# lm_accident_prediction

Ali_Mirzaei

2022-11-30

##library

```
if(!require('dplyr')) install.packages('dplyr')
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
if(!require('mgcv')) install.packages('mgcv')
```

```
## Loading required package: mgcv
```

```
## Loading required package: nlme
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
## This is mgcv 1.8-41. For overview type 'help("mgcv-package")'.
```

```
if(!require('plotly')) install.packages('plotly')
```

```
## Loading required package: plotly
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

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

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
if(!require('GGally')) install.packages('GGally')
```

```
## Loading required package: GGally
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
if(!require('gratia')) install.packages('gratia')
```

```
## Loading required package: gratia
```

```
if(!require('ggeffects')) install.packages('ggeffects')
```

```
## Loading required package: ggeffects
```

```
if(!require('scico')) install.packages('scico')
```

```
## Loading required package: scico
```

```
if(!require('beepr')) install.packages('beepr')
```

```
## Loading required package: beepr
```

```
if(!require('visibly')) devtools::install_github('m-clark/visibly', upgrade = "never")
```

```
## Loading required package: visibly
```

```
if(!require('tidyext')) devtools::install_github('m-clark/tidyext', upgrade = "never")
```

```
## Loading required package: tidyext
```

```
##
## Attaching package: 'tidyext'
```

```
## The following object is masked from 'package:visibly':
##
##     create_prediction_data
```

```
Sys.setlocale(locale = "persian")
```

```
## Warning in Sys.setlocale(locale = "persian"): using locale code page other than
## 65001 ("UTF-8") may cause problems
```

```
## [1] "LC_COLLATE=Persian_Iran.1256;LC_CTYPE=Persian_Iran.1256;LC_MONETARY=Persian_Iran.1256;LC
_NUMERIC=C;LC_TIME=Persian_Iran.1256"
```

```
#install.packages('knitr', dependencies = TRUE)
library(knitr)
library('caret')
```

```
## Loading required package: lattice
```

```
library(dplyr)
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

# data set

```
MF_read_CsV<-function(path,use_UTF8=TRUE,choose_file=FALSE,set_max_overlaps=TRUE){
  if(set_max_overlaps){
    options(ggrepel.max.overlaps = Inf)
  }
  if(choose_file){
    path=file.choose()
  }
  if(use_UTF8){
    data<- read.csv(path,encoding="UTF-8")
  }else{
    data<- read.csv(path)
  }

  return(data)
}

df<- MF_read_CsV(path="C:/Users/Traffic/Desktop/SOHBATZADEH/esfahan-data-98-final.csv")
head(df)
```

```
##        JDATE                DESCRIPTION DISTANCE_VIOLATIONS HEAVY_VEHICLES
## 1 1398-01-01       61026          106390              جشن نوروز/جشن سال نو
## 2 1398-01-02                   64147          119771              عیدنوروز
## 3 1398-01-03                   74003          122579              عیدنوروز
## 4 1398-01-04                   78040          122446              عیدنوروز
## 5 1398-01-05                                         85087          86958
## 6 1398-01-06 83187          63646               روز امید، روز شادباش نویسي
##   IS_HOLIDAY JDAY JMONTH JYEAR SPEED_VIOLATIONS             TAG TOTAL_VEHICLES
## 1          1    1      1  1398            97782 1011333        تعطیلي بلند مدت
## 2          1    2      1  1398           104370 1087166        تعطیلي بلند مدت
## 3          1    3      1  1398           103648 1144199        تعطیلي بلند مدت
## 4          1    4      1  1398           104569 1206069        تعطیلي بلند مدت
## 5          0    5      1  1398            56393            1053166         عادي
## 6          0    6      1  1398            53070            823378          عادي
##   acc_total avg_class1_speed avg_class2_speed avg_class3_speed avg_class4_speed
## 1        16         88.85547         88.04384         81.53555         87.51029
## 2        15         88.89665         88.24069         81.29282         87.29790
## 3        16         88.23374         87.13057         81.66533         87.18464
## 4        16         87.94667         85.78571         80.79592         86.06595
## 5        14         82.63123         78.51480         75.88387         79.74674
## 6        14         81.13687         77.63398         75.12573         80.70641
##   avg_class5_speed avg_spedd85_class1 avg_spedd85_class2 avg_spedd85_class3
## 1         83.74037           95.56849           92.33562           89.21918
## 2         84.39677           95.56849           92.33562           89.21918
## 3         83.57216           95.56849           92.33562           89.21918
## 4         82.23939           95.56849           92.33562           89.21918
## 5         75.77073           95.56849           92.33562           89.21918
## 6         76.29402           95.56849           92.33562           89.21918
##   avg_spedd85_class4 avg_spedd85_class5  class1 class1_speed class2
## 1           96.21233           86.28767  936897        89690  30549
## 2           96.21233           86.28767  985248        95348  33368
## 3           96.21233           86.28767 1045099        94073  38294
## 4           96.21233           86.28767 1050770        94255  39939
## 5           96.21233           86.28767  949578        49931  36973
## 6           96.21233           86.28767  730290        47092  34225
##   class2_speed class3 class3_speed class4 class4_speed class5 class5_speed
## 1         6669   8404          446   8692          207  13381          770
## 2         7696   8421          446   8842          203  13516          677
## 3         8087  10593          484   9643          243  15473          761
## 4         8748  10764          530  10121          239  17216          797
## 5         5307  12423          390  13040          193  24522          572
## 6         4791  13117          406  12788          194  23057          587
##   police_enforcment
## 1              4377
## 2              5492
## 3              5340
## 4              5196
## 5              2376
## 6              2362
```

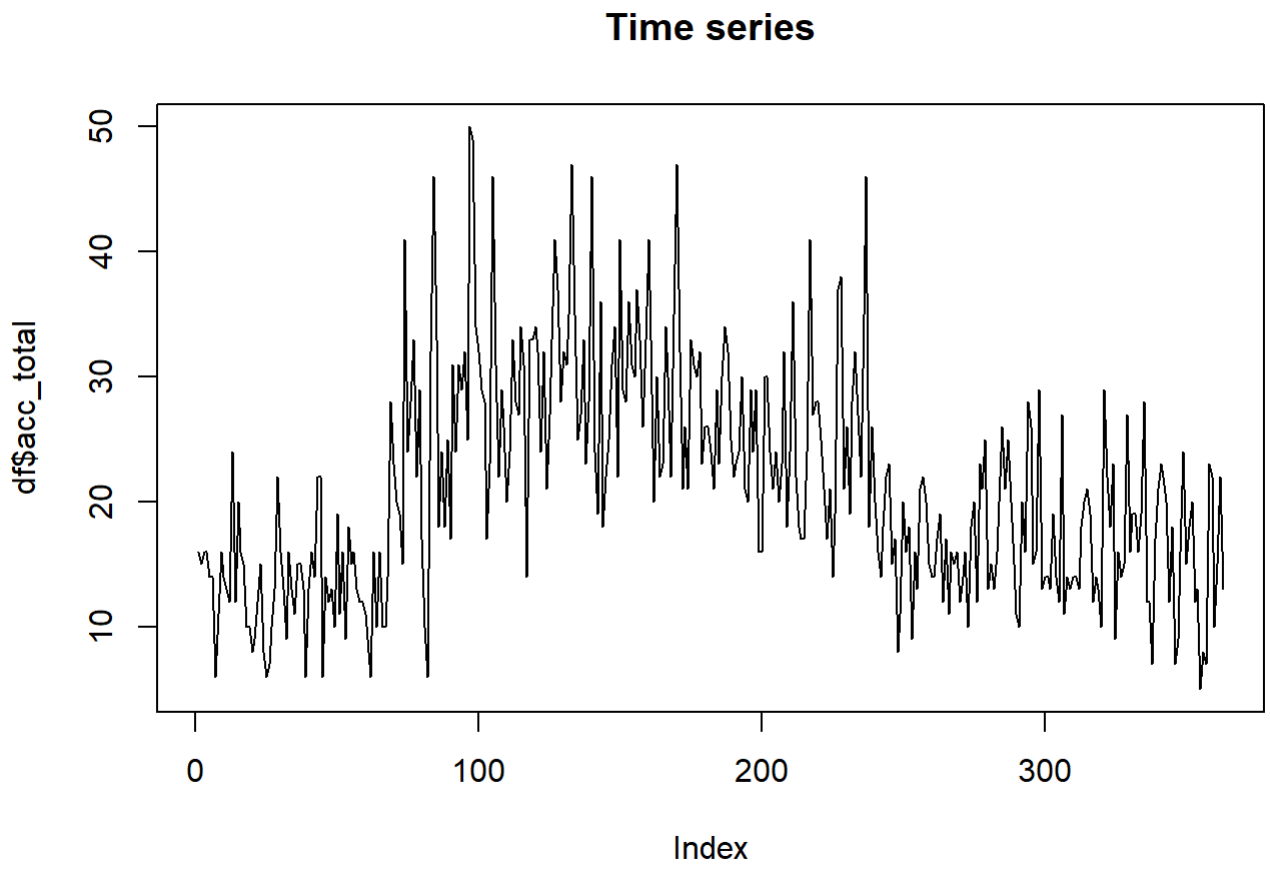
#data explantory

