supervised_analysis

Carlos_Mackenzie

2024-05-08

Summary DF

```
##    cylinders      displacement     horsepower        weight      acceleration
##  Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613   Min.   : 8.00
##  1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225   1st Qu.:13.78
##  Median :4.000   Median :151.0   Median : 93.5   Median :2804   Median :15.50
##  Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978   Mean   :15.54
##  3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615   3rd Qu.:17.02
##  Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140   Max.   :24.80
##    model_year        origin          mpg
##  Min.   :70.00   Min.   :1.000   Min.   : 9.00
##  1st Qu.:73.00   1st Qu.:1.000   1st Qu.:17.00
##  Median :76.00   Median :1.000   Median :22.75
##  Mean   :75.98   Mean   :1.577   Mean   :23.45
##  3rd Qu.:79.00   3rd Qu.:2.000   3rd Qu.:29.00
##  Max.   :82.00   Max.   :3.000   Max.   :46.60
```

Quality

```
## [1] "El número de registros inconsistentes es la base de datos es 0"
```

```
## [1] "El porcentaje de registros inconsistentes es la base de datos es 0"
```

```
##    cylinders      displacement     horsepower        weight      acceleration
##  Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613   Min.   : 8.00
##  1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225   1st Qu.:13.78
##  Median :4.000   Median :151.0   Median : 93.5   Median :2804   Median :15.50
##  Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978   Mean   :15.54
##  3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615   3rd Qu.:17.02
##  Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140   Max.   :24.80
##    model_year        origin          mpg
##  Min.   :70.00   Min.   :1.000   Min.   : 9.00
##  1st Qu.:73.00   1st Qu.:1.000   1st Qu.:17.00
##  Median :76.00   Median :1.000   Median :22.75
##  Mean   :75.98   Mean   :1.577   Mean   :23.45
##  3rd Qu.:79.00   3rd Qu.:2.000   3rd Qu.:29.00
##  Max.   :82.00   Max.   :3.000   Max.   :46.60
```

Dependent Variable Analysis

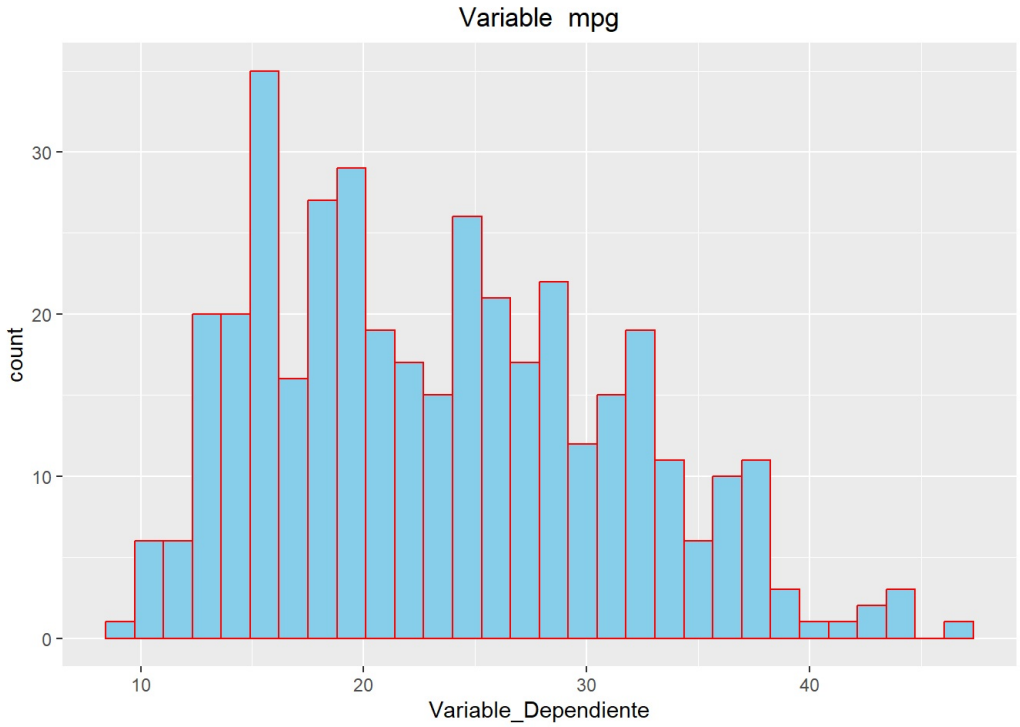
```
## [1] TRUE
```

```
## [1] "numeric"
```

```
## [1] FALSE
```

```
## [1] FALSE
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   17.00   22.75   23.45   29.00   46.60
```

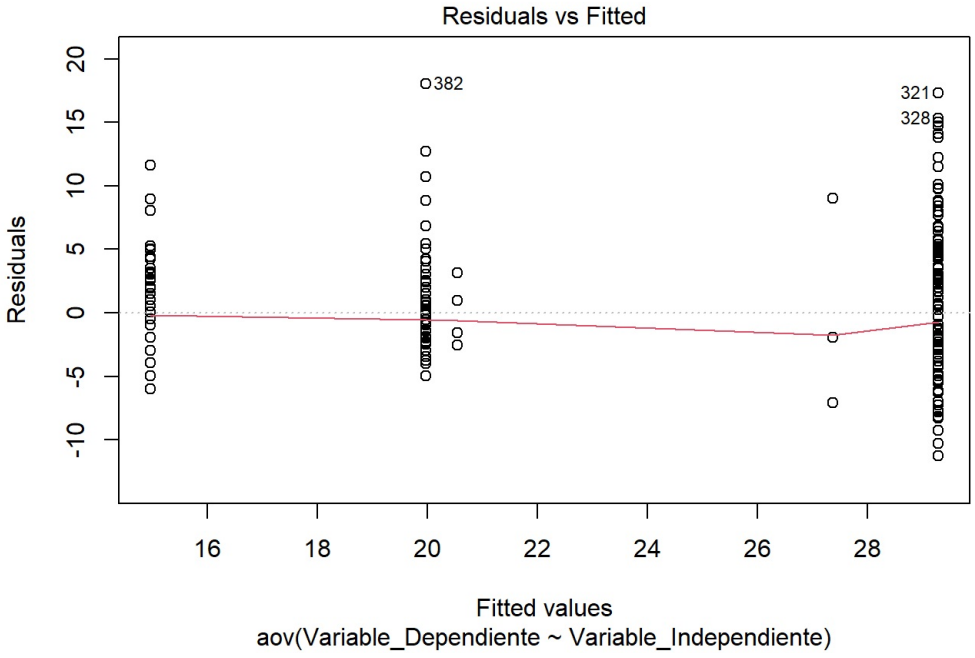
```
## [1] "La variable independiente a evaluar es del tipo"
```

