Code ▾

HW 12 regression in r

Problem 2

Alumno: Daniel Nuño, daniel.nuno@iteso.mx (mailto:daniel.nuno@iteso.mx)

Alumno: David Cisneros

Alumno: Juan Maro Ochoa

Alumno: Rodrigo Huerta

4/18/2022

Problem 2: Application Problems

Note that some details are missing for all the following examples, the problems lack a complete explanation, and the code may need adequate comments. In this form, you must present a proper mathematical formulation, a brief background of the problem (and its bibliographical references) and, a much better explanation.

- The olsrr packeage
    a. Introduction to olsrr (https://olsrr.rsquaredacademy.com/articles/intro.html)
    b. Variable Selection Methods (https://olsrr.rsquaredacademy.com/articles/variable_selection.html)
    c. Residual Diagnostics (https://olsrr.rsquaredacademy.com/articles/residual_diagnostics.html)
    d. Heteroscedasticity (https://olsrr.rsquaredacademy.com/articles/heteroskedasticity.html)
    e. Measures of Influence (https://olsrr.rsquaredacademy.com/articles/influence_measures.html)
    f. Collinearity Diagnostics, Model Fit & Variable Contribution (https://olsrr.rsquaredacademy.com/articles/regression_diagnostics.html)
- The blorr package
    a. A Short Introduction to the blorr Package (https://blorr.rsquaredacademy.com/articles/introduction.html)

Introduction to olsrr

This document is a quick start guide to the tools offered by olsrr. Other vignettes provide more details on specific topics: - Residual Diagnostics: Includes plots to examine residuals to validate OLS assumptions - Variable selection: Different variable selection procedures such as all possible regression, best subset regression, stepwise regression, stepwise forward regression and stepwise backward regression - Heteroskedasticity: Tests for heteroskedasticity include bartlett test, breusch pagan test, score test and f test - Measures of influence: Includes 10 different plots to detect and identify influential observations - Collinearity diagnostics: VIF, Tolerance and condition indices to detect collinearity and plots for assessing mode fit and contributions of variables

This example uses **mtcars** dataset. This dataset contains a subset of the fuel economy data that the EPA makes available on https://fueleconomy.gov/ (https://fueleconomy.gov/). It contains only models which had a new release every year between 1999 and 2008 - this was used as a proxy for the popularity of the car.

A data frame with 234 rows and 11 variables:

- manufacturer: manufacturer name
- model: model name
- displ: engine displacement, in litres
- year: year of manufacture
- cyl: number of cylinders
- trans: type of transmission
- drv: the type of drive train, where f = front-wheel drive, r = rear wheel drive, 4 = 4wd
- cty: city miles per gallon
- hwy: highway miles per gallon
- fl: fuel type
- class: “type” of car

Regression

Hide

```
library(olsrr)
```

```

Attaching package: ‘olsrr’

The following object is masked from ‘package:MASS’:

    cement

The following object is masked from ‘package:datasets’:

    rivers
```

Hide

```
ols_regress(mpg ~ disp + hp + wt + qsec, data = mtcars)
```

```
                        Model Summary
---------------------------------------------------------------
R                       0.914       RMSE                 2.409
R-Squared               0.835       MSE                  6.875
Adj. R-Squared          0.811       Coef. Var           13.051
Pred R-Squared          0.771       AIC                159.070
MAE                     1.858       SBC                167.864
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                              ANOVA
--------------------------------------------------------------------
               Sum of
              Squares        DF    Mean Square      F         Sig.
--------------------------------------------------------------------
Regression    940.412         4        235.103    34.195    0.0000
Residual      185.635        27          6.875
Total        1126.047        31
--------------------------------------------------------------------

                                 Parameter Estimates
------------------------------------------------------------------------------------
----
      model      Beta    Std. Error    Std. Beta      t        Sig      lower     up
per
------------------------------------------------------------------------------------
----
(Intercept)    27.330         8.639                 3.164    0.004      9.604    45.
055
       disp     0.003         0.011        0.055    0.248    0.806     -0.019     0.
025
         hp    -0.019         0.016       -0.212   -1.196    0.242     -0.051     0.
013
         wt    -4.609         1.266       -0.748   -3.641    0.001     -7.206    -2.
012
       qsec     0.544         0.466        0.161    1.166    0.254     -0.413     1.
501
------------------------------------------------------------------------------------
----
```

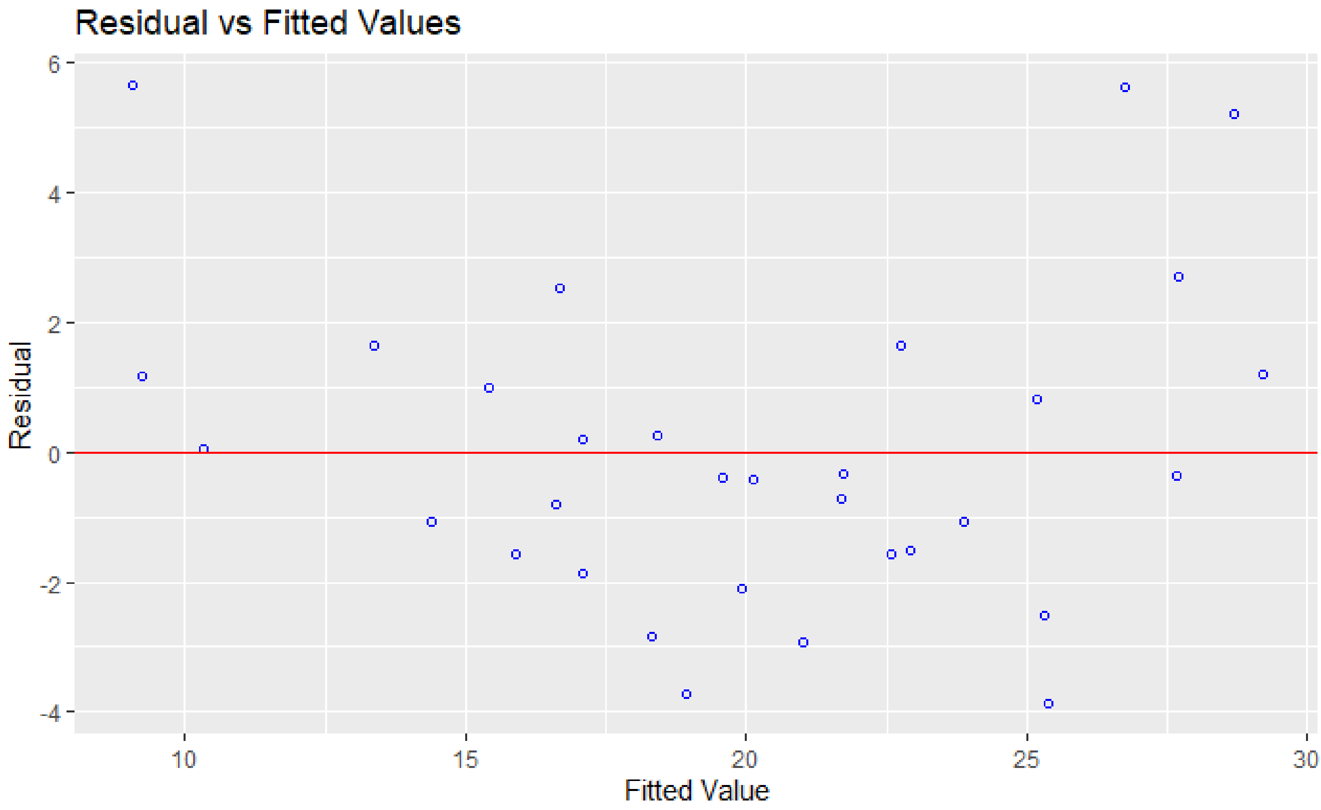
In the presence of interaction terms in the model, the predictors are scaled and centered before computing the standardized betas. `ols_regress()` will detect interaction terms automatically but in case you have created a new variable instead of using the inline function `*`, you can indicate the presence of interaction terms by setting `iterm` to `TRUE` .

Residual vs Fitted Values Plot

Plot to detect non-linearity, unequal error variances, and outliers. Each point is the error in each vector. Red line just marks the 0 to have a visual benchmark.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_resid_fit(model)
```



DFBETAs Panel

DFBETAs measure the difference in each parameter estimate with and without the influential observation. `dfbetas_panel` creates plots to detect influential observations using DFBETAs.

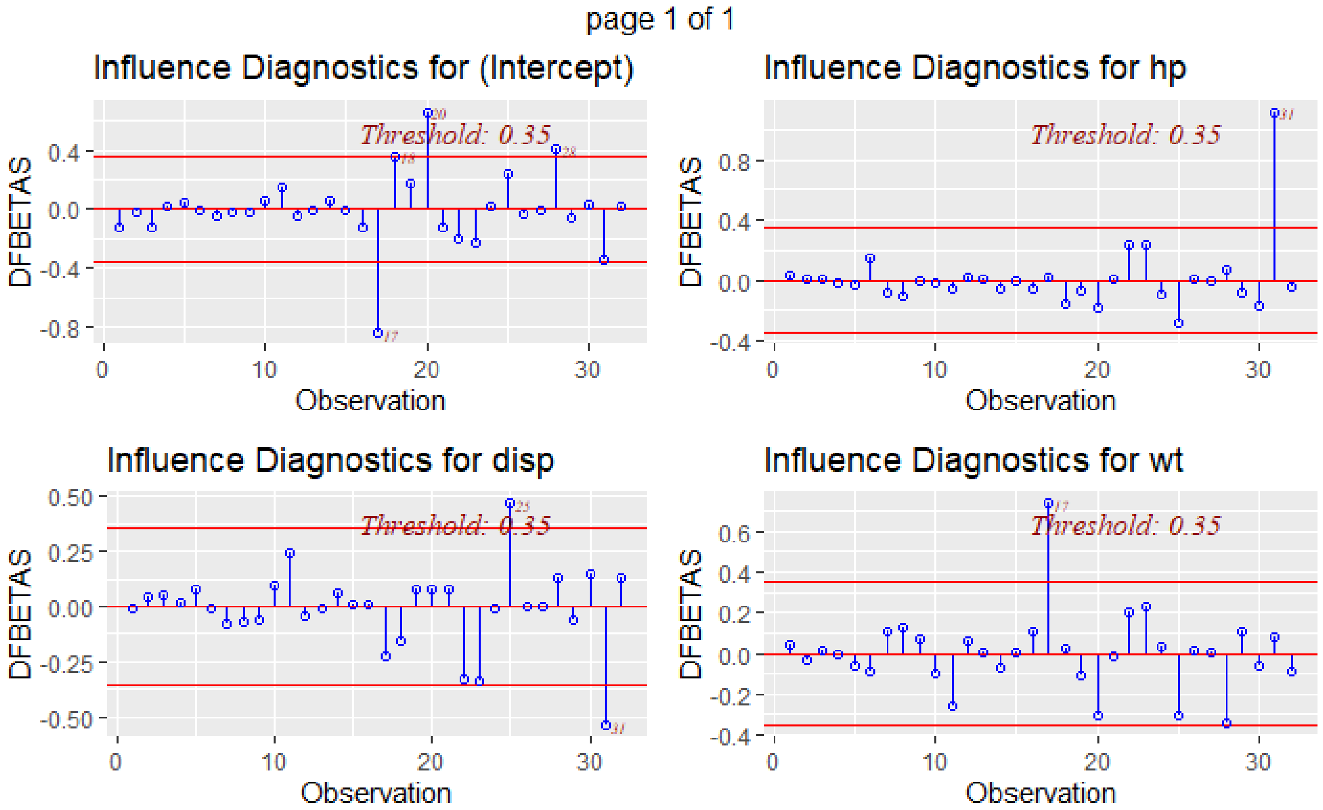
Belsley, Kuh, and Welsch MATH sugirieron una estadística que indica cuanto el coeficiente de regresión estimado b_i cambia, en unidades de desviaciones estándar, si la $i - \acute{e}sima$ observación fuera eliminada. La estadística es

$$DFBETAS_{i,j} = \frac{b_i - b_{j(i)}}{\sqrt{s_i^2 C_{jj}}}$$

Donde MATH es la varianza del coeficiente de regresión b_j calculada sin la $i - ésima$ observación. Un valor grande de DFBETAS$_{j,i}$ indica que la $i - ésima$ observación tiene una considerable influencia sobre el $j - ésimo$ coeficiente de regresión b_j. reference
(http://red.unal.edu.co/cursos/ciencias/2007315/html/un6/cont_12_73.html)

Hide

```
model <- lm(mpg ~ disp + hp + wt, data = mtcars)
ols_plot_dfbetas(model)
```



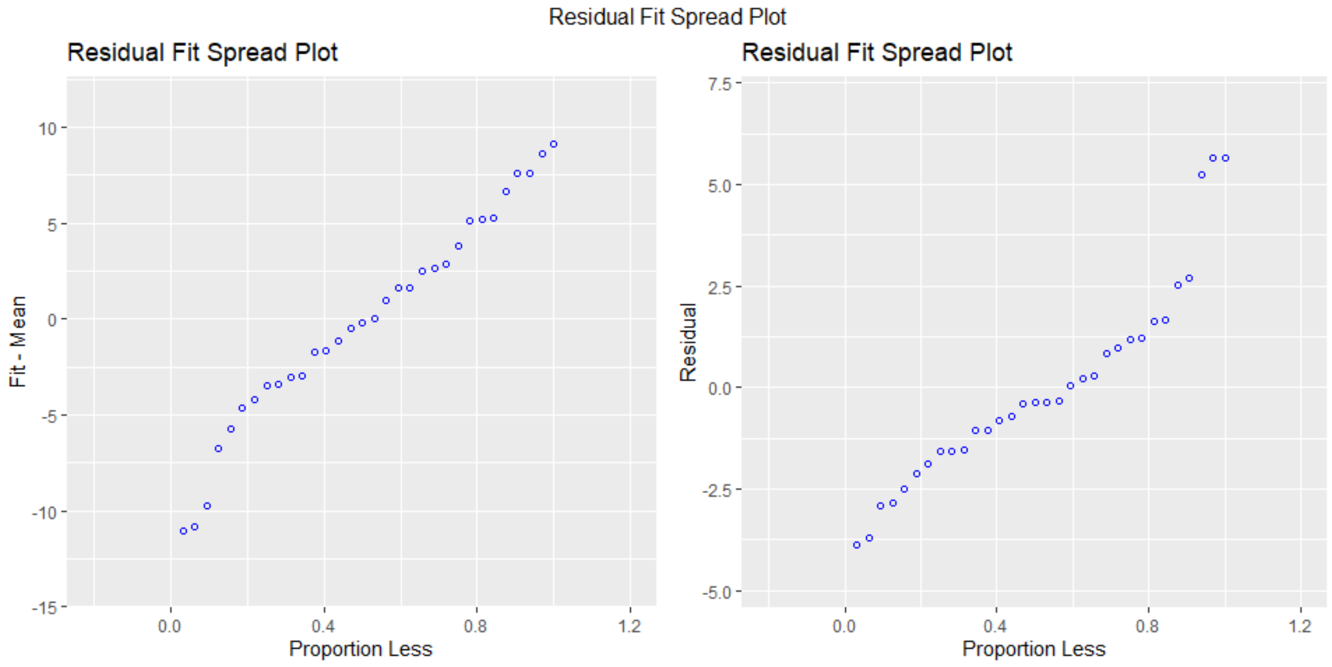
Residual Fit Spread Plot

Plot to detect non-linearity, influential observations and outliers.

Each spread plot is a graph of centered data values plotted against the estimated cumulative probability. Thus, spread plots are similar to a (rotated) plot of the empirical cumulative distribution function. reference (https://blogs.sas.com/content/iml/2013/06/12/interpret-residual-fit-spread-plot.html)

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_resid_fit_spread(model)
```



Breusch Pagan Test

Breusch Pagan test is used to test for herteroskedasticity (non-constant error variance). It tests whether the variance of the errors from a regression is dependent on the values of the independent variables. It is a χ^2 test.

Null hypothesis implies the variance is constant and using an alpha of 0.05 then in this case we reject the null hypothesis because p-value is 0.23

Hide

```
model <- lm(mpg ~ disp + hp + wt + drat, data = mtcars)
ols_test_breusch_pagan(model)
```

```

 Breusch Pagan Test for Heteroskedasticity
 -----------------------------------------
 Ho: the variance is constant
 Ha: the variance is not constant

             Data
 -------------------------------
 Response : mpg
 Variables: fitted values of mpg

       Test Summary
 ---------------------------
 DF            =    1
 Chi2          =    1.429672
 Prob > Chi2   =    0.231818
```

Collinearity Diagnostics

collinearity, in statistics, correlation between predictor variables (or independent variables), such that they express a linear relationship in a regression model. When predictor variables in the same regression model are correlated, they cannot independently predict the value of the dependent variable.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_coll_diag(model)
```

```
Tolerance and Variance Inflation Factor
---------------------------------------
```

Variables <chr>	Tolerance <dbl>	VIF <dbl>
disp	0.1252279	7.985439
hp	0.1935450	5.166758
wt	0.1445726	6.916942
qsec	0.3191708	3.133119

4 rows

```


Eigenvalue and Condition Index
------------------------------
```

Eigenvalue <dbl>	Condition Index <dbl>	intercept <dbl>	disp <dbl>	hp <dbl>	wt <dbl>
4.721487187	1.000000	0.000123237	0.001132468	0.001413094	0.0005253393
0.216562203	4.669260	0.002617424	0.036811051	0.027751289	0.0002096014
0.050416837	9.677242	0.001656551	0.120881424	0.392366164	0.0377028008
0.010104757	21.616057	0.025805998	0.777260487	0.059594623	0.7017528428
0.001429017	57.480524	0.969796790	0.063914571	0.518874831	0.2598094157

5 rows

Stepwise Regression

Build regression model from a set of candidate predictor variables by entering and removing predictors based on p values, in a stepwise manner until there is no variable left to enter or remove any more.

Here p-value and Akaike Information Criterion are used to decide which model is the best in each step of the algorithm.

Variable Selection

Hide

```
# stepwise regression
model <- lm(y ~ ., data = surgical)
ols_step_both_p(model)
```

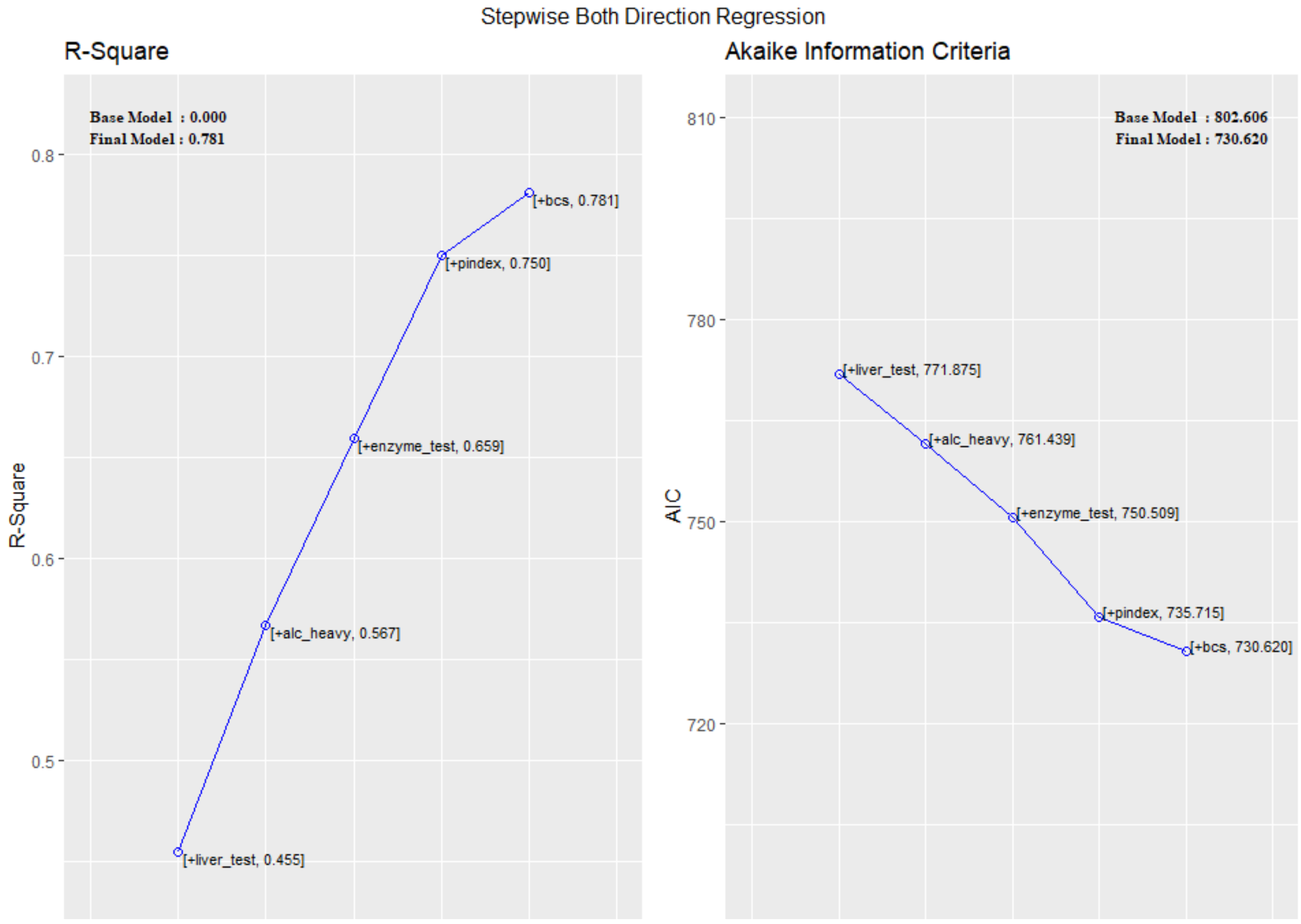
```
                             Stepwise Summary
------------------------------------------------------------------------------
Step    Variable            AIC        SBC       SBIC        R2       Adj. R2
------------------------------------------------------------------------------
 0      Base Model        802.606    806.584    646.794    0.00000    0.00000
 1      liver_test (+)    771.875    777.842    616.009    0.45454    0.44405
 2      alc_heavy (+)     761.439    769.395    605.506    0.56674    0.54975
 3      enzyme_test (+)   750.509    760.454    595.297    0.65900    0.63854
 4      pindex (+)        735.715    747.649    582.943    0.75015    0.72975
 5      bcs (+)           730.620    744.543    579.638    0.78091    0.75808
------------------------------------------------------------------------------

Final Model Output
------------------

                        Model Summary
-------------------------------------------------------------------
R                          0.884       RMSE                184.276
R-Squared                  0.781       MSE               38202.426
Adj. R-Squared             0.758       Coef. Var            27.839
Pred R-Squared             0.700       AIC                 730.620
MAE                      137.656       SBC                 744.543
-------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                              ANOVA
-----------------------------------------------------------------------
                   Sum of
                  Squares        DF    Mean Square      F         Sig.
-----------------------------------------------------------------------
Regression    6535804.090         5    1307160.818    34.217    0.0000
Residual      1833716.447        48      38202.426
Total         8369520.537        53
-----------------------------------------------------------------------

                                  Parameter Estimates
-------------------------------------------------------------------------------------
------------
      model         Beta    Std. Error    Std. Beta      t        Sig         lower
upper
-------------------------------------------------------------------------------------
------------
(Intercept)    -1178.330       208.682                 -5.647    0.000    -1597.914
```

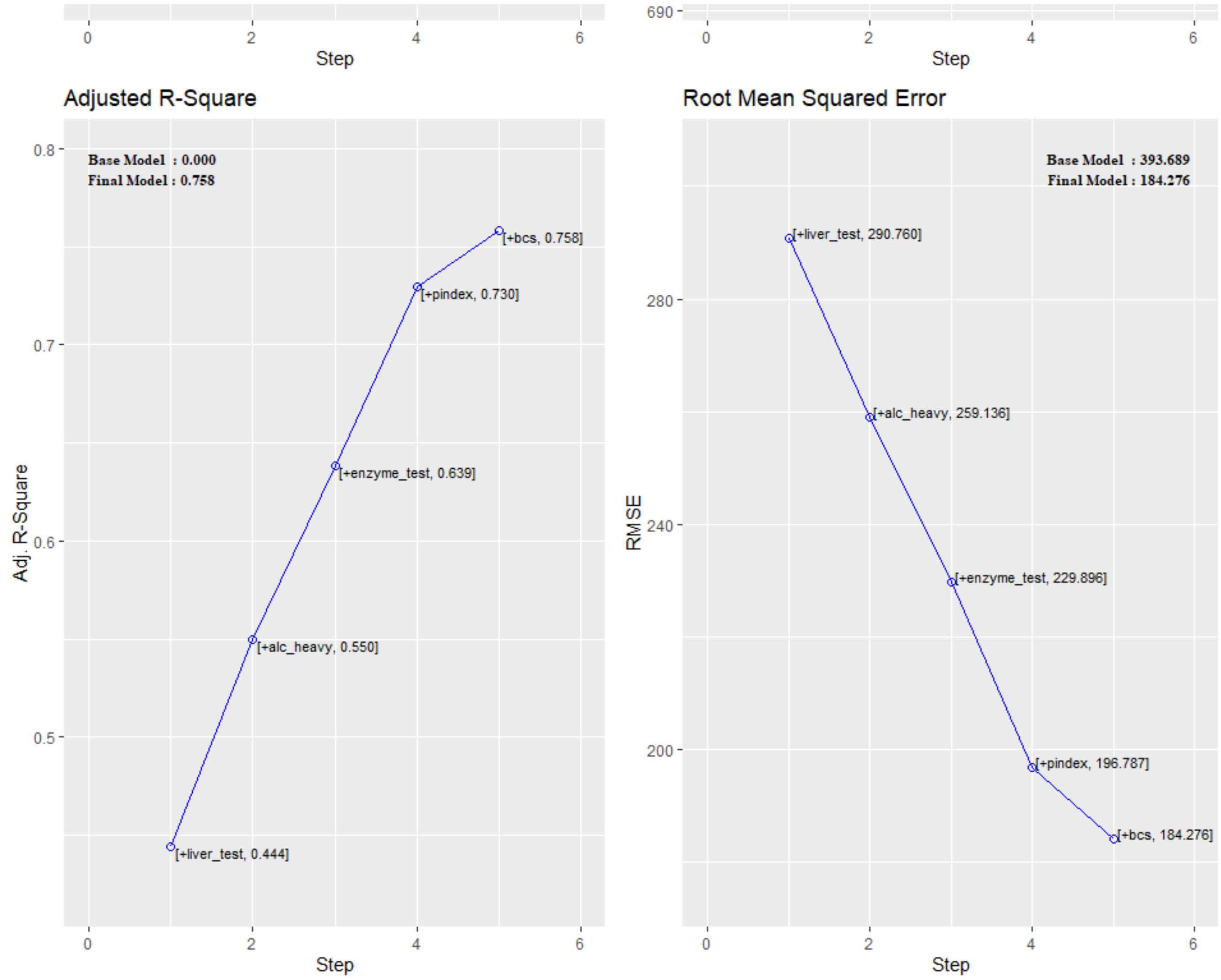
```
-758.746
 liver_test       58.064         40.144         0.156     1.446    0.155      -22.652
138.779
  alc_heavy      317.848         71.634         0.314     4.437    0.000      173.818
461.878
enzyme_test        9.748          1.656         0.521     5.887    0.000        6.419
13.077
     pindex        8.924          1.808         0.380     4.935    0.000        5.288
12.559
        bcs       59.864         23.060         0.241     2.596    0.012       13.498
106.230
--------------------------------------------------------------------------------------
------------
```

Plot

Hide

```
model <- lm(y ~ ., data = surgical)
k <- ols_step_both_p(model)
plot(k)
```


Stepwise Both Direction Regression
R-Square
Base Model : 0.000
Final Model : 0.781
R-Square
0.8
0.7
0.6
0.5
[+liver_test, 0.455]
[+alc_heavy, 0.567]
[+enzyme_test, 0.659]
[+pindex, 0.750]
[+bcs, 0.781]
Akaike Information Criteria
Base Model : 802.606
Final Model : 730.620
AIC
810
780
750
720
[+liver_test, 771.875]
[+alc_heavy, 761.439]
[+enzyme_test, 750.509]
[+pindex, 735.715]
[+bcs, 730.620]



###Stepwise AIC Backward Regression

Build regression model from a set of candidate predictor variables by removing predictors based on Akaike Information Criteria, in a stepwise manner until there is no variable left to remove any more.

###Variable Selection

Hide

```
# stepwise aic backward regression
model <- lm(y ~ ., data = surgical)
k <- ols_step_backward_aic(model)
k
```

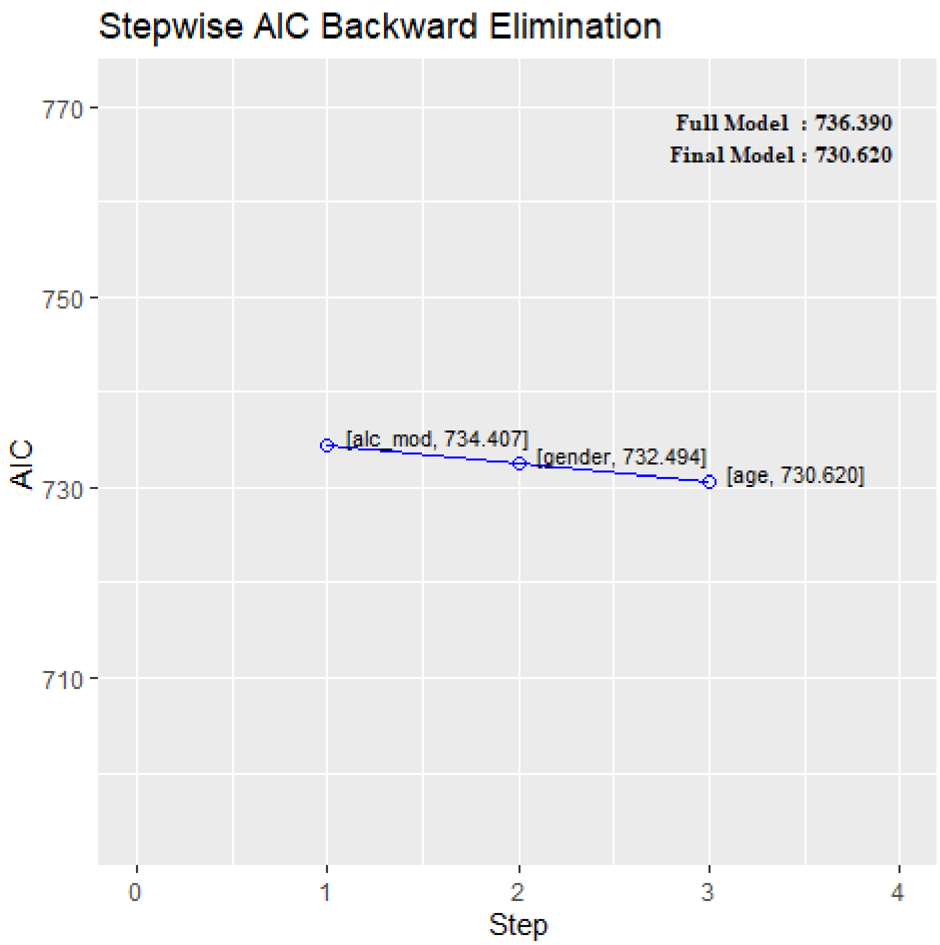
```
                             Stepwise Summary
-------------------------------------------------------------------------
Step    Variable        AIC        SBC       SBIC        R2       Adj. R2
-------------------------------------------------------------------------
 0      Full Model    736.390    756.280    586.665    0.78184    0.74305
 1      alc_mod       734.407    752.308    583.884    0.78177    0.74856
 2      gender        732.494    748.406    581.290    0.78142    0.75351
 3      age           730.620    744.543    578.844    0.78091    0.75808
-------------------------------------------------------------------------

Final Model Output
------------------

                           Model Summary
-------------------------------------------------------------------
R                         0.884       RMSE                 184.276
R-Squared                 0.781       MSE                38202.426
Adj. R-Squared            0.758       Coef. Var             27.839
Pred R-Squared            0.700       AIC                  730.620
MAE                     137.656       SBC                  744.543
-------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                 ANOVA
-----------------------------------------------------------------------
                  Sum of
                 Squares        DF    Mean Square      F          Sig.
-----------------------------------------------------------------------
Regression    6535804.090         5    1307160.818    34.217    0.0000
Residual      1833716.447        48      38202.426
Total         8369520.537        53
-----------------------------------------------------------------------

                                          Parameter Estimates
-------------------------------------------------------------------------------------
------------
      model         Beta    Std. Error    Std. Beta      t         Sig         lower
upper
-------------------------------------------------------------------------------------
------------
(Intercept)    -1178.330       208.682                 -5.647    0.000    -1597.914
-758.746
        bcs       59.864        23.060        0.241     2.596    0.012       13.498
```

```
106.230
     pindex         8.924           1.808           0.380       4.935     0.000           5.288
12.559
enzyme_test         9.748           1.656           0.521       5.887     0.000           6.419
13.077
 liver_test        58.064          40.144           0.156       1.446     0.155         -22.652
138.779
  alc_heavy       317.848          71.634           0.314       4.437     0.000         173.818
461.878
------------------------------------------------------------------------------------
------------
```

Plot

Hide

```
model <- lm(y ~ ., data = surgical)
k <- ols_step_backward_aic(model)
plot(k)
```



Variable Selection Methods

```

Attaching package: ‘ggplot2’

The following object is masked from ‘Auto’:

    mpg


Attaching package: ‘goftest’

The following objects are masked from ‘package:nortest’:

    ad.test, cvm.test
```

Introduction

All Possible Regression

All subset regression tests, all possible subsets of the set of potential independent variables. If there are K potential independent variables (besides the constant), then there are 2^k distinct subsets of them to be tested. For example, if you have 10 candidate independent variables, the number of subsets to be tested is 2^{10}, which is 1024, and if you have 20 candidate variables, the number is 2^{20}, which is more than one million.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_step_all_possible(model)
```