```
glimpse(df)
```

```
## Rows: 363
## Columns: 33
## $ JDATE               <chr> "1398-01-01", "1398-01-02", "1398-01-03", "1398-01~
## $ DESCRIPTION         <chr> "ع" ,"عیدنوروز" ,"عیدنوروز" ,"جشن سال نو~
## $ DISTANCE_VIOLATIONS <int> 106390, 119771, 122579, 122446, 85087, 63646, 1249~
## $ HEAVY_VEHICLES      <int> 61026, 64147, 74003, 78040, 86958, 83187, 104182, ~
## $ IS_HOLIDAY          <int> 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0,~
## $ JDAY                <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,~
## $ JMONTH              <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ JYEAR               <int> 1398, 1398, 1398, 1398, 1398, 1398, 1398, 1398, 13~
## $ SPEED_VIOLATIONS    <int> 97782, 104370, 103648, 104569, 56393, 53070, 10046~
## $ TAG                 <chr> "تعطیلي بلند مدت" ,"تعطیلي بلند مدت" ,"تعطیلي بلند~
## $ TOTAL_VEHICLES      <int> 1011333, 1087166, 1144199, 1206069, 1053166, 82337~
## $ acc_total           <int> 16, 15, 16, 16, 14, 14, 6, 11, 16, 14, 13, 12, 24,~
## $ avg_class1_speed    <dbl> 88.85547, 88.89665, 88.23374, 87.94667, 82.63123, ~
## $ avg_class2_speed    <dbl> 88.04384, 88.24069, 87.13057, 85.78571, 78.51480, ~
## $ avg_class3_speed    <dbl> 81.53555, 81.29282, 81.66533, 80.79592, 75.88387, ~
## $ avg_class4_speed    <dbl> 87.51029, 87.29790, 87.18464, 86.06595, 79.74674, ~
## $ avg_class5_speed    <dbl> 83.74037, 84.39677, 83.57216, 82.23939, 75.77073, ~
## $ avg_spedd85_class1  <dbl> 95.56849, 95.56849, 95.56849, 95.56849, 95.56849, ~
## $ avg_spedd85_class2  <dbl> 92.33562, 92.33562, 92.33562, 92.33562, 92.33562, ~
## $ avg_spedd85_class3  <dbl> 89.21918, 89.21918, 89.21918, 89.21918, 89.21918, ~
## $ avg_spedd85_class4  <dbl> 96.21233, 96.21233, 96.21233, 96.21233, 96.21233, ~
## $ avg_spedd85_class5  <dbl> 86.28767, 86.28767, 86.28767, 86.28767, 86.28767, ~
## $ class1              <int> 936897, 985248, 1045099, 1050770, 949578, 730290, ~
## $ class1_speed        <int> 89690, 95348, 94073, 94255, 49931, 47092, 90296, 8~
## $ class2              <int> 30549, 33368, 38294, 39939, 36973, 34225, 50135, 4~
## $ class2_speed        <int> 6669, 7696, 8087, 8748, 5307, 4791, 8538, 8408, 88~
## $ class3              <int> 8404, 8421, 10593, 10764, 12423, 13117, 15703, 149~
## $ class3_speed        <int> 446, 446, 484, 530, 390, 406, 540, 525, 516, 604, ~
## $ class4              <int> 8692, 8842, 9643, 10121, 13040, 12788, 13877, 1324~
## $ class4_speed        <int> 207, 203, 243, 239, 193, 194, 291, 221, 245, 290, ~
## $ class5              <int> 13381, 13516, 15473, 17216, 24522, 23057, 24467, 2~
## $ class5_speed        <int> 770, 677, 761, 797, 572, 587, 803, 729, 887, 1019,~
## $ police_enforcment   <int> 4377, 5492, 5340, 5196, 2376, 2362, 5221, 5418, 61~
```

#plot Accidents for 365 dayes