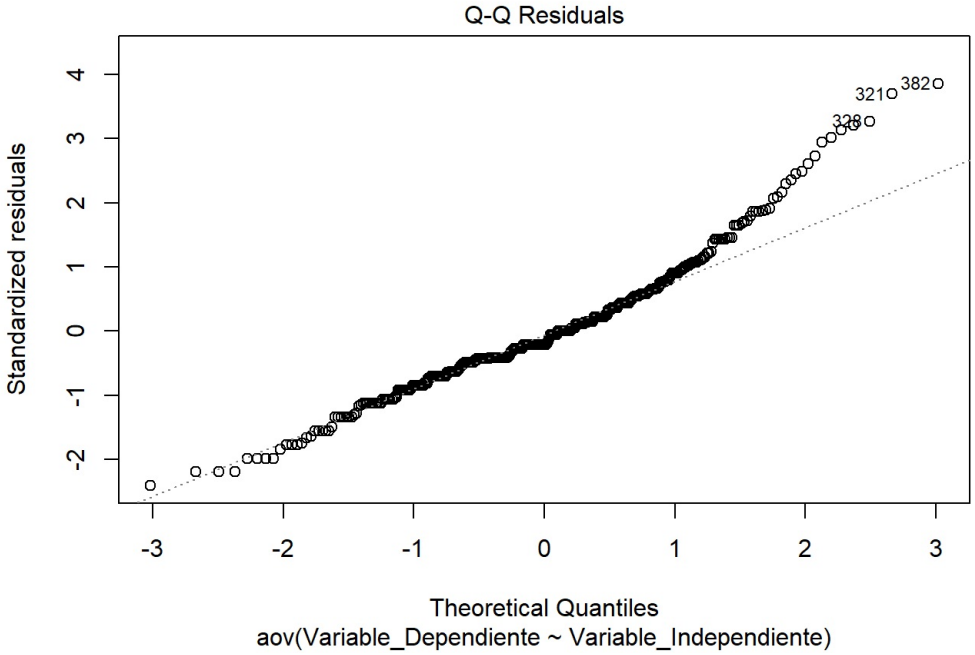
```
## [1] "numerica"
```

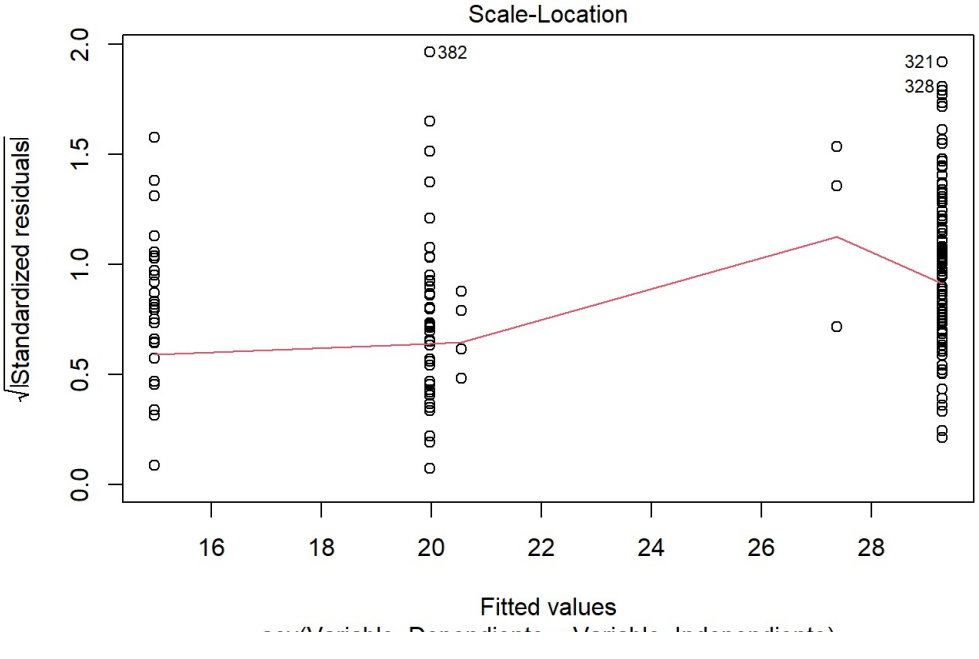
Independent Variable Analysis

```
## [1] "Revision Variable Dependiente  mpg  junto con las variables inpendientes"
```