	Index <int>	N <int>	Predictors <chr>	R-Square <dbl>	Adj. R-Square <dbl>	Mallow's Cp <dbl>
3	1	1	wt	0.7528328	0.7445939	0.70869536
1	2	1	disp	0.7183433	0.7089548	0.67512054
2	3	1	hp	0.6024373	0.5891853	0.50969578
4	4	1	qsec	0.1752963	0.1478062	0.07541973
8	5	2	hp wt	0.8267855	0.8148396	0.78108710
10	6	2	wt qsec	0.8264161	0.8144448	0.77856272
6	7	2	disp wt	0.7809306	0.7658223	0.72532105
5	8	2	disp hp	0.7482402	0.7308774	0.69454380
7	9	2	disp qsec	0.7215598	0.7023571	0.66395284

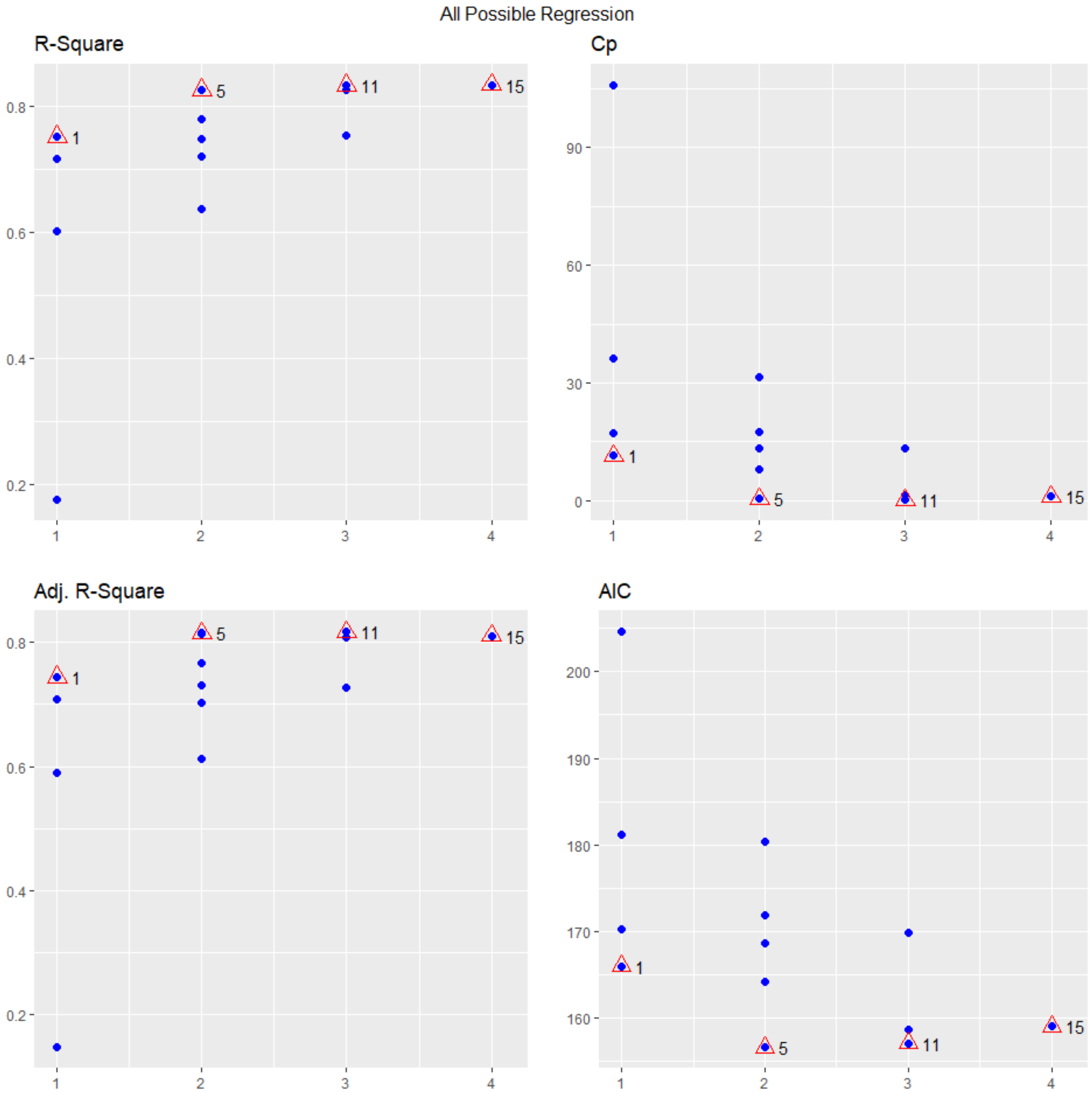
	Index	N	Predictors	R-Square	Adj. R-Square	Mallow's Cp
	<int>	<int>	<chr>	<dbl>	<dbl>	<dbl>
9	10	2	hp qsec	0.6368769	0.6118339	0.52014395

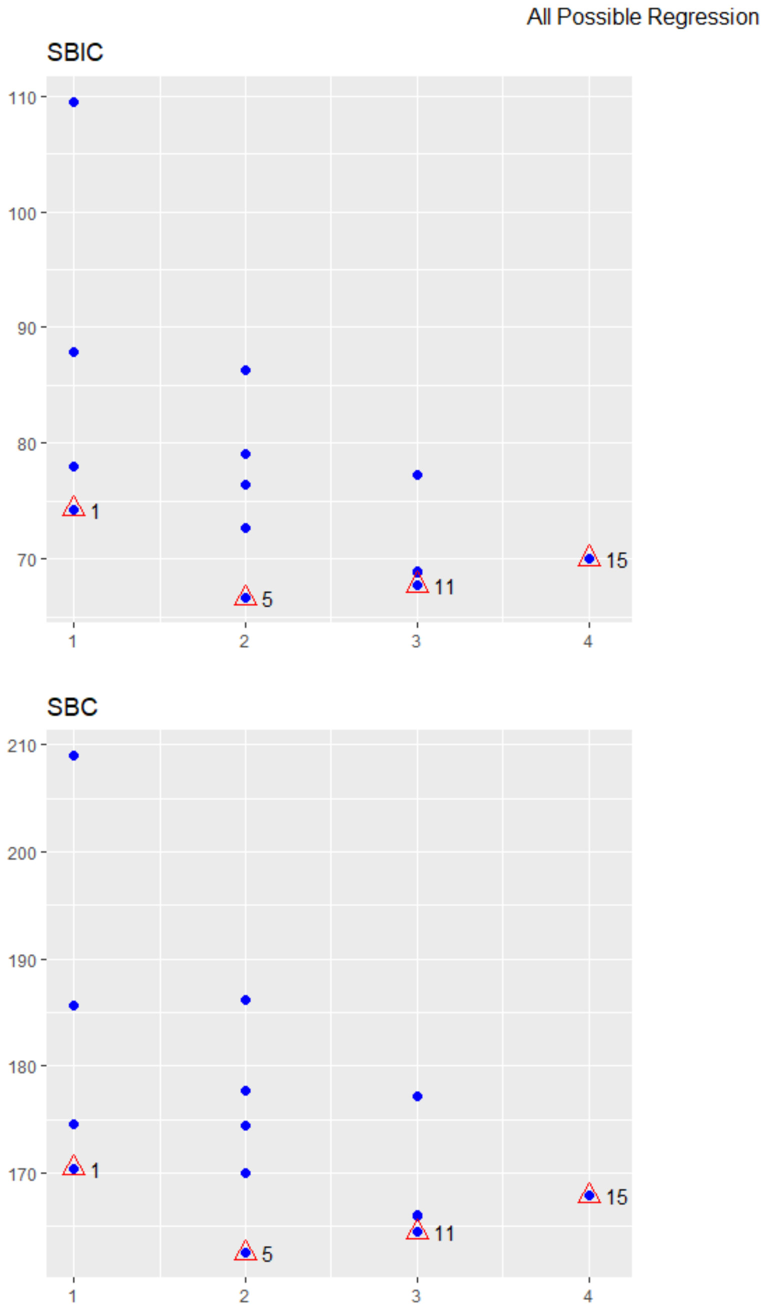
1-10 of 15 rows Previous **1** 2 Next

The `plot` method shows the panel of fit criteria for all possible regression methods.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
k <- ols_step_all_possible(model)
plot(k)
```


All Possible Regression
R-Square
0.2
0.4
0.6
0.8
1
5
11
15
1
2
3
4
Cp
0
30
60
90
1
5
11
15
1
2
3
4
Adj. R-Square
0.2
0.4
0.6
0.8
1
5
11
15
1
2
3
4
AIC
160
170
180
190
200
1
5
11
15
1
2
3
4



Best Subset Regression

Select the subset of predictors that do the best at meeting some well-defined objective criterion, such as having the largest R2 value or the smallest MSE, Mallow’s Cp or AIC.

Hide

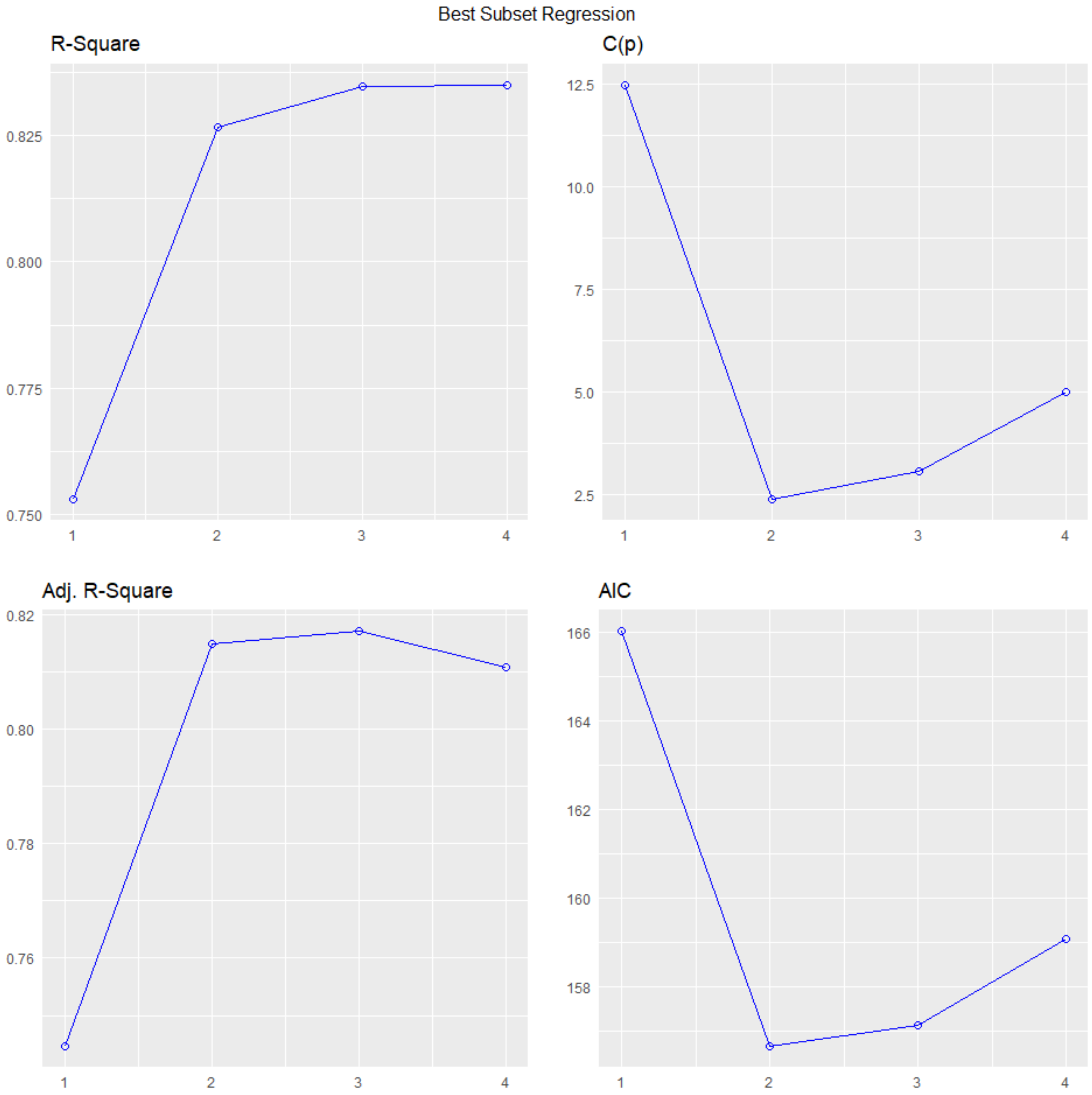
```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_step_best_subset(model)
```

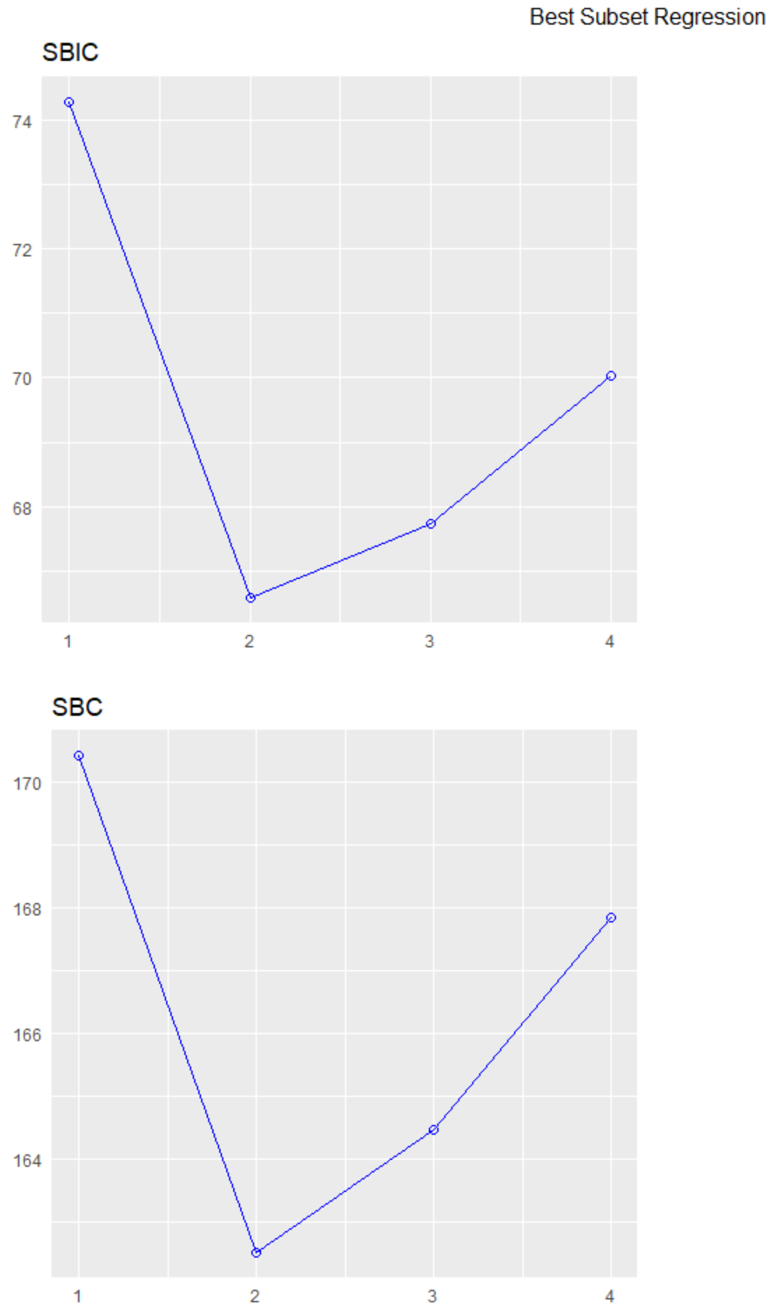
```
   Best Subsets Regression   
------------------------------
Model Index    Predictors
------------------------------
     1         wt                
     2         hp wt             
     3         hp wt qsec        
     4         disp hp wt qsec   
------------------------------

                                                   Subsets Regression Summary                                                   
---------------------------------------------------------------------------------------------------------------------
------------
                       Adj.        Pred                                                                              
Model    R-Square    R-Square    R-Square     C(p)        AIC        SBIC        SBC         MSEP        FPE        HSP      
APC  
---------------------------------------------------------------------------------------------------------------------
------------
  1        0.7528      0.7446      0.7087    12.4809    166.0294    74.2916    170.4266    296.9167    9.8572    0.31
99    0.2801 
  2        0.8268      0.8148      0.7811     2.3690    156.6523    66.5755    162.5153    215.5104    7.3563    0.24
02    0.2091 
  3        0.8348      0.8171       0.782     3.0617    157.1426    67.7238    164.4713    213.1929    7.4756    0.24
61    0.2124 
  4        0.8351      0.8107       0.771     5.0000    159.0696    70.0408    167.8640    220.8882    7.9497    0.26
44    0.2259 
---------------------------------------------------------------------------------------------------------------------
------------
AIC: Akaike Information Criteria 
 SBIC: Sawa's Bayesian Information Criteria 
 SBC: Schwarz Bayesian Criteria 
 MSEP: Estimated error of prediction, assuming multivariate normality 
 FPE: Final Prediction Error 
 HSP: Hocking's Sp 
 APC: Amemiya Prediction Criteria 
```

The `plot` method shows the panel of fit criteria for best subset regression methods.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
k <- ols_step_best_subset(model)
plot(k)
```


Best Subset Regression
R-Square
0.750
0.775
0.800
0.825
1
2
3
4
C(p)
2.5
5.0
7.5
10.0
12.5
1
2
3
4
Adj. R-Square
0.76
0.78
0.80
0.82
1
2
3
4
AIC
158
160
162
164
166
1
2
3
4



Stepwise Forward Regression

Build regression model from a set of candidate predictor variables by entering predictors based on p values, in a stepwise manner until there is no variable left to enter any more. The model should include all the candidate predictor variables. If details is set to `TRUE` , each step is displayed.

Variable Selection

Hide

```
# stepwise forward regression
model <- lm(y ~ ., data = surgical)
ols_step_forward_p(model)
```

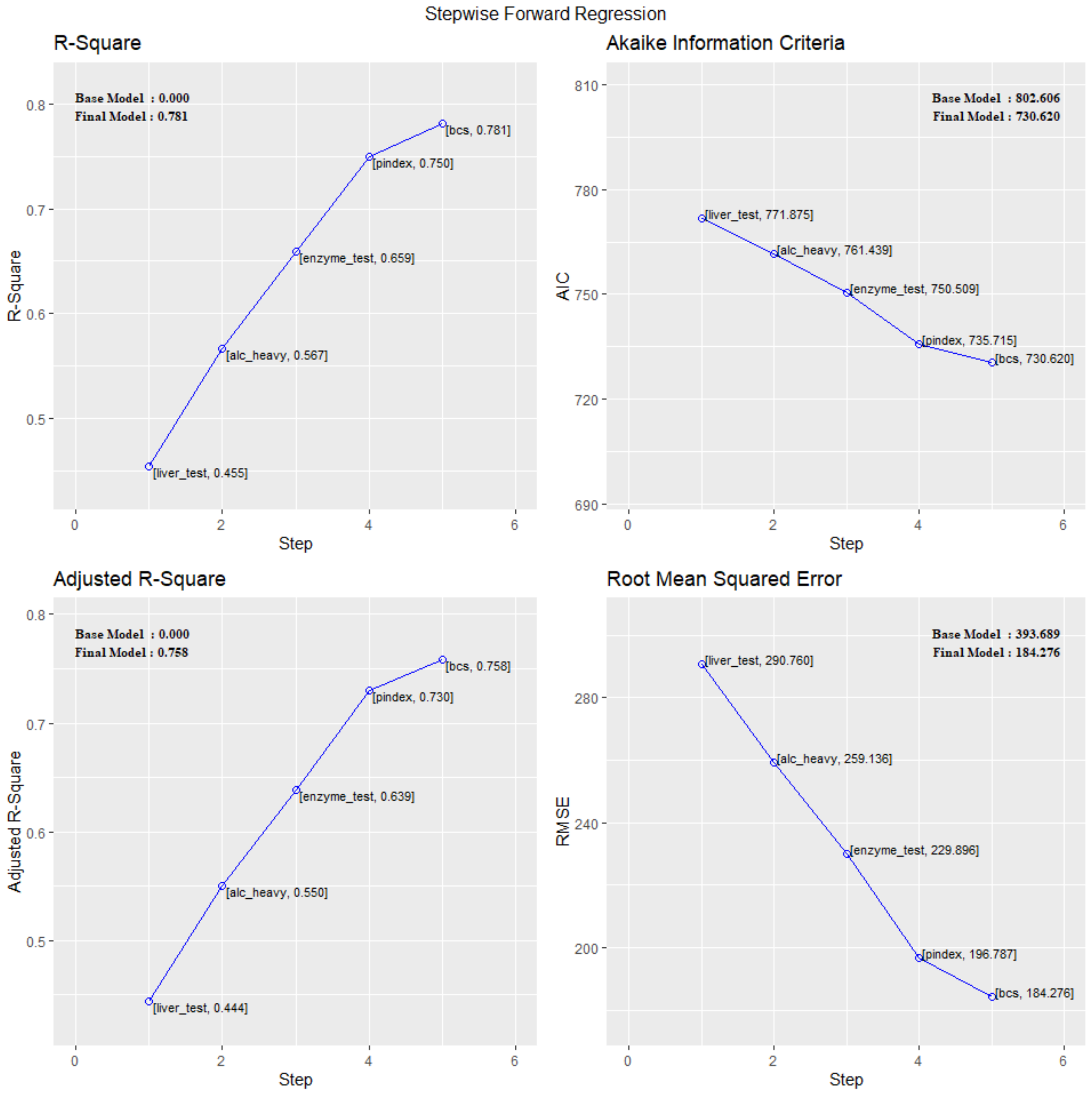
```
                              Stepwise Summary
--------------------------------------------------------------------------
Step    Variable         AIC        SBC       SBIC        R2       Adj. R2
--------------------------------------------------------------------------
 0      Base Model     802.606    806.584    646.794    0.00000    0.00000
 1      liver_test     771.875    777.842    616.009    0.45454    0.44405
 2      alc_heavy      761.439    769.395    605.506    0.56674    0.54975
 3      enzyme_test    750.509    760.454    595.297    0.65900    0.63854
 4      pindex         735.715    747.649    582.943    0.75015    0.72975
 5      bcs            730.620    744.543    579.638    0.78091    0.75808
--------------------------------------------------------------------------

Final Model Output
------------------

                           Model Summary
-------------------------------------------------------------------
R                          0.884       RMSE                184.276
R-Squared                  0.781       MSE               38202.426
Adj. R-Squared             0.758       Coef. Var            27.839
Pred R-Squared             0.700       AIC                 730.620
MAE                      137.656       SBC                 744.543
-------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                 ANOVA
-----------------------------------------------------------------------
                   Sum of
                  Squares        DF    Mean Square      F         Sig.
-----------------------------------------------------------------------
Regression    6535804.090         5    1307160.818    34.217    0.0000
Residual      1833716.447        48      38202.426
Total         8369520.537        53
-----------------------------------------------------------------------

                                      Parameter Estimates
-------------------------------------------------------------------------------------
------------
      model         Beta    Std. Error    Std. Beta      t        Sig         lower
upper
-------------------------------------------------------------------------------------
------------
(Intercept)    -1178.330       208.682                -5.647    0.000    -1597.914
```

```
-758.746
 liver_test       58.064         40.144         0.156     1.446    0.155      -22.652
138.779
  alc_heavy      317.848         71.634         0.314     4.437    0.000      173.818
461.878
enzyme_test        9.748          1.656         0.521     5.887    0.000        6.419
13.077
     pindex        8.924          1.808         0.380     4.935    0.000        5.288
12.559
        bcs       59.864         23.060         0.241     2.596    0.012       13.498
106.230
-------------------------------------------------------------------------------------
------------
```

Plot

Hide

```
model <- lm(y ~ ., data = surgical)
k <- ols_step_forward_p(model)
plot(k)
```



Detailed Output

Hide

```
# stepwise forward regression
model <- lm(y ~ ., data = surgical)
ols_step_forward_p(model, details = TRUE)
```

```
Forward Selection Method
------------------------

Candidate Terms:

1. bcs
2. pindex
3. enzyme_test
4. liver_test
5. age
6. gender
7. alc_mod
8. alc_heavy


Step   => 0
Model  => y ~ 1
R2     => 0

Initiating stepwise selection...

                     Selection Metrics Table
-----------------------------------------------------------------
Predictor      Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
-----------------------------------------------------------------
liver_test      0.00000        0.455             0.444    771.875
enzyme_test     0.00000        0.334             0.322    782.629
pindex          0.00155        0.177             0.161    794.100
alc_heavy       0.00172        0.174             0.158    794.301
bcs             0.01025        0.120             0.103    797.697
alc_mod         0.19286        0.032             0.014    802.828
gender          0.20972        0.030             0.011    802.956
age             0.39073        0.014            -0.005    803.834
-----------------------------------------------------------------

Step      => 1
Selected  => liver_test
Model     => y ~ liver_test
R2        => 0.455

                     Selection Metrics Table
-----------------------------------------------------------------
Predictor      Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
-----------------------------------------------------------------
alc_heavy       0.00065        0.567             0.550    761.439
enzyme_test     0.00089        0.562             0.544    762.077
pindex          0.07087        0.489             0.469    770.387
alc_mod         0.10979        0.481             0.461    771.141
```

```
gender          0.79395        0.455              0.434    773.802
age             0.83908        0.455              0.434    773.831
bcs             0.93062        0.455              0.433    773.867
-----------------------------------------------------------------

Step      => 2
Selected  => alc_heavy
Model     => y ~ liver_test + alc_heavy
R2        => 0.567

                     Selection Metrics Table
-----------------------------------------------------------------
Predictor      Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
-----------------------------------------------------------------
enzyme_test     0.00057        0.659              0.639    750.509
pindex          0.00961        0.622              0.599    756.125
bcs             0.55687        0.570              0.544    763.063
age             0.58269        0.569              0.544    763.110
alc_mod         0.91757        0.567              0.541    763.428
gender          0.93799        0.567              0.541    763.433
-----------------------------------------------------------------

Step      => 3
Selected  => enzyme_test
Model     => y ~ liver_test + alc_heavy + enzyme_test
R2        => 0.659

                    Selection Metrics Table
---------------------------------------------------------------
Predictor    Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
---------------------------------------------------------------
pindex          1e-04        0.750              0.730    735.715
bcs           0.21294        0.670              0.643    750.782
alc_mod       0.75743        0.660              0.632    752.403
age           0.77290        0.660              0.632    752.416
gender        0.99197        0.659              0.631    752.509
---------------------------------------------------------------

Step      => 4
Selected  => pindex
Model     => y ~ liver_test + alc_heavy + enzyme_test + pindex
R2        => 0.75

                    Selection Metrics Table
---------------------------------------------------------------
Predictor    Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
---------------------------------------------------------------
bcs           0.01248        0.781              0.758    730.620
age           0.86220        0.750              0.724    737.680
```

```
gender        0.96390        0.750             0.724    737.712
alc_mod       0.97040        0.750             0.724    737.713
---------------------------------------------------------------

Step      => 5
Selected  => bcs
Model     => y ~ liver_test + alc_heavy + enzyme_test + pindex + bcs
R2        => 0.781

                    Selection Metrics Table
---------------------------------------------------------------
Predictor    Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
---------------------------------------------------------------
age           0.74164        0.781             0.754    732.494
gender        0.80666        0.781             0.753    732.551
alc_mod       0.94086        0.781             0.753    732.614
---------------------------------------------------------------


No more variables to be added.

Variables Selected:

=> liver_test
=> alc_heavy
=> enzyme_test
=> pindex
=> bcs

                              Stepwise Summary
--------------------------------------------------------------------------
Step    Variable         AIC        SBC       SBIC        R2       Adj. R2
--------------------------------------------------------------------------
 0      Base Model     802.606    806.584    646.794    0.00000    0.00000
 1      liver_test     771.875    777.842    616.009    0.45454    0.44405
 2      alc_heavy      761.439    769.395    605.506    0.56674    0.54975
 3      enzyme_test    750.509    760.454    595.297    0.65900    0.63854
 4      pindex         735.715    747.649    582.943    0.75015    0.72975
 5      bcs            730.620    744.543    579.638    0.78091    0.75808
--------------------------------------------------------------------------

Final Model Output
------------------

                           Model Summary
-------------------------------------------------------------------
R                          0.884       RMSE                 184.276
R-Squared                  0.781       MSE                38202.426
Adj. R-Squared             0.758       Coef. Var             27.839
```

```
Pred R-Squared             0.700       AIC                     730.620
MAE                      137.656       SBC                     744.543
-------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                ANOVA
-----------------------------------------------------------------------
                  Sum of
                 Squares        DF    Mean Square      F         Sig.
-----------------------------------------------------------------------
Regression    6535804.090        5    1307160.818    34.217    0.0000
Residual      1833716.447       48      38202.426
Total         8369520.537       53
-----------------------------------------------------------------------

                                     Parameter Estimates
------------------------------------------------------------------------------------
------------
      model          Beta    Std. Error    Std. Beta      t        Sig         lower
upper
------------------------------------------------------------------------------------
------------
(Intercept)    -1178.330       208.682                 -5.647    0.000    -1597.914
-758.746
 liver_test       58.064        40.144        0.156     1.446    0.155      -22.652
138.779
  alc_heavy      317.848        71.634        0.314     4.437    0.000      173.818
461.878
enzyme_test        9.748         1.656        0.521     5.887    0.000        6.419
13.077
     pindex        8.924         1.808        0.380     4.935    0.000        5.288
12.559
        bcs       59.864        23.060        0.241     2.596    0.012       13.498
106.230
------------------------------------------------------------------------------------
------------
```

Stepwise Backward Regression

Build regression model from a set of candidate predictor variables by removing predictors based on p values, in a stepwise manner until there is no variable left to remove any more. The model should include all the candidate predictor variables. If details is set to `TRUE` , each step is displayed.

Variable Selection

Hide

```
# stepwise backward regression
model <- lm(y ~ ., data = surgical)
ols_step_backward_p(model)
```

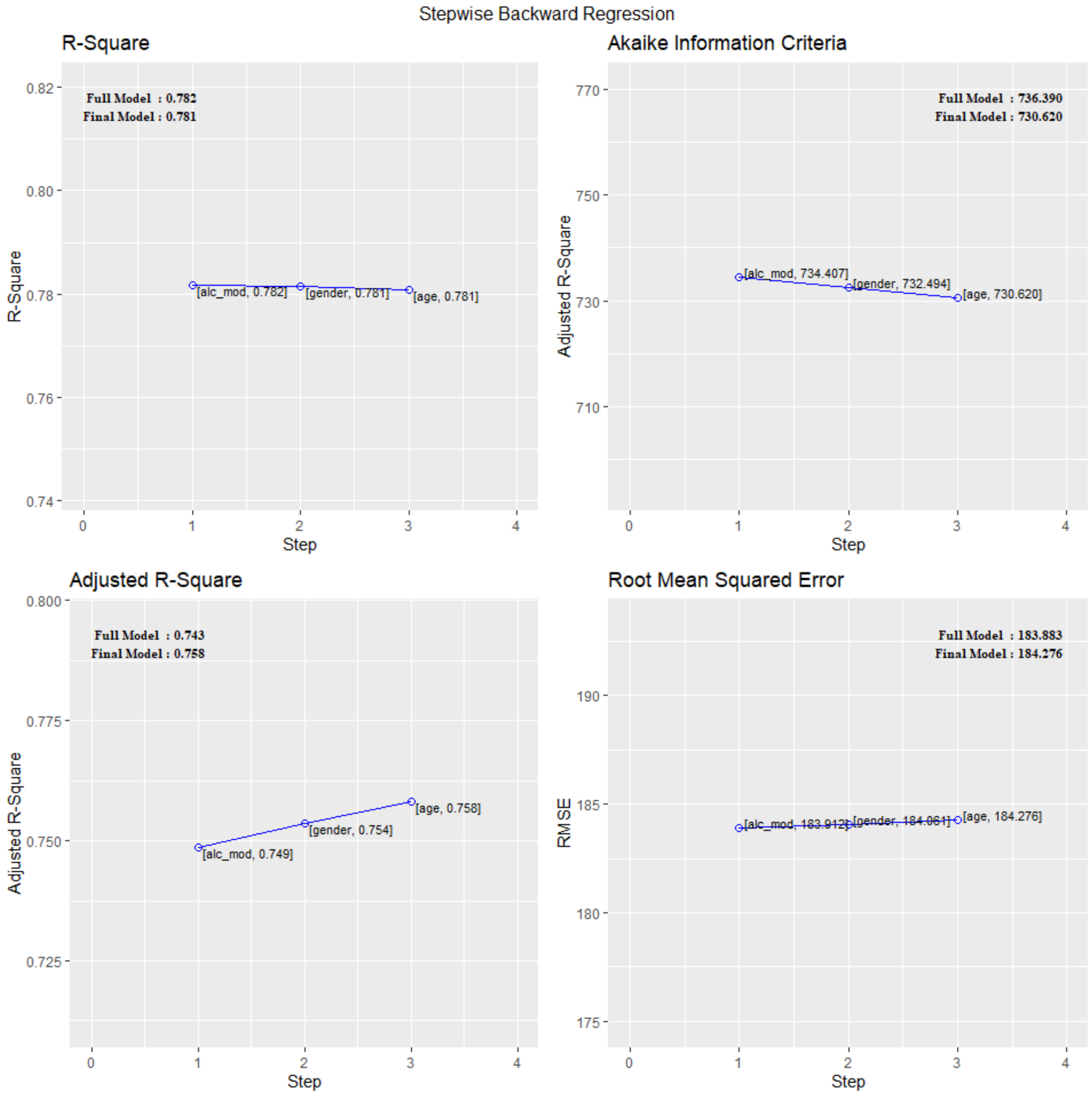
```
                             Stepwise Summary
-------------------------------------------------------------------------
Step    Variable        AIC        SBC       SBIC        R2       Adj. R2
-------------------------------------------------------------------------
 0      Full Model    736.390    756.280    586.665    0.78184    0.74305
 1      alc_mod       734.407    752.308    584.276    0.78177    0.74856
 2      gender        732.494    748.406    581.938    0.78142    0.75351
 3      age           730.620    744.543    579.638    0.78091    0.75808
-------------------------------------------------------------------------

Final Model Output
------------------

                           Model Summary
-------------------------------------------------------------------
R                           0.884       RMSE               184.276
R-Squared                   0.781       MSE              38202.426
Adj. R-Squared              0.758       Coef. Var           27.839
Pred R-Squared              0.700       AIC                730.620
MAE                       137.656       SBC                744.543
-------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                 ANOVA
-----------------------------------------------------------------------
                  Sum of
                 Squares        DF    Mean Square      F          Sig.
-----------------------------------------------------------------------
Regression    6535804.090        5    1307160.818    34.217    0.0000
Residual      1833716.447       48      38202.426
Total         8369520.537       53
-----------------------------------------------------------------------

                                        Parameter Estimates
-------------------------------------------------------------------------------------
------------
      model         Beta    Std. Error    Std. Beta      t        Sig         lower
upper
-------------------------------------------------------------------------------------
------------
(Intercept)    -1178.330       208.682                -5.647    0.000    -1597.914
-758.746
        bcs       59.864        23.060        0.241     2.596    0.012       13.498
```

```
106.230
     pindex        8.924          1.808          0.380     4.935    0.000         5.288
12.559
enzyme_test        9.748          1.656          0.521     5.887    0.000         6.419
13.077
 liver_test       58.064         40.144          0.156     1.446    0.155       -22.652
138.779
  alc_heavy      317.848         71.634          0.314     4.437    0.000       173.818
461.878
-------------------------------------------------------------------------------------
------------
```

Plot

Hide

```
model <- lm(y ~ ., data = surgical)
k <- ols_step_backward_p(model)
plot(k)
```



Detailed Output

Hide

```
# stepwise backward regression
model <- lm(y ~ ., data = surgical)
ols_step_backward_p(model, details = TRUE)
```

```
Backward Elimination Method
---------------------------

Candidate Terms:

1. bcs
2. pindex
3. enzyme_test
4. liver_test
5. age
6. gender
7. alc_mod
8. alc_heavy


Step   => 0
Model  => y ~ bcs + pindex + enzyme_test + liver_test + age + gender + alc_mod + alc
_heavy
R2     => 0.782

Initiating stepwise selection...

Step     => 1
Removed  => alc_mod
Model    => y ~ bcs + pindex + enzyme_test + liver_test + age + gender + alc_heavy
R2       => 0.78177

Step     => 2
Removed  => gender
Model    => y ~ bcs + pindex + enzyme_test + liver_test + age + alc_heavy
R2       => 0.78142

Step     => 3
Removed  => age
Model    => y ~ bcs + pindex + enzyme_test + liver_test + alc_heavy
R2       => 0.78091


No more variables to be removed.

Variables Removed:

=> alc_mod
=> gender
=> age

                             Stepwise Summary
-------------------------------------------------------------------------
```

```
Step    Variable        AIC        SBC       SBIC        R2       Adj. R2
--------------------------------------------------------------------------
 0      Full Model    736.390    756.280    586.665    0.78184    0.74305
 1      alc_mod       734.407    752.308    584.276    0.78177    0.74856
 2      gender        732.494    748.406    581.938    0.78142    0.75351
 3      age           730.620    744.543    579.638    0.78091    0.75808
--------------------------------------------------------------------------

Final Model Output
------------------

                           Model Summary
--------------------------------------------------------------------
R                          0.884       RMSE                 184.276
R-Squared                  0.781       MSE                38202.426
Adj. R-Squared             0.758       Coef. Var             27.839
Pred R-Squared             0.700       AIC                  730.620
MAE                      137.656       SBC                  744.543
--------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                 ANOVA
------------------------------------------------------------------------
                  Sum of
                 Squares        DF    Mean Square      F         Sig.
------------------------------------------------------------------------
Regression    6535804.090        5    1307160.818    34.217    0.0000
Residual      1833716.447       48      38202.426
Total         8369520.537       53
------------------------------------------------------------------------

                                      Parameter Estimates
------------------------------------------------------------------------------------
------------
      model          Beta    Std. Error    Std. Beta      t        Sig         lower
upper
------------------------------------------------------------------------------------
------------
(Intercept)    -1178.330       208.682                 -5.647    0.000    -1597.914
-758.746
        bcs       59.864        23.060         0.241    2.596    0.012       13.498
106.230
     pindex        8.924         1.808         0.380    4.935    0.000        5.288
12.559
enzyme_test        9.748         1.656         0.521    5.887    0.000        6.419
```

```
13.077
 liver_test       58.064        40.144        0.156     1.446    0.155      -22.652
138.779
  alc_heavy      317.848        71.634        0.314     4.437    0.000      173.818
461.878
-------------------------------------------------------------------------------------
------------
```

Stepwise Regression

Build regression model from a set of candidate predictor variables by entering and removing predictors based on p values, in a stepwise manner until there is no variable left to enter or remove any more. The model should include all the candidate predictor variables. If details is set to `TRUE`, each step is displayed.

