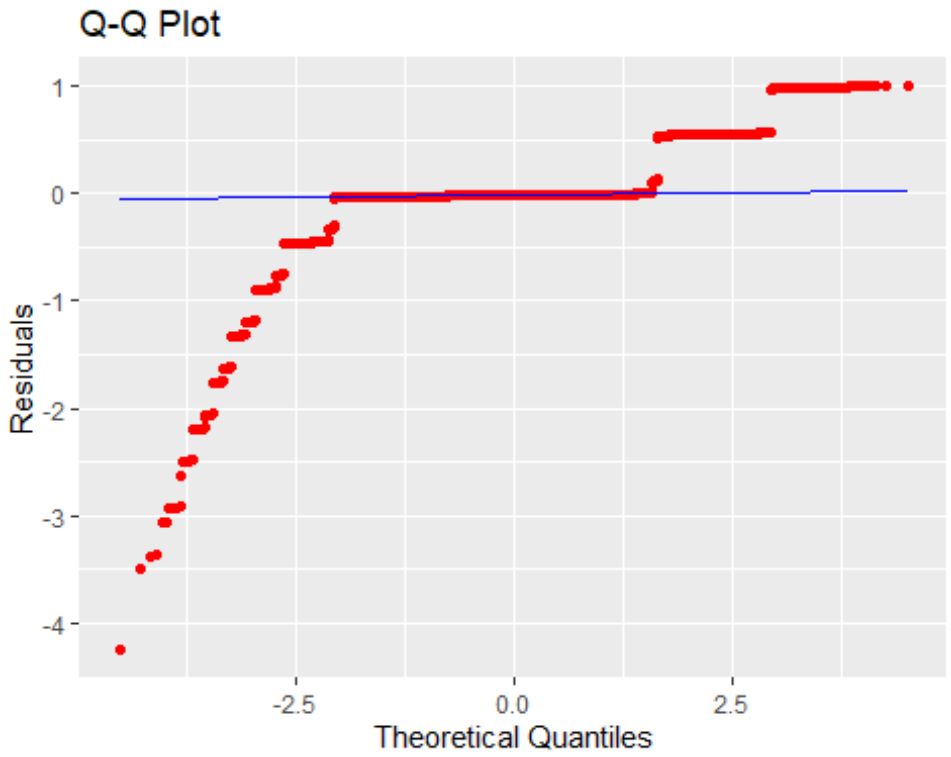
```
##  (8) Residue                 :   709   Max.   :3.000   Max.   :34.00
##
##          V2121B                                V2120            V2119
##  (1) Yes    :   575   (1) Yes (public housing)   :  3333   Min.   :1.00
##  (2) No     :158448   (2) No (not public housing): 44116   1st Qu.:2.00
##  (8) Residue:    10   (7) Item Blank             :     0   Median :2.00
##                       (8) Residue                :    63   Mean   :1.99
##                       NA's                       :111521   3rd Qu.:2.00
##                                                            Max.   :8.00
##
##             V2107                V2106            V2006
##  (1) Yes       :     0   (1) Yes    :     0   Min.   :1.000
##  (2) No        :     0   (2) No     :     0   1st Qu.:1.000
##  (3) Don't know:     0   (8) Residue:     0   Median :1.000
##  (8) Residue   :     0   NA's       :159033   Mean   :1.169
##  NA's          :159033                       3rd Qu.:1.000
##                                              Max.   :7.000
##
##           V2143                    V2025            V2025B          V2026
##  (1) Urban   :     0   (1) Yes       : 28267   Min.   :1.000   Min.   :
1.00
##  (2) Suburban:     0   (2) No        :  8787   1st Qu.:2.000   1st
Qu.:10.00
##  (3) Rural   :     0   (3) Don't know:   193   Median :2.000   Median
:14.00
##  (8) Residue :     0   (7) Item blank:     0   Mean   :1.935   Mean
:38.45
##  NA's        :159033   (8) Residue   :  2438   3rd Qu.:2.000   3rd
Qu.:98.00
##                        NA's          :119348   Max.   :2.000   Max.
:98.00
##
##                V2031            V2033           V2034           V2036
##  (01) White only :     0   Min.   :12.00   Min.   :1.000   Min.   :1.000
##  (02) Black only :     0   1st Qu.:36.00   1st Qu.:1.000   1st Qu.:2.000
##  (06) White-Black:     0   Median :48.00   Median :1.000   Median :2.000
##  (21) Other only :     0   Mean   :49.37   Mean   :2.292   Mean   :1.794
##  (22) White-Other:     0   3rd Qu.:61.00   3rd Qu.:3.000   3rd Qu.:2.000
##  (Other)         :     0   Max.   :90.00   Max.   :8.000   Max.   :2.000
##  NA's            :159033
##      V2038                    V2043              V2046
##  Min.   : 0.00   (1) Married      :84562   (1) Yes    :  1786
##  1st Qu.:28.00   (2) Widowed      :15977   (2) No     :124863
##  Median :40.00   (3) Divorced     :22838   (8) Residue:   562
##  Mean   :34.69   (4) Separated    : 4490   NA's       : 31822
##  3rd Qu.:42.00   (5) Never married:29290
##  Max.   :98.00   (8) Residue      : 1876
##
##      V2073             V2074           V2077              V2078
##  Min.   : 0.0000   Min.   :1.000   Min.   :  0.0000   Min.   :0.000
```

```