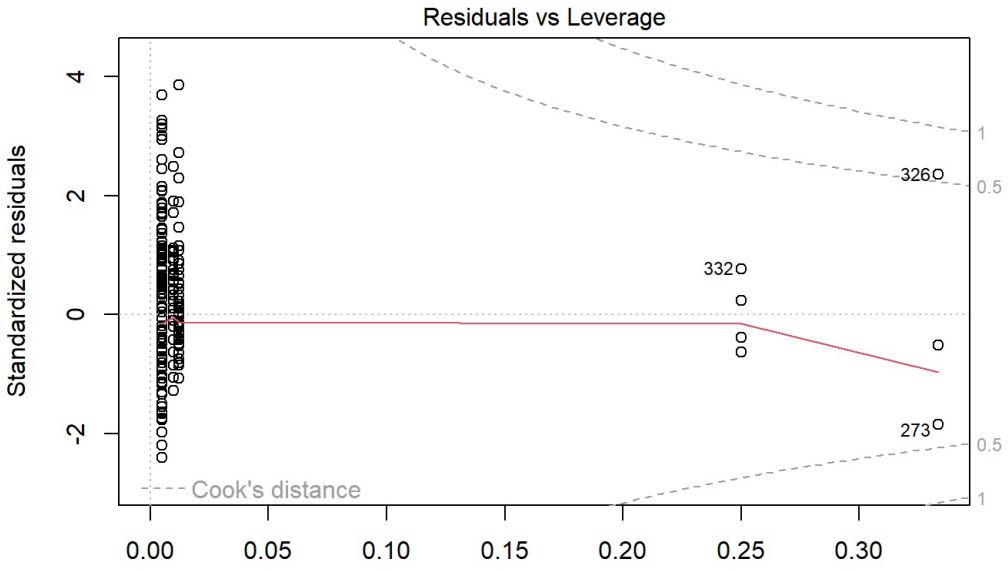
```
##                         Df Sum Sq Mean Sq F value Pr(>F)
## Variable_Independiente   4  15275    3819     173 <2e-16 ***
## Residuals              387   8544      22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Residuals vs Fitted
Residuals
Fitted values
aov(Variable_Dependiente ~ Variable_Independiente)
382
321
328

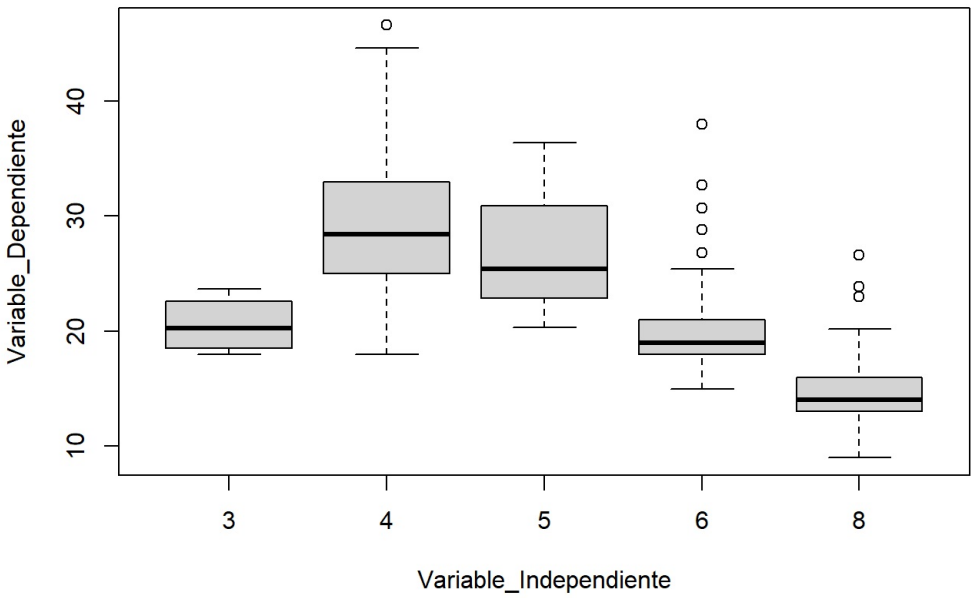

Q-Q Residuals
Standardized residuals
Theoretical Quantiles
aov(Variable_Dependiente ~ Variable_Independiente)
382
321
328


Scale-Location
√|Standardized residuals|
Fitted values
382
321
328

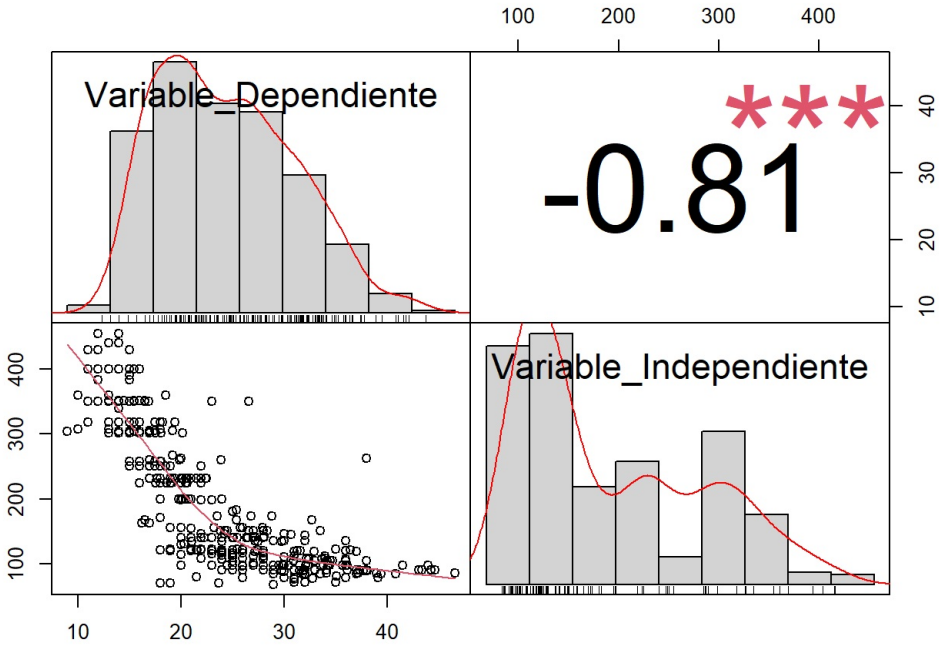
aov(Variable_Dependiente ~ Variable_Independiente)



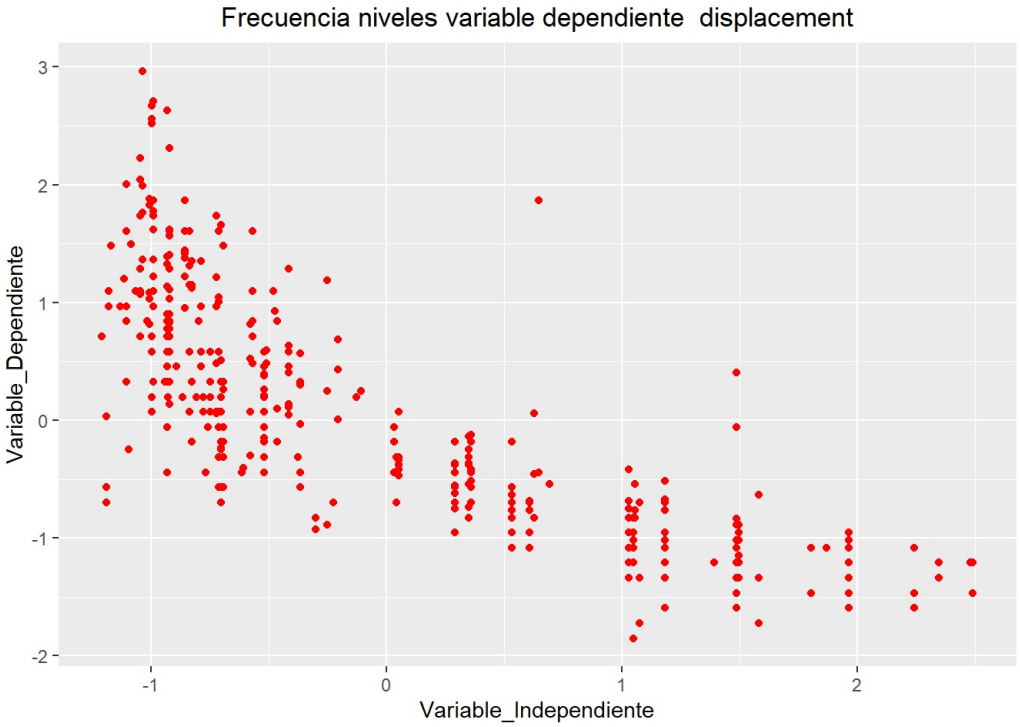
aov(Variable_Dependiente ~ Variable_Independiente)

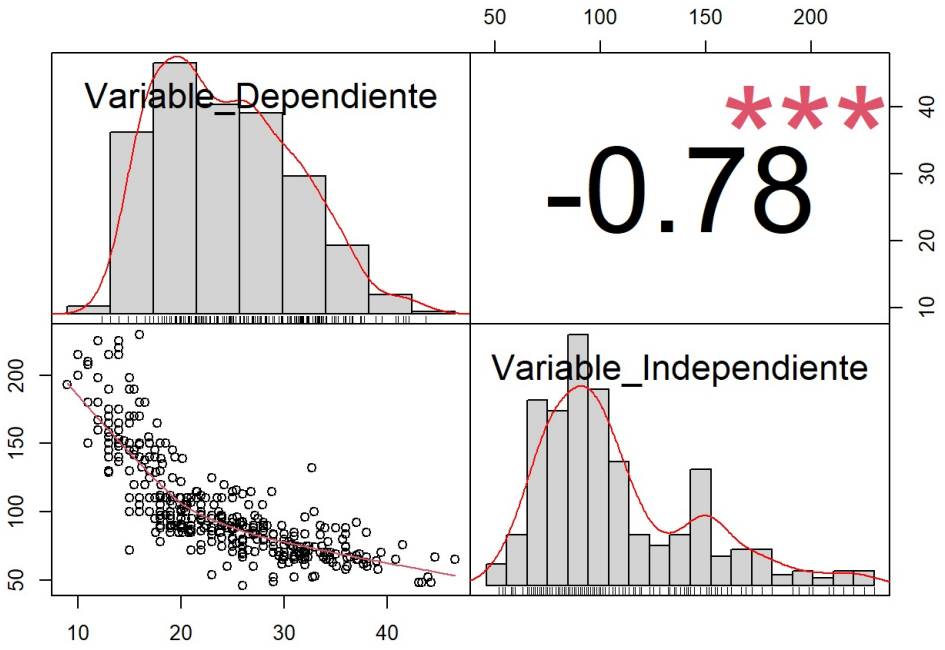