Variable Selection

Hide

```
# stepwise regression
model <- lm(y ~ ., data = surgical)
ols_step_both_p(model)
```

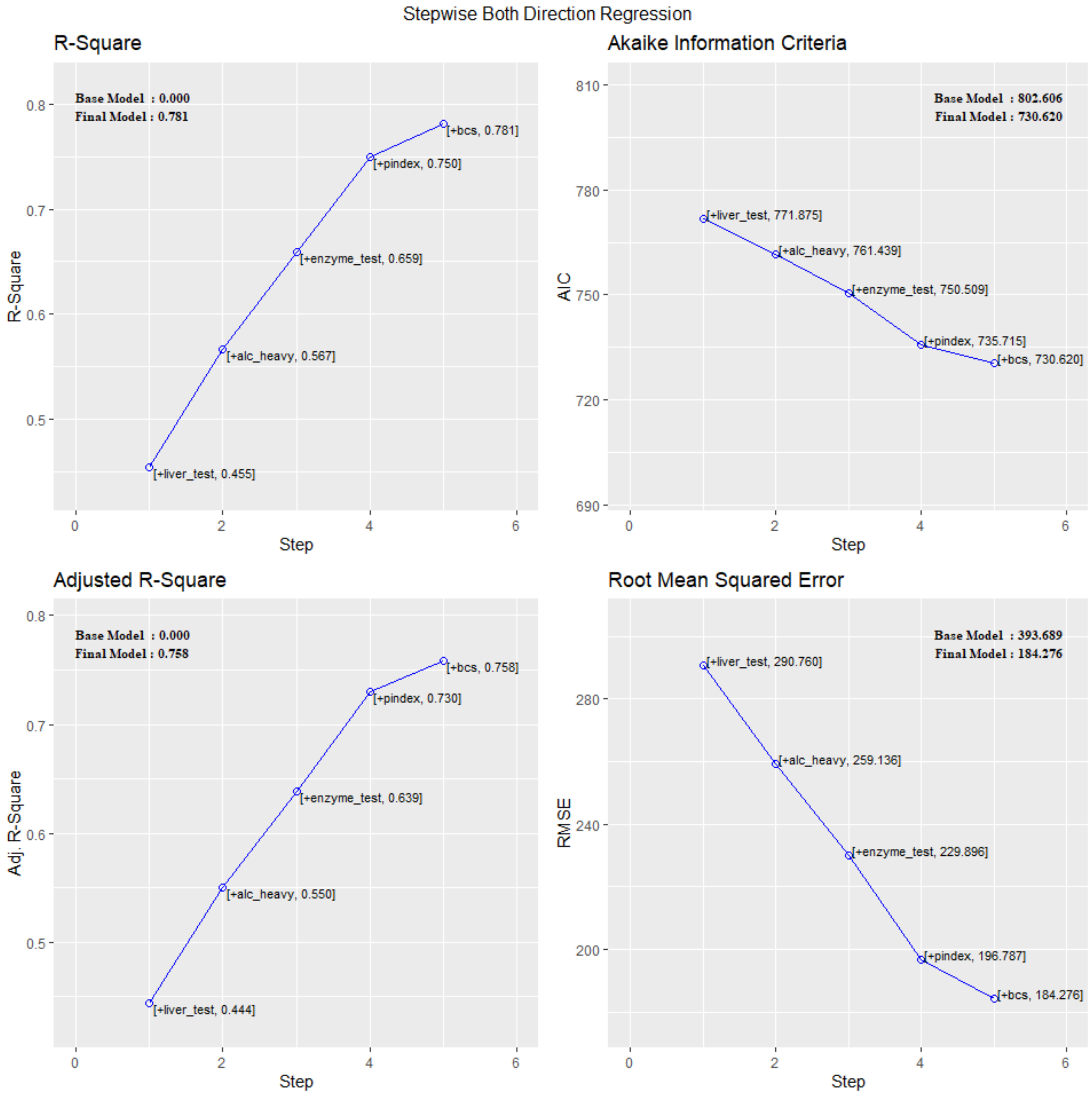
```
                                Stepwise Summary
------------------------------------------------------------------------------
Step    Variable             AIC        SBC       SBIC        R2       Adj. R2
------------------------------------------------------------------------------
 0      Base Model         802.606    806.584    646.794    0.00000    0.00000
 1      liver_test (+)     771.875    777.842    616.009    0.45454    0.44405
 2      alc_heavy (+)      761.439    769.395    605.506    0.56674    0.54975
 3      enzyme_test (+)    750.509    760.454    595.297    0.65900    0.63854
 4      pindex (+)         735.715    747.649    582.943    0.75015    0.72975
 5      bcs (+)            730.620    744.543    579.638    0.78091    0.75808
------------------------------------------------------------------------------

Final Model Output
------------------

                           Model Summary
-------------------------------------------------------------------
R                          0.884       RMSE                184.276
R-Squared                  0.781       MSE               38202.426
Adj. R-Squared             0.758       Coef. Var            27.839
Pred R-Squared             0.700       AIC                 730.620
MAE                      137.656       SBC                 744.543
-------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                 ANOVA
-----------------------------------------------------------------------
                   Sum of
                  Squares        DF    Mean Square      F         Sig.
-----------------------------------------------------------------------
Regression    6535804.090         5    1307160.818    34.217    0.0000
Residual      1833716.447        48      38202.426
Total         8369520.537        53
-----------------------------------------------------------------------

                                      Parameter Estimates
------------------------------------------------------------------------------------
------------
      model         Beta    Std. Error    Std. Beta      t        Sig         lower
upper
------------------------------------------------------------------------------------
------------
(Intercept)    -1178.330       208.682                 -5.647    0.000    -1597.914
```

```
-758.746
 liver_test       58.064         40.144         0.156      1.446    0.155      -22.652
138.779
  alc_heavy      317.848         71.634         0.314      4.437    0.000      173.818
461.878
enzyme_test        9.748          1.656         0.521      5.887    0.000        6.419
13.077
     pindex        8.924          1.808         0.380      4.935    0.000        5.288
12.559
        bcs       59.864         23.060         0.241      2.596    0.012       13.498
106.230
----------------------------------------------------------------------------------
------------
```

Plot

Hide

```
model <- lm(y ~ ., data = surgical)
k <- ols_step_both_p(model)
plot(k)
```



Detailed Output

Hide

```
# stepwise regression
model <- lm(y ~ ., data = surgical)
ols_step_both_p(model, details = TRUE)
```

```
Stepwise Selection Method
-------------------------

Candidate Terms:

1. bcs
2. pindex
3. enzyme_test
4. liver_test
5. age
6. gender
7. alc_mod
8. alc_heavy


Step   => 0
Model  => y ~ 1
R2     => 0

Initiating stepwise selection...

Step      => 1
Selected  => liver_test
Model     => y ~ liver_test
R2        => 0.455

Step      => 2
Selected  => alc_heavy
Model     => y ~ liver_test + alc_heavy
R2        => 0.567

Step      => 3
Selected  => enzyme_test
Model     => y ~ liver_test + alc_heavy + enzyme_test
R2        => 0.659

Step      => 4
Selected  => pindex
Model     => y ~ liver_test + alc_heavy + enzyme_test + pindex
R2        => 0.75

Step      => 5
Selected  => bcs
Model     => y ~ liver_test + alc_heavy + enzyme_test + pindex + bcs
R2        => 0.781


No more variables to be added or removed.
```

```
                                Stepwise Summary
------------------------------------------------------------------------------
Step    Variable             AIC        SBC       SBIC        R2       Adj. R2
------------------------------------------------------------------------------
 0      Base Model         802.606    806.584    646.794    0.00000    0.00000
 1      liver_test (+)     771.875    777.842    616.009    0.45454    0.44405
 2      alc_heavy (+)      761.439    769.395    605.506    0.56674    0.54975
 3      enzyme_test (+)    750.509    760.454    595.297    0.65900    0.63854
 4      pindex (+)         735.715    747.649    582.943    0.75015    0.72975
 5      bcs (+)            730.620    744.543    579.638    0.78091    0.75808
------------------------------------------------------------------------------

Final Model Output
------------------

                           Model Summary
-------------------------------------------------------------------
R                           0.884       RMSE                184.276
R-Squared                   0.781       MSE               38202.426
Adj. R-Squared              0.758       Coef. Var            27.839
Pred R-Squared              0.700       AIC                 730.620
MAE                       137.656       SBC                 744.543
-------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                 ANOVA
-----------------------------------------------------------------------
                   Sum of
                  Squares        DF    Mean Square      F         Sig.
-----------------------------------------------------------------------
Regression    6535804.090         5    1307160.818    34.217    0.0000
Residual      1833716.447        48      38202.426
Total         8369520.537        53
-----------------------------------------------------------------------

                                       Parameter Estimates
-------------------------------------------------------------------------------------
------------
      model         Beta    Std. Error    Std. Beta      t        Sig         lower
upper
-------------------------------------------------------------------------------------
------------
(Intercept)    -1178.330       208.682                 -5.647    0.000    -1597.914
-758.746
```

```
 liver_test       58.064         40.144         0.156      1.446    0.155      -22.652
138.779
  alc_heavy      317.848         71.634         0.314      4.437    0.000      173.818
461.878
enzyme_test        9.748          1.656         0.521      5.887    0.000        6.419
13.077
     pindex        8.924          1.808         0.380      4.935    0.000        5.288
12.559
        bcs       59.864         23.060         0.241      2.596    0.012       13.498
106.230
-------------------------------------------------------------------------------------
------------
```

Stepwise AIC Forward Regression

Build regression model from a set of candidate predictor variables by entering predictors based on Akaike Information Criteria, in a stepwise manner until there is no variable left to enter any more.

The model should include all the candidate predictor variables. If details is set to `TRUE` , each step is displayed.

Variable Selection

Hide

```
# stepwise aic forward regression
model <- lm(y ~ ., data = surgical)
ols_step_forward_aic(model)
```

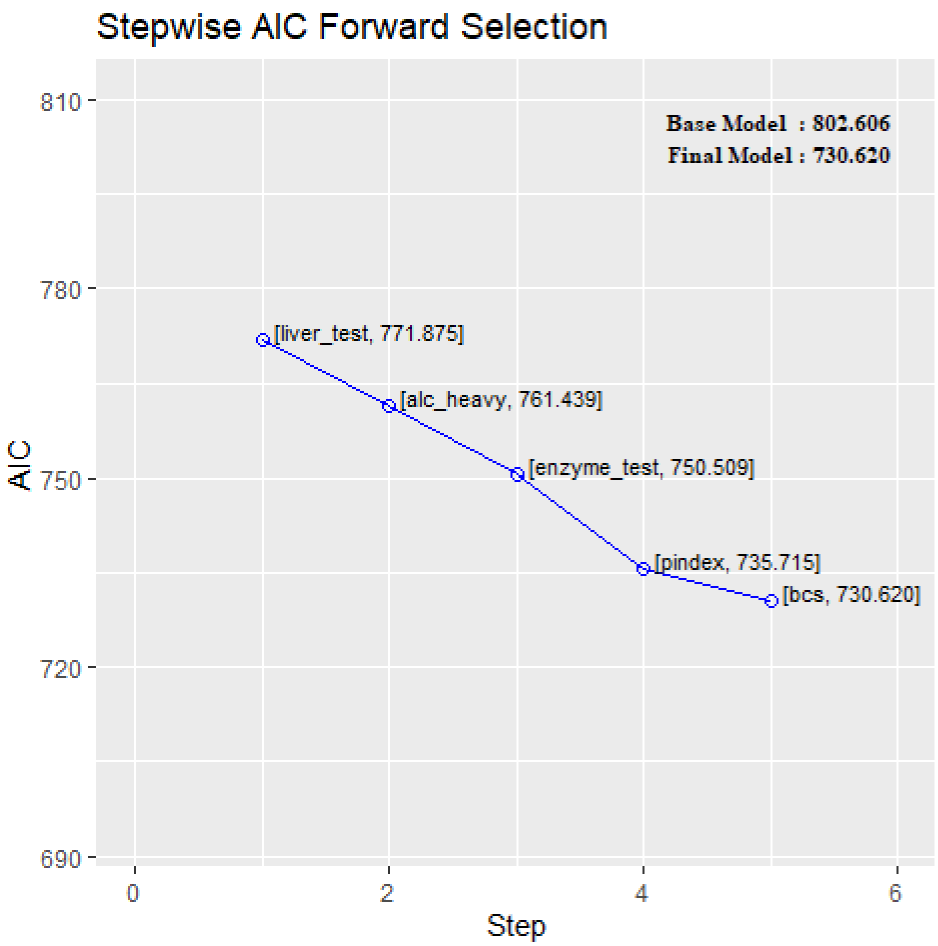
```
                              Stepwise Summary
--------------------------------------------------------------------------
Step    Variable         AIC        SBC        SBIC         R2       Adj. R2
--------------------------------------------------------------------------
 0      Base Model     802.606    806.584    646.794    0.00000    0.00000
 1      liver_test     771.875    777.842    616.009    0.45454    0.44405
 2      alc_heavy      761.439    769.395    605.506    0.56674    0.54975
 3      enzyme_test    750.509    760.454    595.297    0.65900    0.63854
 4      pindex         735.715    747.649    582.943    0.75015    0.72975
 5      bcs            730.620    744.543    579.638    0.78091    0.75808
--------------------------------------------------------------------------

Final Model Output
------------------

                           Model Summary
-------------------------------------------------------------------
R                          0.884       RMSE                 184.276
R-Squared                  0.781       MSE                38202.426
Adj. R-Squared             0.758       Coef. Var             27.839
Pred R-Squared             0.700       AIC                  730.620
MAE                      137.656       SBC                  744.543
-------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                 ANOVA
-----------------------------------------------------------------------
                   Sum of
                  Squares        DF    Mean Square      F         Sig.
-----------------------------------------------------------------------
Regression    6535804.090         5    1307160.818    34.217    0.0000
Residual      1833716.447        48      38202.426
Total         8369520.537        53
-----------------------------------------------------------------------

                                      Parameter Estimates
------------------------------------------------------------------------------------
------------
      model         Beta    Std. Error    Std. Beta      t         Sig         lower
upper
------------------------------------------------------------------------------------
------------
(Intercept)    -1178.330       208.682                 -5.647    0.000    -1597.914
```

```
-758.746
 liver_test       58.064         40.144          0.156       1.446      0.155        -22.652
138.779
  alc_heavy      317.848         71.634          0.314       4.437      0.000        173.818
461.878
enzyme_test        9.748          1.656          0.521       5.887      0.000          6.419
13.077
     pindex        8.924          1.808          0.380       4.935      0.000          5.288
12.559
        bcs       59.864         23.060          0.241       2.596      0.012         13.498
106.230
-------------------------------------------------------------------------------------
------------
```

Plot

Hide

```
model <- lm(y ~ ., data = surgical)
k <- ols_step_forward_aic(model)
plot(k)
```



Detailed Output

Hide

```
# stepwise aic forward regression
model <- lm(y ~ ., data = surgical)
ols_step_forward_aic(model, details = TRUE)
```

```
Forward Selection Method
------------------------

Candidate Terms:

1. bcs
2. pindex
3. enzyme_test
4. liver_test
5. age
6. gender
7. alc_mod
8. alc_heavy


Step     => 0
Model    => y ~ 1
AIC      => 802.606

Initiating stepwise selection...

                       Table: Adding New Variables
-------------------------------------------------------------------------
Predictor      DF      AIC        SBC       SBIC        R2       Adj. R2
-------------------------------------------------------------------------
liver_test      1    771.875    777.842    616.009    0.45454     0.44405
enzyme_test     1    782.629    788.596    626.220    0.33435     0.32154
pindex          1    794.100    800.067    637.196    0.17680     0.16097
alc_heavy       1    794.301    800.268    637.389    0.17373     0.15784
bcs             1    797.697    803.664    640.655    0.12010     0.10318
alc_mod         1    802.828    808.795    645.601    0.03239     0.01378
gender          1    802.956    808.923    645.725    0.03009     0.01143
age             1    803.834    809.801    646.572    0.01420    -0.00476
-------------------------------------------------------------------------

Step     => 1
Added    => liver_test
Model    => y ~ liver_test
AIC      => 771.8753

                      Table: Adding New Variables
------------------------------------------------------------------------
Predictor      DF      AIC        SBC       SBIC        R2       Adj. R2
------------------------------------------------------------------------
alc_heavy       1    761.439    769.395    605.506    0.56674    0.54975
enzyme_test     1    762.077    770.033    606.090    0.56159    0.54440
pindex          1    770.387    778.343    613.737    0.48866    0.46861
alc_mod         1    771.141    779.097    614.435    0.48147    0.46113
```

```
gender          1    773.802    781.758    616.901    0.45528    0.43391
age             1    773.831    781.787    616.928    0.45498    0.43361
bcs             1    773.867    781.823    616.961    0.45462    0.43323
------------------------------------------------------------------------

Step     => 2
Added    => alc_heavy
Model    => y ~ liver_test + alc_heavy
AIC      => 761.4394

                      Table: Adding New Variables
------------------------------------------------------------------------
Predictor      DF      AIC        SBC       SBIC        R2       Adj. R2
------------------------------------------------------------------------
enzyme_test     1    750.509    760.454    595.297    0.65900    0.63854
pindex          1    756.125    766.070    600.225    0.62163    0.59892
bcs             1    763.063    773.008    606.379    0.56975    0.54394
age             1    763.110    773.055    606.421    0.56938    0.54354
alc_mod         1    763.428    773.373    606.704    0.56683    0.54084
gender          1    763.433    773.378    606.709    0.56679    0.54080
------------------------------------------------------------------------

Step     => 3
Added    => enzyme_test
Model    => y ~ liver_test + alc_heavy + enzyme_test
AIC      => 750.5089

                     Table: Adding New Variables
----------------------------------------------------------------------
Predictor    DF      AIC        SBC       SBIC        R2       Adj. R2
----------------------------------------------------------------------
pindex        1    735.715    747.649    582.943    0.75015    0.72975
bcs           1    750.782    762.716    595.377    0.66973    0.64277
alc_mod       1    752.403    764.337    596.743    0.65967    0.63189
age           1    752.416    764.350    596.755    0.65959    0.63180
gender        1    752.509    764.443    596.833    0.65900    0.63116
----------------------------------------------------------------------

Step     => 4
Added    => pindex
Model    => y ~ liver_test + alc_heavy + enzyme_test + pindex
AIC      => 735.7146

                     Table: Adding New Variables
----------------------------------------------------------------------
Predictor    DF      AIC        SBC       SBIC        R2       Adj. R2
----------------------------------------------------------------------
bcs           1    730.620    744.543    579.638    0.78091    0.75808
age           1    737.680    751.603    585.012    0.75030    0.72429
```

```
gender        1    737.712    751.635    585.036    0.75016    0.72413
alc_mod       1    737.713    751.636    585.037    0.75015    0.72413
----------------------------------------------------------------------

Step     => 5
Added    => bcs
Model    => y ~ liver_test + alc_heavy + enzyme_test + pindex + bcs
AIC      => 730.6204

                     Table: Adding New Variables
----------------------------------------------------------------------
Predictor    DF      AIC        SBC       SBIC        R2       Adj. R2
----------------------------------------------------------------------
age           1    732.494    748.406    581.938    0.78142    0.75351
gender        1    732.551    748.463    581.978    0.78119    0.75325
alc_mod       1    732.614    748.526    582.023    0.78093    0.75297
----------------------------------------------------------------------


No more variables to be added.

Variables Selected:

=> liver_test
=> alc_heavy
=> enzyme_test
=> pindex
=> bcs

                              Stepwise Summary
--------------------------------------------------------------------------
Step    Variable         AIC        SBC       SBIC        R2       Adj. R2
--------------------------------------------------------------------------
 0      Base Model     802.606    806.584    646.794    0.00000    0.00000
 1      liver_test     771.875    777.842    616.009    0.45454    0.44405
 2      alc_heavy      761.439    769.395    605.506    0.56674    0.54975
 3      enzyme_test    750.509    760.454    595.297    0.65900    0.63854
 4      pindex         735.715    747.649    582.943    0.75015    0.72975
 5      bcs            730.620    744.543    579.638    0.78091    0.75808
--------------------------------------------------------------------------

Final Model Output
------------------

                           Model Summary
-------------------------------------------------------------------
R                         0.884       RMSE                 184.276
R-Squared                 0.781       MSE                38202.426
Adj. R-Squared            0.758       Coef. Var             27.839
```

```
Pred R-Squared            0.700       AIC                  730.620
MAE                     137.656       SBC                  744.543
-------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                 ANOVA
-----------------------------------------------------------------------
                   Sum of
                  Squares        DF    Mean Square      F         Sig.
-----------------------------------------------------------------------
Regression    6535804.090         5    1307160.818    34.217    0.0000
Residual      1833716.447        48      38202.426
Total         8369520.537        53
-----------------------------------------------------------------------

                                      Parameter Estimates
------------------------------------------------------------------------------------
------------
      model         Beta    Std. Error    Std. Beta      t        Sig         lower
upper
------------------------------------------------------------------------------------
------------
(Intercept)    -1178.330       208.682                 -5.647    0.000    -1597.914
-758.746
 liver_test       58.064        40.144        0.156     1.446    0.155      -22.652
138.779
  alc_heavy      317.848        71.634        0.314     4.437    0.000      173.818
461.878
enzyme_test        9.748         1.656        0.521     5.887    0.000        6.419
13.077
     pindex        8.924         1.808        0.380     4.935    0.000        5.288
12.559
        bcs       59.864        23.060        0.241     2.596    0.012       13.498
106.230
------------------------------------------------------------------------------------
------------
```

Stepwise AIC Backward Regression

Build regression model from a set of candidate predictor variables by removing predictors based on Akaike Information Criteria, in a stepwise manner until there is no variable left to remove any more. The model should include all the candidate predictor variables. If details is set to `TRUE`, each step is displayed.

Variable Selection

Hide

```
# stepwise aic backward regression
model <- lm(y ~ ., data = surgical)
k <- ols_step_backward_aic(model)
k
```

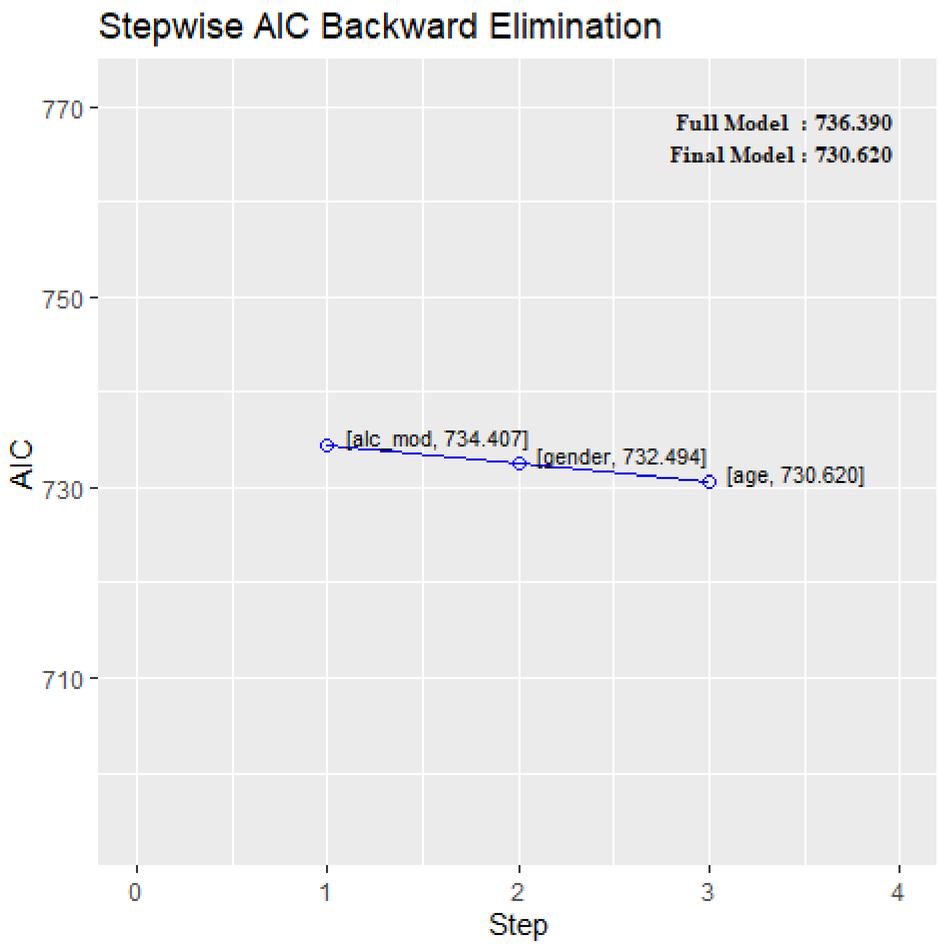
```
                             Stepwise Summary
-------------------------------------------------------------------------
Step    Variable        AIC        SBC       SBIC        R2       Adj. R2
-------------------------------------------------------------------------
 0      Full Model    736.390    756.280    586.665    0.78184    0.74305
 1      alc_mod       734.407    752.308    583.884    0.78177    0.74856
 2      gender        732.494    748.406    581.290    0.78142    0.75351
 3      age           730.620    744.543    578.844    0.78091    0.75808
-------------------------------------------------------------------------

Final Model Output
------------------

                           Model Summary
-------------------------------------------------------------------
R                         0.884       RMSE                 184.276
R-Squared                 0.781       MSE                38202.426
Adj. R-Squared            0.758       Coef. Var             27.839
Pred R-Squared            0.700       AIC                  730.620
MAE                     137.656       SBC                  744.543
-------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                 ANOVA
-----------------------------------------------------------------------
                 Sum of
                Squares        DF    Mean Square      F          Sig.
-----------------------------------------------------------------------
Regression    6535804.090         5    1307160.818    34.217    0.0000
Residual      1833716.447        48      38202.426
Total         8369520.537        53
-----------------------------------------------------------------------

                                      Parameter Estimates
------------------------------------------------------------------------------------
------------
      model         Beta    Std. Error    Std. Beta      t        Sig        lower
upper
------------------------------------------------------------------------------------
------------
(Intercept)    -1178.330       208.682                -5.647    0.000    -1597.914
-758.746
        bcs       59.864        23.060        0.241    2.596    0.012       13.498
```

```
106.230
     pindex        8.924         1.808         0.380      4.935     0.000         5.288
12.559
enzyme_test        9.748         1.656         0.521      5.887     0.000         6.419
13.077
 liver_test       58.064        40.144         0.156      1.446     0.155       -22.652
138.779
  alc_heavy      317.848        71.634         0.314      4.437     0.000       173.818
461.878
------------------------------------------------------------------------------------
------------
```

Plot

Hide

```
model <- lm(y ~ ., data = surgical)
k <- ols_step_backward_aic(model)
plot(k)
```



Detailed Output

Hide

```
# stepwise aic backward regression
model <- lm(y ~ ., data = surgical)
ols_step_backward_aic(model, details = TRUE)
```

```
Backward Elimination Method
---------------------------

Candidate Terms:

1. bcs
2. pindex
3. enzyme_test
4. liver_test
5. age
6. gender
7. alc_mod
8. alc_heavy


Step     => 0
Model    => y ~ bcs + pindex + enzyme_test + liver_test + age + gender + alc_mod + a
lc_heavy
AIC      => 736.3899

Initiating stepwise selection...

                   Table: Removing Existing Variables
------------------------------------------------------------------------
Predictor      DF      AIC        SBC       SBIC        R2       Adj. R2
------------------------------------------------------------------------
alc_mod         1    734.407    752.308    584.276    0.78177    0.74856
gender          1    734.478    752.379    584.323    0.78148    0.74823
age             1    734.544    752.445    584.367    0.78121    0.74792
liver_test      1    735.878    753.779    585.255    0.77574    0.74162
bcs             1    741.677    759.577    589.203    0.75032    0.71233
alc_heavy       1    749.210    767.111    594.541    0.71294    0.66926
pindex          1    756.624    774.525    600.014    0.67070    0.62059
enzyme_test     1    763.557    781.458    605.318    0.62559    0.56861
------------------------------------------------------------------------

Step     => 1
Removed  => alc_mod
Model    => y ~ bcs + pindex + enzyme_test + liver_test + age + gender + alc_heavy
AIC      => 734.4068

                   Table: Removing Existing Variables
------------------------------------------------------------------------
Predictor      DF      AIC        SBC       SBIC        R2       Adj. R2
------------------------------------------------------------------------
gender          1    732.494    748.406    581.938    0.78142    0.75351
age             1    732.551    748.463    581.978    0.78119    0.75325
liver_test      1    733.921    749.833    582.951    0.77556    0.74691
```

```
bcs              1    739.677    755.589    587.106    0.75032    0.71845
alc_heavy        1    750.486    766.398    595.217    0.69499    0.65605
pindex           1    754.759    770.671    598.530    0.66987    0.62773
enzyme_test      1    761.595    777.507    603.950    0.62532    0.57749
------------------------------------------------------------------------

Step     => 2
Removed  => gender
Model    => y ~ bcs + pindex + enzyme_test + liver_test + age + alc_heavy
AIC      => 732.4942

                   Table: Removing Existing Variables
------------------------------------------------------------------------
Predictor      DF      AIC        SBC        SBIC        R2       Adj. R2
------------------------------------------------------------------------
age             1    730.620    744.543    579.638    0.78091    0.75808
liver_test      1    732.339    746.262    580.934    0.77382    0.75026
bcs             1    737.680    751.603    585.012    0.75030    0.72429
alc_heavy       1    748.486    762.409    593.500    0.69499    0.66322
pindex          1    752.777    766.700    596.959    0.66976    0.63536
enzyme_test     1    759.596    773.518    602.553    0.62532    0.58629
------------------------------------------------------------------------

Step     => 3
Removed  => age
Model    => y ~ bcs + pindex + enzyme_test + liver_test + alc_heavy
AIC      => 730.6204

                   Table: Removing Existing Variables
------------------------------------------------------------------------
Predictor      DF      AIC        SBC        SBIC        R2       Adj. R2
------------------------------------------------------------------------
liver_test      1    730.924    742.858    579.087    0.77136    0.75269
bcs             1    735.715    747.649    582.943    0.75015    0.72975
alc_heavy       1    747.181    759.114    592.362    0.69104    0.66582
pindex          1    750.782    762.716    595.377    0.66973    0.64277
enzyme_test     1    757.971    769.905    601.477    0.62270    0.59190
------------------------------------------------------------------------


No more variables to be removed.

Variables Removed:

=> alc_mod
=> gender
=> age

                             Stepwise Summary
```

```
-------------------------------------------------------------------------
Step    Variable        AIC        SBC       SBIC        R2       Adj. R2
-------------------------------------------------------------------------
 0      Full Model    736.390    756.280    586.665    0.78184    0.74305
 1      alc_mod       734.407    752.308    583.884    0.78177    0.74856
 2      gender        732.494    748.406    581.290    0.78142    0.75351
 3      age           730.620    744.543    578.844    0.78091    0.75808
-------------------------------------------------------------------------

Final Model Output
------------------

                           Model Summary
-------------------------------------------------------------------
R                         0.884       RMSE                 184.276
R-Squared                 0.781       MSE                38202.426
Adj. R-Squared            0.758       Coef. Var             27.839
Pred R-Squared            0.700       AIC                  730.620
MAE                     137.656       SBC                  744.543
-------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                 ANOVA
-----------------------------------------------------------------------
                   Sum of
                  Squares        DF    Mean Square      F          Sig.
-----------------------------------------------------------------------
Regression    6535804.090         5    1307160.818    34.217    0.0000
Residual      1833716.447        48      38202.426
Total         8369520.537        53
-----------------------------------------------------------------------

                                      Parameter Estimates
------------------------------------------------------------------------------------
------------
      model         Beta    Std. Error    Std. Beta      t        Sig         lower
upper
------------------------------------------------------------------------------------
------------
(Intercept)    -1178.330      208.682                   -5.647    0.000    -1597.914
-758.746
        bcs       59.864       23.060        0.241     2.596    0.012       13.498
106.230
     pindex        8.924        1.808        0.380     4.935    0.000        5.288
12.559
```

```
enzyme_test        9.748          1.656         0.521      5.887    0.000         6.419
13.077
 liver_test       58.064         40.144         0.156      1.446    0.155       -22.652
138.779
  alc_heavy      317.848         71.634         0.314      4.437    0.000       173.818
461.878
-------------------------------------------------------------------------------------
------------
```

Stepwise AIC Regression

Build regression model from a set of candidate predictor variables by entering and removing predictors based on Akaike Information Criteria, in a stepwise manner until there is no variable left to enter or remove any more. The model should include all the candidate predictor variables. If details is set to `TRUE` , each step is displayed.