```
plot(df$acc_total, main = "Time series",type = "l")
```




#lm

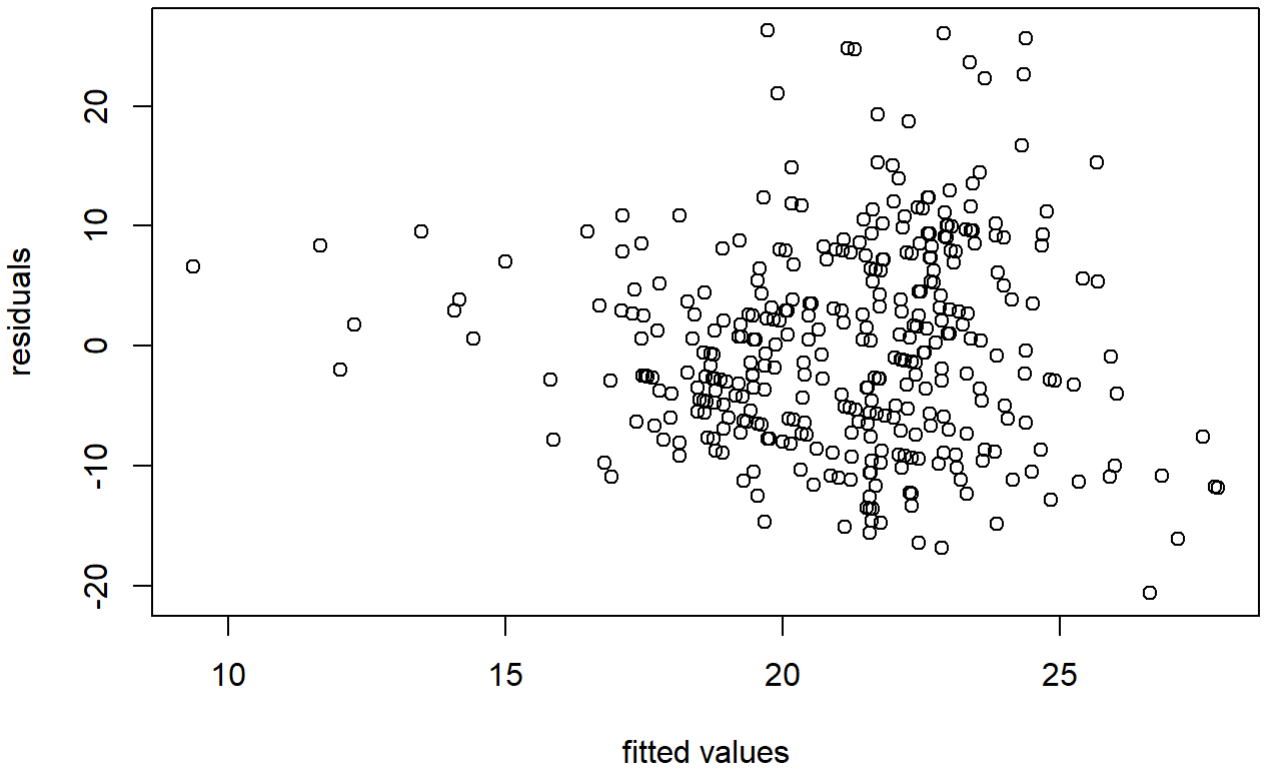
model0

```
modl0 <- lm(acc_total ~
              TOTAL_VEHICLES, data=df)
summary(modl0 )
```

```
##
## Call:
## lm(formula = acc_total ~ TOTAL_VEHICLES, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.6387  -6.3694  -0.7062   6.2448  26.2769
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.090e+00  2.313e+00   3.498 0.000527 ***
## TOTAL_VEHICLES 1.639e-05  2.850e-06   5.750  1.9e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.534 on 361 degrees of freedom
## Multiple R-squared:  0.0839, Adjusted R-squared:  0.08136
## F-statistic: 33.06 on 1 and 361 DF,  p-value: 1.904e-08
```

#residuals

```
plot(fitted(modl0),residuals(modl0), xlab="fitted values",ylab="residuals")
```




#confidence interval for coefficients

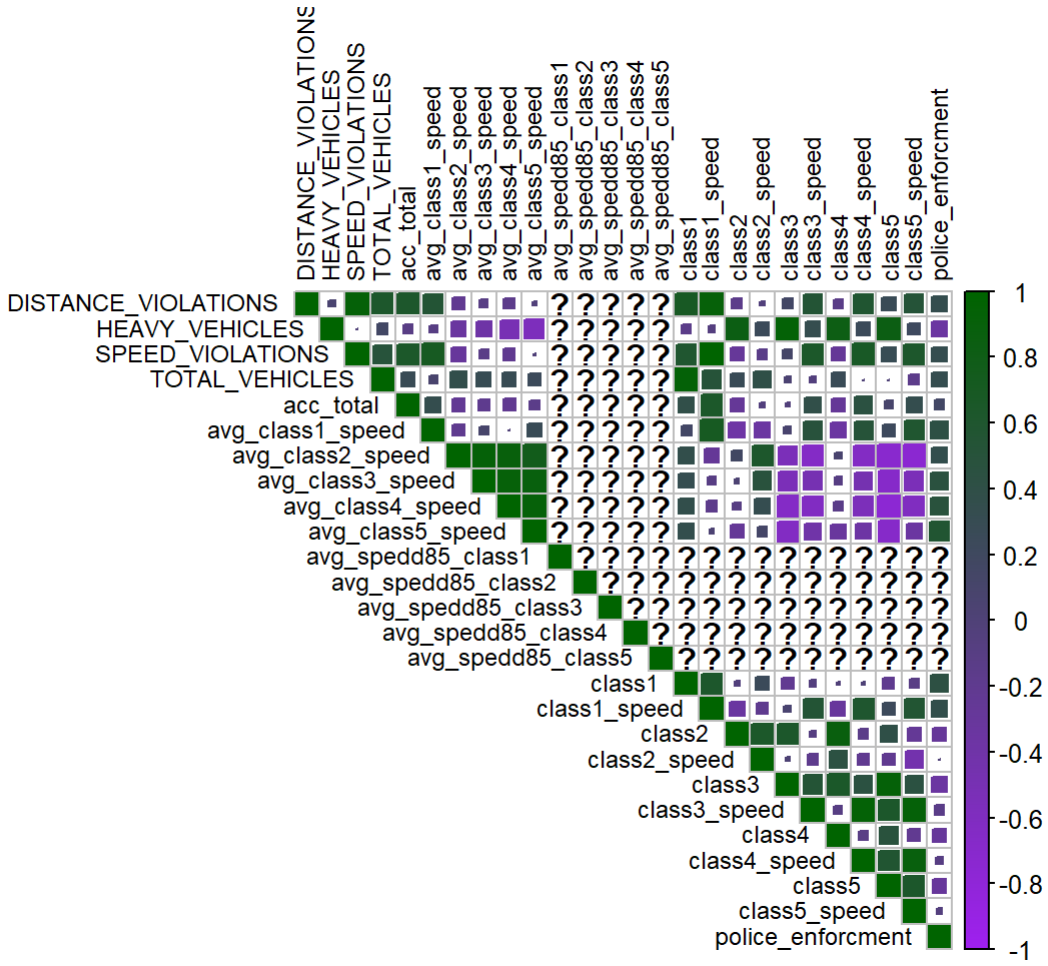
```
confint(modl0, 'TOTAL_VEHICLES', level=0.95)
```

```
##                       2.5 %       97.5 %
## TOTAL_VEHICLES 1.078398e-05 2.199496e-05
```

#pairs plot