```
## $stats
##       [,1]  [,2]  [,3] [,4] [,5]
## [1,] 18.00 18.00 20.30 15.0  9.0
## [2,] 18.50 25.00 22.85 18.0 13.0
## [3,] 20.25 28.40 25.40 19.0 14.0
## [4,] 22.60 32.95 30.90 21.0 16.0
## [5,] 23.70 44.60 36.40 25.4 20.2
##
## $n
## [1]   4 199   3  83 103
##
## $conf
##        [,1]     [,2]     [,3]     [,4]     [,5]
## [1,] 17.011 27.50957 18.05668 18.47972 13.53295
## [2,] 23.489 29.29043 32.74332 19.52028 14.46705
##
## $out
## [1] 46.6 28.8 26.8 32.7 30.7 38.0 23.0 23.9 26.6
##
## $group
## [1] 2 4 4 4 4 4 5 5 5
##
## $names
## [1] "3" "4" "5" "6" "8"
```

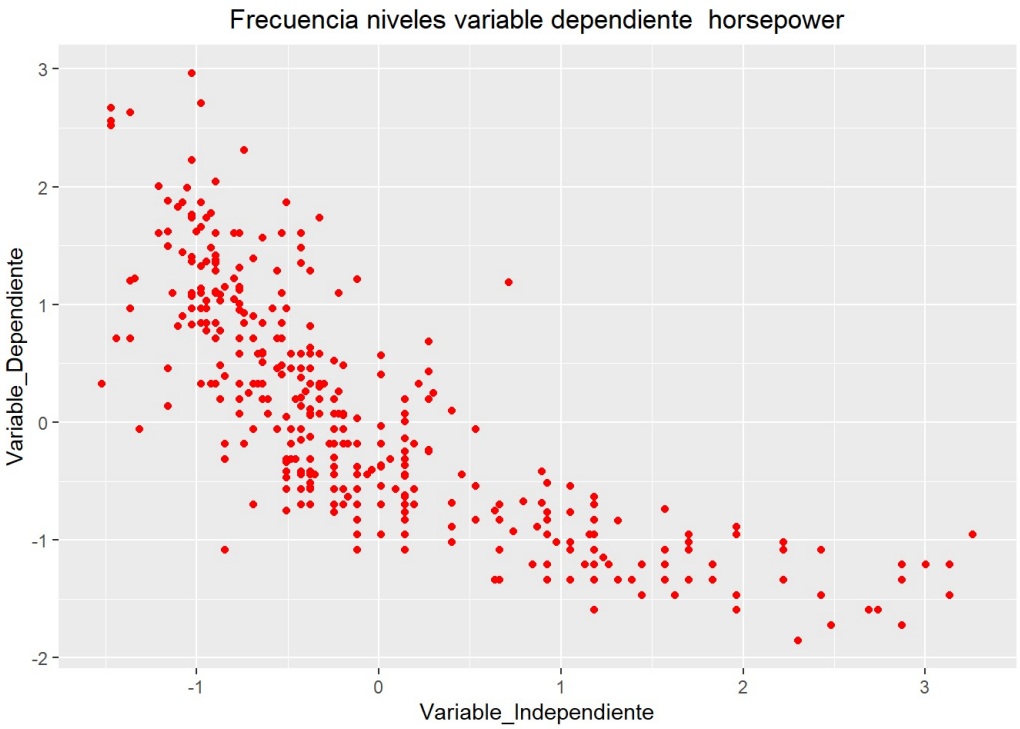


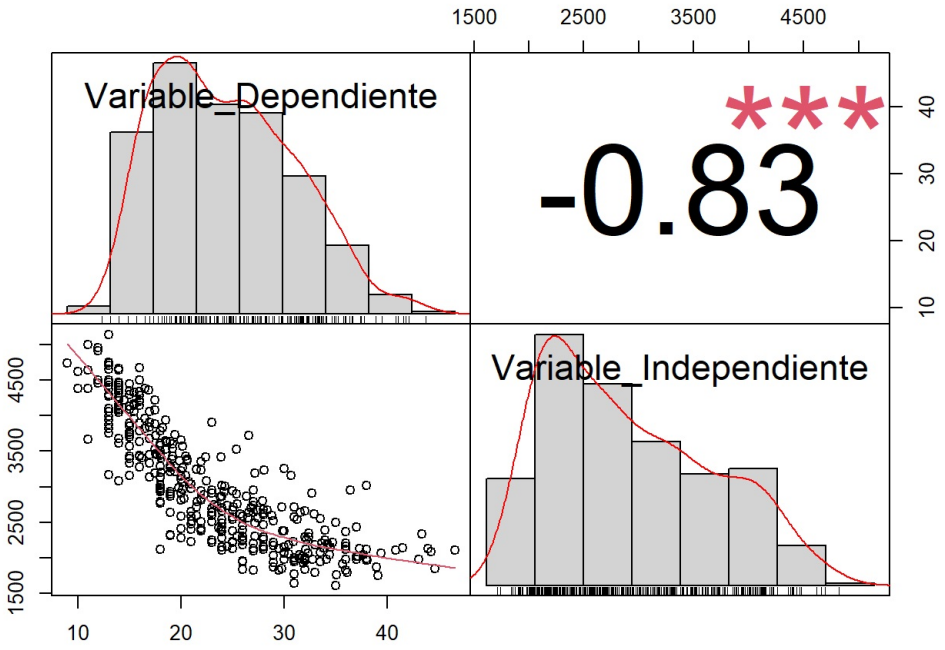