##  1st Qu.: 0.0000   1st Qu.:2.000   1st Qu.:  0.0000   1st Qu.:1.000
##  Median : 0.0000   Median :2.000   Median :  0.0000   Median :2.000
##  Mean   : 0.1022   Mean   :1.952   Mean   :  0.6763   Mean   :1.958
##  3rd Qu.: 0.0000   3rd Qu.:2.000   3rd Qu.:  0.0000   3rd Qu.:3.000
##  Max.   :12.0000   Max.   :8.000   Max.   :998.0000   Max.   :8.000
##
##      V2080              VFLAG           VFlag
##  Min.   :  0.0000   Min.   :1.000   Min.   :0.0000
##  1st Qu.:  0.0000   1st Qu.:1.000   1st Qu.:0.0000
##  Median :  0.0000   Median :1.000   Median :0.0000
##  Mean   :  0.8317   Mean   :1.063   Mean   :0.0628
##  3rd Qu.:  0.0000   3rd Qu.:1.000   3rd Qu.:0.0000
##  Max.   :998.0000   Max.   :2.000   Max.   :1.0000
##

# V2120, V2107, V2106, V2025, V2031, V2046 are removed as a majority of
values are NA

lmod0<-lm(VFlag ~  V2135 + V2132 + V2129 + V2122 + V2121B  + V2119 + V2006 +
V2025B + V2026 + V2033 + V2034 + V2036 + V2038 + V2043 + V2073 + V2074 +
V2077 + V2078 + V2080, data=train)
lmod<-ols_step_backward_p(lmod0, prem=0.05)$model
summary(lmod)

##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2431 -0.0227 -0.0181 -0.0128  0.9999
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       3.666e-02  5.988e-03   6.122 9.27e-10
***
## V2132(1) College/university        1.054e-02  5.213e-03   2.021  0.04325
*
## V2132(2) Trade school              6.427e-03  1.083e-02   0.594  0.55275
## V2132(3) Vocational school         1.086e-02  7.939e-03   1.368  0.17118
## V2132(4) None of the above schools  6.930e-03  4.989e-03   1.389  0.16482
## V2132(8) Residue                  -8.094e-03  7.804e-03  -1.037  0.29968
## V2129                             -5.656e-03  5.871e-04  -9.633  < 2e-16
***
## V2026                             -4.292e-05  9.855e-06  -4.356 1.33e-05
***
## V2033                             -1.880e-04  2.456e-05  -7.657 1.91e-14
***
## V2073                              4.343e-01  9.530e-04 455.662  < 2e-16
```

```
***
## V2074                              -4.689e-03  1.511e-03  -3.104  0.00191
**
## V2078                               2.718e-03  3.306e-04   8.222  < 2e-16
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1589 on 159021 degrees of freedom
## Multiple R-squared:  0.5708, Adjusted R-squared:  0.5708
## F-statistic: 1.923e+04 on 11 and 159021 DF,  p-value: < 2.2e-16

summary(lmod$fitted.values)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.01477  0.01431  0.01920  0.06280  0.02407  5.24306

# The fitted values outbound the 0 to 1 range for the response

# Anderson-Darling Test. H0: Distribution is normal
ad.test(lmod$residuals)

##
##  Anderson-Darling normality test
##
## data:  lmod$residuals
## A = 43915, p-value < 2.2e-16

#Reject the null. Residuals are not normal
#Shapiro-Wilk Test is bound to a maximum of 5000 responses

# Heteroscadicty
ggplot(lmod, aes(x = .fitted, y = .resid)) + geom_point(col="red") +
geom_hline(yintercept = 0, col="blue") +labs(title='Residual vs. Fitted
Values Plot', x='Fitted Values', y='Residuals')
```




```
# Q-Q plot
ggplot(lmod, aes(sample=residuals(lmod)))+ labs(title="Q-Q
Plot",x="Theoretical Quantiles" , y="Residuals") + stat_qq(col="red")+
stat_qq_line(col="blue")
```