```
corrplot(
  cor(df %>% select(-JDATE,-JYEAR, -TAG,
                    -JMONTH, -JDAY, -JDATE,
                    -IS_HOLIDAY, -DESCRIPTION,
                   )),
  method = 'square',
  type = 'upper',
  tl.col = 'black',
  tl.cex = 0.75, tl.srt = 90,
  col = colorRampPalette(c('purple', 'dark green'))(200)
)
```

```
## Warning in cor(df %>% select(-JDATE, -JYEAR, -TAG, -JMONTH, -JDAY, -JDATE, : the
## standard deviation is zero
```



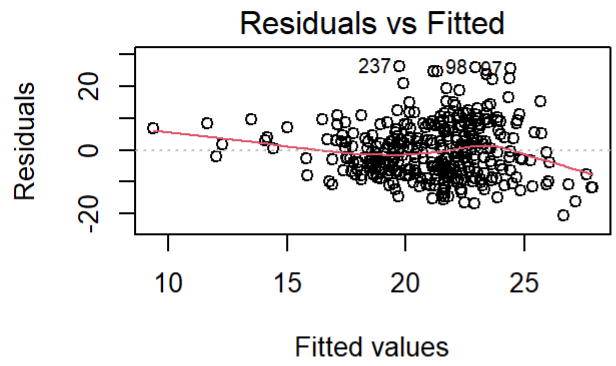

#model matrix

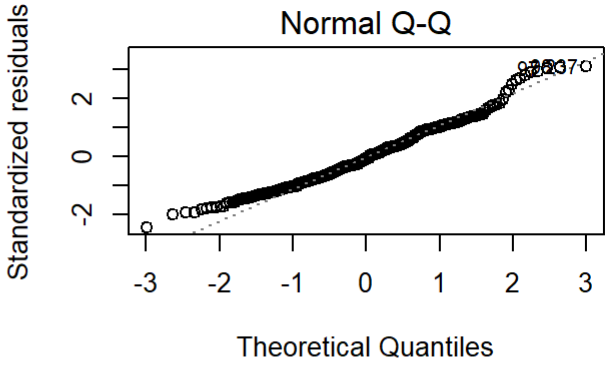
```
head(model.matrix(modl0))
```

```
##   (Intercept) TOTAL_VEHICLES
## 1           1        1011333
## 2           1        1087166
## 3           1        1144199
## 4           1        1206069
## 5           1        1053166
## 6           1         823378
```

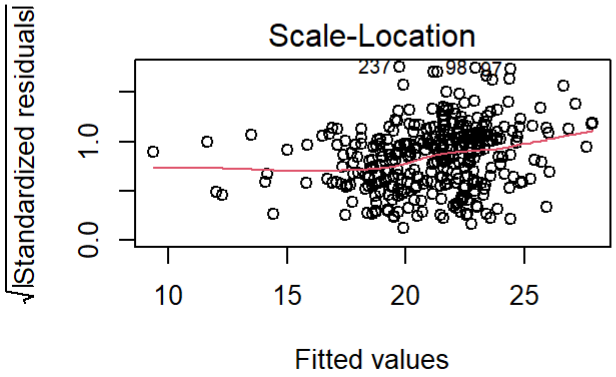
#model plot

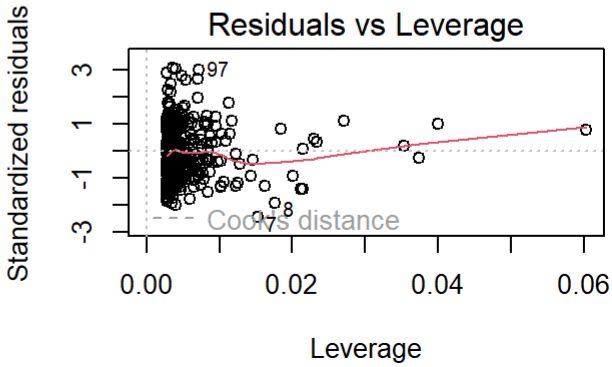
```
par(mfrow=c(2,2)) # split the graphics device into 4 panels
plot(modl0) # (uses plot.lm as modl0 is class 'lm')
```




#AIC




```
AIC(modl0)
```

```
## [1] 2590.695
```

#model selection ##lm modl1

```
modl1 <- lm(acc_total ~
              SPEED_VIOLATIONS, data=df)
summary(modl1 )
```

```
##
## Call:
## lm(formula = acc_total ~ SPEED_VIOLATIONS, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.3461  -4.5460  -0.8298   4.6666  22.8627
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.580e-01  1.311e+00   0.731    0.466
## SPEED_VIOLATIONS 1.859e-04  1.162e-05  15.992   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.821 on 361 degrees of freedom
## Multiple R-squared:  0.4147, Adjusted R-squared:  0.4131
## F-statistic: 255.8 on 1 and 361 DF,  p-value: < 2.2e-16
```

##AIC FOR TWO MODEL

```
AIC(modl1,modl0)
```

```
##       df      AIC
## modl1  3 2428.082
## modl0  3 2590.695
```

# Specifying an 80-20 train-test split

```
train_idx = createDataPartition(df$SPEED_VIOLATIONS, p = .8, list = F)
train = df[train_idx, ]
test = df[-train_idx, ]
print(head(test))
```

```
##         JDATE                          DESCRIPTION DISTANCE_VIOLATIONS
## 1  1398-01-01                106390              جشن نوروز/جشن سال نو
## 4  1398-01-04                                122446              عیدنوروز
## 8  1398-01-08                                                    120339
## 19 1398-01-19          83311                فروردین روز،جشن فروردینگان
## 22 1398-01-22 98278               ولادت امام زین العابدین علیه السلام
## 24 1398-01-24                                                    72757
##    HEAVY_VEHICLES IS_HOLIDAY JDAY JMONTH JYEAR SPEED_VIOLATIONS            TAG
## 1           61026          1    1      1  1398            97782 تعطیلي بلند مدت
## 4           78040          1    4      1  1398           104569 تعطیلي بلند مدت
## 8          101717          0    8      1  1398            95557           عادي
## 19         212501          0   19      1  1398            58462           عادي
## 22         201590          0   22      1  1398            71847           عادي
## 24         195500          0   24      1  1398            55354           عادي
##    TOTAL_VEHICLES acc_total avg_class1_speed avg_class2_speed avg_class3_speed
## 1         1011333        16         88.85547         88.04384         81.53555
## 4         1206069        16         87.94667         85.78571         80.79592
## 8         1161622        11         87.94379         83.01725         78.83649
## 19         828836        10         85.61723         75.30593         73.89143
## 22         923445        12         86.91019         76.89606         74.39518
## 24         819724         8         86.78557         77.17442         74.39028
##    avg_class4_speed avg_class5_speed avg_spedd85_class1 avg_spedd85_class2
## 1          87.51029         83.74037           95.56849           92.33562
## 4          86.06595         82.23939           95.56849           92.33562
## 8          83.38092         79.16589           95.56849           92.33562
## 19         77.21821         74.19520           95.56849           92.33562
## 22         77.77529         74.48671           95.56849           92.33562
## 24         77.96785         75.14482           95.56849           92.33562
##    avg_spedd85_class3 avg_spedd85_class4 avg_spedd85_class5  class1
## 1            89.21918           96.21233           86.28767  936897
## 4            89.21918           96.21233           86.28767 1050770
## 8            89.21918           96.21233           86.28767 1033487
## 19           89.21918           96.21233           86.28767  613624
## 22           89.21918           96.21233           86.28767  719423
## 24           89.21918           96.21233           86.28767  564010
##    class1_speed class2 class2_speed class3 class3_speed class4 class4_speed
## 1         89690  30549         6669   8404          446   8692          207
## 4         94255  39939         8748  10764          530  10121          239
## 8         85674  47555         8408  14937          525  13248          221
## 19        49909  73151         5373  42181         1099  27737          530
## 22        62543  70381         6212  37922          878  25938          450
## 24        47322  67701         5395  37147          709  26225          377
##    class5 class5_speed police_enforcment
## 1   13381          770              4377
## 4   17216          797              5196
## 8   25977          729              5418
## 19  69432         1551              2399
## 22  67349         1764              2861
## 24  64427         1551              2942
```

# Declaring the trainControl function