```
## NULL
```


Frecuencia niveles variable dependiente displacement
Variable_Dependiente
Variable_Independiente
3
2
1
0
-1
-2
-1
0
1
2

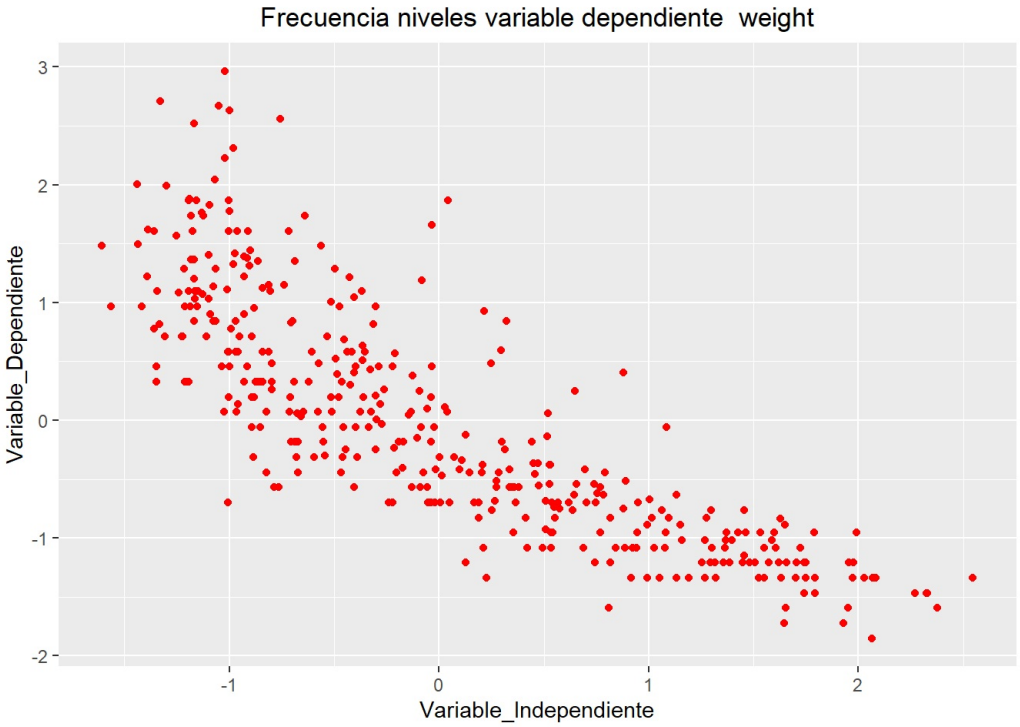

Variable_Dependiente
-0.78
***
Variable_Independiente
50
100
150
200
10
20
30
40

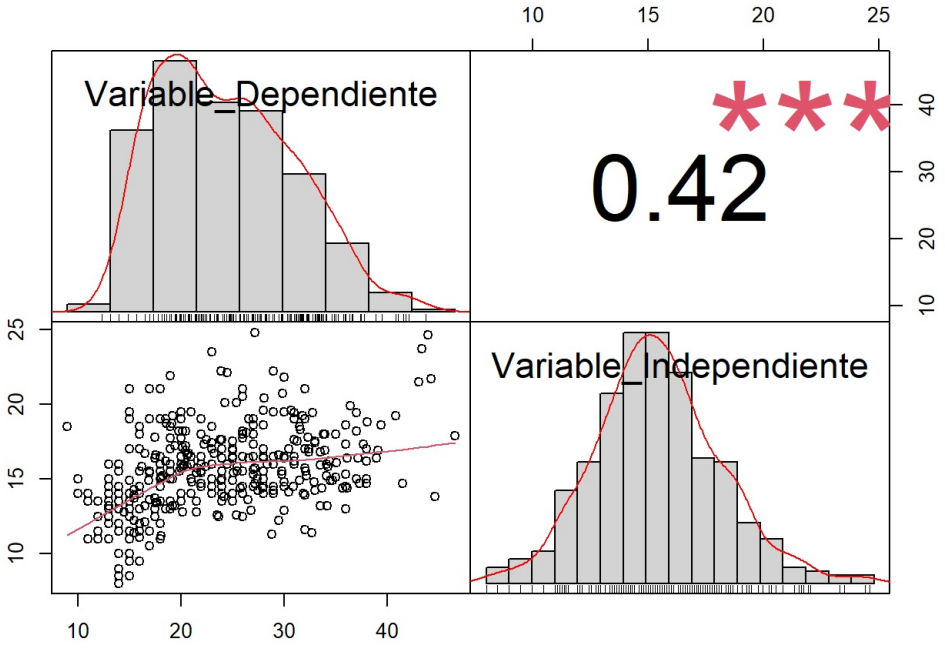
```
## NULL
```


Frecuencia niveles variable dependiente horsepower
Variable_Dependiente
Variable_Independiente
3
2
1
0
-1
-2
-1
0
1
2
3

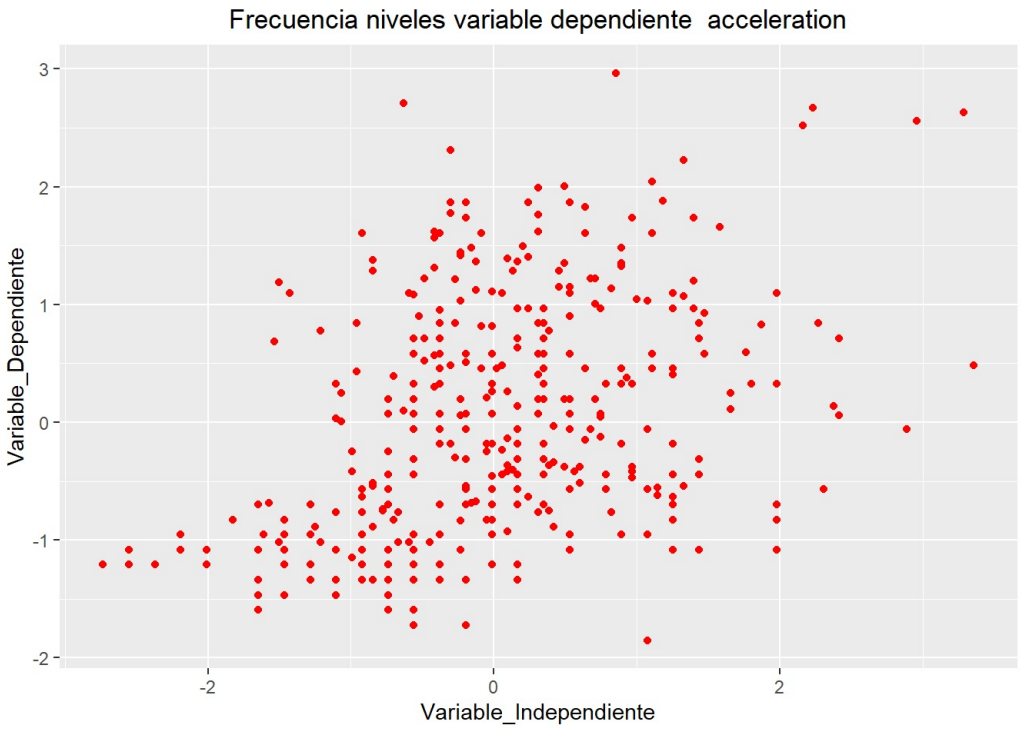

Variable_Dependiente
Variable_Independiente
-0.83
***
1500
2500
3500
4500
10
20
30
40

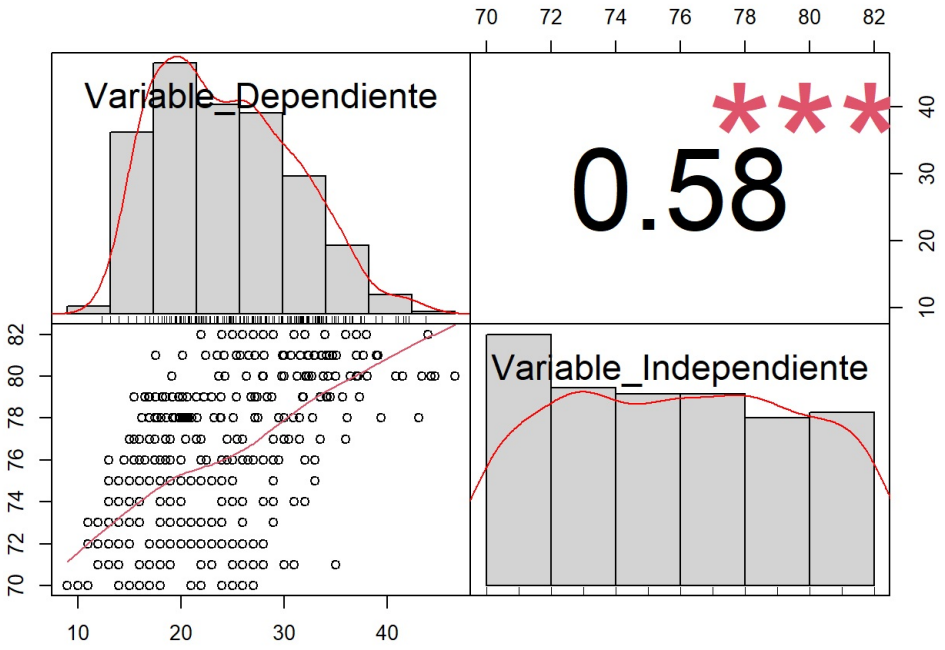
```
## NULL
```


Frecuencia niveles variable dependiente weight
Variable_Dependiente
Variable_Independiente
3
2
1
0
-1
-2
-1
0
1
2

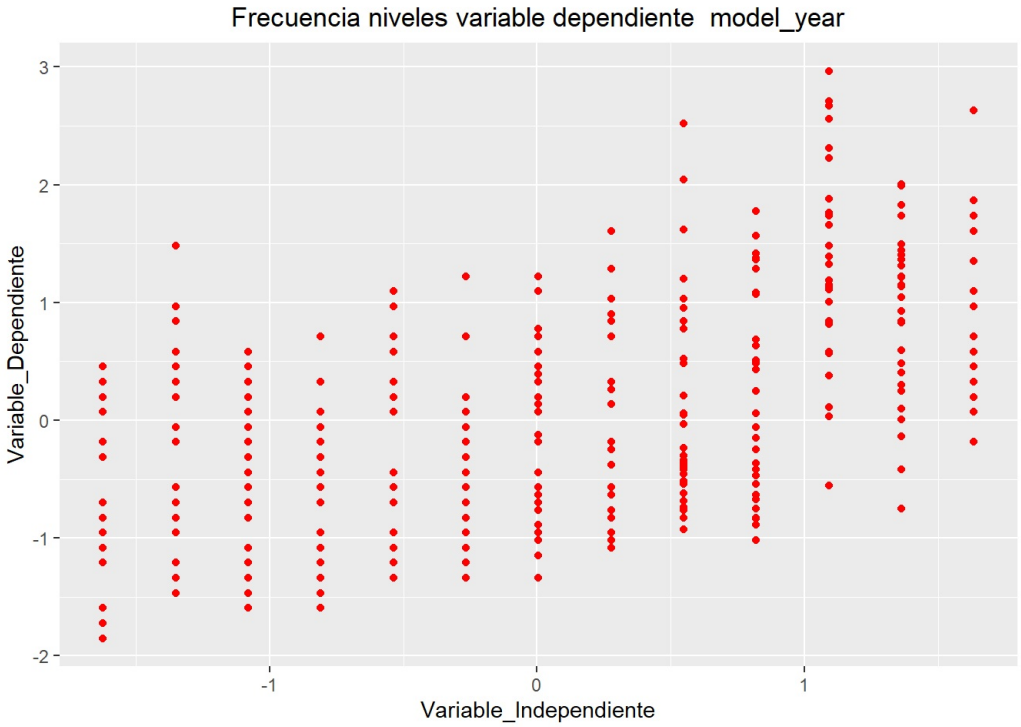

Variable_Dependiente
0.42
***
Variable_Independiente
10
15
20
25
10
20
30
40