```
# There is Heteroscadicity and the residuals are not normally distributed.
Violates assumption of Ordinary Least Squares regression
```

Creating Binaries(SLR significant)

```
dt$motorGe1<-ifelse(dt$V2078 >= 1,1,0) # Motor Vehicles Owned greater than 1
dt$timesbrkIn<-dt$V2077 # No. of times broken-in
dt$busFadd<-ifelse(dt$V2074 == 1,1,0) # Business from household address
dt$ppEduL5<-ifelse(dt$V2038 < 5,1,0) # Principal Person Education Less than
5th grade
dt$age<-dt$V2033 # Principal Person age
dt$MSA<-ifelse(dt$V2129 == 3, 0,1) # Metropolitan Statistical Area
dt$famstr1<-ifelse((dt$V2122 %in% c(8,15,19,23)),1,0) # Family Structure
Cluster 1
dt$famstr2<-ifelse((dt$V2122 %in% c(25,26,27,28,29,30,31,32)),1,0) # Family
Structure Cluster 2

test0<-subset(dt, YEAR>=2006 & YEAR < 2007) # Non-crisis Before
test1<-subset(dt, YEAR>=2010 & YEAR < 2012) # Non-crisis After
train<-subset(dt, YEAR>=2007 & YEAR < 2009) # Crisis (Mortgage crisis)
val<-subset(dt, YEAR>=2020 & YEAR < 2022)   # Validation (COVID-19)
```

Logistic Regression

```
model00<-glm(VFlag ~ 1, family=binomial("logit"),data=train)
model0<-glm(VFlag ~ V2135 + V2132 + MSA + famstr1 + famstr2 + V2121B  + V2119
+ V2006 + V2025B + V2026 + age + V2034 + V2036 + ppEduL5 + V2043 + V2073 +
busFadd + timesbrkIn + motorGe1 + V2080, family=binomial("logit"),data=train)
lrtest(model00,model0) # comparing null and full model
```

```
## Likelihood ratio test for MLE method
## Chi-squared 29 d.f. =  52857.96 , P value =  0

# p-value = 0. Reject H0. Thus Full model is more significant

summary(model0)

##
## Call:
## glm(formula = VFlag ~ V2135 + V2132 + MSA + famstr1 + famstr2 +
##     V2121B + V2119 + V2006 + V2025B + V2026 + age + V2034 + V2036 +
##     ppEduL5 + V2043 + V2073 + busFadd + timesbrkIn + motorGe1 +
##     V2080, family = binomial("logit"), data = train)
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -5.419e+00  5.820e-01  -9.311  < 2e-16
***
## V2135                            -7.365e-05  1.075e-03  -0.068  0.94539
## V2132(1) College/university       1.284e-01  2.531e-01   0.507  0.61210
## V2132(2) Trade school             1.635e-01  4.510e-01   0.363  0.71691
## V2132(3) Vocational school        1.841e-01  3.528e-01   0.522  0.60167
## V2132(4) None of the above schools  1.405e-01  2.460e-01   0.571  0.56790
## V2132(8) Residue                 -1.828e-01  4.883e-01  -0.374  0.70815
## MSA                               1.420e-01  5.204e-02   2.728  0.00636
**
## famstr1                          -4.100e-02  6.461e-02  -0.635  0.52571
## famstr2                          -1.824e-01  9.721e-02  -1.876  0.06065
.
## V2121B(2) No                      2.359e-01  2.807e-01   0.840  0.40068
## V2121B(8) Residue                 1.315e+00  2.102e+00   0.626  0.53160
## V2119                            -4.305e-01  1.684e-01  -2.556  0.01058
*
## V2006                             5.975e-03  3.643e-02   0.164  0.86974
## V2025B                            3.485e-02  7.922e-02   0.440  0.65996
## V2026                            -4.537e-04  4.840e-04  -0.937  0.34858
## age                              -1.282e-03  1.533e-03  -0.836  0.40289
## V2034                            -1.437e-01  1.621e-01  -0.886  0.37546
## V2036                             1.269e-01  9.664e-02   1.313  0.18920
## ppEduL5                           7.333e-02  1.890e-01   0.388  0.69796
## V2043(2) Widowed                  3.824e-01  1.974e-01   1.937  0.05273
.