```
train_ctrl = trainControl(
  method  = "cv", #Specifying Cross validation
  number  = 2, # Specifying 5-fold
  verboseIter = TRUE, # So that each iteration you get an update of the progress
  classProbs = TRUE # So that you can obtain the probabilities for each example
)
rf_model = train(
  acc_total ~SPEED_VIOLATIONS, # Specifying the response variable and the feature variables
  method = "rf", # Specifying the model to use
  data = train,
  trControl = train_ctrl
)
```

```
## Warning in train.default(x, y, weights = w, ...): cannnot compute class
## probabilities for regression
```

```
## + Fold1: mtry=2
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range
```

```
## - Fold1: mtry=2
## + Fold2: mtry=2
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range
```

```
## - Fold2: mtry=2
## Aggregating results
## Fitting final model on full training set
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range
```

# Get the predictions of your model in the test set

```
predictions = predict(rf_model, newdata = test)
print(head(predictions))
```

```
##        1        4        8       19       22       24
## 20.64730 16.84053 10.69817 11.95203 20.23193 11.34197
```

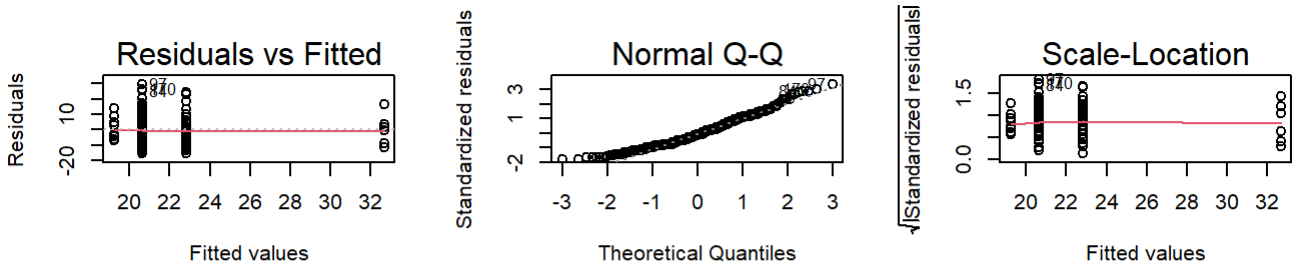
# factor variables

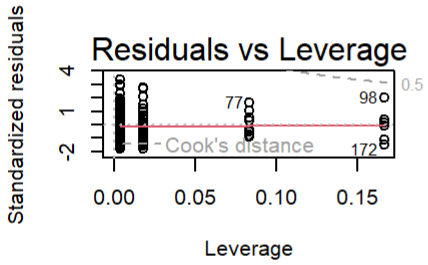
Notice how R reports an intercept parameter and parameters for the two treatment levels, but, in order to obtain an identifiable model, it has not included a parameter for the control level of the group factor. TAG is a factor variable in my data.

```
df$TAG <- as.factor(df$TAG)
fit <- lm(acc_total ~ TAG , data=df)
summary(fit)
```

```
##
## Call:
## lm(formula = acc_total ~ TAG, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.8214  -6.6471  -0.8214   5.7658  29.3529
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           22.821      1.172  19.468  < 2e-16 ***
## TAG0.20144  1.280-  2.791      3.571-    تعطيلي بلند مدت
## TAG0.00936  2.613   3.768      9.845    تعطيلي كوتاه مدت **
## TAG0.09045  1.698-  1.281      2.174-               عادي .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.773 on 359 degrees of freedom
## Multiple R-squared:  0.03724,    Adjusted R-squared:  0.02919
## F-statistic: 4.628 on 3 and 359 DF,  p-value: 0.003434
```