Variable Selection

Hide

```
# stepwise aic regression
model <- lm(y ~ ., data = surgical)
ols_step_both_aic(model)
```

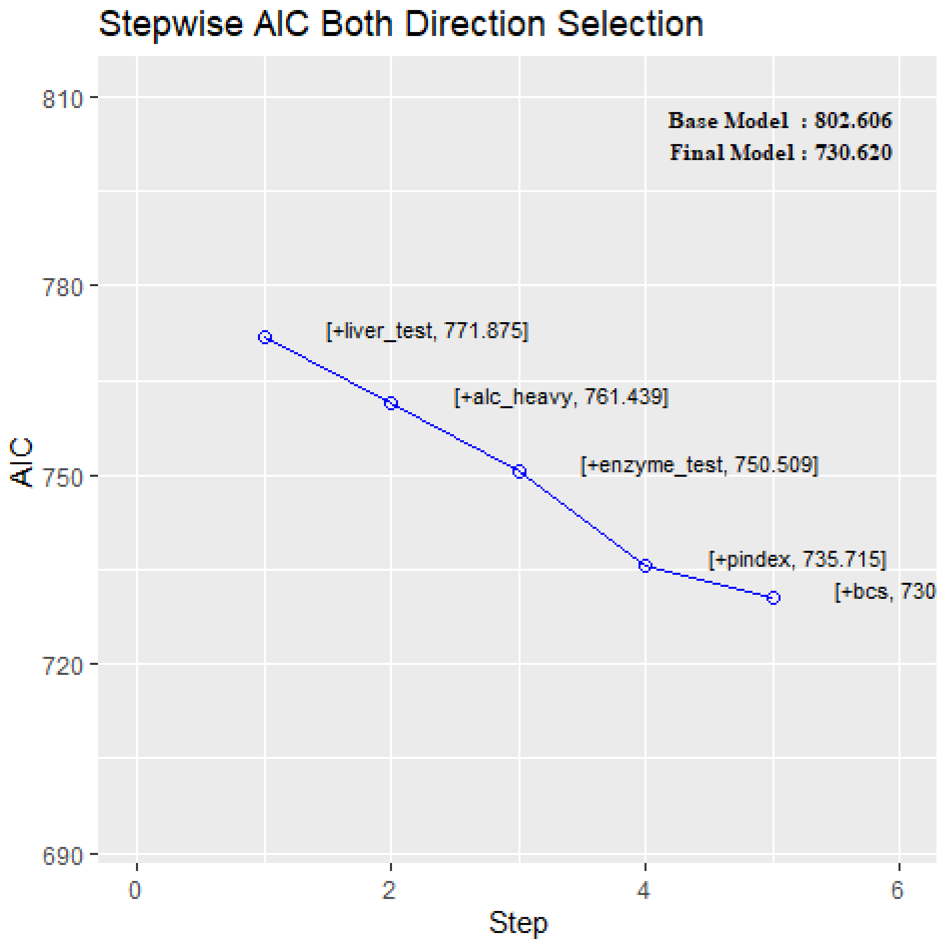
```
                              Stepwise Summary
------------------------------------------------------------------------------
Step    Variable             AIC        SBC       SBIC        R2       Adj. R2
------------------------------------------------------------------------------
 0      Base Model         802.606    806.584    646.794    0.00000    0.00000
 1      liver_test (+)     771.875    777.842    616.009    0.45454    0.44405
 2      alc_heavy (+)      761.439    769.395    605.506    0.56674    0.54975
 3      enzyme_test (+)    750.509    760.454    595.297    0.65900    0.63854
 4      pindex (+)         735.715    747.649    582.943    0.75015    0.72975
 5      bcs (+)            730.620    744.543    579.638    0.78091    0.75808
------------------------------------------------------------------------------

Final Model Output
------------------

                         Model Summary
-------------------------------------------------------------------
R                          0.884       RMSE                184.276
R-Squared                  0.781       MSE               38202.426
Adj. R-Squared             0.758       Coef. Var            27.839
Pred R-Squared             0.700       AIC                 730.620
MAE                      137.656       SBC                 744.543
-------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                ANOVA
-----------------------------------------------------------------------
                  Sum of
                 Squares        DF    Mean Square      F         Sig.
-----------------------------------------------------------------------
Regression    6535804.090        5    1307160.818    34.217    0.0000
Residual      1833716.447       48      38202.426
Total         8369520.537       53
-----------------------------------------------------------------------

                                     Parameter Estimates
-------------------------------------------------------------------------------------
------------
      model         Beta    Std. Error    Std. Beta      t        Sig         lower
upper
-------------------------------------------------------------------------------------
------------
(Intercept)    -1178.330       208.682                 -5.647    0.000    -1597.914
```

```
-758.746
 liver_test       58.064         40.144          0.156       1.446     0.155        -22.652
138.779
  alc_heavy      317.848         71.634          0.314       4.437     0.000        173.818
461.878
enzyme_test        9.748          1.656          0.521       5.887     0.000          6.419
13.077
     pindex        8.924          1.808          0.380       4.935     0.000          5.288
12.559
        bcs       59.864         23.060          0.241       2.596     0.012         13.498
106.230
-----------------------------------------------------------------------------------
------------
```

Plot

Hide

```
model <- lm(y ~ ., data = surgical)
k <- ols_step_both_aic(model)
plot(k)
```



Detailed Output

Hide

```
# stepwise aic regression
model <- lm(y ~ ., data = surgical)
ols_step_both_aic(model, details = TRUE)
```

```
Stepwise Selection Method
-------------------------

Candidate Terms:

1. bcs
2. pindex
3. enzyme_test
4. liver_test
5. age
6. gender
7. alc_mod
8. alc_heavy


Step     => 0
Model    => y ~ 1
AIC      => 802.606

Initiating stepwise selection...

                       Table: Adding New Variables
-------------------------------------------------------------------------
Predictor      DF      AIC        SBC       SBIC        R2       Adj. R2
-------------------------------------------------------------------------
bcs             1    797.697    803.664    640.655    0.12010     0.10318
pindex          1    794.100    800.067    637.196    0.17680     0.16097
enzyme_test     1    782.629    788.596    626.220    0.33435     0.32154
liver_test      1    771.875    777.842    616.009    0.45454     0.44405
age             1    803.834    809.801    646.572    0.01420    -0.00476
gender          1    802.956    808.923    645.725    0.03009     0.01143
alc_mod         1    802.828    808.795    645.601    0.03239     0.01378
alc_heavy       1    794.301    800.268    637.389    0.17373     0.15784
-------------------------------------------------------------------------

Step     => 1
Added    => liver_test
Model    => y ~ liver_test
AIC      => 771.8753

                      Table: Adding New Variables
------------------------------------------------------------------------
Predictor      DF      AIC        SBC       SBIC        R2       Adj. R2
------------------------------------------------------------------------
bcs             1    773.867    781.823    616.961    0.45462    0.43323
pindex          1    770.387    778.343    613.737    0.48866    0.46861
enzyme_test     1    762.077    770.033    606.090    0.56159    0.54440
age             1    773.831    781.787    616.928    0.45498    0.43361
```

```
gender          1    773.802    781.758    616.901    0.45528    0.43391
alc_mod         1    771.141    779.097    614.435    0.48147    0.46113
alc_heavy       1    761.439    769.395    605.506    0.56674    0.54975
------------------------------------------------------------------------

Step     => 2
Added    => alc_heavy
Model    => y ~ liver_test + alc_heavy
AIC      => 761.4394

                  Table: Removing Existing Variables
-----------------------------------------------------------------------
Predictor     DF      AIC        SBC       SBIC        R2       Adj. R2
-----------------------------------------------------------------------
liver_test     1    794.301    800.268    637.389    0.17373    0.15784
alc_heavy      1    771.875    777.842    616.009    0.45454    0.44405
-----------------------------------------------------------------------

                      Table: Adding New Variables
------------------------------------------------------------------------
Predictor      DF      AIC        SBC       SBIC        R2       Adj. R2
------------------------------------------------------------------------
bcs             1    763.063    773.008    606.379    0.56975    0.54394
pindex          1    756.125    766.070    600.225    0.62163    0.59892
enzyme_test     1    750.509    760.454    595.297    0.65900    0.63854
age             1    763.110    773.055    606.421    0.56938    0.54354
gender          1    763.433    773.378    606.709    0.56679    0.54080
alc_mod         1    763.428    773.373    606.704    0.56683    0.54084
------------------------------------------------------------------------

Step     => 3
Added    => enzyme_test
Model    => y ~ liver_test + alc_heavy + enzyme_test
AIC      => 750.5089

                   Table: Removing Existing Variables
------------------------------------------------------------------------
Predictor      DF      AIC        SBC       SBIC        R2       Adj. R2
------------------------------------------------------------------------
liver_test      1    773.555    781.511    616.671    0.45777    0.43650
alc_heavy       1    762.077    770.033    606.090    0.56159    0.54440
enzyme_test     1    761.439    769.395    605.506    0.56674    0.54975
------------------------------------------------------------------------


                     Table: Adding New Variables
----------------------------------------------------------------------
Predictor    DF      AIC        SBC       SBIC        R2       Adj. R2
----------------------------------------------------------------------
bcs           1    750.782    762.716    595.377    0.66973    0.64277
```

```
pindex        1    735.715    747.649    582.943    0.75015    0.72975
age           1    752.416    764.350    596.755    0.65959    0.63180
gender        1    752.509    764.443    596.833    0.65900    0.63116
alc_mod       1    752.403    764.337    596.743    0.65967    0.63189
----------------------------------------------------------------------

Step     => 4
Added    => pindex
Model    => y ~ liver_test + alc_heavy + enzyme_test + pindex
AIC      => 735.7146

                   Table: Removing Existing Variables
------------------------------------------------------------------------
Predictor      DF      AIC        SBC       SBIC        R2       Adj. R2
------------------------------------------------------------------------
liver_test      1    748.167    758.112    593.257    0.67347    0.65388
alc_heavy       1    755.099    765.044    599.321    0.62875    0.60647
enzyme_test     1    756.125    766.070    600.225    0.62163    0.59892
pindex          1    750.509    760.454    595.297    0.65900    0.63854
------------------------------------------------------------------------

                     Table: Adding New Variables
----------------------------------------------------------------------
Predictor    DF      AIC        SBC       SBIC        R2       Adj. R2
----------------------------------------------------------------------
bcs           1    730.620    744.543    579.638    0.78091    0.75808
age           1    737.680    751.603    585.012    0.75030    0.72429
gender        1    737.712    751.635    585.036    0.75016    0.72413
alc_mod       1    737.713    751.636    585.037    0.75015    0.72413
----------------------------------------------------------------------

Step     => 5
Added    => bcs
Model    => y ~ liver_test + alc_heavy + enzyme_test + pindex + bcs
AIC      => 730.6204

                   Table: Removing Existing Variables
------------------------------------------------------------------------
Predictor      DF      AIC        SBC       SBIC        R2       Adj. R2
------------------------------------------------------------------------
liver_test      1    730.924    742.858    579.087    0.77136    0.75269
alc_heavy       1    747.181    759.114    592.362    0.69104    0.66582
enzyme_test     1    757.971    769.905    601.477    0.62270    0.59190
pindex          1    750.782    762.716    595.377    0.66973    0.64277
bcs             1    735.715    747.649    582.943    0.75015    0.72975
------------------------------------------------------------------------

                     Table: Adding New Variables
----------------------------------------------------------------------
```

```
Predictor    DF      AIC        SBC       SBIC        R2       Adj. R2
----------------------------------------------------------------------
age           1    732.494    748.406    581.938    0.78142    0.75351
gender        1    732.551    748.463    581.978    0.78119    0.75325
alc_mod       1    732.614    748.526    582.023    0.78093    0.75297
----------------------------------------------------------------------


No more variables to be added or removed.

Variables Selected:

=> liver_test
=> alc_heavy
=> enzyme_test
=> pindex
=> bcs

                                Stepwise Summary
------------------------------------------------------------------------------
Step    Variable             AIC        SBC       SBIC        R2       Adj. R2
------------------------------------------------------------------------------
 0      Base Model         802.606    806.584    646.794    0.00000    0.00000
 1      liver_test (+)     771.875    777.842    616.009    0.45454    0.44405
 2      alc_heavy (+)      761.439    769.395    605.506    0.56674    0.54975
 3      enzyme_test (+)    750.509    760.454    595.297    0.65900    0.63854
 4      pindex (+)         735.715    747.649    582.943    0.75015    0.72975
 5      bcs (+)            730.620    744.543    579.638    0.78091    0.75808
------------------------------------------------------------------------------

Final Model Output
------------------

                           Model Summary
-------------------------------------------------------------------
R                         0.884       RMSE                 184.276
R-Squared                 0.781       MSE                38202.426
Adj. R-Squared            0.758       Coef. Var             27.839
Pred R-Squared            0.700       AIC                  730.620
MAE                     137.656       SBC                  744.543
-------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                 ANOVA
-----------------------------------------------------------------------
```

```
                    Sum of
                   Squares        DF    Mean Square      F         Sig.
----------------------------------------------------------------------
Regression    6535804.090         5     1307160.818    34.217    0.0000
Residual      1833716.447        48       38202.426
Total         8369520.537        53
----------------------------------------------------------------------

                                      Parameter Estimates
-----------------------------------------------------------------------------------------
------------
      model         Beta    Std. Error    Std. Beta      t        Sig         lower
upper
-----------------------------------------------------------------------------------------
------------
(Intercept)    -1178.330       208.682                 -5.647    0.000    -1597.914
-758.746
 liver_test       58.064        40.144        0.156     1.446    0.155      -22.652
138.779
  alc_heavy      317.848        71.634        0.314     4.437    0.000      173.818
461.878
enzyme_test        9.748         1.656        0.521     5.887    0.000        6.419
13.077
     pindex        8.924         1.808        0.380     4.935    0.000        5.288
12.559
        bcs       59.864        23.060        0.241     2.596    0.012       13.498
106.230
-----------------------------------------------------------------------------------------
------------
```

Notes on stepwise

A fundamental problem with stepwise regression is that some real explanatory variables that have causal effects on the dependent variable may happen to not be statistically significant, while nuisance variables may be coincidentally significant. As a result, the model may fit the data well in-sample, but do poorly out-of-sample.

Many Big-Data researchers believe that, the larger the number of possible explanatory variables, the more useful is stepwise regression for selecting explanatory variables. The reality is that stepwise regression is less effective the larger the number of potential explanatory variables. Stepwise regression does not solve the Big-Data problem of too many explanatory variables. Big Data exacerbates the failings of stepwise regression. reference (https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0143-6)

Residual Diagnostics

Introduction

olsrr offers tools for detecting violation of standard regression assumptions. Here we take a look at residual diagnostics. The standard regression assumptions include the following about residuals/errors:

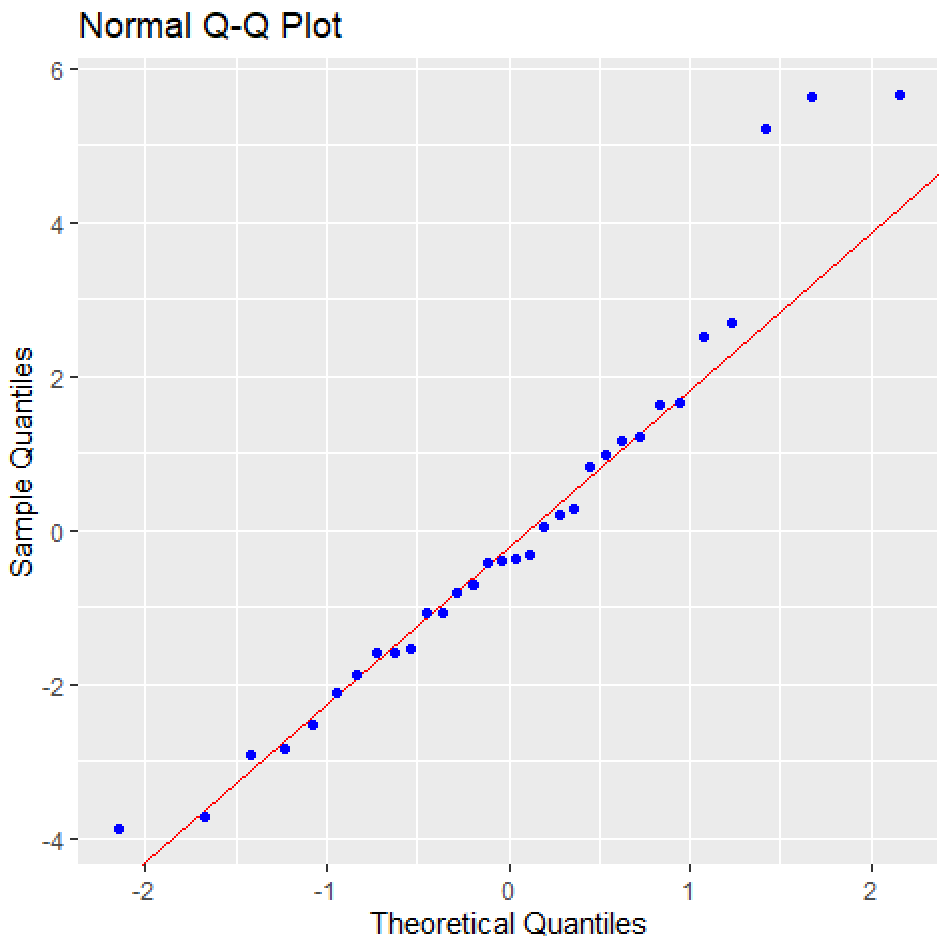
- The error has a normal distribution (normality assumption).
- The errors have mean zero.
- The errors have same but unknown variance (homoscedasticity assumption).
- The error are independent of each other (independent errors assumption).

Residual QQ Plot

Graph for detecting violation of normality assumption.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_resid_qq(model)
```



Residual Normality Test

Test for detecting violation of normality assumption.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_test_normality(model)
```

```
-----------------------------------------------
       Test             Statistic       pvalue  
-----------------------------------------------
Shapiro-Wilk              0.9366         0.0600 
Kolmogorov-Smirnov        0.1152         0.7464 
Cramer-von Mises          2.8122         0.0000 
Anderson-Darling          0.5859         0.1188 
-----------------------------------------------
```

Correlation between observed residuals and expected residuals under normality.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_test_correlation(model)
```

```
[1] 0.970066
```

Residual vs Fitted Values Plot

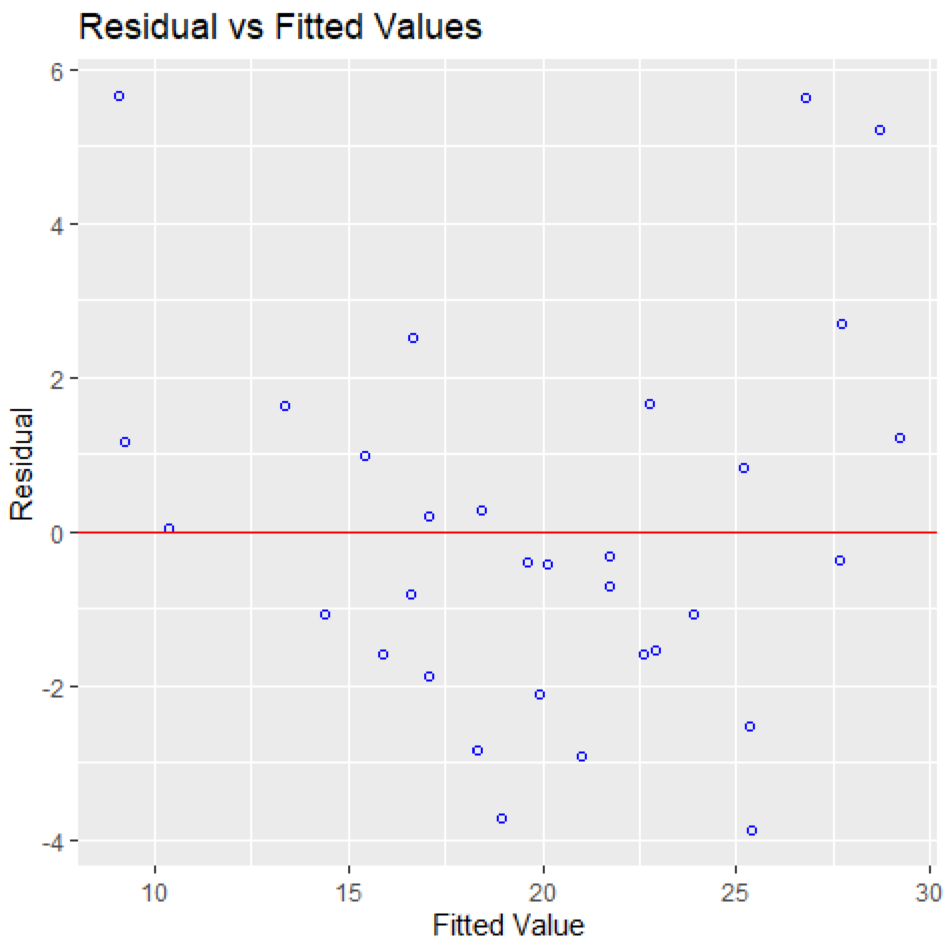
It is a scatter plot of residuals on the y axis and fitted values on the x axis to detect non-linearity, unequal error variances, and outliers.

Characteristics of a well behaved residual vs fitted plot:

- The residuals spread randomly around the 0 line indicating that the relationship is linear.
- The residuals form an approximate horizontal band around the 0 line indicating homogeneity of error variance.
- No one residual is visibly away from the random pattern of the residuals indicating that there are no outliers.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_resid_fit(model)
```

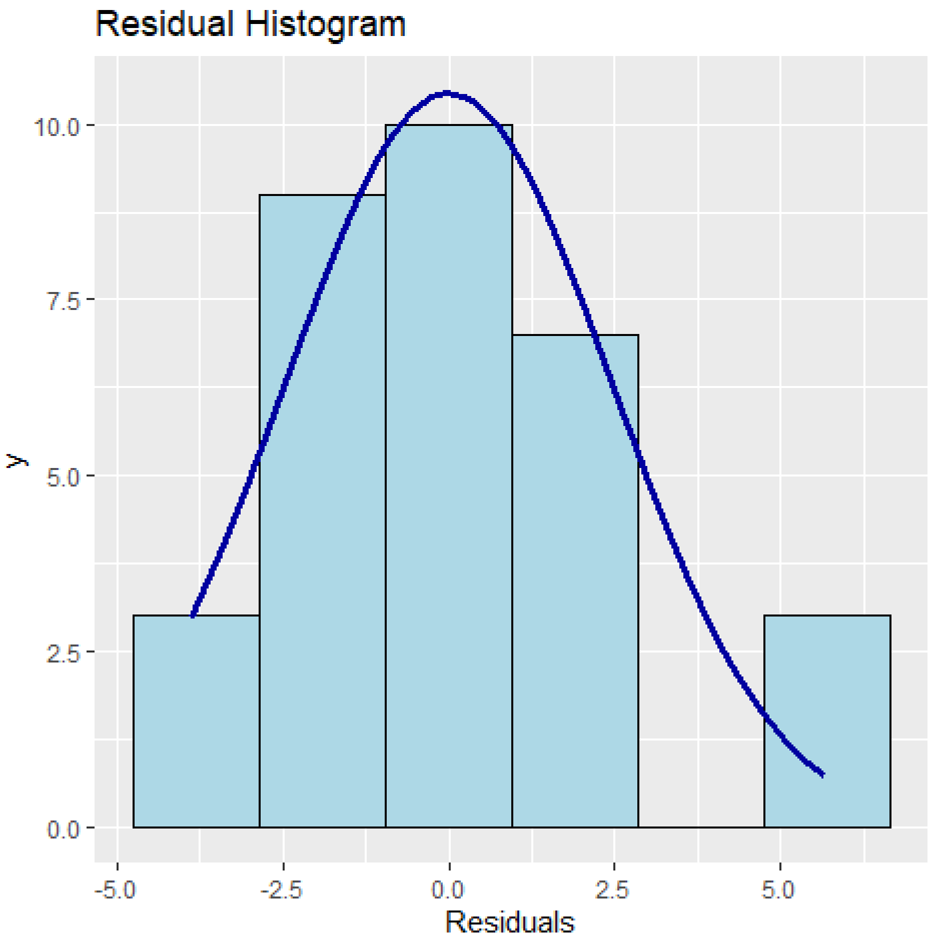


Residual Histogram

Histogram of residuals for detecting violation of normality assumption.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_resid_hist(model)
```



Heteroscedasticity

Introduction

One of the assumptions made about residuals/errors in OLS regression is that the errors have the same but unknown variance. This is known as constant variance or homoscedasticity. When this assumption is violated, the problem is known as heteroscedasticity.

Consequences of Heteroscedasticity

- The OLS estimators and regression predictions based on them remains unbiased and consistent.
- The OLS estimators are no longer the BLUE (Best Linear Unbiased Estimators) because they are no longer efficient, so the regression predictions will be inefficient too.
- Because of the inconsistency of the covariance matrix of the estimated regression coefficients, the tests of hypotheses, (t-test, F-test) are no longer valid.

olsrr provides the following 4 tests for detecting heteroscedasticity:

- Bartlett Test
- Breusch Pagan Test
- Score Test
- F Test

Bartlett Test

Bartlett’s test is used to test if variances across samples is equal. It is sensitive to departures from normality. The Levene test is an alternative test that is less sensitive to departures from normality.

You can perform the test using 2 continuous variables, one continuous and one grouping variable, a formula or a linear model.

Use grouping variable

Hide

```
ols_test_bartlett(hsb, 'read', group_var = 'female')
```

```

    Bartlett's Test of Homogenity of Variances
------------------------------------------------
Ho: Variances are equal across groups
Ha: Variances are unequal for atleast two groups

        Test Summary
 ----------------------------
 DF            =    1
 Chi2          =    0.1866579
 Prob > Chi2   =    0.6657129
```

Using variables

Hide

```
ols_test_bartlett(hsb, 'read', 'write')
```

```

    Bartlett's Test of Homogenity of Variances
------------------------------------------------
Ho: Variances are equal across groups
Ha: Variances are unequal for atleast two groups

        Data
 ---------------------
 Variables: read write

        Test Summary
 ----------------------------
 DF            =    1
 Chi2          =    1.222871
 Prob > Chi2   =    0.2687979
```

Breusch Pagan Test

Breusch Pagan Test was introduced by Trevor Breusch and Adrian Pagan in 1979. It is used to test for heteroskedasticity in a linear regression model and assumes that the error terms are normally distributed. It tests whether the variance of the errors from a regression is dependent on the values of the independent variables. It is a χ^2 test.

You can perform the test using the fitted values of the model, the predictors in the model and a subset of the independent variables. It includes options to perform multiple tests and p value adjustments. The options for p value adjustments include Bonferroni, Sidak and Holm's method.

Use fitted values of the model

Hide

```
model <- lm(mpg ~ disp + hp + wt + drat, data = mtcars)
ols_test_breusch_pagan(model)
```

```

 Breusch Pagan Test for Heteroskedasticity
 -----------------------------------------
 Ho: the variance is constant
 Ha: the variance is not constant

             Data
 -------------------------------
 Response : mpg
 Variables: fitted values of mpg

       Test Summary
 ---------------------------
 DF            =    1
 Chi2          =    1.429672
 Prob > Chi2   =    0.231818
```

Use independent variables of the model

Hide

```
model <- lm(mpg ~ disp + hp + wt + drat, data = mtcars)
ols_test_breusch_pagan(model, rhs = TRUE)
```

```

Breusch Pagan Test for Heteroskedasticity
-----------------------------------------
Ho: the variance is constant
Ha: the variance is not constant

          Data
--------------------------
Response : mpg
Variables: disp hp wt drat

       Test Summary
----------------------------
DF            =    4
Chi2          =    1.513808
Prob > Chi2   =    0.8241927
```

Use independent variables of the model and perform multiple tests

Hide

```
model <- lm(mpg ~ disp + hp + wt + drat, data = mtcars)
ols_test_breusch_pagan(model, rhs = TRUE, multiple = TRUE)
```

```

Breusch Pagan Test for Heteroskedasticity
-----------------------------------------
Ho: the variance is constant
Ha: the variance is not constant

          Data
--------------------------
Response : mpg
Variables: disp hp wt drat

       Test Summary (Unadjusted p values)
----------------------------------------------
 Variable           chi2       df        p
----------------------------------------------
 disp             1.2355345     1    0.2663334
 hp               0.9209878     1    0.3372157
 wt               1.2529988     1    0.2629805
 drat             1.1668486     1    0.2800497
----------------------------------------------
 simultaneous     1.5138083     4    0.8241927
----------------------------------------------
```

Bonferroni p value Adjustment

Hide

```
model <- lm(mpg ~ disp + hp + wt + drat, data = mtcars)
ols_test_breusch_pagan(model, rhs = TRUE, multiple = TRUE, p.adj = 'bonferroni')
```

```

 Breusch Pagan Test for Heteroskedasticity
 -----------------------------------------
 Ho: the variance is constant
 Ha: the variance is not constant

           Data
 --------------------------
 Response : mpg
 Variables: disp hp wt drat

       Test Summary (Bonferroni p values)
 ----------------------------------------------
  Variable           chi2       df        p
 ----------------------------------------------
  disp             1.2355345     1    1.0000000
  hp               0.9209878     1    1.0000000
  wt               1.2529988     1    1.0000000
  drat             1.1668486     1    1.0000000
 ----------------------------------------------
  simultaneous     1.5138083     4    0.8241927
 ----------------------------------------------
```

Sidak p value Adjustment

Hide

```
model <- lm(mpg ~ disp + hp + wt + drat, data = mtcars)
ols_test_breusch_pagan(model, rhs = TRUE, multiple = TRUE, p.adj = 'sidak')
```

```

Breusch Pagan Test for Heteroskedasticity
-----------------------------------------
Ho: the variance is constant
Ha: the variance is not constant

          Data
--------------------------
Response : mpg
Variables: disp hp wt drat

         Test Summary (Sidak p values)
----------------------------------------------
 Variable           chi2       df        p
----------------------------------------------
 disp             1.2355345     1    0.7102690
 hp               0.9209878     1    0.8070305
 wt               1.2529988     1    0.7049362
 drat             1.1668486     1    0.7313356
----------------------------------------------
 simultaneous     1.5138083     4    0.8241927
----------------------------------------------
```

Holm’s p value Adjustment

Hide

```
model <- lm(mpg ~ disp + hp + wt + drat, data = mtcars)
ols_test_breusch_pagan(model, rhs = TRUE, multiple = TRUE, p.adj = 'holm')
```

```

 Breusch Pagan Test for Heteroskedasticity
 -----------------------------------------
 Ho: the variance is constant            
 Ha: the variance is not constant        

           Data            
 --------------------------
 Response : mpg 
 Variables: disp hp wt drat 

          Test Summary (Holm's p values)
 ----------------------------------------------
  Variable           chi2       df        p    
 ----------------------------------------------
  disp             1.2355345     1    0.7990002 
  hp               0.9209878     1    0.3372157 
  wt               1.2529988     1    1.0000000 
  drat             1.1668486     1    0.5600994 
 ----------------------------------------------
  simultaneous     1.5138083     4    0.8241927 
 ----------------------------------------------
```

Score Test

Test for heteroskedasticity under the assumption that the errors are independent and identically distributed (i.i.d.). You can perform the test using the fitted values of the model, the predictors in the model and a subset of the independent variables.

Use fitted values of the model

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_test_score(model)
```

```

 Score Test for Heteroskedasticity
 ---------------------------------
 Ho: Variance is homogenous
 Ha: Variance is not homogenous

 Variables: fitted values of mpg

        Test Summary
 -----------------------------
 DF            =    1
 Chi2          =    0.5163959
 Prob > Chi2   =    0.4723832
```

Use independent variables of the model

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_test_score(model, rhs = TRUE)
```

```

 Score Test for Heteroskedasticity
 ---------------------------------
 Ho: Variance is homogenous
 Ha: Variance is not homogenous

 Variables: disp hp wt qsec

        Test Summary
 -----------------------------
 DF            =    4
 Chi2          =    2.039404
 Prob > Chi2   =    0.7285114
```

Specify variables

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_test_score(model, vars = c('disp', 'hp'))
```

```
Score Test for Heteroskedasticity
---------------------------------
Ho: Variance is homogenous
Ha: Variance is not homogenous

Variables: disp hp

       Test Summary
----------------------------
DF            =    2
Chi2          =    0.9983196
Prob > Chi2   =    0.6070405
```

F Test

F Test for heteroskedasticity under the assumption that the errors are independent and identically distributed (i.i.d.). You can perform the test using the fitted values of the model, the predictors in the model and a subset of the independent variables.