## V2043(3) Divorced                 2.909e-01  3.352e-01   0.868  0.38554
## V2043(4) Separated                2.112e-01  4.986e-01   0.424  0.67189
## V2043(5) Never married            6.338e-01  6.498e-01   0.975  0.32933
## V2043(8) Residue                  1.159e+00  1.150e+00   1.008  0.31361
## V2073                             6.016e+00  4.259e-02 141.238  < 2e-16
***
## busFadd                           9.283e-02  7.788e-02   1.192  0.23326
## timesbrkIn                        8.709e-04  1.158e-03   0.752  0.45217
```

```
## motorGe1                          5.599e-01  7.865e-02   7.119 1.09e-12
***
## V2080                            -1.060e-03  1.020e-03  -1.040  0.29849
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74623  on 159032  degrees of freedom
## Residual deviance: 21765  on 159003  degrees of freedom
## AIC: 21825
##
## Number of Fisher Scoring iterations: 8

# Binaries for the significant Vars: V2119 & V2073 and for potential vars:
V2025B, V2026, V2006, V2034 & V2036
```

Creating Binaries(Logistic Significant)

```
dt$noUniF<-ifelse(dt$V2119 == 2,1,0) # No University Attended
dt$crimInc<-ifelse(dt$V2073 > 0,1,0) # Criminal Incident
dt$noResAc<-ifelse(dt$V2025B == 2, 1,0) # No restricted access to household
dt$repHous<-ifelse(dt$V2006 > 1,1,0) # Replacement Household
dt$IncGe35k<-ifelse(dt$V2026 < 11, 0, 1) # Household income greater than
$35,000
dt$widowed<-ifelse(dt$V2034 == 2, 1,0) # Primary person is widowed
dt$female<-ifelse(dt$V2036 == 2, 1,0)  # Primary person is female

test0<-subset(dt, YEAR>=2006 & YEAR < 2007) # Non-crisis Before
test1<-subset(dt, YEAR>=2010 & YEAR < 2012) # Non-crisis After
train<-subset(dt, YEAR>=2007 & YEAR < 2009) # Crisis (Mortgage crisis)
val<-subset(dt, YEAR>=2020 & YEAR < 2022)   # Validation (COVID-19)
```

Stepwise Logistic

```
model1<-stepwiseLogit(VFlag ~ noUniF + MSA + famstr1 + famstr2  + repHous +
noResAc + IncGe35k + age + widowed + female + ppEduL5 + crimInc + busFadd +
timesbrkIn + motorGe1 , data=train, selection="backward", select="AIC",
sigMethod = "LRT")
model1

##     Table 1. Summary of Parameters
## ═══════════════════════════════════════
##          Paramters            Value
## ═══════════════════════════════════════
## Response Variable           VFlag
## Included Variable           NULL
## Selection Method            backward
## Select Criterion            AIC
## Variable significance test  LRT
## Multicollinearity Terms     NULL
## Intercept                   1
```

```
##
##                                                    Table 2. Variables
Type
##
```

class	variable
numeric	VFlag noUniF MSA famstr1 famstr2 repHous noResAc IncGe35k age widowed female ppEduL5 crimInc busFadd timesbrkIn motorGe1

```
##
##                     Table 3. Process of Selection
```

Step	EnteredEffect	RemovedEffect	DF	NumberIn	AIC
1			16	16	15872.0458731687
2		repHous	1	15	15870.045878359
3		timesbrkIn	1	14	15868.0584588474
4		famstr1	1	13	15866.1353113192
5		ppEduL5	1	12	15864.2684908948
6		noResAc	1	11	15862.4311543795
7		busFadd	1	10	15860.9793542074
8		age	1	9	15859.6294800089
9		MSA	1	8	15858.5412647206

```
##
##                                  Table 4. Selected Varaibles
##
```

variables1	variables2	variables3	variables4	variables5	variables6	variables7	variables8
noUniF	famstr2	IncGe35k	widowed	female	crimInc	motorGe1	

```
##
##                          Table 5. Coefficients of the Selected Variables
##
```

Variable	Estimate	StdError	t.value

```
P.value
## ------------------------------------------------------------------------
------------------------
## (Intercept)  -6.503587499113      0.220026888010731   -29.5581488149567
5.16014679442623e-192
## noUniF       -0.493547242597409   0.201745738537012   -2.44638249202406
0.0144297826312012