```
## NULL
```

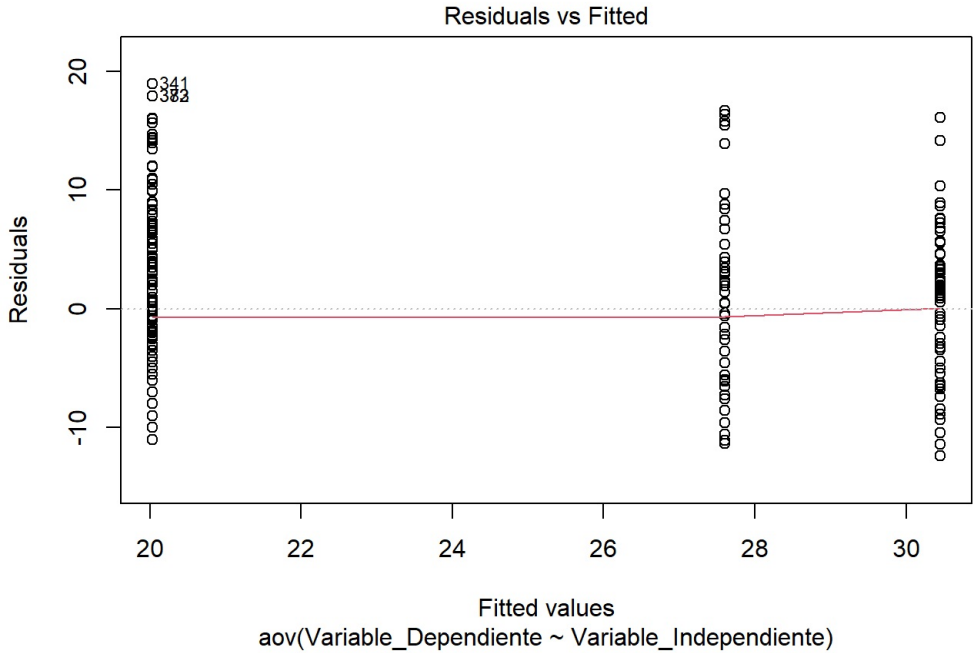

Frecuencia niveles variable dependiente acceleration
Variable_Dependiente
Variable_Independiente
3
2
1
0
-1
-2
-2
0
2

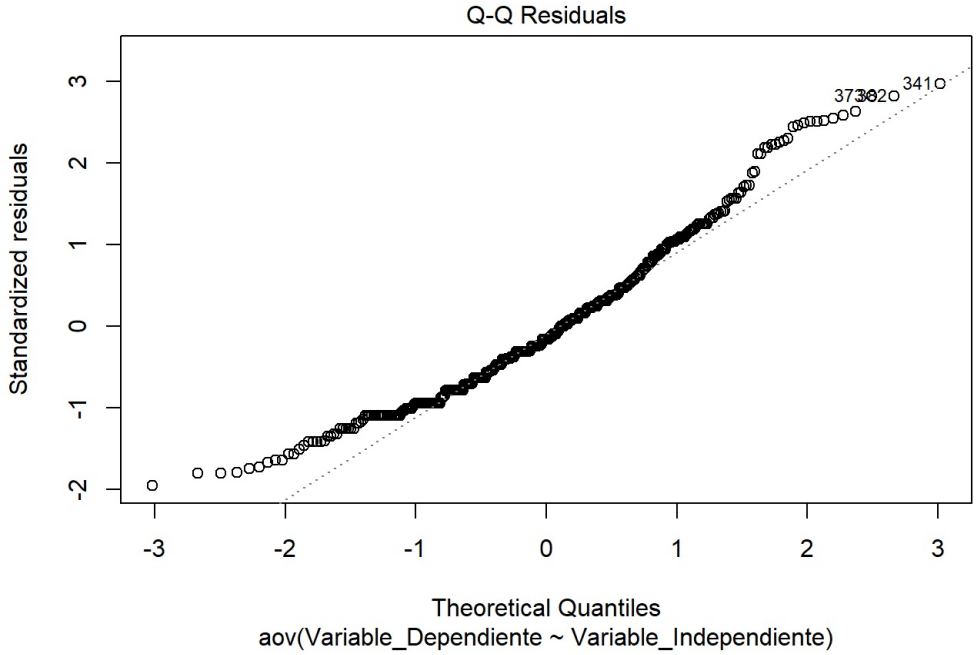

Variable_Dependiente
0.58
***
Variable_Independiente
70
72
74
76
78
80
82
10
20
30
40

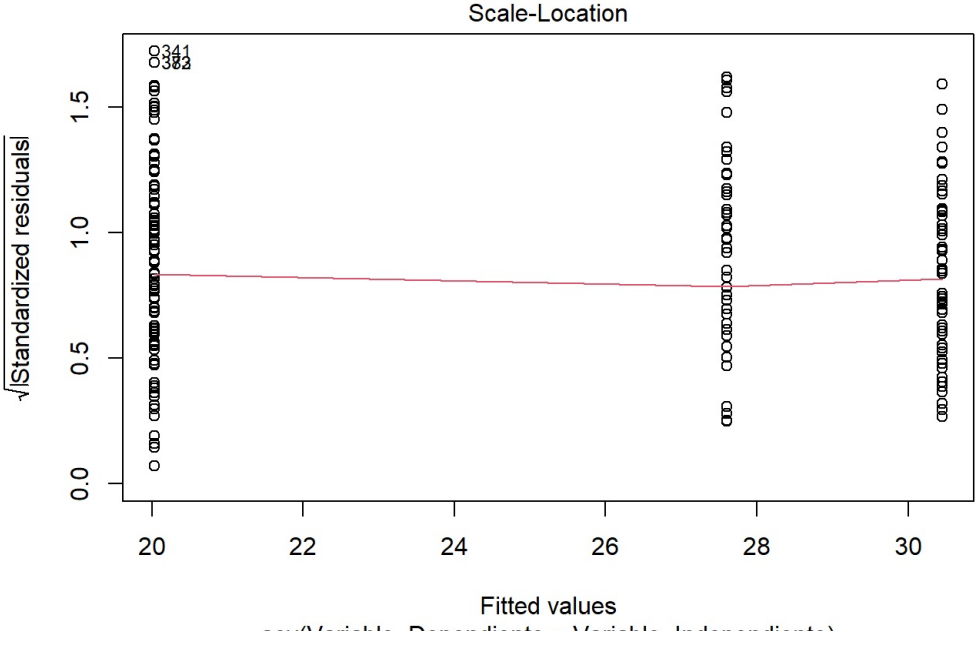
```
## NULL
```



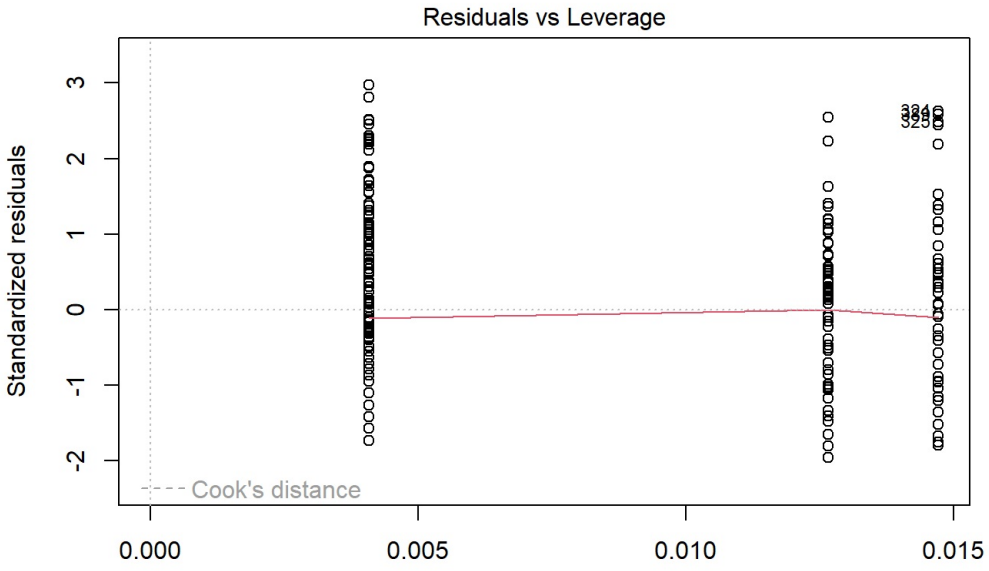