```
par(mfrow=c(3,3))
plot(fit) # Then R will show you four diagnostic plots one by one:1. Residuals vs Fitted,2. Norm
al Q-Q,3. Scale-Location,4. Residuals vs Leverage
```




##

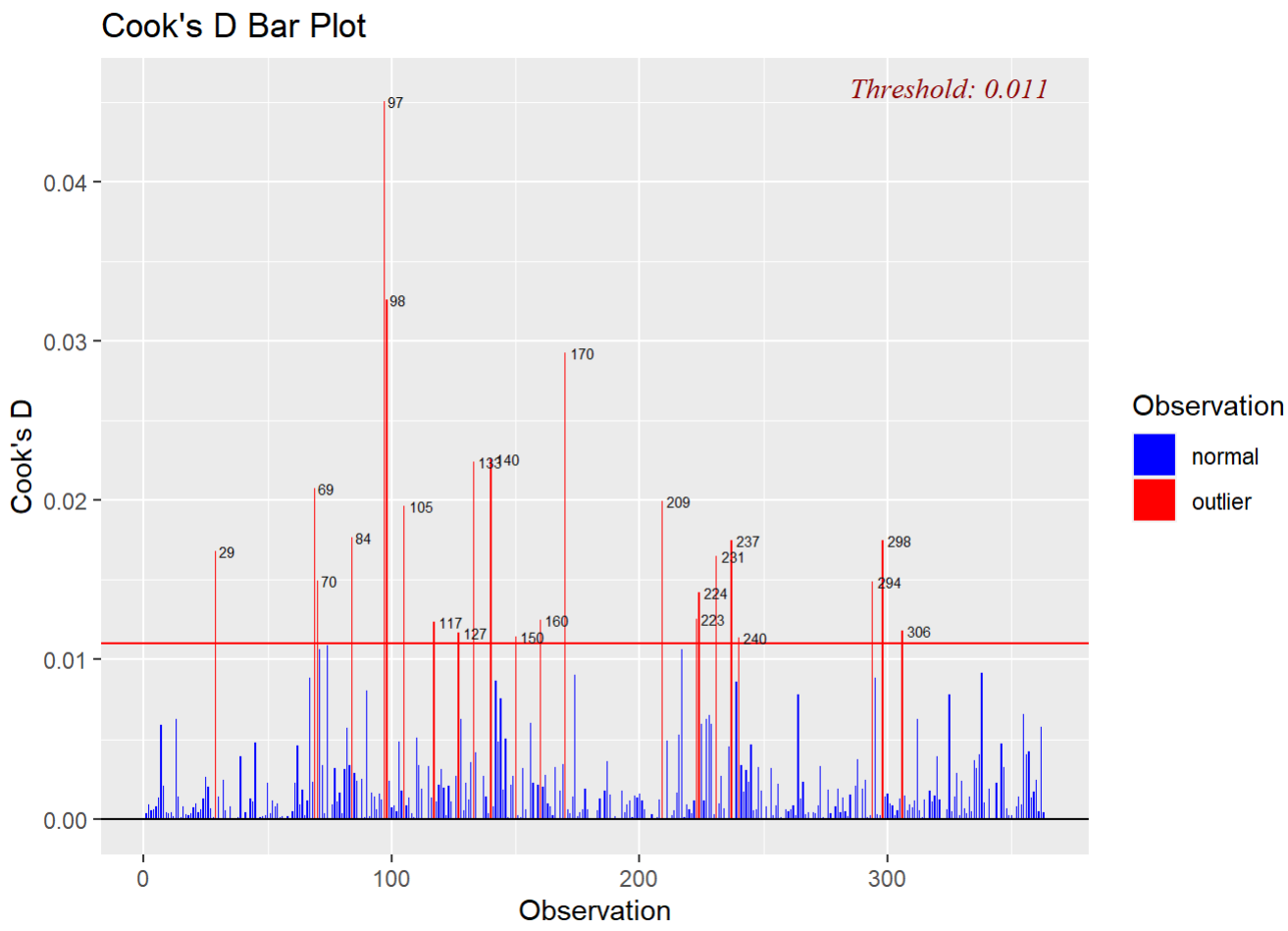
Measures of Influence

```
library(olsrr)
```

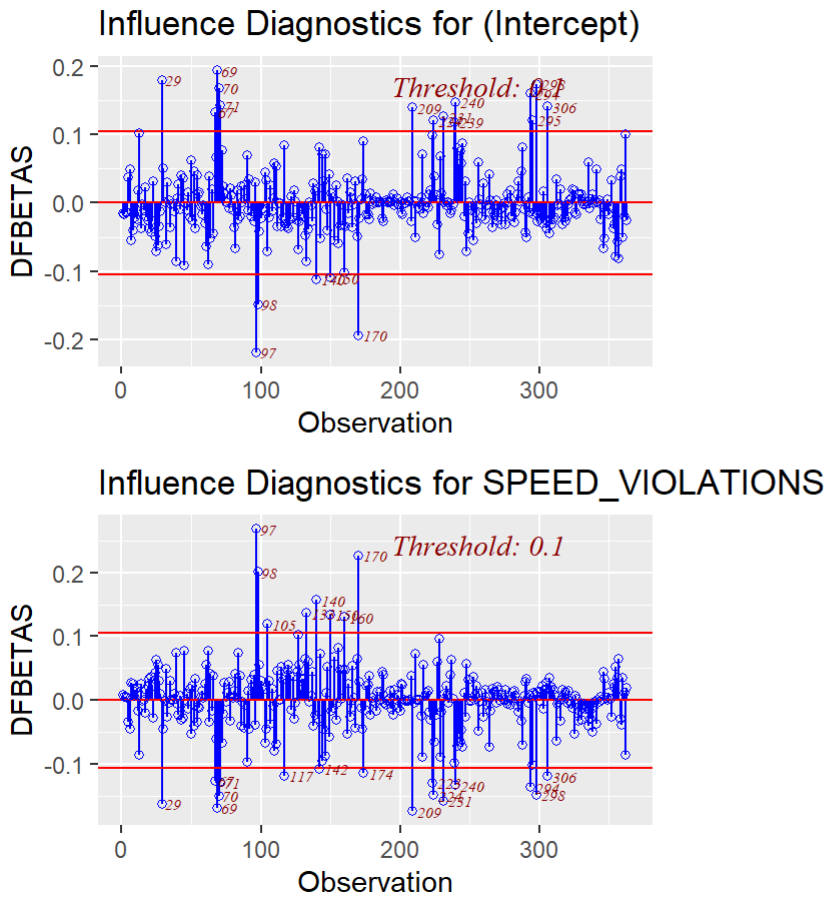
```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
par(mfrow=c(1,6))
ols_plot_cooksd_bar(modl1)
```



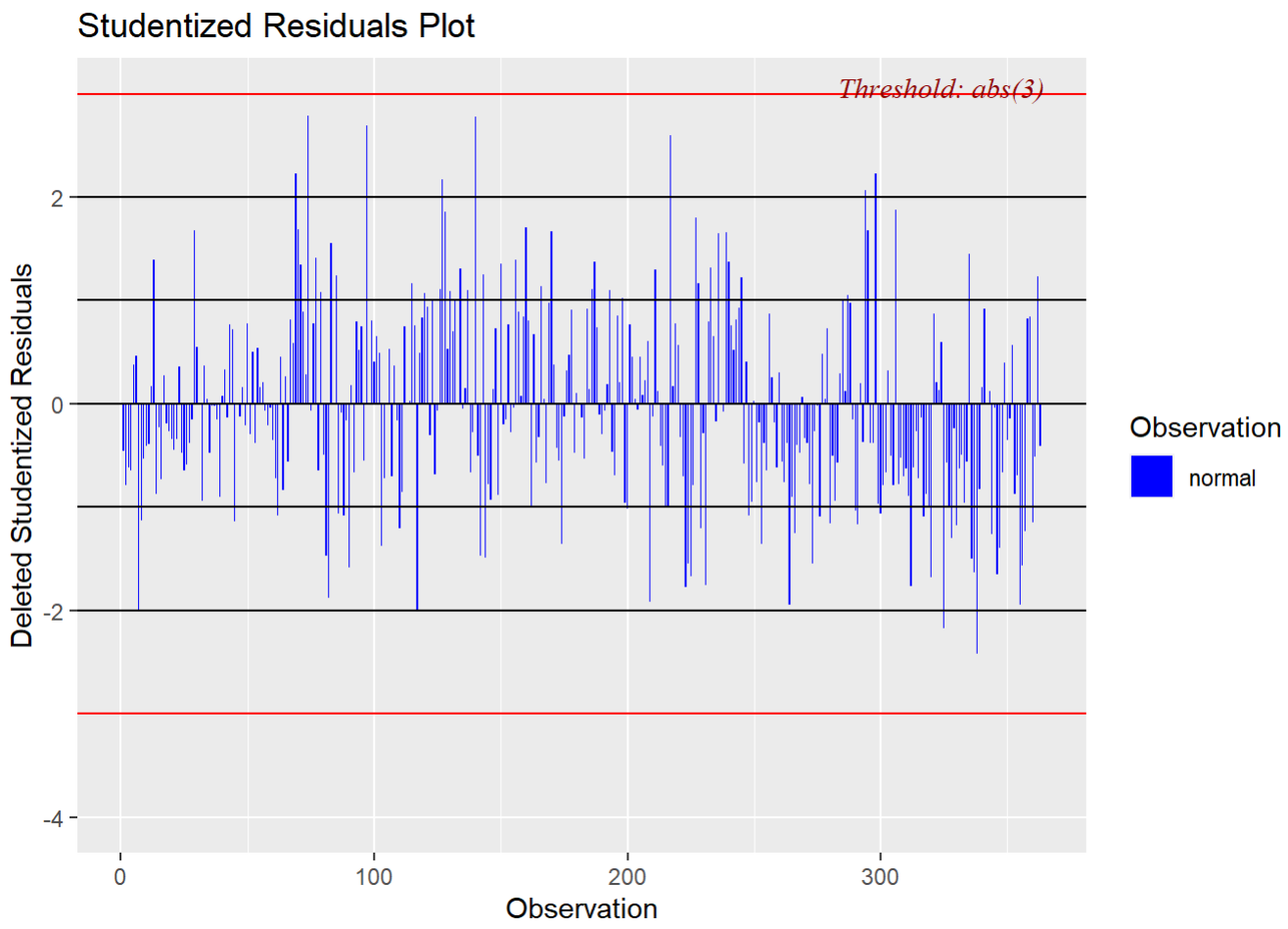