Use fitted values of the model

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_test_f(model)
```

```
F Test for Heteroskedasticity
-----------------------------
Ho: Variance is homogenous
Ha: Variance is not homogenous

Variables: fitted values of mpg

     Test Summary
-------------------------
Num DF     =    1
Den DF     =    30
F          =    0.4920617
Prob > F   =    0.4884154
```

Use independent variables of the model

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_test_f(model, rhs = TRUE)
```

```

 F Test for Heteroskedasticity
 -----------------------------
 Ho: Variance is homogenous
 Ha: Variance is not homogenous

 Variables: disp hp wt qsec

      Test Summary
 -------------------------
 Num DF     =    4
 Den DF     =    27
 F          =    0.4594694
 Prob > F   =    0.7647271
```

Specify variables

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_test_f(model, vars = c('disp', 'hp'))
```

```

 F Test for Heteroskedasticity
 -----------------------------
 Ho: Variance is homogenous
 Ha: Variance is not homogenous

 Variables: disp hp

      Test Summary
 -------------------------
 Num DF     =    2
 Den DF     =    29
 F          =    0.4669306
 Prob > F   =    0.631555
```

Measures of Influence

Introduction

It is possible for a single observation to have a great influence on the results of a regression analysis. It is therefore important to detect influential observations and to take them into consideration when interpreting the results.

olsrr offers the following tools to detect influential observations:

- Cook's D Bar Plot
- Cook's D Chart
- DFBETAs Panel
- DFFITs Plot
- Studentized Residual Plot
- Standardized Residual Chart
- Studentized Residuals vs Leverage Plot
- Deleted Studentized Residual vs Fitted Values Plot
- Hadi Plot
- Potential Residual Plot

Cook's D Bar Plot

Bar Plot of Cook's distance to detect observations that strongly influence fitted values of the model. Cook's distance was introduced by American statistician R Dennis Cook in 1977. It is used to identify influential data points. It depends on both the residual and leverage i.e it takes it account both the **x** value and **y** value of the observation.

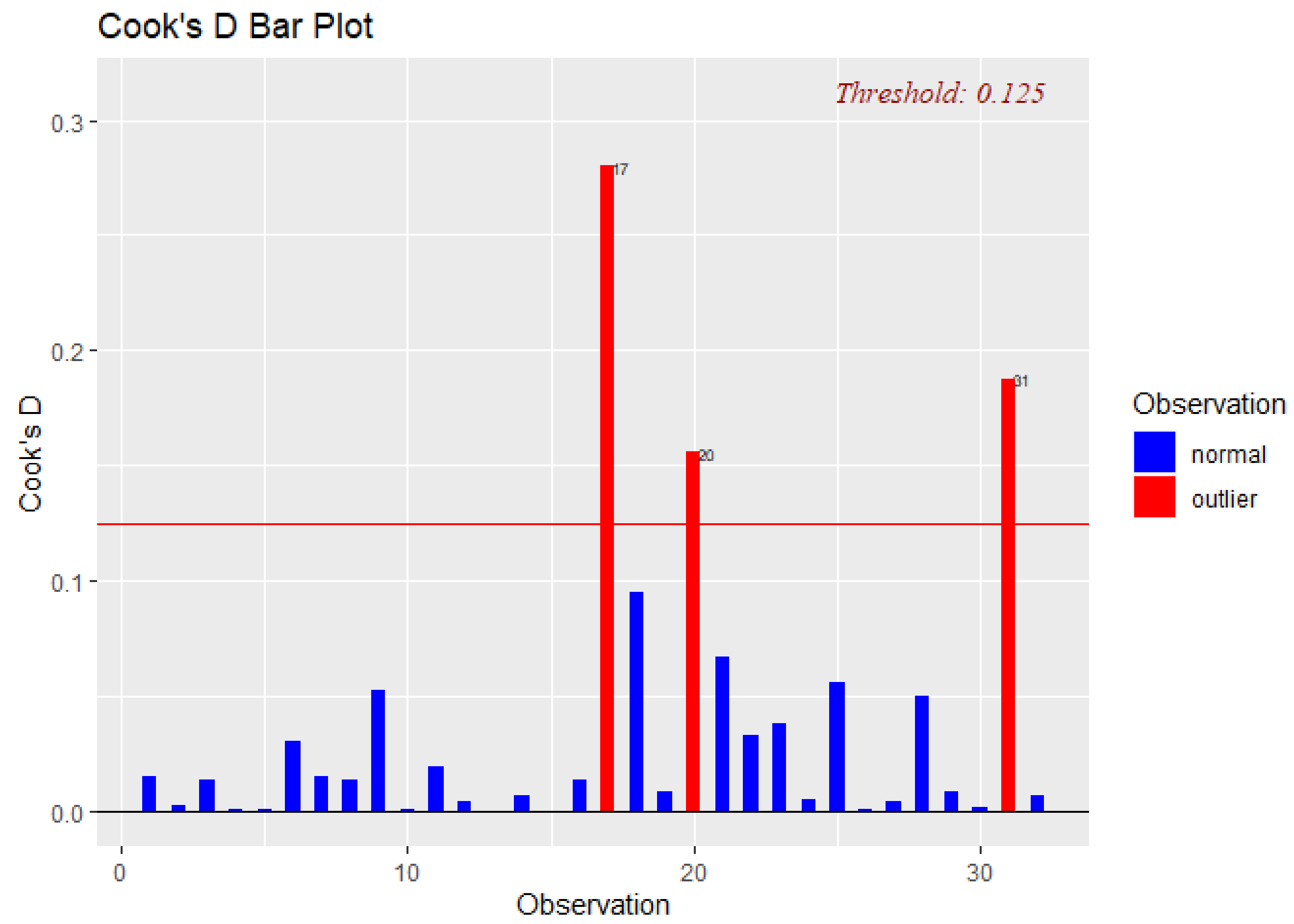
Steps to compute Cook's distance:

- delete observations one at a time.
- refit the regression model on remaining $(n-1)$ observations
- examine how much all of the fitted values change when the ith observation is deleted.

A data point having a large cook's d indicates that the data point strongly influences the fitted values.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_cooksd_bar(model)
```

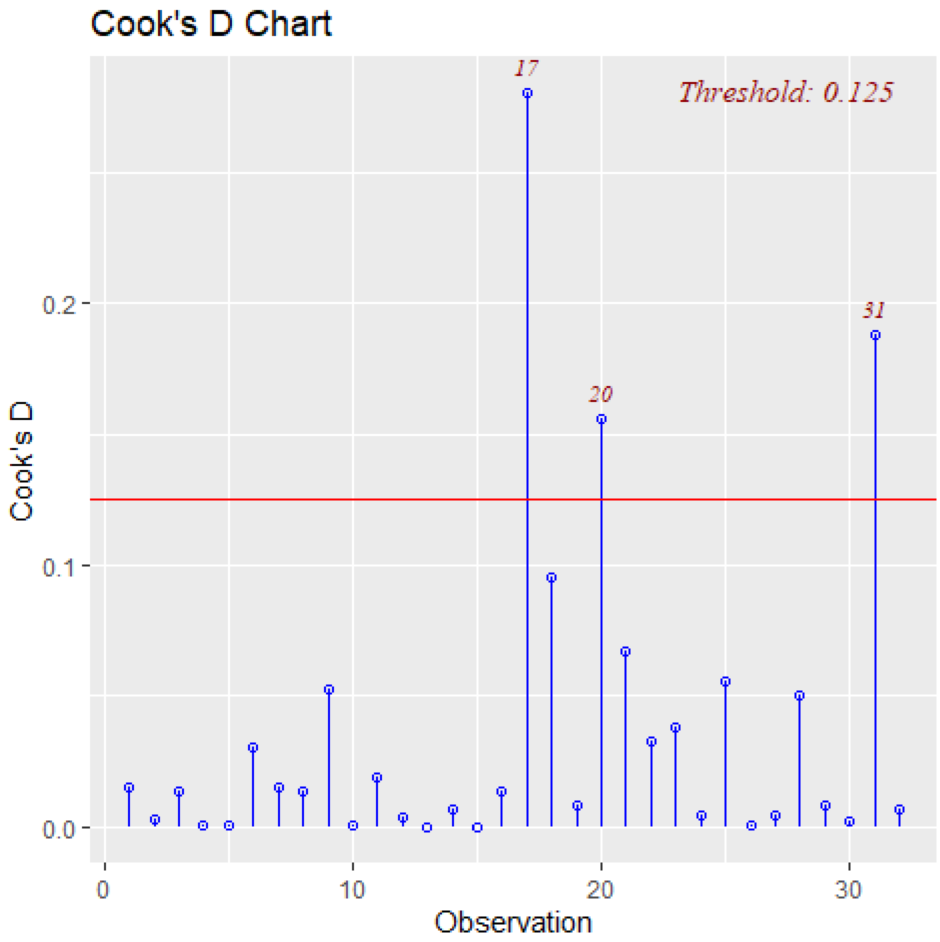


Cook’s D Chart

Chart of Cook’s distance to detect observations that strongly influence fitted values of the model.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_cooksd_chart(model)
```

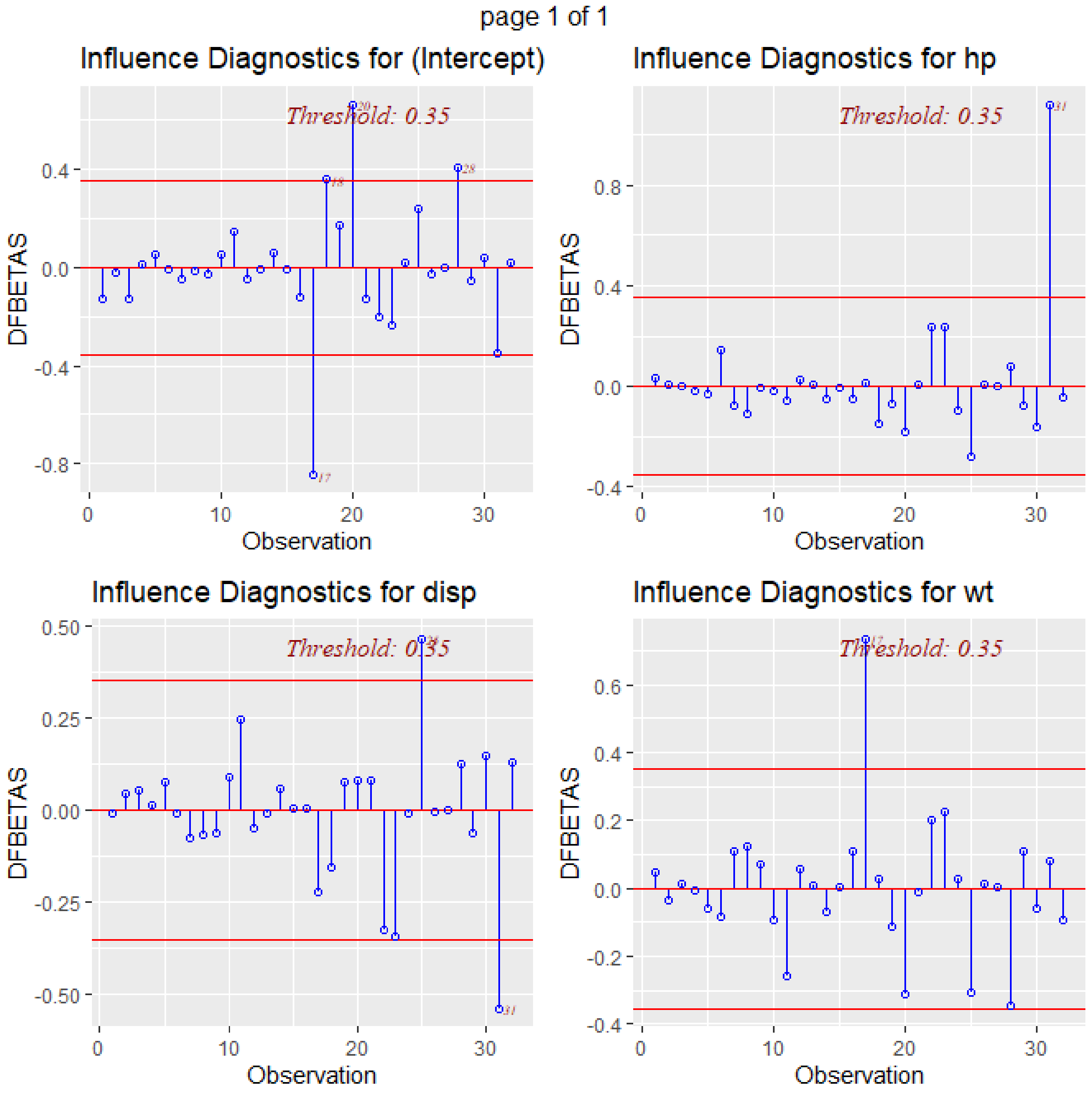


DFBETAs Panel

DFBETA measures the difference in each parameter estimate with and without the influential point. There is a DFBETA for each data point i.e if there are n observations and k variables, there will be $n * k$ DFBETAs. In general, large values of DFBETAS indicate observations that are influential in estimating a given parameter. Belsley, Kuh, and Welsch recommend 2 as a general cutoff value to indicate influential observations and $\frac{2}{\sqrt{n}}$ as a size-adjusted cutoff.

Hide

```
model <- lm(mpg ~ disp + hp + wt, data = mtcars)
ols_plot_dfbetas(model)
```



DFFITS Plot

Proposed by Welsch and Kuh (1977). It is the scaled difference between the i^{th} fitted value obtained from the full data and the i^{th} fitted value obtained by deleting the i^{th} observation. DFFIT - difference in fits, is used to identify influential data points. It quantifies the number of standard deviations that the fitted value changes when the ith data point is omitted.

Steps to compute DFFITs:

- delete observations one at a time.
- refit the regression model on remaining $ {n - 1} $ observations
- examine how much all of the fitted values change when the ith observation is deleted.

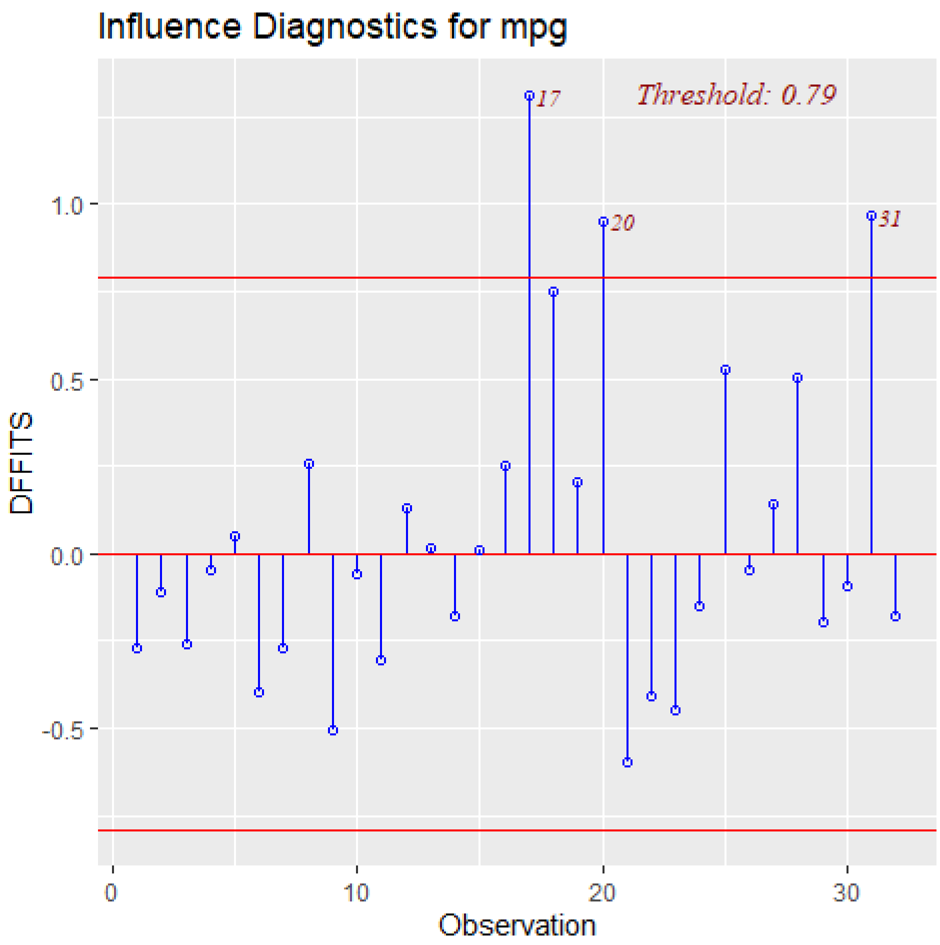
An observation is deemed influential if the absolute value of its DFFITS value is greater than:

$$2 * \frac{\sqrt{(p+1)}}{(n-p-1)}$$

where n is the number of observations and p is the number of predictors including intercept.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_dffits(model)
```

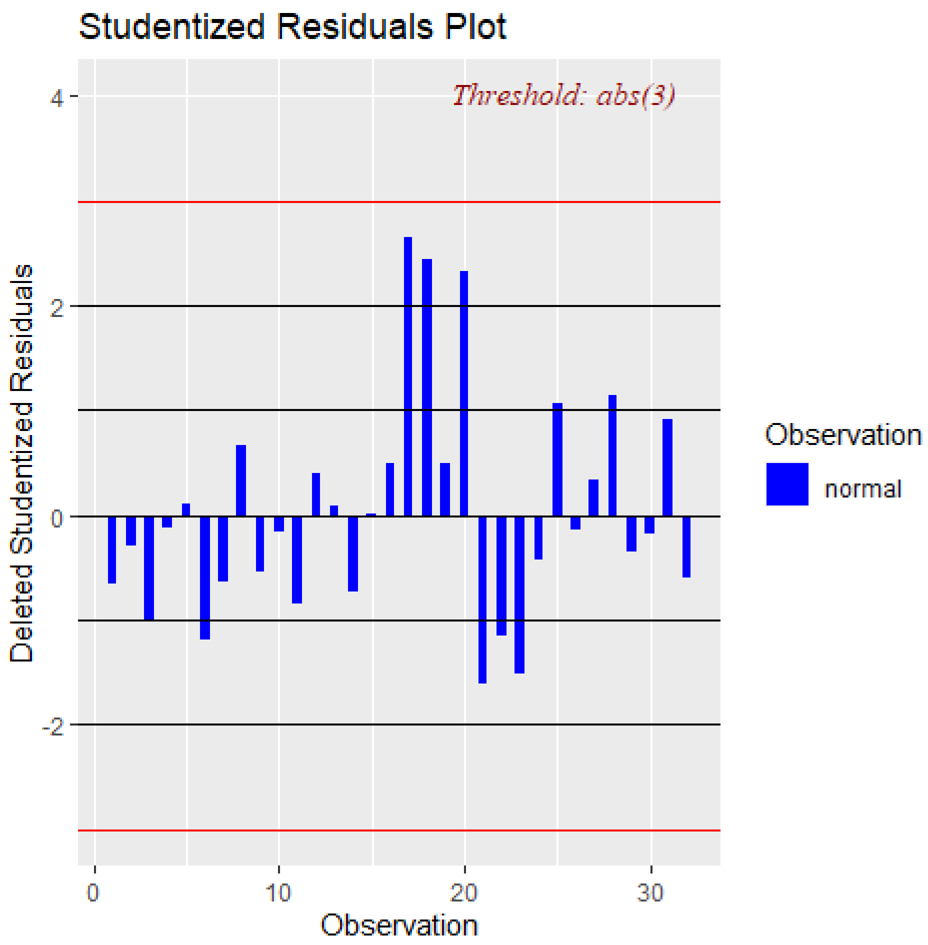


Studentized Residual Plot

Plot for detecting outliers. Studentized deleted residuals (or externally studentized residuals) is the deleted residual divided by its estimated standard deviation. Studentized residuals are going to be more effective for detecting outlying Y observations than standardized residuals. If an observation has an externally studentized residual that is larger than 3 (in absolute value) we can call it an outlier.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_resid_stud(model)
```

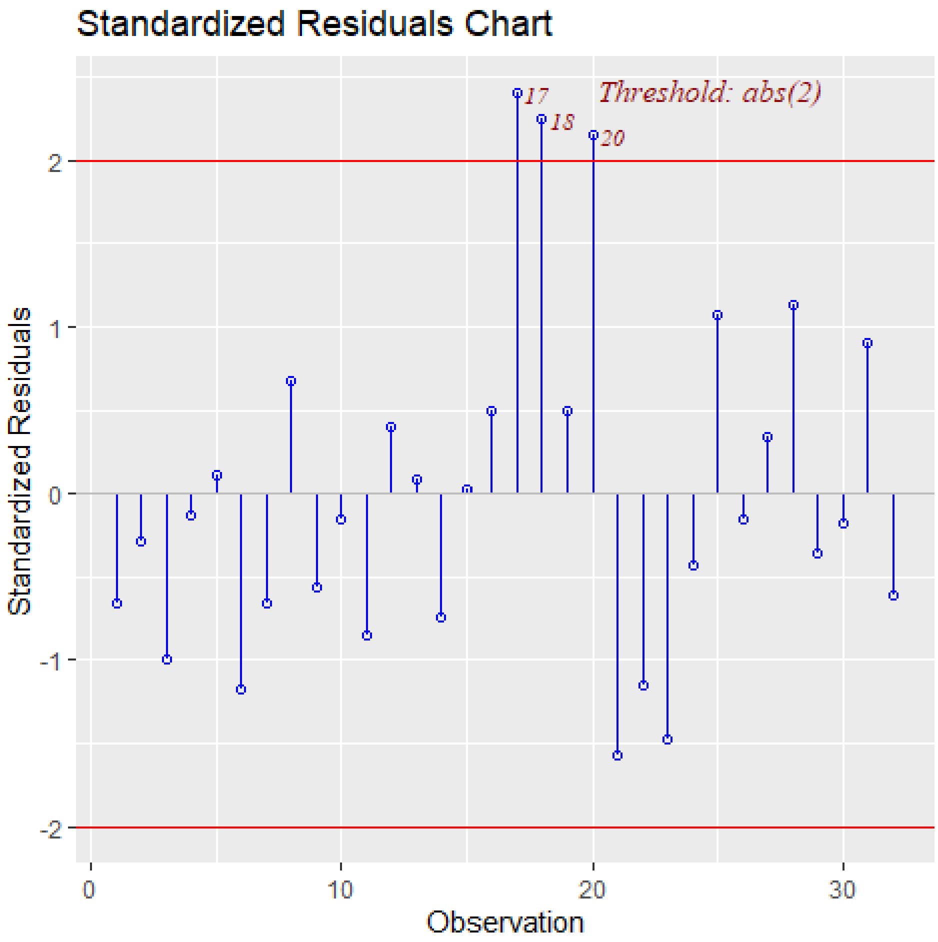


Standardized Residual Chart

Chart for detecting outliers. Standardized residual (internally studentized) is the residual divided by estimated standard deviation.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_resid_stand(model)
```

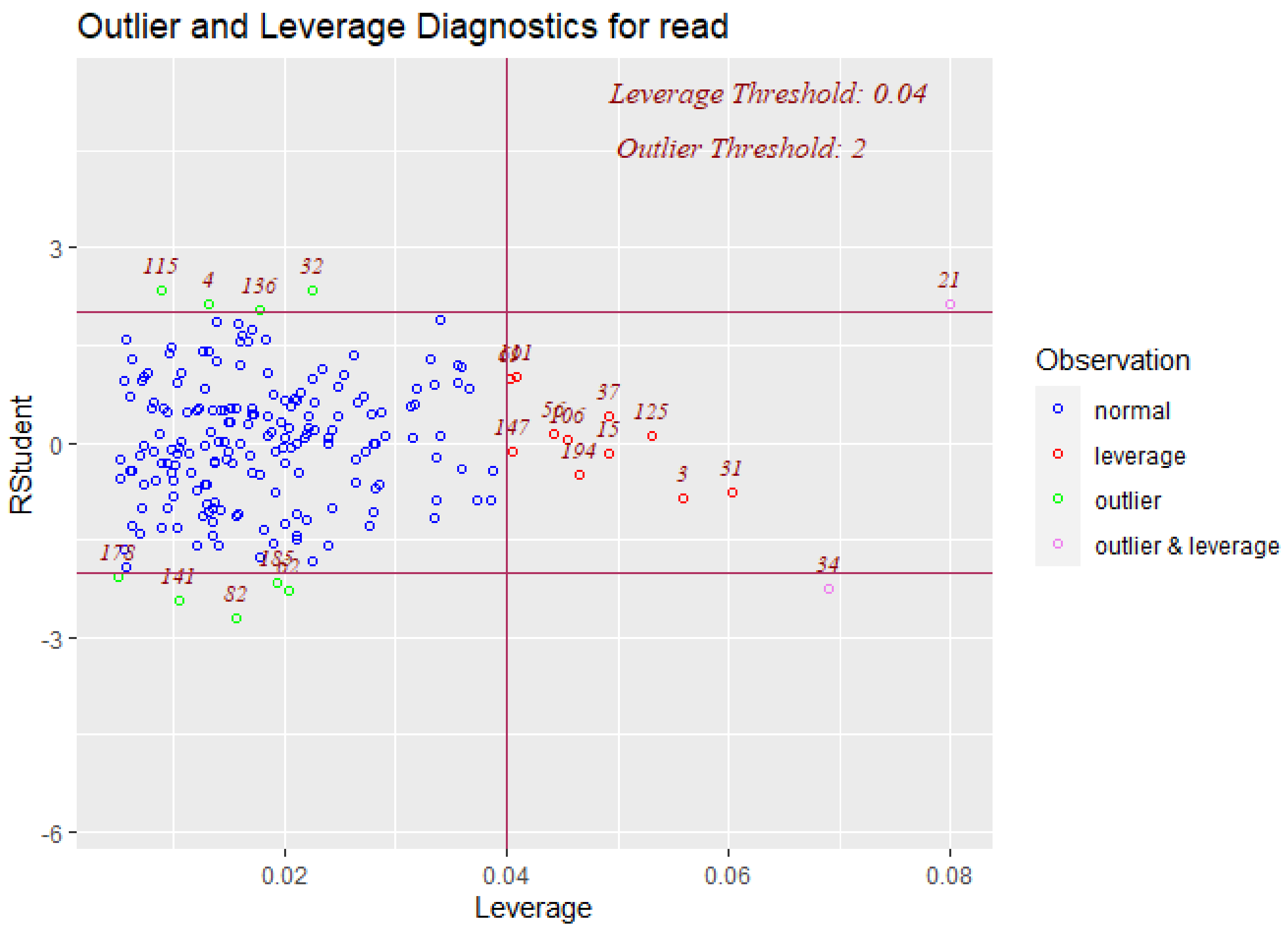


Studentized Residuals vs Leverage Plot

Graph for detecting influential observations.

Hide

```
model <- lm(read ~ write + math + science, data = hsb)
ols_plot_resid_lev(model)
```

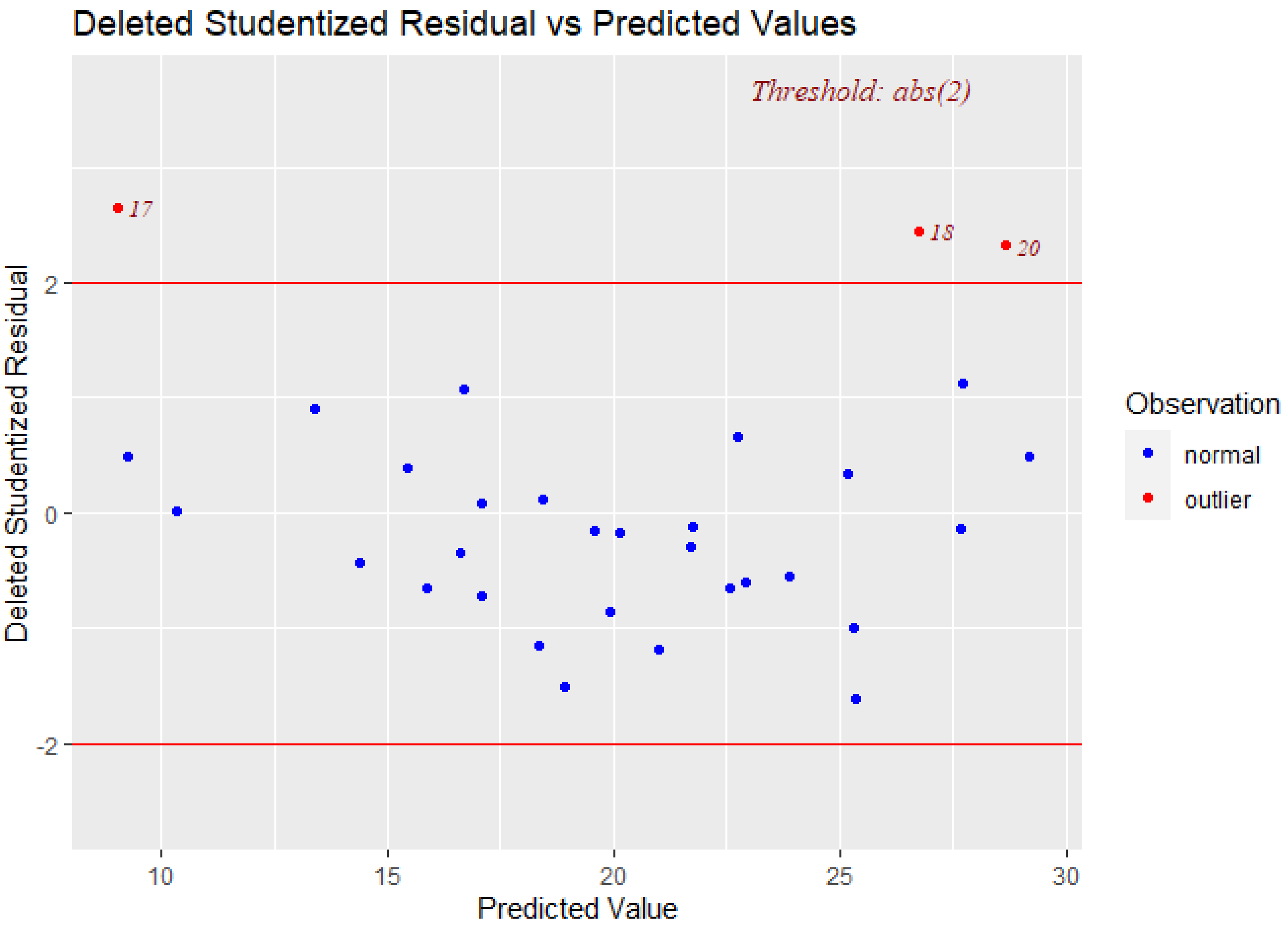


Deleted Studentized Residual vs Fitted Values Plot

Graph for detecting outliers.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_resid_stud_fit(model)
```

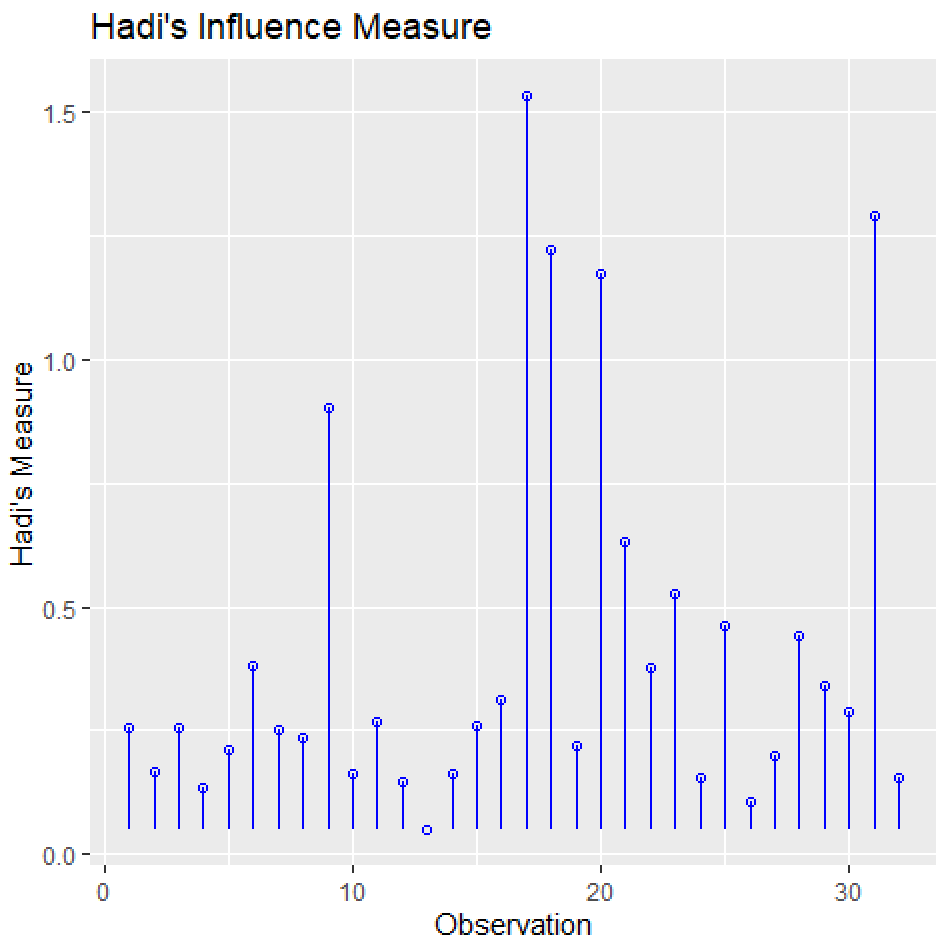


Hadi Plot

Hadi’s measure of influence based on the fact that influential observations can be present in either the response variable or in the predictors or both. The plot is used to detect influential observations based on Hadi’s measure.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_hadi(model)
```

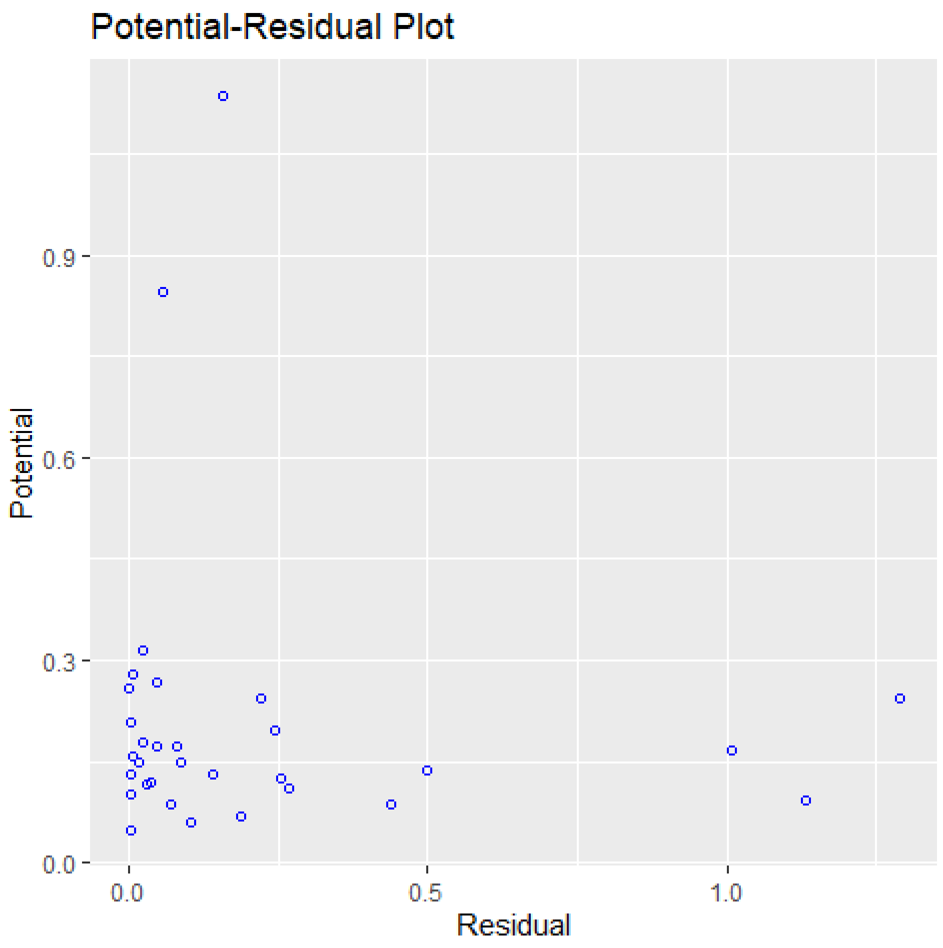


Potential Residual Plot

Plot to aid in classifying unusual observations as high-leverage points, outliers, or a combination of both.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_resid_pot(model)
```



Collinearity Diagnostics, Model Fit & Variable Contribution

Collinearity Diagnostics

Collinearity implies two variables are near perfect linear combinations of one another. Multicollinearity involves more than two variables. In the presence of multicollinearity, regression estimates are unstable and have high standard errors.

VIF

Variance inflation factors measure the inflation in the variances of the parameter estimates due to collinearities that exist among the predictors. It is a measure of how much the variance of the estimated regression coefficient β_k is "inflated" by the existence of correlation among the predictor variables in the model. A VIF of 1 means that there is no correlation among the kth predictor and the remaining predictor variables, and hence the variance of β_k is not inflated at all. The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

Steps to calculate VIF:

- Regress the k^{th} predictor on rest of the predictors in the model.
- Compute the R^2_k

$$VIF = \frac{1}{1 - R_k^2} = \frac{1}{Tolerance}$$

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_vif_tol(model)
```

Variables <chr>	Tolerance <dbl>	VIF <dbl>
disp	0.1252279	7.985439
hp	0.1935450	5.166758
wt	0.1445726	6.916942
qsec	0.3191708	3.133119

4 rows

Tolerance

Percent of variance in the predictor that cannot be accounted for by other predictors.