```
##                          Df Sum Sq Mean Sq F value Pr(>F)
## Variable_Independiente    2   7904    3952    96.6 <2e-16 ***
## Residuals               389  15915      41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Residuals vs Fitted
Residuals
Fitted values
aov(Variable_Dependiente ~ Variable_Independiente)

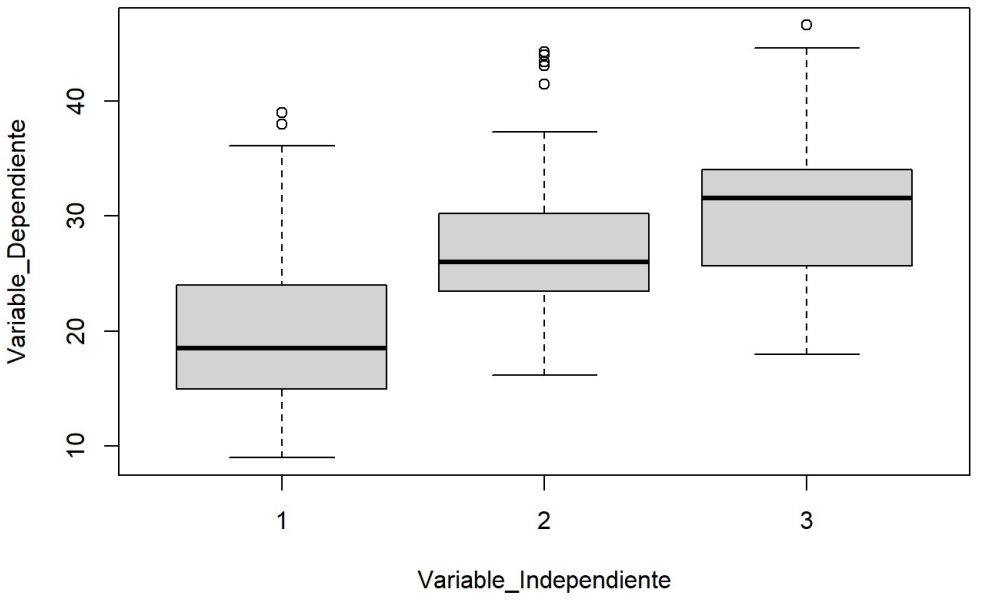

Q-Q Residuals
Standardized residuals
Theoretical Quantiles
aov(Variable_Dependiente ~ Variable_Independiente)


Scale-Location
√|Standardized residuals|
Fitted values

aov(Variable_Dependiente ~ Variable_Independiente)



aov(Variable_Dependiente ~ Variable_Independiente)



```
## $stats
##      [,1]  [,2]  [,3]
## [1,]  9.0 16.20 18.00
## [2,] 15.0 23.50 25.70
## [3,] 18.5 26.00 31.60
## [4,] 24.0 30.25 34.05
## [5,] 36.1 37.30 44.60
##
## $n
## [1] 245  68  79
##
## $conf
##          [,1]     [,2]     [,3]
## [1,] 17.59152 24.70668 30.11567
## [2,] 19.40848 27.29332 33.08433
##
## $out
## [1] 39.0 38.0 38.0 43.1 41.5 44.3 43.4 44.0 46.6
##
## $group
## [1] 1 1 1 2 2 2 2 2 3
##
## $names
## [1] "1" "2" "3"
```

```
##   num_variable       nombre     tipo Incluir
## 1            1    cylinders   factor      Si
## 2            2 displacement numerica      Si
## 3            3   horsepower numerica      Si
## 4            4       weight numerica      Si
## 5            5 acceleration numerica      Si
## 6            6   model_year numerica      Si
## 7            7       origin   factor      Si
```

```
##   num_variable       nombre     tipo Incluir
## 1            1    cylinders   factor      Si
## 2            2 displacement numerica      Si
## 3            3   horsepower numerica      Si
## 4            4       weight numerica      Si
## 5            5 acceleration numerica      Si
## 6            6   model_year numerica      Si
## 7            7       origin   factor      Si
```

```
## [1] "mpg"
```

```
## [1] "variables sin colinealidad: "
## [1] "cylinders"            "acceleration"         "model_year"
## [4] "origin"               "Variable_Dependiente"
```