## famstr2      -0.224236927893962   0.0539882953476861  -4.15343597070199
3.27519887092142e-05
## IncGe35k     -0.0790012683677412  0.0497858497597644  -1.58682173245917
0.112553025289203
## widowed      0.387153804187029    0.0940024243382983  4.11855127048352
3.81261749370966e-05
## female       0.179271155597058    0.0561456874841517  3.19296394131181
0.00140820530019598
## crimInc      7.78214690436798     0.0672719935896395  115.681823729489   0
## motorGe1     0.610825145696425    0.0802593048995493  7.61064584923728
2.72729605169467e-14
##
================================================================================
==================
```

```
x<-
data.matrix(train[,c('noUniF','famstr2','IncGe35k','widowed','female','crimIn
c','motorGe1')])
round(cor(x),2) # correlation matrix
```

```
##          noUniF famstr2 IncGe35k widowed female crimInc motorGe1
## noUniF     1.00   -0.01     0.05    0.01   0.03   -0.01     0.04
## famstr2   -0.01    1.00    -0.21    0.36   0.33    0.01    -0.20
## IncGe35k   0.05   -0.21     1.00   -0.15   0.08   -0.03     0.19
## widowed    0.01    0.36    -0.15    1.00   0.00   -0.04    -0.17
## female     0.03    0.33     0.08    0.00   1.00   -0.02     0.05
## crimInc   -0.01    0.01    -0.03   -0.04  -0.02    1.00     0.01
## motorGe1   0.04   -0.20     0.19   -0.17   0.05    0.01     1.00
```

```
#Independent variables are not correlated
```

Cross Validation and Accuracy

```
# 5-fold cross validation
train_control <- trainControl(method = "cv", number = 5)

# Using the selected variables
modelcv <- train(VFlag ~ noUniF + famstr2 + IncGe35k + widowed + female +
crimInc + motorGe1, data =train, trControl = train_control, method = "glm",
family=binomial(logit))
modelcv
```

```
## Generalized Linear Model
##
## 159033 samples
```

```
##      7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 127226, 127227, 127227, 127226, 127226
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   0.117267  0.7663095  0.02750274
```

```
vif(modelcv$finalModel)
```

```
##   noUniF  famstr2 IncGe35k  widowed   female  crimInc motorGe1
## 1.008876 1.386311 1.123912 1.088233 1.230207 1.021334 1.095154
```

```
# Calculating Accuracy as (no. of correct predictions/Total)

train$pred<-modelcv$finalModel$fitted.values
train$Fpred<-ifelse(train$pred>0.5,1,0)
train$flag<-ifelse(train$Fpred==train$VFlag,1,0)
# % Accuracy in Training
100*mean(train$flag)
```

```
## [1] 98.30853
```

```
test0$pred<-predict(modelcv,newdata = test0)
test0$Fpred<-ifelse(test0$pred>0.5,1,0)
test0$flag<-ifelse(test0$Fpred==test0$VFlag,1,0)
# % Accuracy in Before crisis
100*mean(test0$flag)
```

```
## [1] 97.54932
```

```
test1$pred<-predict(modelcv,newdata = test1)
test1$Fpred<-ifelse(test1$pred>0.5,1,0)
test1$flag<-ifelse(test1$Fpred==test1$VFlag,1,0)
# % Accuracy in After crisis
100*mean(test1$flag)
```

```
## [1] 99.03555
```

```
val$pred<-predict(modelcv,newdata = val)
val$Fpred<-ifelse(val$pred>0.5,1,0)
val$flag<-ifelse(val$Fpred==val$VFlag,1,0)
# % Accuracy in Validation
100*mean(val$flag)
```

```
## [1] 99.23249
```

Comparisons

```
# Fitting a model to before and after crisis time period using the variables
selected from training set(Mortgage crisis)
```

```r
modelTest0<-glm(VFlag ~ noUniF + famstr2 + IncGe35k + widowed + female +
crimInc + motorGe1, family=binomial(logit), data =test0)
modelTest1<-glm(VFlag ~ noUniF + famstr2 + IncGe35k + widowed + female +
crimInc + motorGe1, family=binomial(logit), data =test1)
modelVal<-glm(VFlag ~ noUniF + famstr2 + IncGe35k + widowed + female +
crimInc + motorGe1, family=binomial(logit), data =val)
summary(modelTest0)
```

```