Steps to calculate tolerance:

- Regress the k^{th} predictor on rest of the predictors in the model.
- Compute the R_k^2

$$Tolerance = 1 - R_k^2$$

Condition Index

Most multivariate statistical approaches involve decomposing a correlation matrix into linear combinations of variables. The linear combinations are chosen so that the first combination has the largest possible variance (subject to some restrictions we won't discuss), the second combination has the next largest variance, subject to being uncorrelated with the first, the third has the largest possible variance, subject to being uncorrelated with the first and second, and so forth. The variance of each of these linear combinations is called an eigenvalue. Collinearity is spotted by finding 2 or more variables that have large proportions of variance (.50 or more) that correspond to large condition indices. A rule of thumb is to label as large those condition indices in the range of 30 or larger.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_eigen_cindex(model)
```

Eigenvalue <dbl>	Condition Index <dbl>	intercept <dbl>	disp <dbl>	hp <dbl>	wt <dbl>

Eigenvalue <dbl>	Condition Index <dbl>	intercept <dbl>	disp <dbl>	hp <dbl>	wt <dbl>
4.721487187	1.000000	0.000123237	0.001132468	0.001413094	0.0005253393
0.216562203	4.669260	0.002617424	0.036811051	0.027751289	0.0002096014
0.050416837	9.677242	0.001656551	0.120881424	0.392366164	0.0377028008
0.010104757	21.616057	0.025805998	0.777260487	0.059594623	0.7017528428
0.001429017	57.480524	0.969796790	0.063914571	0.518874831	0.2598094157

5 rows

Collinearity Diagnostics

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_coll_diag(model)
```

```
Tolerance and Variance Inflation Factor
---------------------------------------
```

Variables <chr>	Tolerance <dbl>	VIF <dbl>
disp	0.1252279	7.985439
hp	0.1935450	5.166758
wt	0.1445726	6.916942
qsec	0.3191708	3.133119

4 rows

```


Eigenvalue and Condition Index
------------------------------
```

Eigenvalue <dbl>	Condition Index <dbl>	intercept <dbl>	disp <dbl>	hp <dbl>	wt <dbl>
4.721487187	1.000000	0.000123237	0.001132468	0.001413094	0.0005253393
0.216562203	4.669260	0.002617424	0.036811051	0.027751289	0.0002096014
0.050416837	9.677242	0.001656551	0.120881424	0.392366164	0.0377028008

Eigenvalue <dbl>	Condition Index <dbl>	intercept <dbl>	disp <dbl>	hp <dbl>	wt <dbl>
0.010104757	21.616057	0.025805998	0.777260487	0.059594623	0.7017528428
0.001429017	57.480524	0.969796790	0.063914571	0.518874831	0.2598094157

5 rows

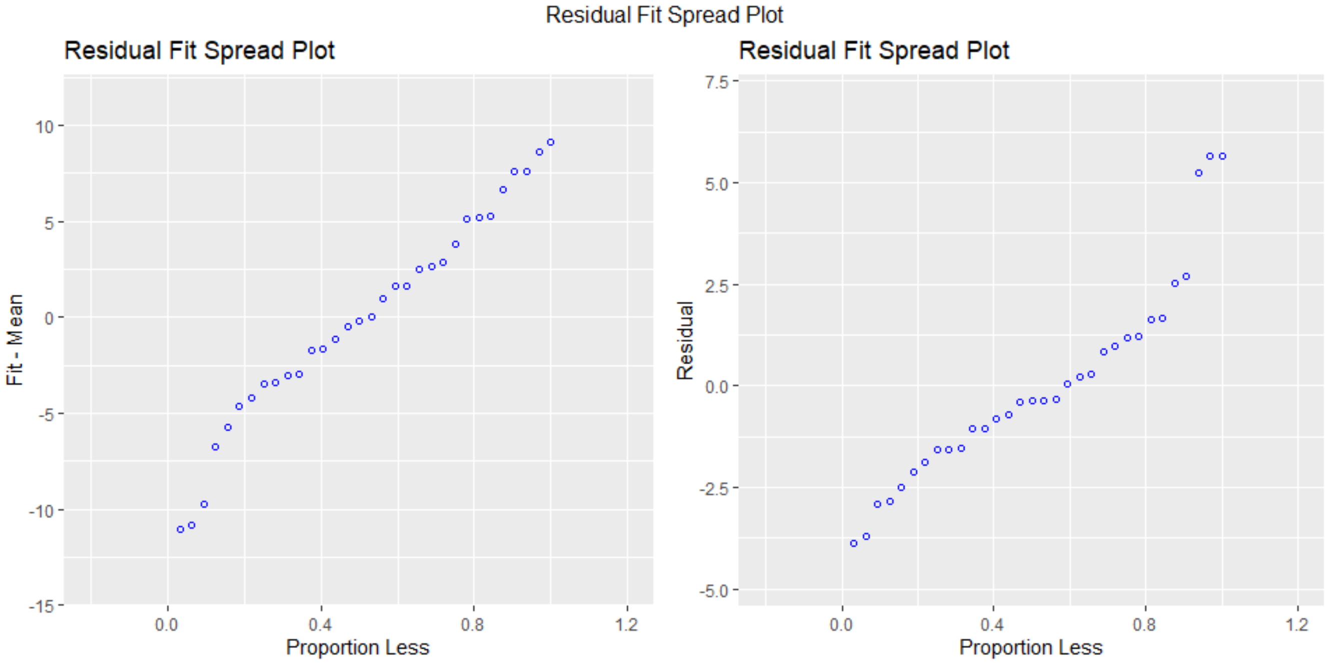
Model Fit Assessment

Residual Fit Spread Plot

Plot to detect non-linearity, influential observations and outliers. Consists of side-by-side quantile plots of the centered fit and the residuals. It shows how much variation in the data is explained by the fit and how much remains in the residuals. For inappropriate models, the spread of the residuals in such a plot is often greater than the spread of the centered fit.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_resid_fit_spread(model)
```



Part & Partial Correlations

Correlations

Relative importance of independent variables in determining **Y**. How much each variable uniquely contributes to R^2 over and above that which can be accounted for by the other predictors.

Zero Order

Pearson correlation coefficient between the dependent variable and the independent variables.

Part

Unique contribution of independent variables. How much R^2 will decrease if that variable is removed from the model?

Partial

How much of the variance in **Y**, which is not estimated by the other independent variables in the model, is estimated by the specific variable?

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_correlations(model)
```

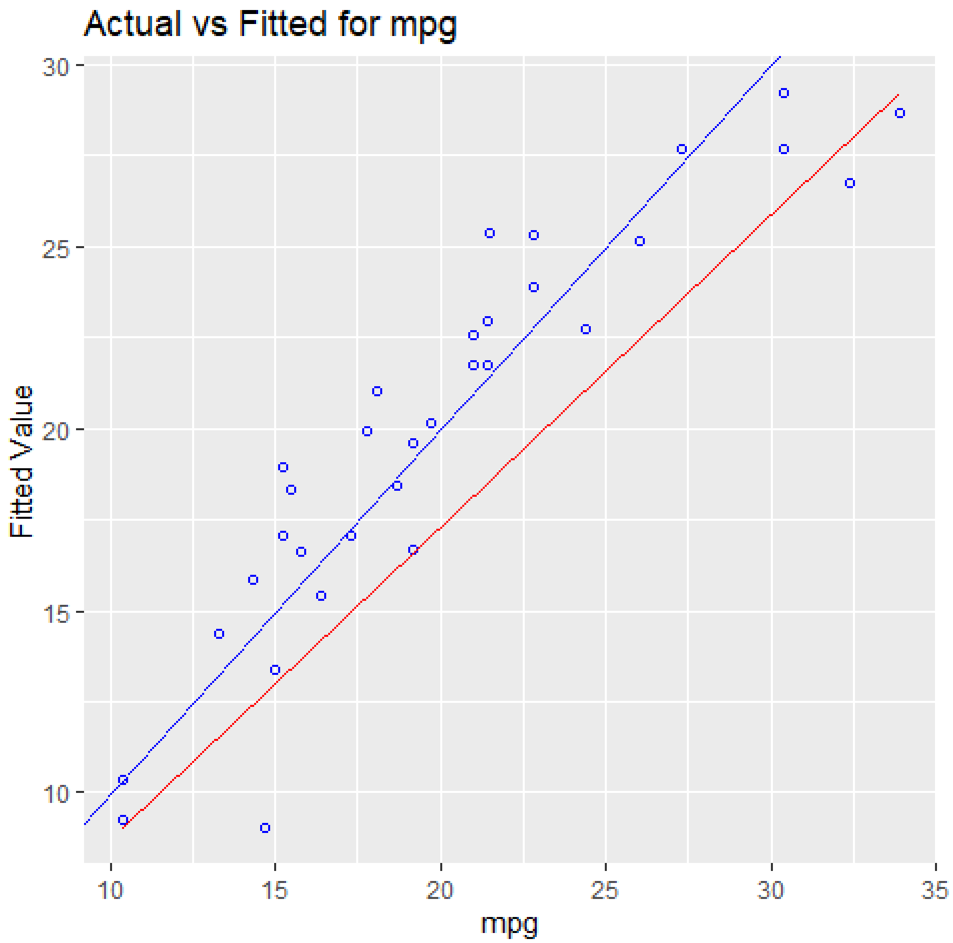
```
                Correlations
-------------------------------------------
Variable    Zero Order    Partial     Part
-------------------------------------------
disp            -0.848      0.048    0.019
hp              -0.776     -0.224   -0.093
wt              -0.868     -0.574   -0.285
qsec             0.419      0.219    0.091
-------------------------------------------
```

Observed vs Predicted Plot

Plot of observed vs fitted values to assess the fit of the model. Ideally, all your points should be close to a regressed diagonal line. Draw such a diagonal line within your graph and check out where the points lie. If your model had a high R Square, all the points would be close to this diagonal line. The lower the R Square, the weaker the Goodness of fit of your model, the more foggy or dispersed your points are from this diagonal line.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_obs_fit(model)
```



Lack of Fit F Test

Assess how much of the error in prediction is due to lack of model fit. The residual sum of squares resulting from a regression can be decomposed into 2 components:

- Due to lack of fit
- Due to random variation

If most of the error is due to lack of fit and not just random error, the model should be discarded and a new model must be built. The lack of fit F test works only with simple linear regression. Moreover, it is important that the data contains repeat observations i.e. replicates for at least one of the values of the predictor x. This test generally only applies to datasets with plenty of replicates.

Hide

```
model <- lm(mpg ~ disp, data = mtcars)
ols_pure_error_anova(model)
```

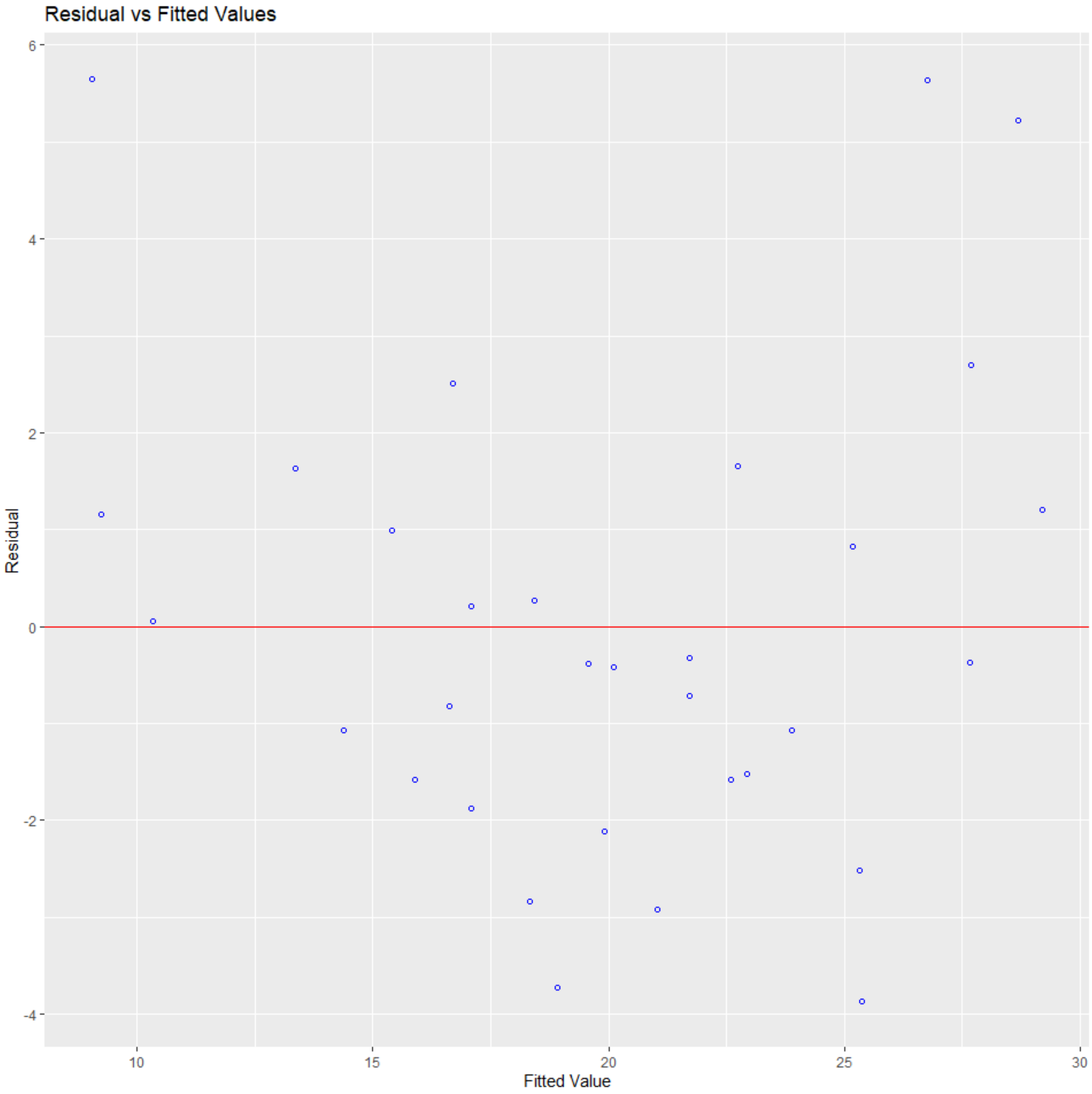
```
Lack of Fit F Test
-----------------
Response :   mpg
Predictor:   disp