```
ols_plot_dfbetas(modl1)
```



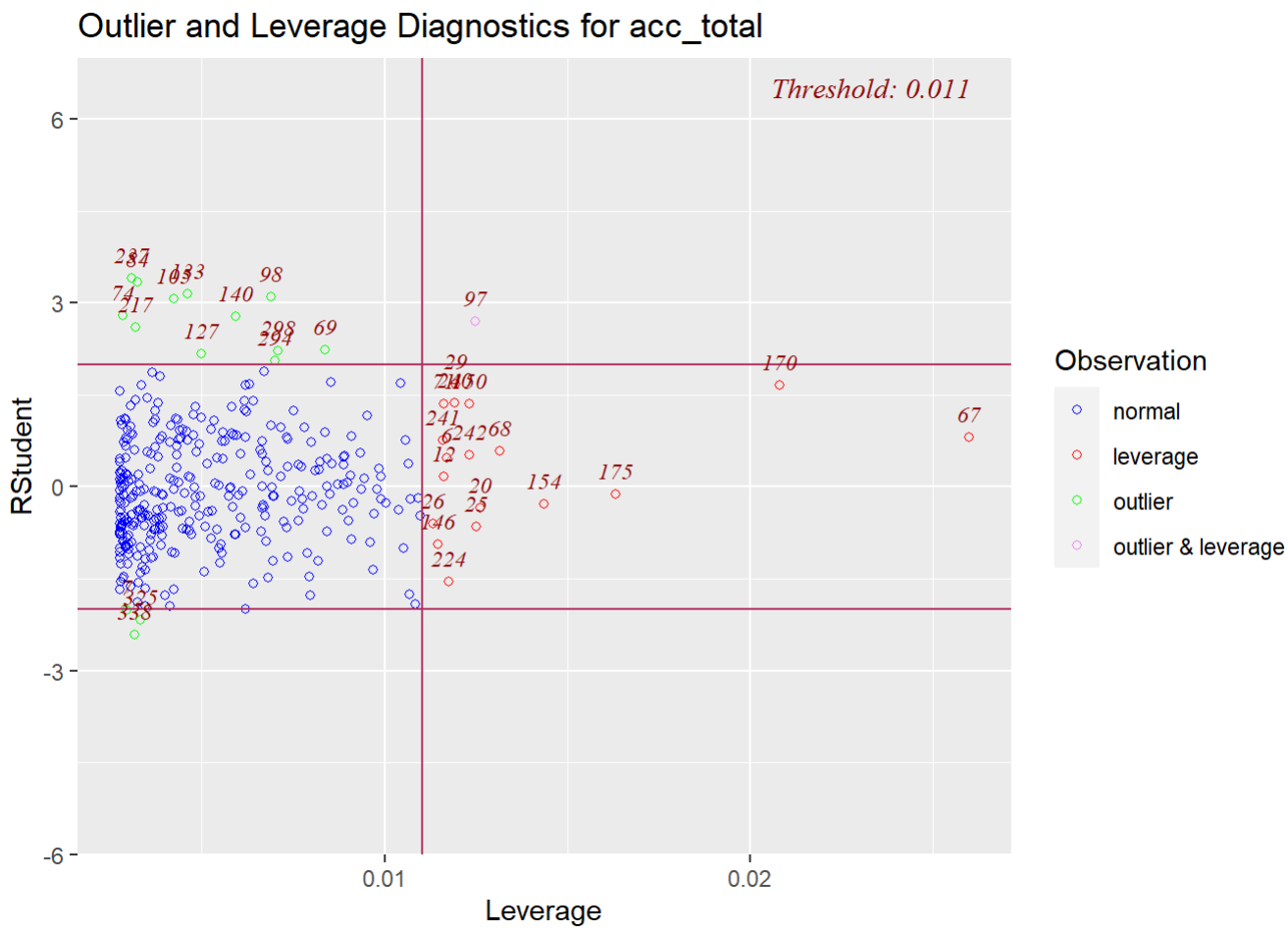




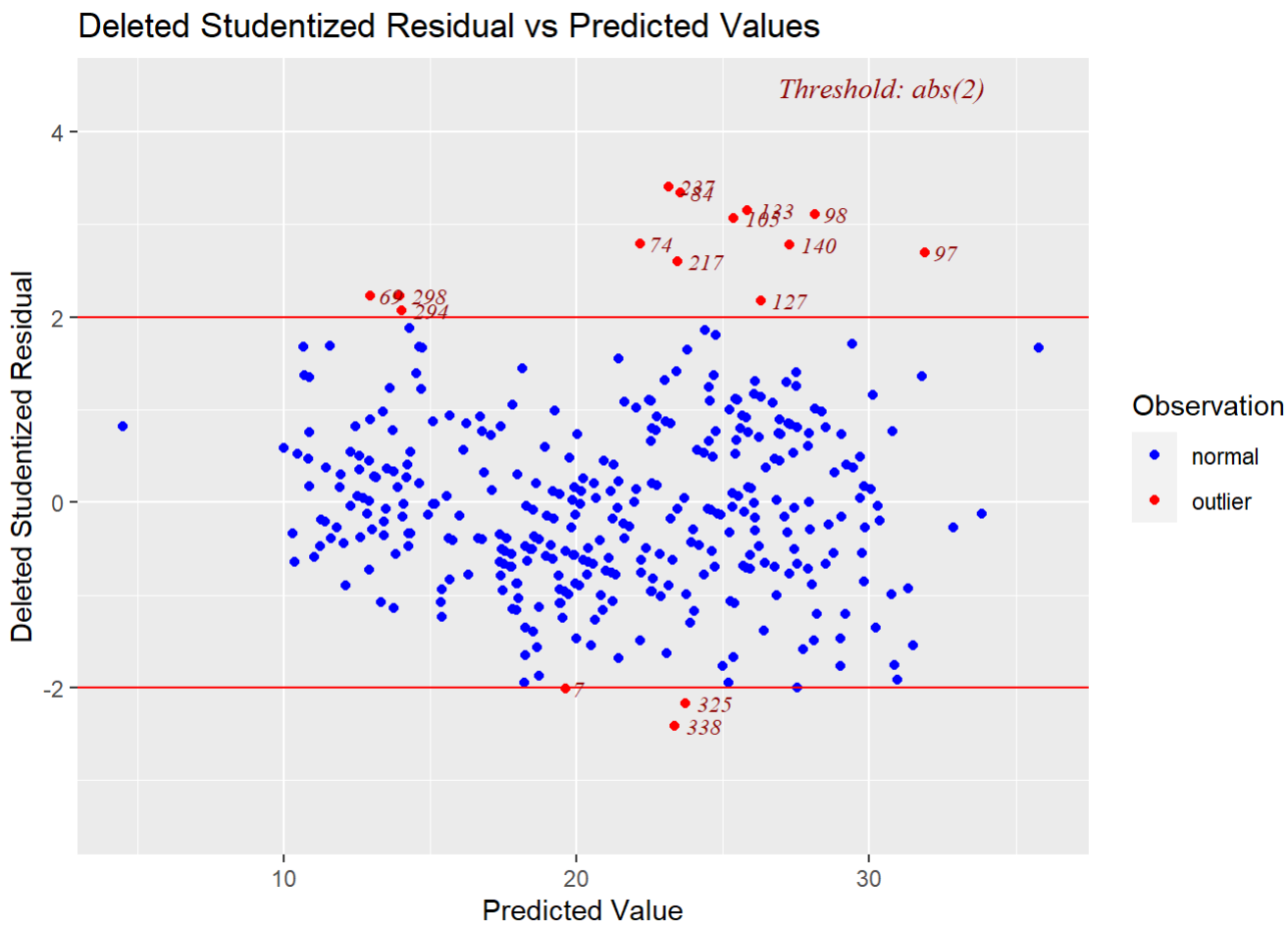
```
ols_plot_resid_stud(modl1)
```



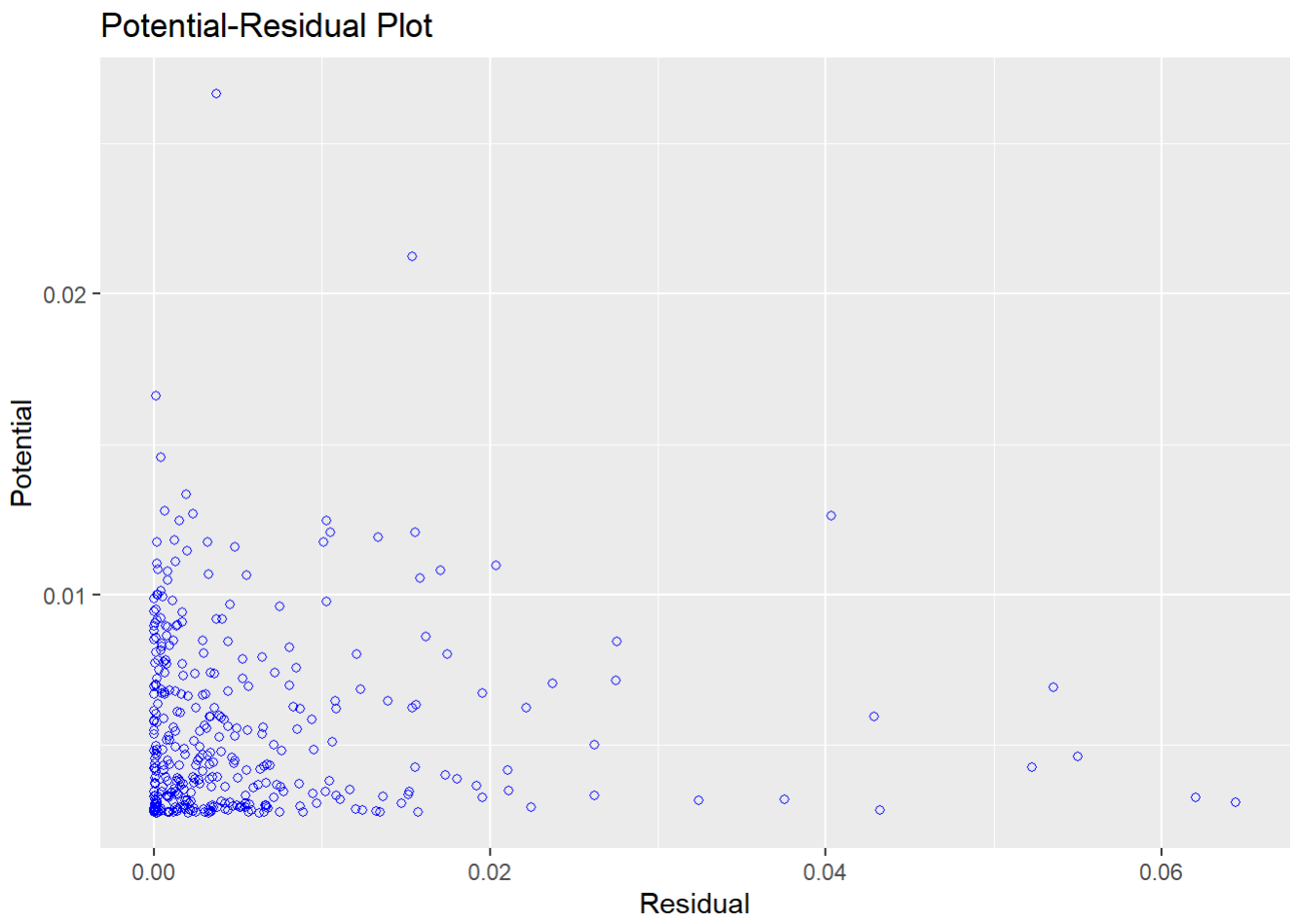

```
ols_plot_resid_lev(modl1)
```




```
ols_plot_resid_stud_fit(modl1)
```




```
ols_plot_resid_pot(modl1)
```




#confint

```
confint(modl1, 'SPEED_VIOLATIONS', level=0.95)
```

```
##                        2.5 %       97.5 %
## SPEED_VIOLATIONS 0.00016303 0.0002087466
```