Models

```
## Linear Regression
##
## 392 samples
##   4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (3 fold)
## Summary of sample sizes: 261, 261, 262
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   4.031343  0.7343273  3.064676
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
##   intercept     RMSE  Rsquared      MAE   RMSESD RsquaredSD     MAESD
## 1      TRUE 4.031343 0.7343273 3.064676 0.266246 0.03140484 0.1537732
```

```
##    user  system elapsed
##    0.10    0.00    0.29
```

```
## k-Nearest Neighbors
##
## 392 samples
##   4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (3 fold)
## Summary of sample sizes: 262, 262, 260
## Resampling results across tuning parameters:
##
##   k  RMSE      Rsquared   MAE
##   5  3.678185  0.7821011  2.754404
##   7  3.644010  0.7865681  2.739622
##   9  3.684857  0.7817904  2.801920
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 7.
```

```
##   k     RMSE  Rsquared      MAE    RMSESD RsquaredSD      MAESD
## 1 5 3.678185 0.7821011 2.754404 0.4215774 0.03945549 0.09439188
## 2 7 3.644010 0.7865681 2.739622 0.2161018 0.01797218 0.10515139
## 3 9 3.684857 0.7817904 2.801920 0.2029236 0.01460103 0.09182213
```

```
##    user  system elapsed
##    0.16    0.00    0.21
```

```
## k-Nearest Neighbors
##
## 392 samples
##   4 predictor
##
## Pre-processing: centered (4), scaled (4)
## Resampling: Cross-Validated (3 fold)
## Summary of sample sizes: 262, 262, 260
## Resampling results across tuning parameters:
##
##   k  RMSE      Rsquared   MAE
##   5  4.175114  0.7261802  2.952598
##   7  4.078585  0.7376388  2.883442
##   9  4.024494  0.7457379  2.841101
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 9.
```

```
##   k     RMSE  Rsquared      MAE    RMSESD RsquaredSD     MAESD
## 1 5 4.175114 0.7261802 2.952598 0.4789630 0.05921917 0.3809556
## 2 7 4.078585 0.7376388 2.883442 0.5025602 0.05591307 0.4387239
## 3 9 4.024494 0.7457379 2.841101 0.4285189 0.04854955 0.3944046
```

```
##    user  system elapsed
##    0.15    0.00    0.23
```

```
## k-Nearest Neighbors
##
## 392 samples
##   4 predictor
##
## Pre-processing: centered (4), scaled (4)
## Resampling: Cross-Validated (3 fold)
## Summary of sample sizes: 261, 261, 262
## Resampling results across tuning parameters:
##
##   k    RMSE      Rsquared   MAE
##    10  3.812703  0.7636373  2.773160
##    20  3.709460  0.7778916  2.731727
##    50  4.047461  0.7477476  2.934400
##   100  4.715341  0.6940427  3.583003
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 20.
```

```
##     k     RMSE  Rsquared      MAE    RMSESD RsquaredSD     MAESD
## 1  10 3.812703 0.7636373 2.773160 0.1162774 0.01988567 0.1388443
## 2  20 3.709460 0.7778916 2.731727 0.1493090 0.01458970 0.1067199
## 3  50 4.047461 0.7477476 2.934400 0.2921118 0.01319650 0.2252628
## 4 100 4.715341 0.6940427 3.583003 0.3770373 0.02046333 0.2961538
```

```
##    user  system elapsed
##    0.17    0.00    0.25
```

```
## CART
##
## 392 samples
##   4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (3 fold)
## Summary of sample sizes: 261, 260, 263
## Resampling results across tuning parameters:
##
##   maxdepth  RMSE      Rsquared   MAE
##   1         5.145352  0.5649264  3.981206
##   2         4.669218  0.6436030  3.566033
##   3         3.866841  0.7545544  2.903941
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was maxdepth = 3.
```

```
##   maxdepth     RMSE  Rsquared      MAE    RMSESD RsquaredSD     MAESD
## 1        1 5.145352 0.5649264 3.981206 0.1657347 0.01992867 0.1418133
## 2        2 4.669218 0.6436030 3.566033 0.2587949 0.03683630 0.2921242
## 3        3 3.866841 0.7545544 2.903941 0.2328663 0.03482522 0.2462075
```

```
##    user  system elapsed
##     0.2     0.0     0.3
```

```
## CART
##
## 392 samples
##   4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (3 fold)
## Summary of sample sizes: 262, 262, 260
## Resampling results across tuning parameters:
##
##   maxdepth  RMSE      Rsquared   MAE
##    1        5.050923  0.5825953  3.946139
##    3        4.300565  0.6982880  3.203943
##    5        4.197841  0.7134244  3.087400
##   10        4.093410  0.7286272  3.038019
##   20        4.093410  0.7286272  3.038019
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was maxdepth = 10.
```

```
##   maxdepth     RMSE  Rsquared
## 1        1 5.050923 0.5825953
## 2        3 4.300565 0.6982880
## 3        5 4.197841 0.7134244
## 4       10 4.093410 0.7286272
## 5       20 4.093410 0.7286272
```

```
##    user  system elapsed
##    0.09    0.03    0.23
```

```
## Multivariate Adaptive Regression Spline
##
## 392 samples
##   4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (3 fold)
## Summary of sample sizes: 261, 262, 261
## Resampling results across tuning parameters:
##
##   degree  nprune  RMSE      Rsquared   MAE
##   1        2      5.159514  0.5640957  4.010201
##   1        5      3.968999  0.7424467  2.963539
##   1       10      3.823100  0.7594914  2.874214
##   1       15      3.823100  0.7594914  2.874214
##   2        2      4.879201  0.6108456  3.669229
##   2        5      4.026604  0.7353546  2.939173
##   2       10      3.838395  0.7589035  2.770735
##   2       15      3.838395  0.7589035  2.770735
##   3        2      5.090901  0.5756679  3.846417
##   3        5      3.942758  0.7467626  2.902105
##   3       10      3.729413  0.7747115  2.726847
##   3       15      3.716457  0.7762913  2.713580
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were nprune = 15 and degree = 3.
```

```
##    degree nprune     RMSE  Rsquared      MAE    RMSESD RsquaredSD      MAESD
## 1       1      2 5.159514 0.5640957 4.010201 0.5526436 0.09486998 0.41041003
## 5       2      2 4.879201 0.6108456 3.669229 0.2157390 0.03832628 0.19503994
## 9       3      2 5.090901 0.5756679 3.846417 0.5787235 0.09914538 0.50100502
## 2       1      5 3.968999 0.7424467 2.963539 0.1956691 0.02548987 0.02433286
## 6       2      5 4.026604 0.7353546 2.939173 0.1417920 0.02319250 0.07805820
## 10      3      5 3.942758 0.7467626 2.902105 0.2311092 0.03166700 0.15984469
## 3       1     10 3.823100 0.7594914 2.874214 0.2405268 0.03287434 0.19891744
## 7       2     10 3.838395 0.7589035 2.770735 0.3122852 0.04401616 0.26024195
## 11      3     10 3.729413 0.7747115 2.726847 0.2061261 0.02913808 0.25254296
## 4       1     15 3.823100 0.7594914 2.874214 0.2405268 0.03287434 0.19891744
## 8       2     15 3.838395 0.7589035 2.770735 0.3122852 0.04401616 0.26024195
## 12      3     15 3.716457 0.7762913 2.713580 0.1862172 0.02655609 0.23271098
```

```
##    user  system elapsed
##    0.62    0.00    1.05
```

```
## Random Forest
##
## 392 samples
##   4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (3 fold)
## Summary of sample sizes: 260, 262, 262
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared   MAE
##   2     3.616173  0.7919566  2.681204
##   3     3.704553  0.7842054  2.729694
##   4     3.792032  0.7760431  2.765694
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 2.
```

```
##   mtry     RMSE  Rsquared      MAE    RMSESD RsquaredSD     MAESD
## 1    2 3.616173 0.7919566 2.681204 0.4174678 0.02679269 0.2526268
## 2    3 3.704553 0.7842054 2.729694 0.4512967 0.03656871 0.2719742
## 3    4 3.792032 0.7760431 2.765694 0.5041682 0.04370216 0.2978098
```

```
##    user  system elapsed
##    1.89    0.08    2.99
```

```
## [1] "xgbTree"
## [1] "Regression"
##    eta max_depth gamma colsample_bytree min_child_weight subsample nrounds
## 10 0.3         1     0              0.8                1      0.50      50
## 12 0.3         1     0              0.8                1      0.50     150
## 17 0.3         1     0              0.8                1      1.00     100
## 18 0.3         1     0              0.8                1      1.00     150
## 16 0.3         1     0              0.8                1      1.00      50
## 13 0.3         1     0              0.8                1      0.75      50
## 11 0.3         1     0              0.8                1      0.50     100
## 67 0.4         1     0              0.8                1      0.75      50
## 58 0.4         1     0              0.6                1      0.75      50
## 14 0.3         1     0              0.8                1      0.75     100
##        RMSE  Rsquared      MAE    RMSESD  RsquaredSD      MAESD
## 10 3.724506 0.7767230 2.819333 0.1202830 0.005926799 0.04096251
## 12 3.718521 0.7759875 2.826727 0.2418589 0.015721693 0.13087897
## 17 3.744662 0.7742664 2.812245 0.1715206 0.007279682 0.07349908
## 18 3.752120 0.7730660 2.817402 0.1878343 0.010316390 0.07894048
## 16 3.761934 0.7730225 2.814028 0.1877777 0.008429065 0.06410188
## 13 3.761874 0.7715125 2.864474 0.1835315 0.013107648 0.03424426
## 11 3.779048 0.7714008 2.872469 0.1869592 0.011344422 0.07995420
## 67 3.762520 0.7709713 2.839924 0.1344634 0.009520170 0.01761765
## 58 3.766380 0.7707004 2.876996 0.2257469 0.013487500 0.08959507
## 14 3.760697 0.7706260 2.886832 0.1476525 0.009911142 0.06262230
```

```
##    user  system elapsed
##    3.32    2.59   26.21
```