                      Analysis of Variance Table
----------------------------------------------------------------------
                DF     Sum Sq     Mean Sq     F Value        Pr(>F)
----------------------------------------------------------------------
disp             1    808.8885    808.8885    314.0095    1.934413e-17
Residual        30    317.1587    10.57196
 Lack of fit    25    304.2787    12.17115    4.724824      0.04563623
 Pure Error      5       12.88       2.576
----------------------------------------------------------------------
```

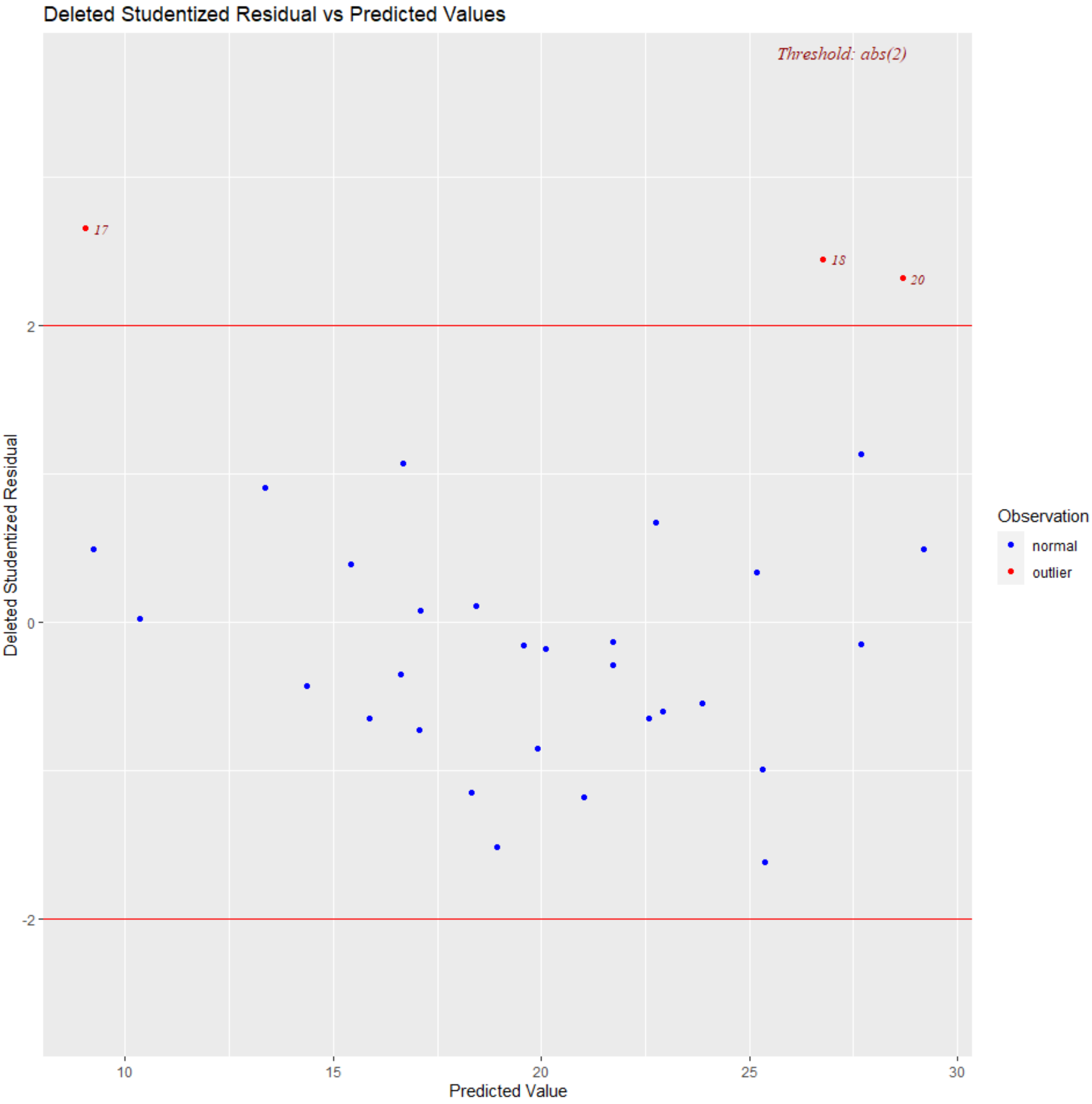
Diagnostics Panel

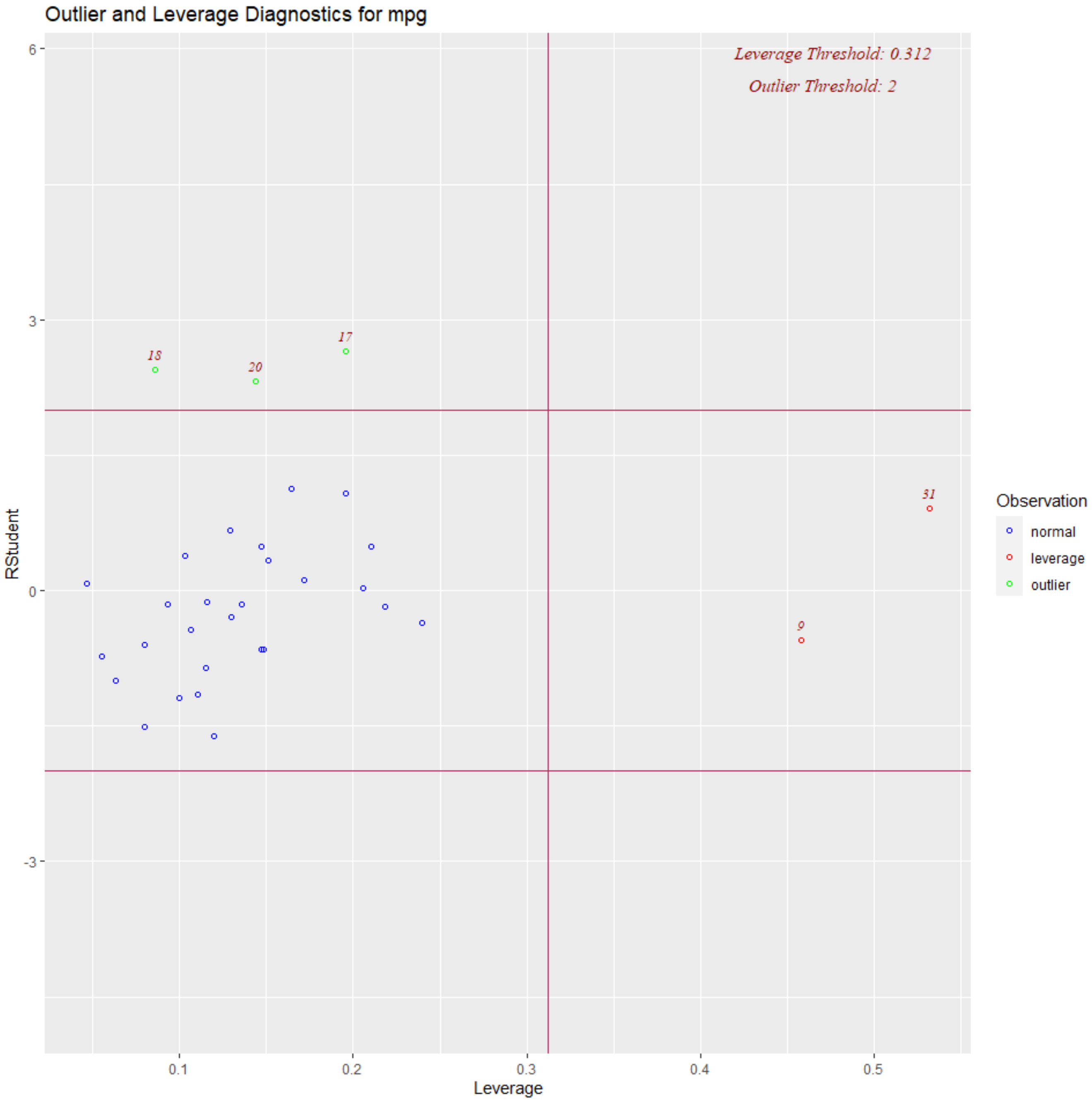
Panel of plots for regression diagnostics

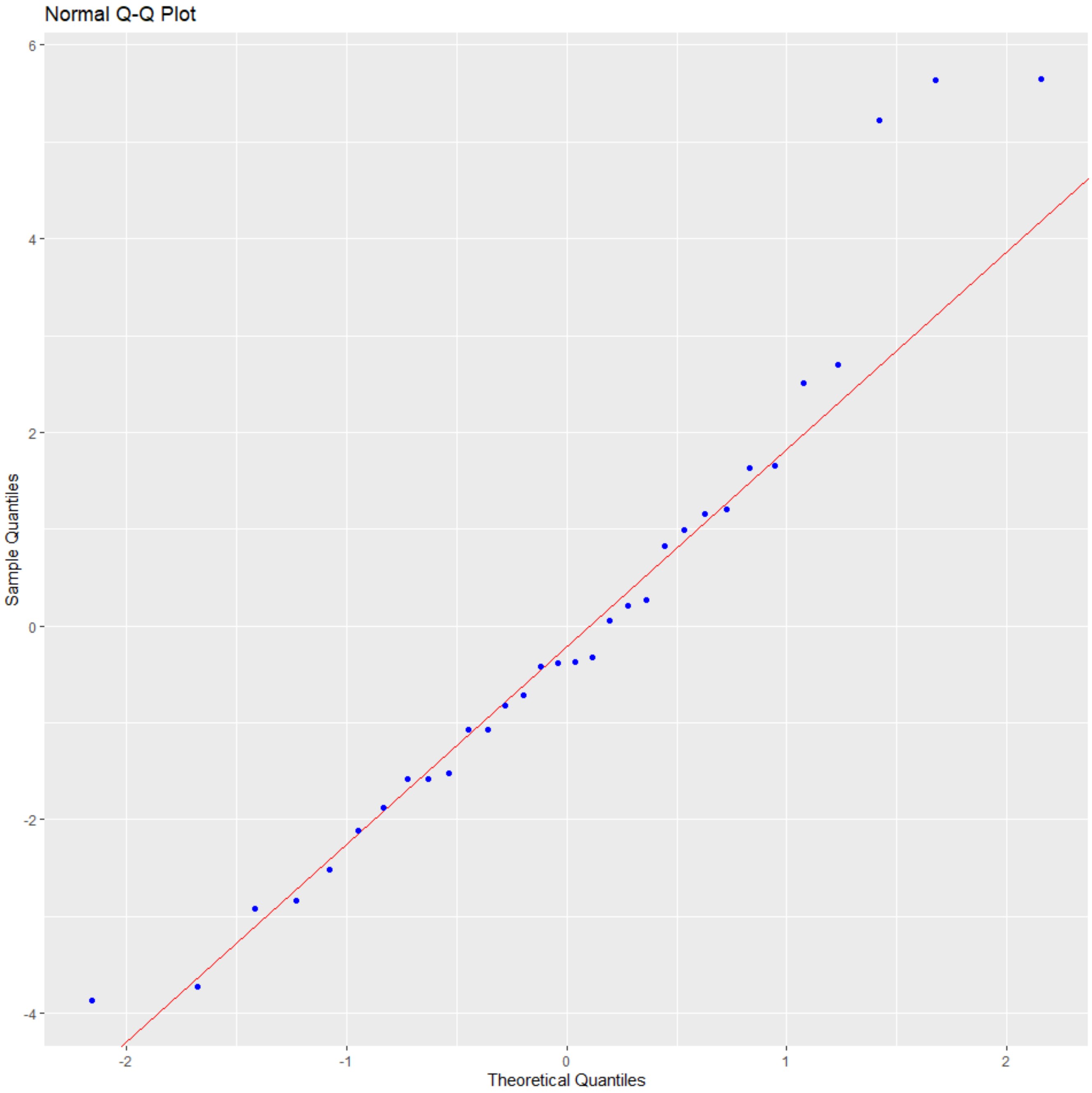
Hide

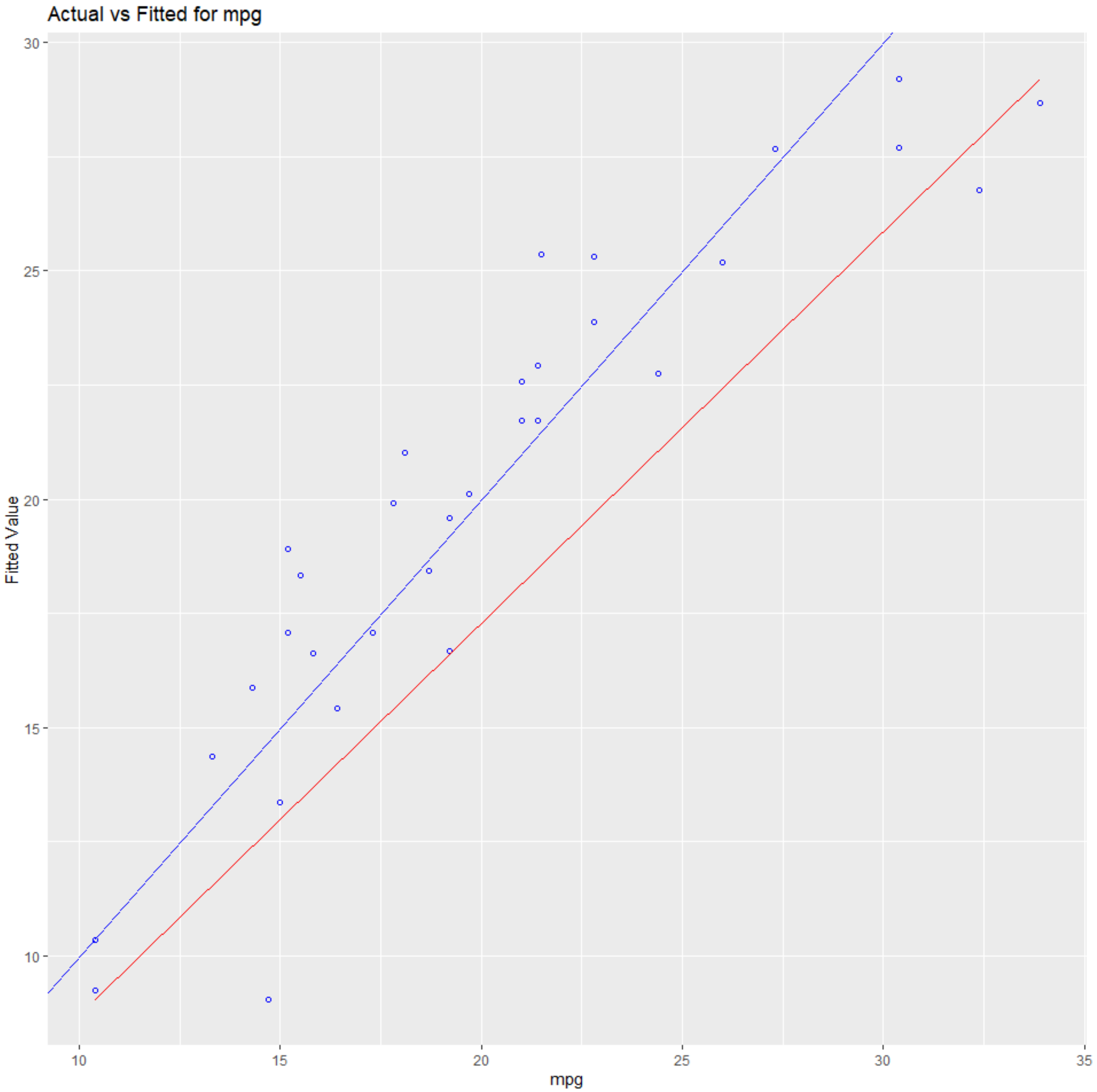
```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_diagnostics(model)
```

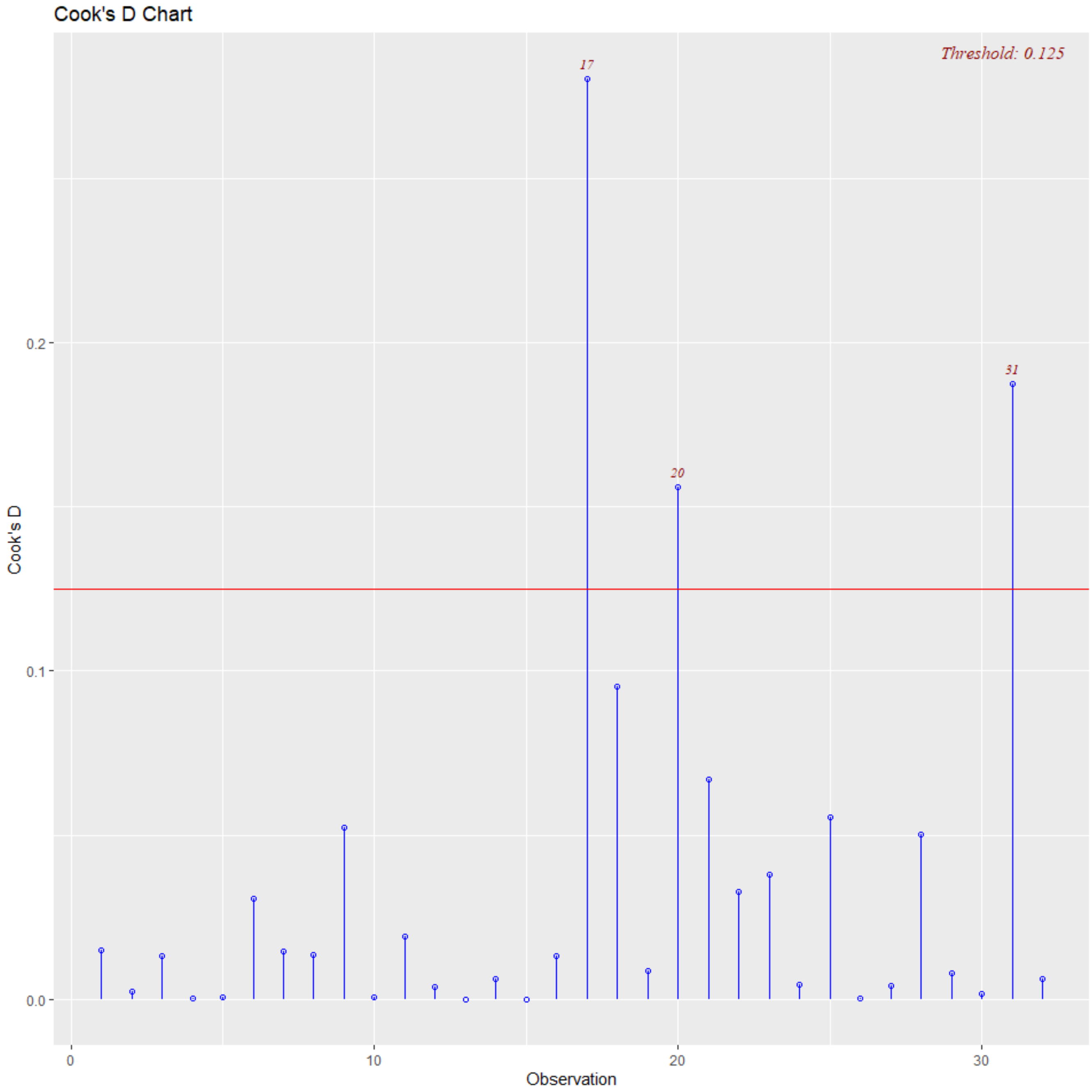

Residual vs Fitted Values
Residual
6
4
2
0
-2
-4
Fitted Value
10
15
20
25
30

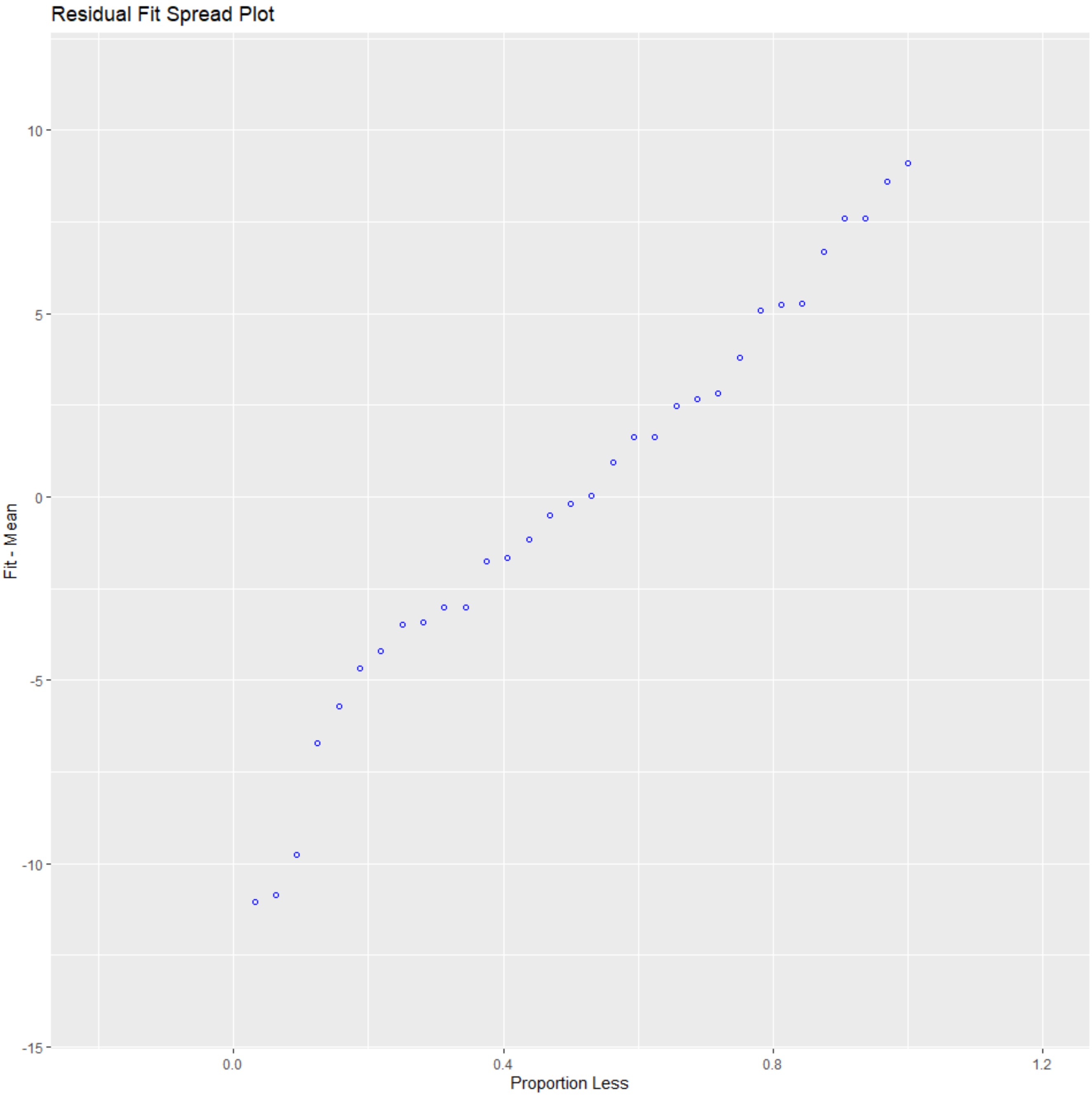

Deleted Studentized Residual vs Predicted Values
Threshold: abs(2)
17
18
20
2
0
-2
Deleted Studentized Residual
10
15
20
25
30
Predicted Value
Observation
● normal
● outlier

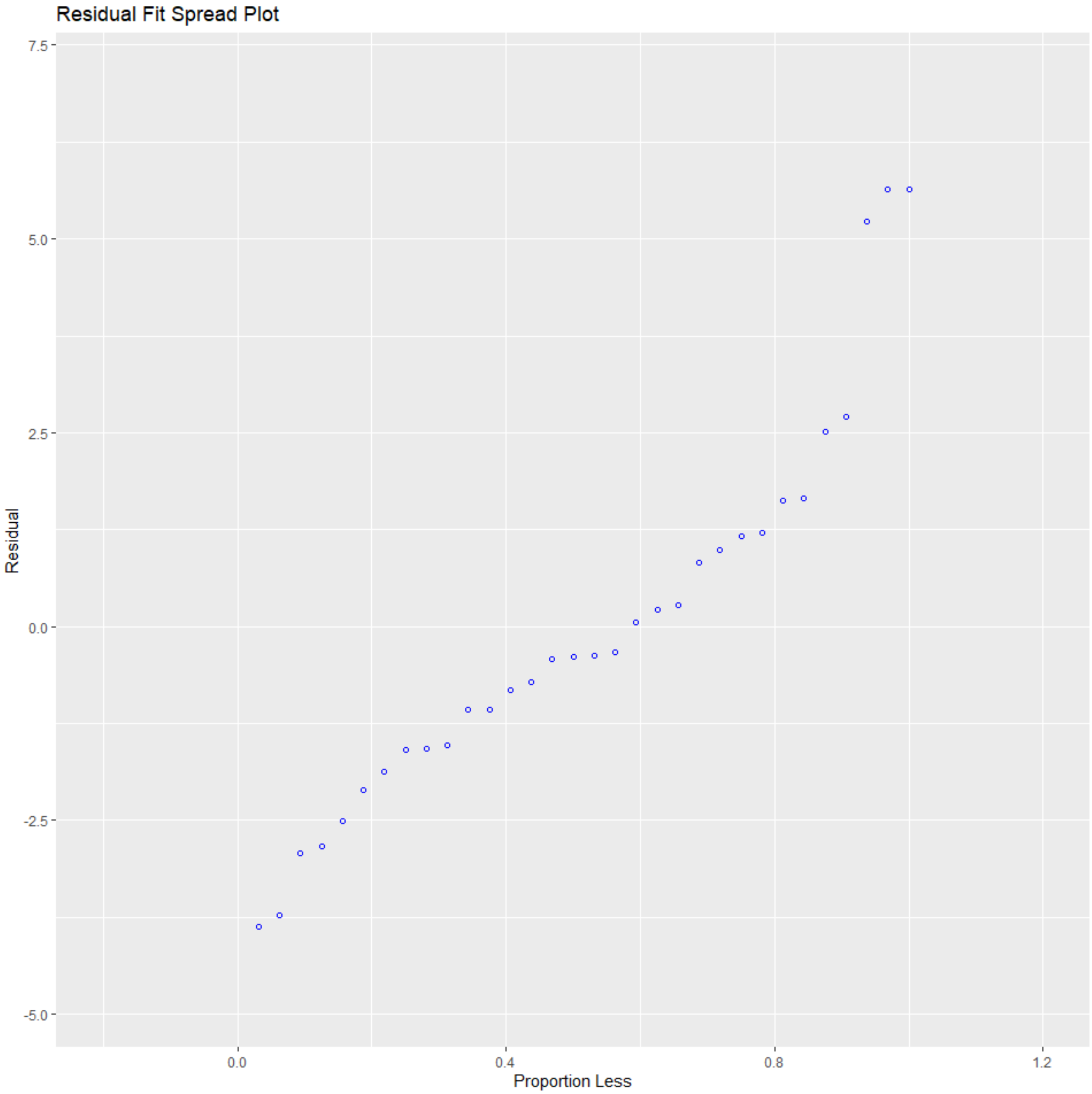

Outlier and Leverage Diagnostics for mpg
Leverage Threshold: 0.312
Outlier Threshold: 2
RStudent
Leverage
6
3
0
-3
0.1
0.2
0.3
0.4
0.5
18
20
17
31
9
Observation
normal
leverage
outlier

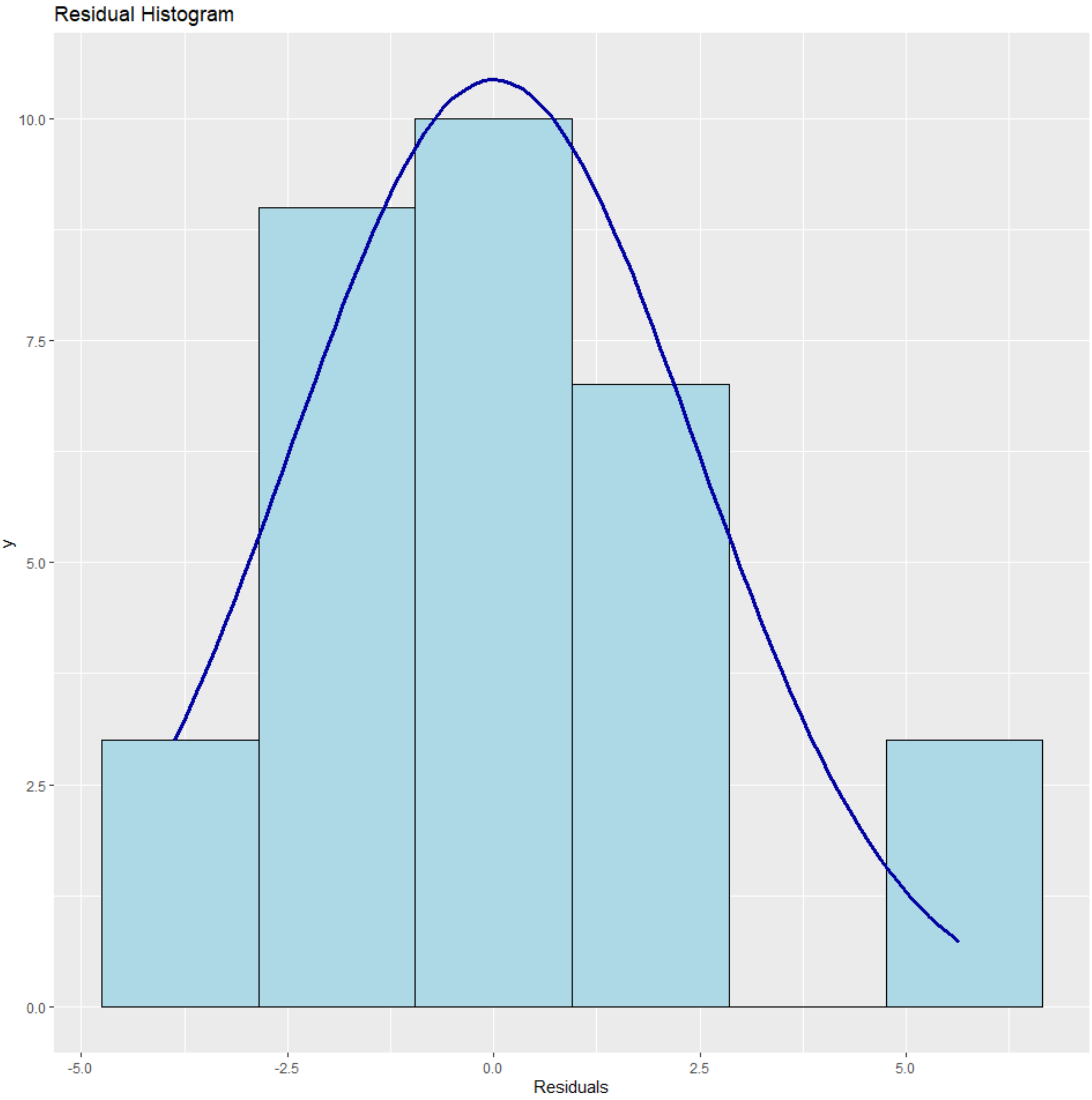

Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
-2
-4
-2
-1
0
1
2

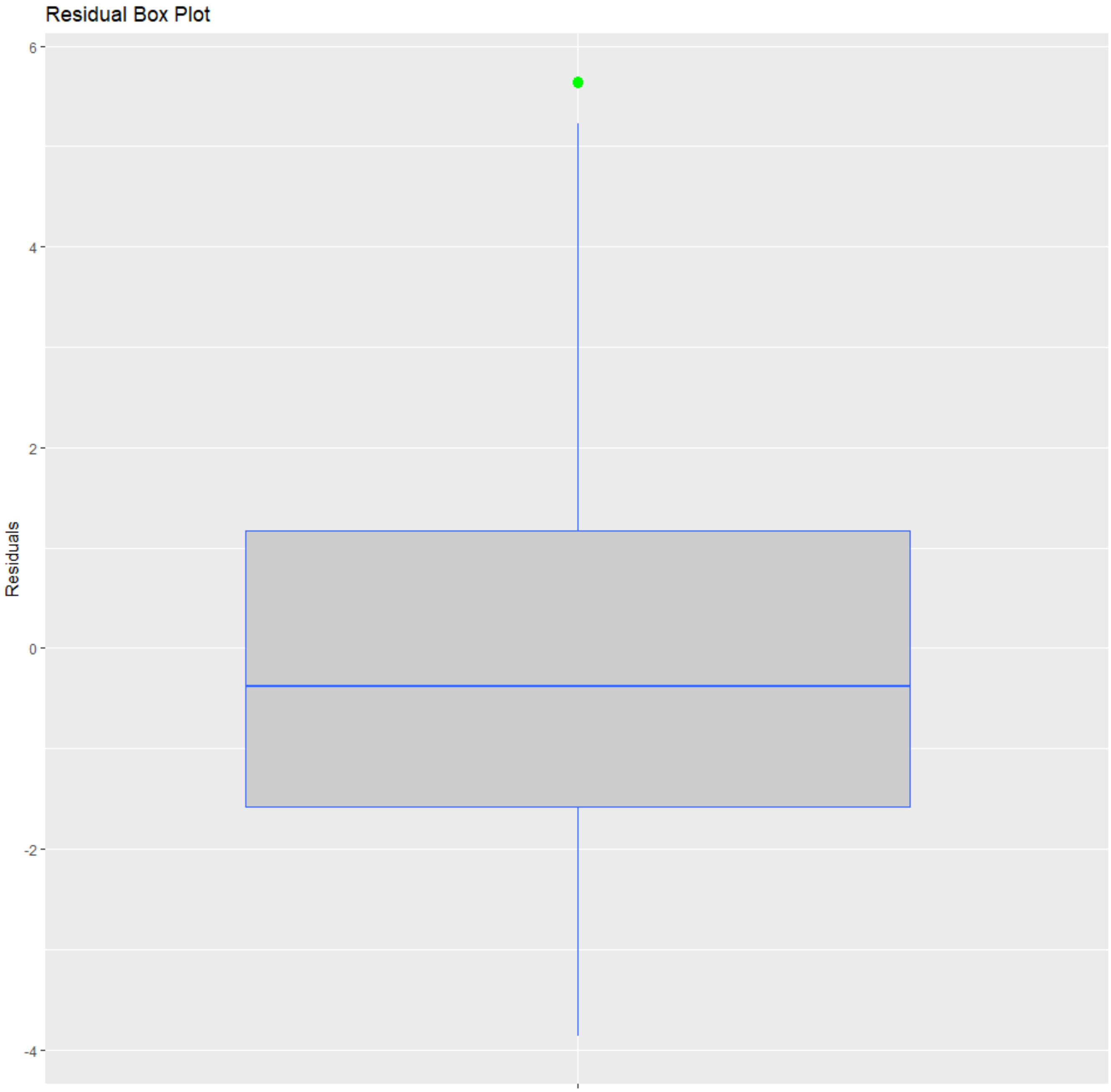

Actual vs Fitted for mpg
Fitted Value
30
25
20
15
10
mpg
10
15
20
25
30
35

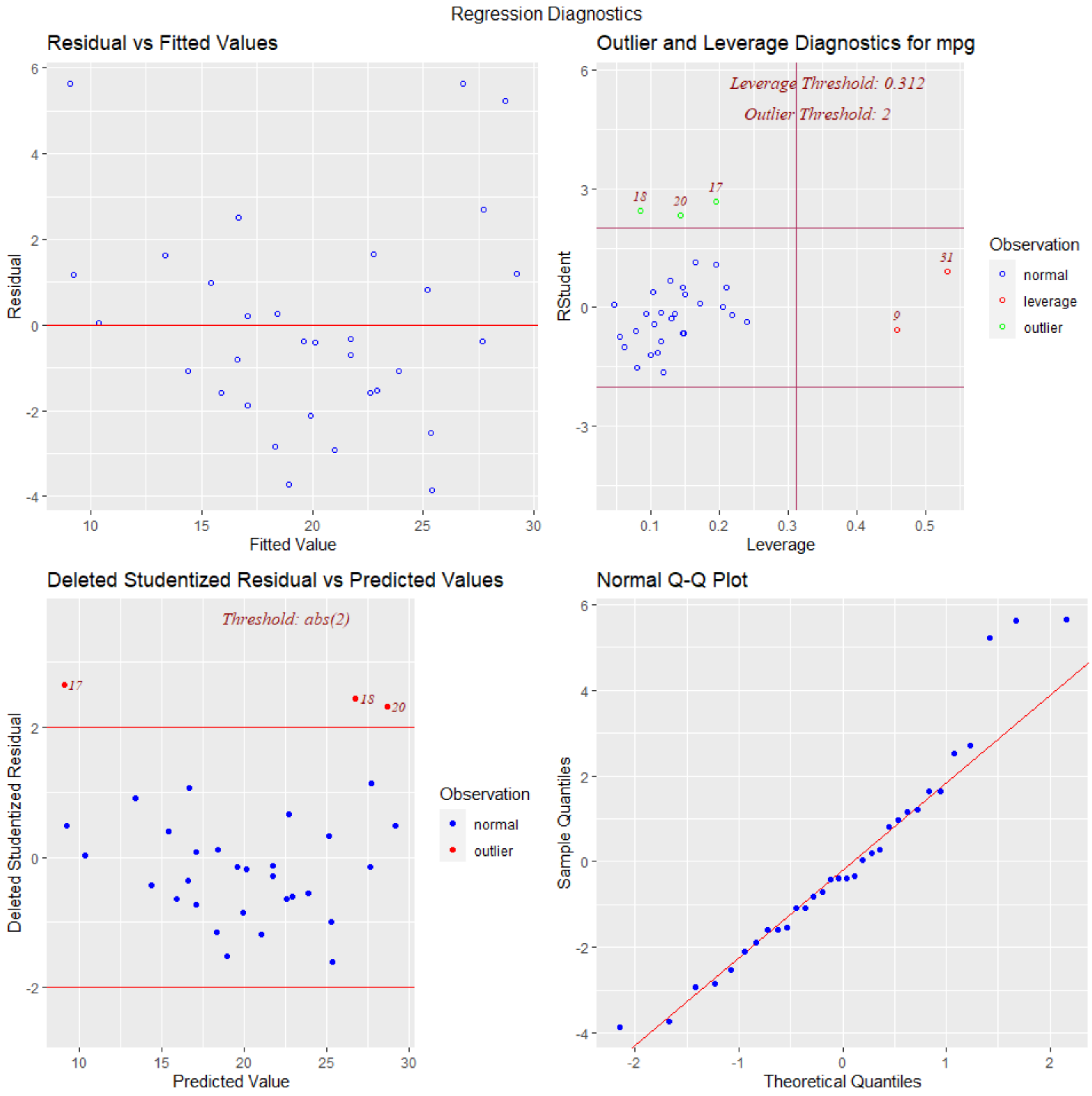

Cook's D Chart
Threshold: 0.125
Cook's D
0.0
0.1
0.2
Observation
0
10
20
30
17
20
31

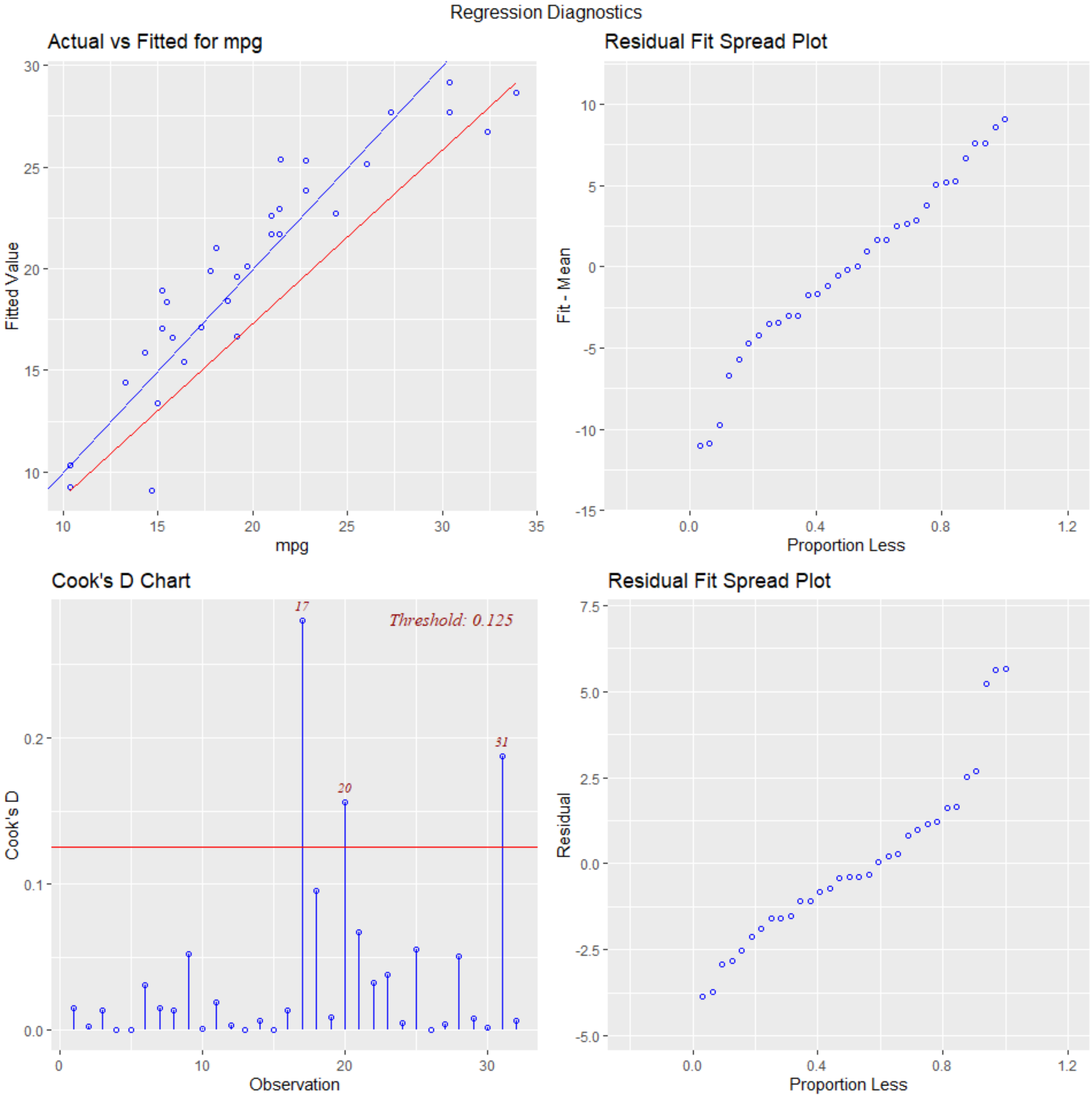

Residual Fit Spread Plot
10
5
0
-5
-10
-15
Fit - Mean
0.0
0.4
0.8
1.2
Proportion Less

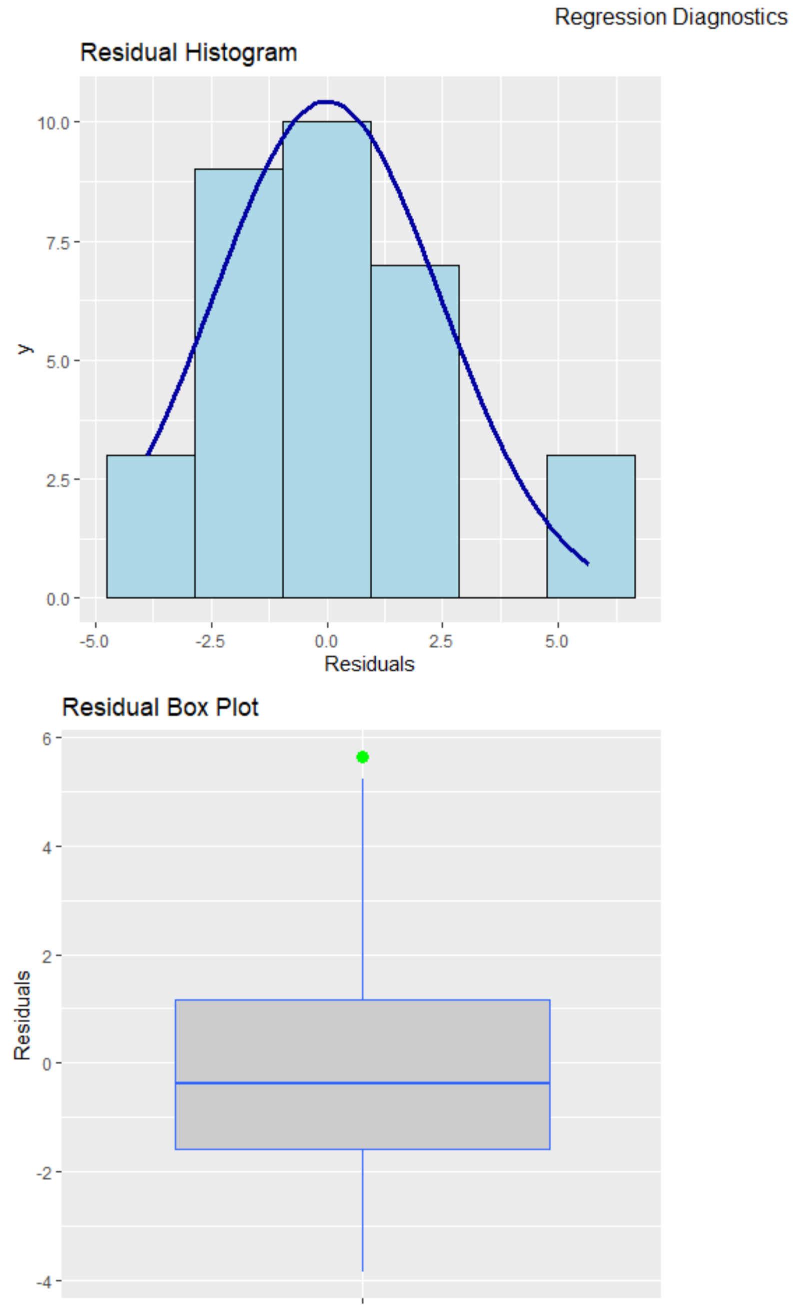

Residual Fit Spread Plot
Residual
7.5
5.0
2.5
0.0
-2.5
-5.0
0.0
0.4
0.8
1.2
Proportion Less


Residual Histogram
y
10.0
7.5
5.0
2.5
0.0
-5.0
-2.5
0.0
2.5
5.0
Residuals


Residual Box Plot
Residuals
6
4
2
0
-2
-4


Regression Diagnostics
Residual vs Fitted Values
Residual
Fitted Value
6
4
2
0
-2
-4
10
15
20
25
30
Outlier and Leverage Diagnostics for mpg
Leverage Threshold: 0.312
Outlier Threshold: 2
RStudent
Leverage
6
3
0
-3
0.1
0.2
0.3
0.4
0.5
18
20
17
31
9
Observation
normal
leverage
outlier
Deleted Studentized Residual vs Predicted Values
Threshold: abs(2)
Deleted Studentized Residual
Predicted Value
17
18
20
2
0
-2
10
15
20
25
30
Observation
normal
outlier
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
-2
-4
-2
-1
0
1
2


Regression Diagnostics
Actual vs Fitted for mpg
Fitted Value
30
25
20
15
10
10
15
20
25
30
35
mpg
Residual Fit Spread Plot
Fit - Mean
10
5
0
-5
-10
-15
0.0
0.4
0.8
1.2
Proportion Less
Cook's D Chart
Cook's D
17
Threshold: 0.125
31
20
0.2
0.1
0.0
0
10
20
30
Observation
Residual Fit Spread Plot
Residual
7.5
5.0
2.5
0.0
-2.5
-5.0
0.0
0.4
0.8
1.2
Proportion Less



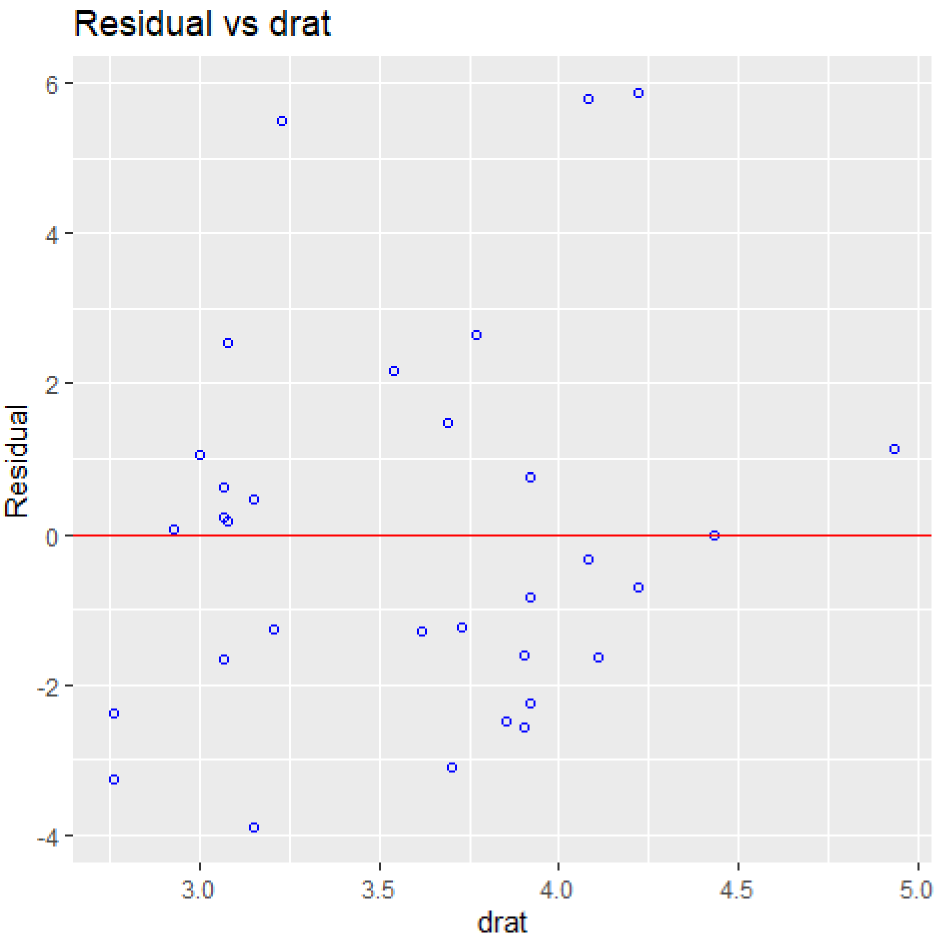
Variable Contributions

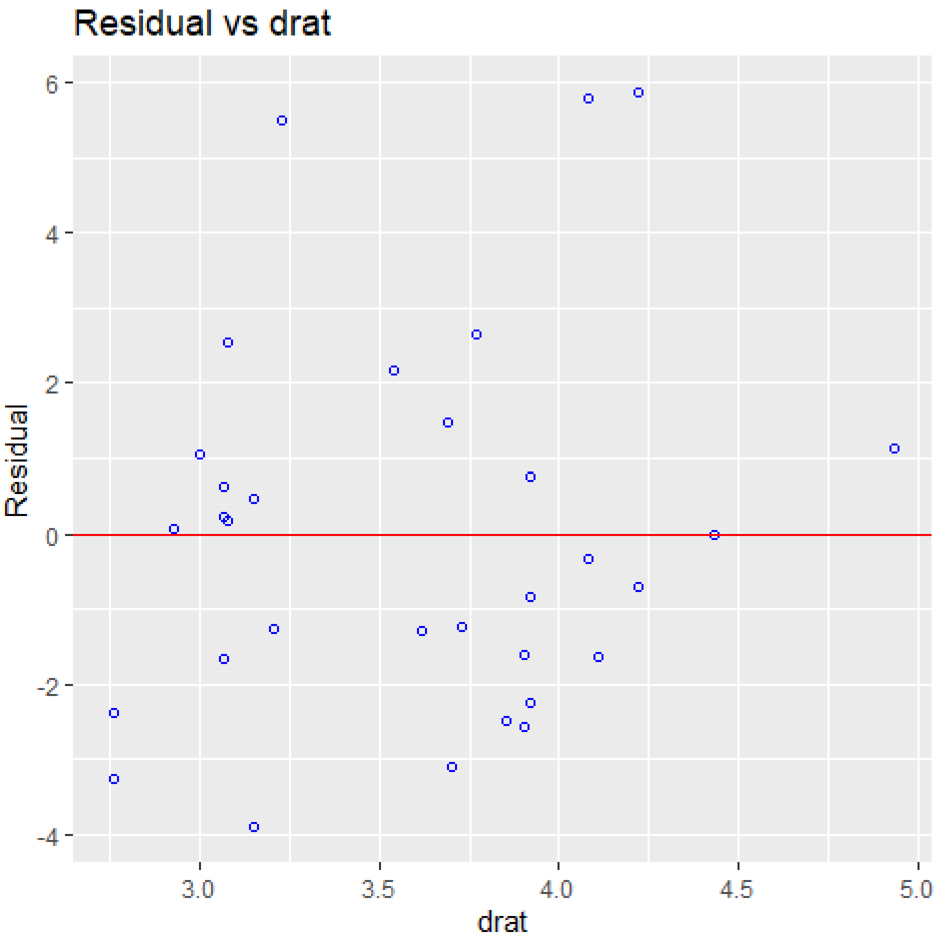
Residual vs Regressor Plots

Graph to determine whether we should add a new predictor to the model already containing other predictors. The residuals from the model is regressed on the new predictor and if the plot shows non random pattern, you should consider adding the new predictor to the model.

Hide