##
## Call:
## glm(formula = VFlag ~ noUniF + famstr2 + IncGe35k + widowed +
##     female + crimInc + motorGe1, family = binomial(logit), data = test0)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.67307    0.20303 -32.867  < 2e-16 ***
## noUniF       0.14089    0.17032   0.827  0.40812
## famstr2     -0.14263    0.06581  -2.167  0.03021 *
## IncGe35k     0.02264    0.05850   0.387  0.69870
## widowed      0.37071    0.11800   3.142  0.00168 **
## female       0.18457    0.07027   2.627  0.00862 **
## crimInc      7.07009    0.07855  90.003  < 2e-16 ***
## motorGe1     0.51542    0.10386   4.963 6.96e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 39173  on 75978  degrees of freedom
## Residual deviance: 10242  on 75971  degrees of freedom
## AIC: 10258
##
## Number of Fisher Scoring iterations: 8
```

```r
summary(modelcv$finalModel)
```

```
##
## Call:
## NULL
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.50359    0.22003 -29.558  < 2e-16 ***
## noUniF      -0.49355    0.20175  -2.446  0.01443 *
## famstr2     -0.22424    0.05399  -4.153 3.28e-05 ***
## IncGe35k    -0.07900    0.04979  -1.587  0.11255
## widowed      0.38715    0.09400   4.119 3.81e-05 ***
## female       0.17927    0.05615   3.193  0.00141 **
## crimInc      7.78215    0.06727 115.682  < 2e-16 ***
## motorGe1     0.61083    0.08026   7.611 2.73e-14 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74623  on 159032  degrees of freedom
## Residual deviance: 15843  on 159025  degrees of freedom
## AIC: 15859
##
## Number of Fisher Scoring iterations: 9
```

```
summary(modelTest1)
```

```
##
## Call:
## glm(formula = VFlag ~ noUniF + famstr2 + IncGe35k + widowed +
##     female + crimInc + motorGe1, family = binomial(logit), data = test1)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -23.48979  123.71669  -0.190 0.849413
## noUniF        0.17009    0.21993   0.773 0.439289
## famstr2      -0.37794    0.06753  -5.597 2.18e-08 ***
## IncGe35k      0.01906    0.06103   0.312 0.754880
## widowed       0.61652    0.14064   4.384 1.17e-05 ***
## female        0.27563    0.07275   3.788 0.000152 ***
## crimInc      24.47153  123.71649   0.198 0.843199
## motorGe1      0.59514    0.08878   6.704 2.03e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 72319.0  on 161749  degrees of freedom
## Residual deviance:  8888.5  on 161742  degrees of freedom
## AIC: 8904.5
##
## Number of Fisher Scoring iterations: 21
```

```
summary(modelVal)
```

```
##
## Call:
## glm(formula = VFlag ~ noUniF + famstr2 + IncGe35k + widowed +
##     female + crimInc + motorGe1, family = binomial(logit), data = val)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -24.31627  151.27952  -0.161  0.87230
## noUniF        0.16904    0.18809   0.899  0.36881
## famstr2      -0.17575    0.05857  -3.001  0.00269 **
```

```
## IncGe35k      0.12885    0.05560   2.317  0.02048 *
## widowed       0.34772    0.10738   3.238  0.00120 **
## female        0.14568    0.06050   2.408  0.01603 *
## crimInc      25.24053  151.27939   0.167  0.86749
## motorGe1      0.40146    0.08021   5.005 5.58e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 94395  on 288464  degrees of freedom
## Residual deviance: 11932  on 288457  degrees of freedom
## AIC: 11948
##
## Number of Fisher Scoring iterations: 22

# The effect of the crisis can be assessed from the coefficients of noUniF
and IncGe35K. During the period of crisis the sign is reversed when compared
to the time before and after.
# If the principal person attended University then the chance of household
victimization is increased during the crisis while it would follows an
opposite trend otherwise.
# Household income greater than $35,000 sees a higher risk of household
victimization during the non-crisis periods.
```