```
model <- lm(mpg ~ disp + hp + wt, data = mtcars)
ols_plot_resid_regressor(model, 'drat')
```


Residual vs drat
6
4
2
0
-2
-4
Residual
3.0
3.5
4.0
4.5
5.0
drat


Residual vs drat
6
4
2
0
-2
-4
Residual
3.0
3.5
4.0
4.5
5.0
drat

Added Variable Plot

Added variable plot provides information about the marginal importance of a predictor variable X_k, given the other predictor variables already in the model. It shows the marginal importance of the variable in reducing the residual variability.

The added variable plot was introduced by Mosteller and Tukey (1977). It enables us to visualize the regression coefficient of a new variable being considered to be included in a model. The plot can be constructed for each predictor variable.

Let us assume we want to test the effect of adding/removing variable X from a model. Let the response variable of the model be Y

Steps to construct an added variable plot:

- Regress Y on all variables other than X and store the residuals (Y residuals).
- Regress X on all the other variables included in the model (X residuals).
- Construct a scatter plot of Y residuals and X residuals.

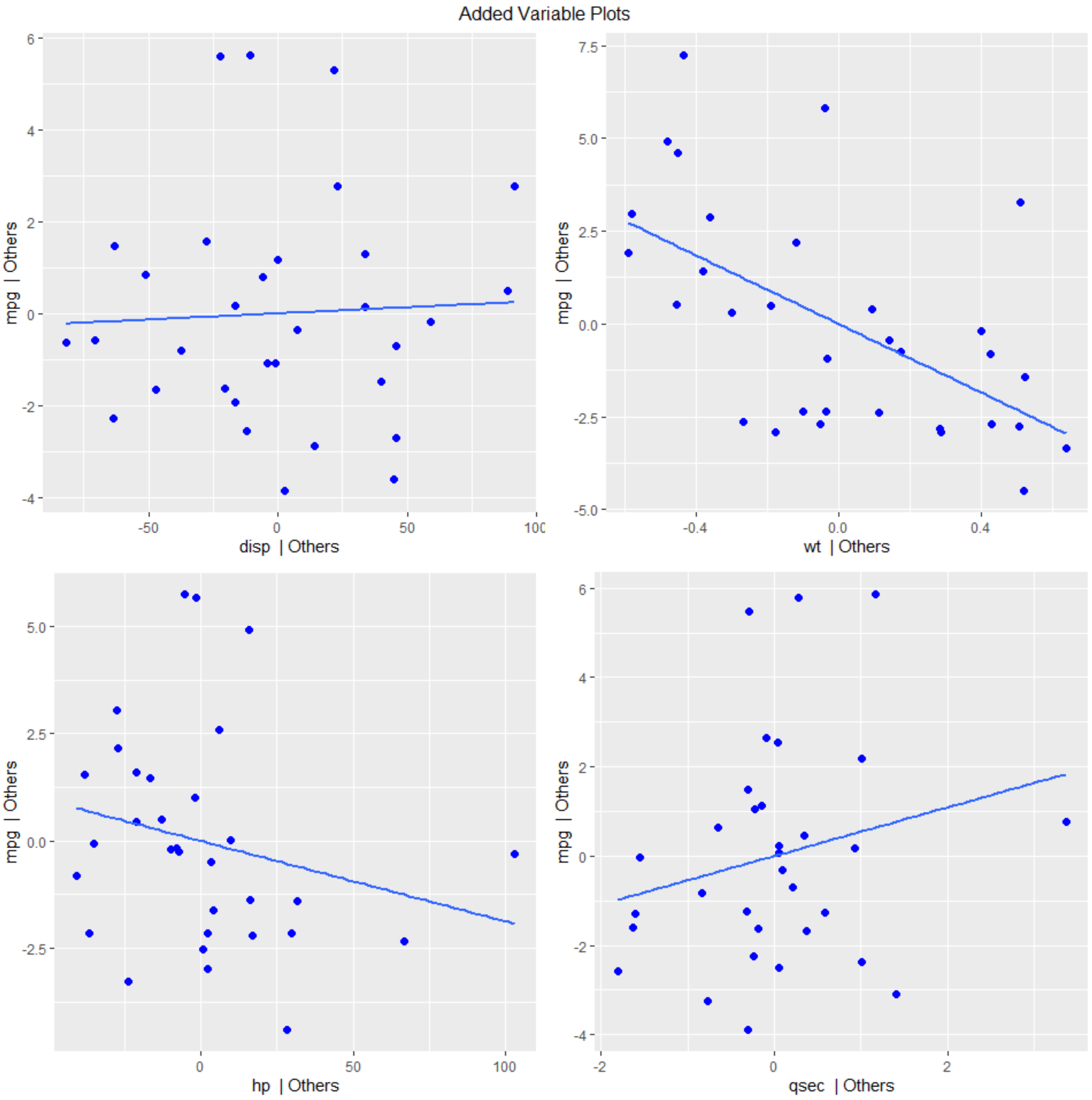
What do the Y and X residuals represent? The Y residuals represent the part of **Y** not explained by all the variables other than X. The X residuals represent the part of **X** not explained by other variables. The slope of the line fitted to the points in the added variable plot is equal to the regression coefficient when **Y** is regressed on all variables including **X**.

A strong linear relationship in the added variable plot indicates the increased importance of the contribution of **X** to the model already containing the other predictors.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_added_variable(model)
```

```
`geom_smooth()` using formula 'y ~ x'
`geom_smooth()` using formula 'y ~ x'
`geom_smooth()` using formula 'y ~ x'
`geom_smooth()` using formula 'y ~ x'
```



Residual Plus Component Plot

The residual plus component plot was introduced by Ezekeil (1924). It was called as Partial Residual Plot by Larsen and McCleary (1972). Hadi and Chatterjee (2012) called it the residual plus component plot.

Steps to construct the plot:

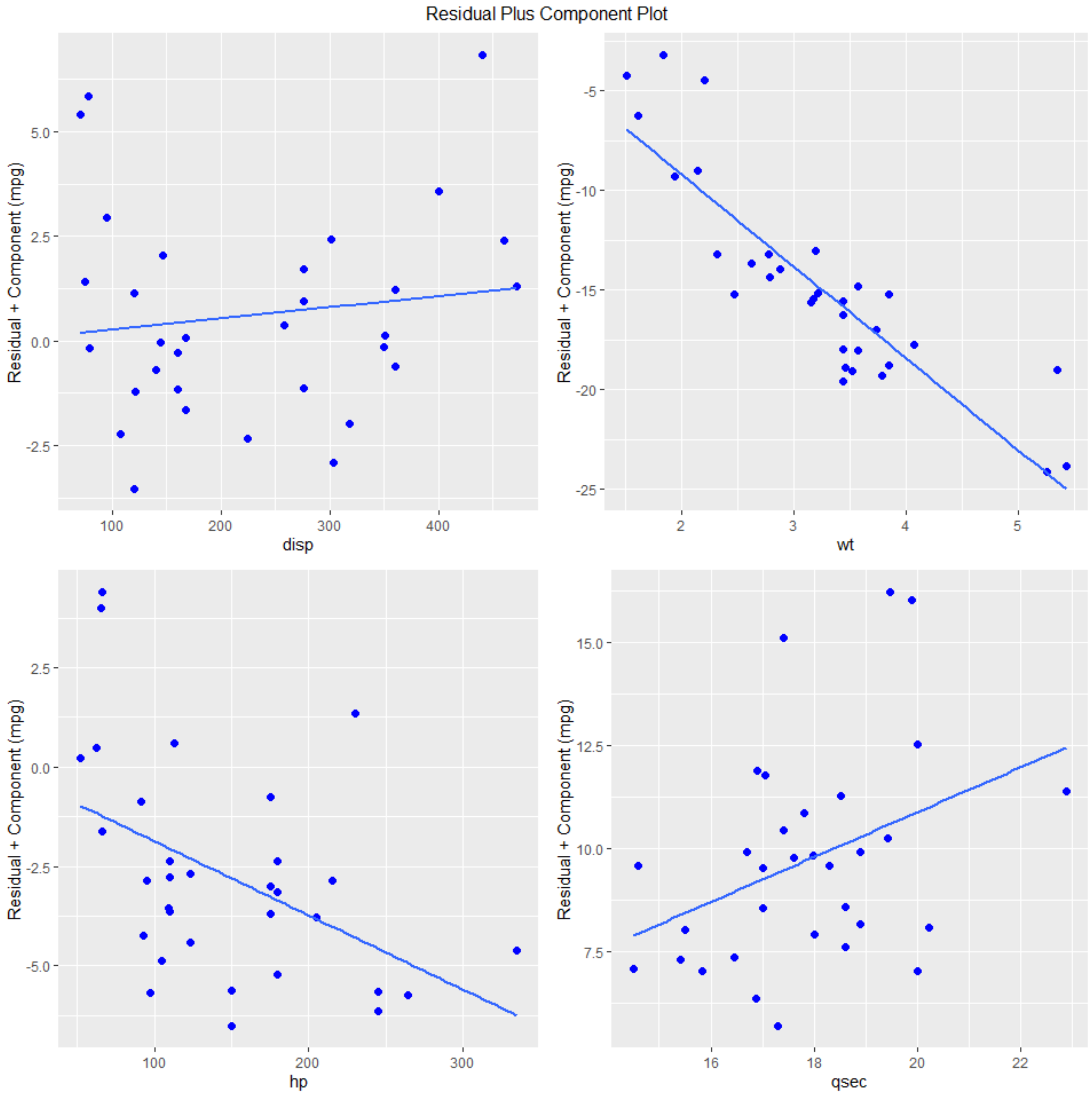
- Regress **Y** on all variables including **X** and store the residuals (**e**).
- Multiply **e** with regression coefficient of **X** (**eX**).
- Construct scatter plot of **eX** and **X**

The residual plus component plot indicates whether any non-linearity is present in the relationship between **Y** and **X** and can suggest possible transformationsfor linearizing the data.

Hide

```
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
ols_plot_comp_plus_resid(model)
```

```
`geom_smooth()` using formula 'y ~ x'
`geom_smooth()` using formula 'y ~ x'
`geom_smooth()` using formula 'y ~ x'
`geom_smooth()` using formula 'y ~ x'
```



A Short Introduction to the blorr Package

Introduction

The blorr package offers tools for building and validating binary logistic regression models. It is most suitable for beginner/intermediate R users and those who teach statistics using R. The APIis very simple and most of the functions take either a `data.frame` / `tibble` or a `model` as input. **blorr** useconsistent prefix **blr_** for easy tab completion.

Installation

You can install **blorr** using:

Hide

```
install.packages("blorr")
```

The documentation of the package can be found at https://blorr.rsquaredacademy.com (https://blorr.rsquaredacademy.com). This vignette gives a quick tour of the package.

Libraries

The following libraries are used in the examples in the vignette:

Hide

```
library(blorr)
```

```
Warning: package ‘blorr’ was built under R version 4.1.3
Registered S3 method overwritten by 'data.table':
  method           from
  print.data.table
```

Hide

```
library(magrittr)
```

Data

To demonstrate the features of blorr, we will use the bank marketing data set. The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (‘yes’) or not (‘no’) subscribed. It contains a random sample (~4k) of the original data set which can be found at https://archive.ics.uci.edu/ml/datasets/bank+marketing (https://archive.ics.uci.edu/ml/datasets/bank+marketing).

Bivariate Analysis

Let us begin with careful bivariate analysis of each possible variable and the outcome variable. We will use information value and likelihood ratio chi square test for selecting the initial set of predictors for our model. The bivariateanalysis is currently avialable for categorical predictors only.

Hide

```
blr_bivariate_analysis(bank_marketing, y, job, marital, education, default,
  housing, loan, contact, poutcome)
```

```
                          Bivariate Analysis
----------------------------------------------------------------------
Variable     Information Value    LR Chi Square    LR DF    LR p-value
----------------------------------------------------------------------
   job             0.16              75.2690        11        0.0000
 marital           0.05              21.6821         2        0.0000
education          0.05              25.0466         3        0.0000
 default           0.02              6.0405          1        0.0140
 housing           0.16              72.2813         1        0.0000
  loan             0.06              26.6615         1        0.0000
 contact           0.31             124.3834         2        0.0000
poutcome           0.53             270.6450         3        0.0000
----------------------------------------------------------------------
```

Weight of Evidence & Information Value

Weight of evidence (WoE) is used to assess the relative risk of diferent attributes for a characteristic and as a means to transform characteristics into variables. It is also a very useful tool for binning. The WoE for any group with average odds is zero. A negative WoE indicates that the proportion of defaults is higher for that attribute than the overall proportion and indicates higher risk.

The information value is used to rank order variables in terms of their predictive power. A high information value indicates a high ability to discriminate. Values for the information value will always be positive and may be above 3 when assessing highly predictive characteristics. Characteristics with information values less than 0:10 are typically viewed as weak, while values over 0.30 are sought after.

Hide

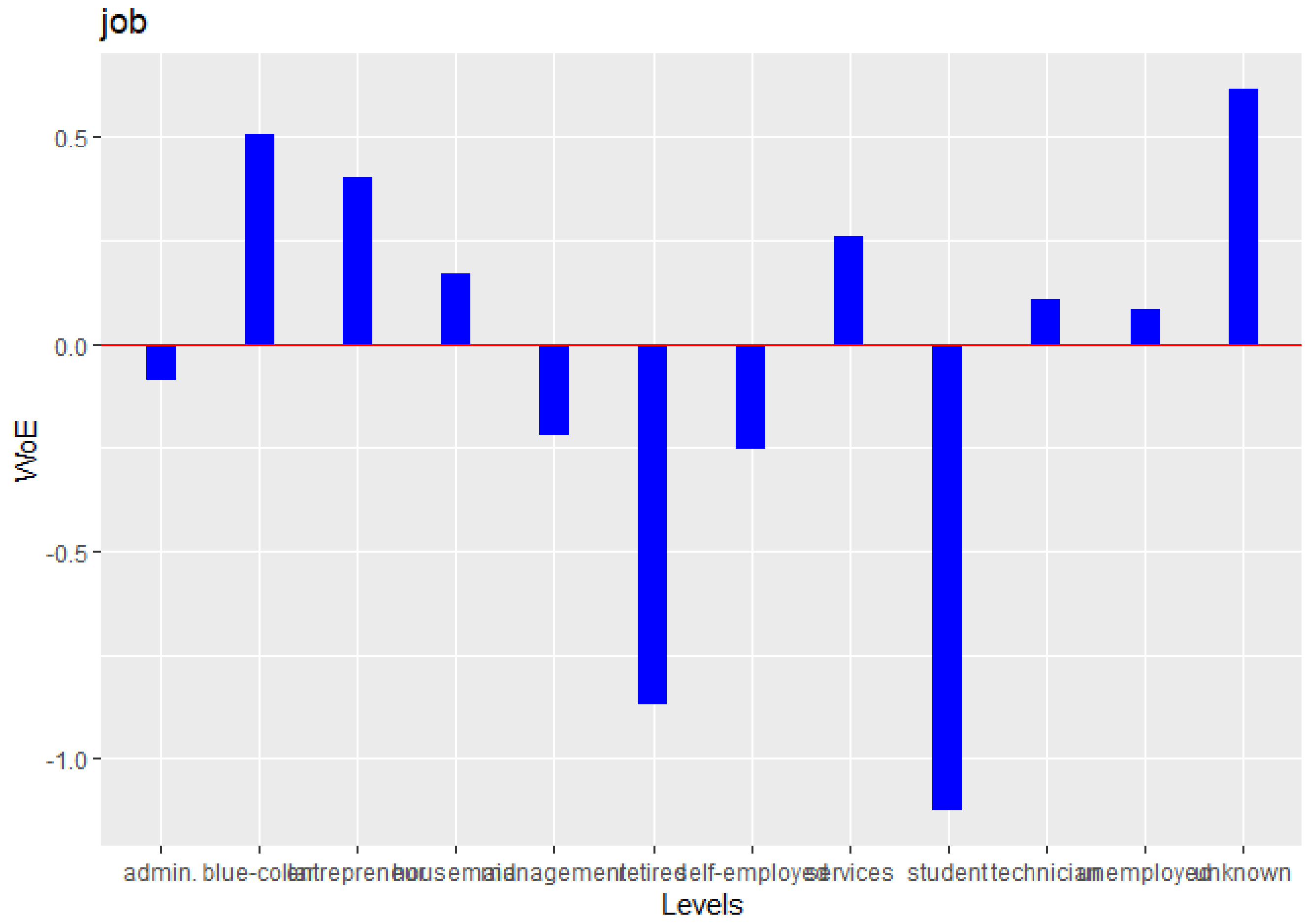
```
blr_woe_iv(bank_marketing, job, y)
```

```
                               Weight of Evidence
--------------------------------------------------------------------------------
   levels        count_0s    count_1s    dist_0s    dist_1s        woe      iv
--------------------------------------------------------------------------------
  management        809         130         0.20       0.25      -0.22     0.01
  technician        682          79         0.17       0.15       0.11     0.00
entrepreneur        139          12         0.03       0.02       0.40     0.00
  blue-collar       937          73         0.23       0.14       0.51     0.05
    unknown          29          2          0.01       0.00       0.61     0.00
    retired         152          47         0.04       0.09      -0.87     0.05
    admin.          433          61         0.11       0.12      -0.09     0.00
   services         392          39         0.10       0.08       0.26     0.01
self-employed       132          22         0.03       0.04      -0.26     0.00
  unemployed        126          15         0.03       0.03       0.08     0.00
   housemaid        110          12         0.03       0.02       0.17     0.00
    student          63          25         0.02       0.05      -1.13     0.04
--------------------------------------------------------------------------------


      Information Value
-----------------------------
Variable    Information Value
-----------------------------
  job             0.1594
-----------------------------
```

Plot

Hide

```
k <- blr_woe_iv(bank_marketing, job, y)
plot(k)
```



Multiple Variables

We can generate the weight of evidence and information value for multiple variables using `blr_woe_iv_stats()`.

Hide

```
blr_woe_iv_stats(bank_marketing, y, job, marital, education)
```

```
Variable: job

                              Weight of Evidence
--------------------------------------------------------------------------------
   levels        count_0s    count_1s    dist_0s    dist_1s        woe      iv
--------------------------------------------------------------------------------
 management         809         130         0.20       0.25      -0.22     0.01
 technician         682          79         0.17       0.15       0.11     0.00
entrepreneur        139          12         0.03       0.02       0.40     0.00
 blue-collar        937          73         0.23       0.14       0.51     0.05
   unknown           29           2         0.01       0.00       0.61     0.00
   retired          152          47         0.04       0.09      -0.87     0.05
   admin.           433          61         0.11       0.12      -0.09     0.00
  services          392          39         0.10       0.08       0.26     0.01
self-employed       132          22         0.03       0.04      -0.26     0.00
 unemployed         126          15         0.03       0.03       0.08     0.00
  housemaid         110          12         0.03       0.02       0.17     0.00
   student           63          25         0.02       0.05      -1.13     0.04
--------------------------------------------------------------------------------

      Information Value
-----------------------------
Variable    Information Value
-----------------------------
  job            0.1594
-----------------------------


Variable: marital

                           Weight of Evidence
---------------------------------------------------------------------------
 levels     count_0s    count_1s    dist_0s    dist_1s        woe      iv
---------------------------------------------------------------------------
married       2467         273         0.62       0.53       0.15     0.01
 single       1079         191         0.27       0.37      -0.32     0.03
divorced      458           53         0.11       0.10       0.11     0.00
---------------------------------------------------------------------------

      Information Value
-----------------------------
Variable    Information Value
-----------------------------
marital          0.0464
-----------------------------


Variable: education
```

```
                             Weight of Evidence
----------------------------------------------------------------------------
 levels      count_0s    count_1s    dist_0s    dist_1s        woe      iv
----------------------------------------------------------------------------
tertiary       1104         195         0.28       0.38      -0.31    0.03
secondary      2121         231         0.53       0.45       0.17    0.01
 unknown       154           25         0.04       0.05      -0.23    0.00
 primary       625           66         0.16       0.13       0.20    0.01
----------------------------------------------------------------------------


      Information Value
------------------------------
Variable     Information Value
------------------------------
education         0.0539
------------------------------
```

`blr_woe_iv()` and `blr_woe_iv_stats()` are currently available for categorical predictors only.

Stepwise Selection

For the initial/ first cut model, all the independent variables are put into the model. Our goal is to include a limited number of independent variables (5-15) which are all significant, without sacrificing too much on the model performance. The rationale behind not-including too many variables is that the model would be over fitted and would become unstable when tested on the validation sample. The variable reduction is done using forward or backward or stepwise variable selection procedures. We will use `blr_step_aic_both()` to shortlist predictors for our model.

Model

Hide

```
model <- glm(y ~ ., data = bank_marketing, family = binomial(link = 'logit'))
```

Selection Summary

Hide

```
blr_step_aic_both(model)
```

```
Stepwise Selection Method
-------------------------

Candidate Terms:

1 . age
2 . job
3 . marital
4 . education
5 . default
6 . balance
7 . housing
8 . loan
9 . contact
10 . day
11 . month
12 . duration
13 . campaign
14 . pdays
15 . previous
16 . poutcome


Variables Entered/Removed:

- duration added
- poutcome added
- month added
- contact added
- housing added
- loan added
- campaign added
- marital added
- education added
- age added

No more variables to be added or removed.

                     Stepwise Summary
---------------------------------------------------------
Variable      Method       AIC         BIC       Deviance
---------------------------------------------------------
duration     addition    2674.384    2687.217    2670.384
poutcome     addition    2396.014    2428.097    2386.014
month        addition    2274.109    2376.773    2242.109
contact      addition    2207.884    2323.381    2171.884
housing      addition    2184.550    2306.463    2146.550
loan         addition    2171.972    2300.302    2131.972
```

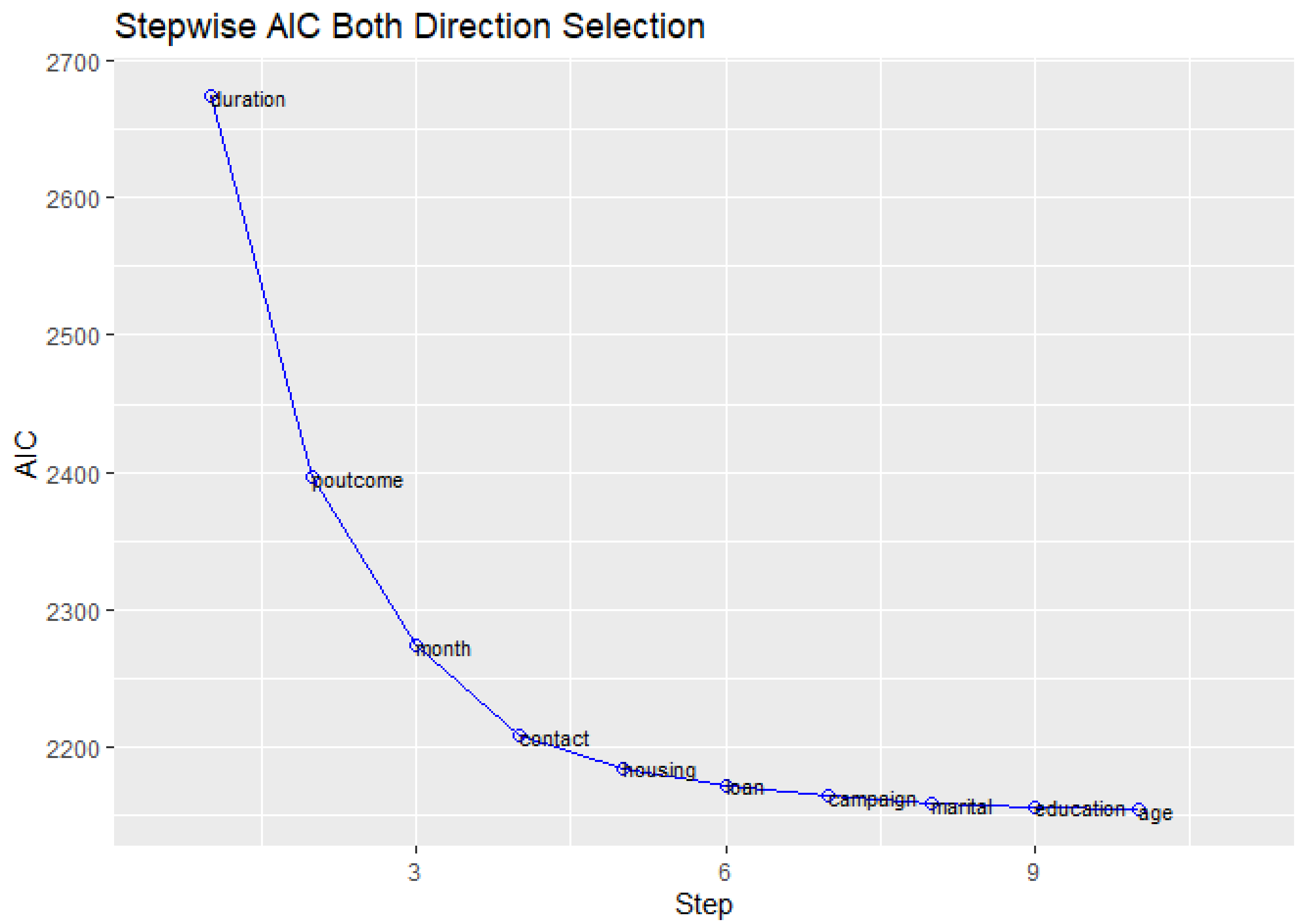
```
campaign     addition    2164.164    2298.910    2122.164
marital      addition    2158.524    2306.103    2112.524
education    addition    2155.837    2322.666    2103.837
age          addition    2154.272    2327.517    2100.272
---------------------------------------------------------
```

Plot

Hide

```
model %>%
  blr_step_aic_both() %>%
  plot()
```

```
Stepwise Selection Method
-------------------------

Candidate Terms:

1 . age
2 . job
3 . marital
4 . education
5 . default
6 . balance
7 . housing
8 . loan
9 . contact
10 . day
11 . month
12 . duration
13 . campaign
14 . pdays
15 . previous
16 . poutcome


Variables Entered/Removed:

- duration added
- poutcome added
- month added
- contact added
- housing added
- loan added
- campaign added
- marital added
- education added
- age added

No more variables to be added or removed.
```



Regression Output

Model

We can use bivariate analysis and stepwise selection procedure to shortlist predictors and build the model using the `glm()` . The predictors used in the below model are for illustration purposes and not necessarily shortlisted from the bivariate analysis and variable selection procedures.

Hide

```
model <- glm(y ~  age + duration + previous + housing + default +
             loan + poutcome + job + marital, data = bank_marketing,
             family = binomial(link = 'logit'))
```

Use `blr_regress()` to generate comprehensive regression output. It accepts either of the following

- model built using `glm()`
- model formula and data

Using Model

Let us look at the output generated from `blr_regress()` :

Hide

```
blr_regress(model)
```

```
- Creating model overview.
- Creating response profile.
- Extracting maximum likelihood estimates.
- Estimating concordant and discordant pairs.
                             Model Overview
------------------------------------------------------------------------
Data Set    Resp Var    Obs.    Df. Model    Df. Residual    Convergence
------------------------------------------------------------------------
  data         y        4521      4520           4498           TRUE
------------------------------------------------------------------------


                    Response Summary
--------------------------------------------------------
Outcome        Frequency        Outcome        Frequency
--------------------------------------------------------
   0             4004              1              517
--------------------------------------------------------


                     Maximum Likelihood Estimates
-----------------------------------------------------------------------
   Parameter        DF    Estimate    Std. Error    z value     Pr(>|z|)
-----------------------------------------------------------------------
  (Intercept)       1      -5.1347        0.3728    -13.7729      0.0000
      age           1       0.0096        0.0067      1.4299      0.1528
    duration        1       0.0042         2e-04     20.7853      0.0000
    previous        1      -0.0357        0.0392     -0.9089      0.3634
   housingno        1       0.7894        0.1232      6.4098      0.0000
   defaultyes       1      -0.8691        0.6919     -1.2562      0.2091
     loanno         1       0.6598        0.1945      3.3925       7e-04
poutcomefailure     1       0.6085        0.2012      3.0248      0.0025
 poutcomeother      1       1.1354        0.2700      4.2057      0.0000
poutcomesuccess     1       3.2481        0.2462     13.1913      0.0000
 jobtechnician      1      -0.2713        0.1806     -1.5019      0.1331
jobentrepreneur     1      -0.7041        0.3809     -1.8486      0.0645
 jobblue-collar     1      -0.6132        0.1867     -3.2851      0.0010
   jobunknown       1      -0.9932        0.8226     -1.2073      0.2273
   jobretired       1       0.3197        0.2729      1.1713      0.2415
   jobadmin.        1       0.1120        0.2001      0.5599      0.5755
  jobservices       1      -0.1750        0.2265     -0.7728      0.4397
jobself-employed    1      -0.1408        0.3009     -0.4680      0.6398
 jobunemployed      1      -0.6581        0.3432     -1.9174      0.0552
  jobhousemaid      1      -0.7456        0.3932     -1.8963      0.0579
   jobstudent       1       0.1927        0.3433      0.5613      0.5746
 maritalsingle      1       0.5451        0.1387      3.9299       1e-04
maritaldivorced     1      -0.1989        0.1986     -1.0012      0.3167
-----------------------------------------------------------------------

 Association of Predicted Probabilities and Observed Responses
```

```
---------------------------------------------------------------
% Concordant          0.8886          Somers' D        0.7773
% Discordant          0.1114          Gamma            0.7773
% Tied                0.0000          Tau-a            0.1575
Pairs                2070068          c                0.8886
---------------------------------------------------------------
```

If you want to examine the odds ratio estimates, set `odd_conf_limit` to `TRUE`. The odds ratio estimates are not explicitly computed as we observed considerable increase in computation time when dealing with large data sets.

Using Formula

Let us use the model formula and the data set to generate the above results.

Hide

```
blr_regress(y ~  age + duration + previous + housing + default +
             loan + poutcome + job + marital, data = bank_marketing)
```

```
- Creating model overview.
- Creating response profile.
- Extracting maximum likelihood estimates.
- Estimating concordant and discordant pairs.
                             Model Overview
------------------------------------------------------------------------
Data Set    Resp Var    Obs.    Df. Model    Df. Residual    Convergence
------------------------------------------------------------------------
  data         y        4521      4520           4498           TRUE
------------------------------------------------------------------------


                    Response Summary
--------------------------------------------------------
Outcome        Frequency        Outcome        Frequency
--------------------------------------------------------
   0             4004              1              517
--------------------------------------------------------


                     Maximum Likelihood Estimates
-----------------------------------------------------------------------
   Parameter        DF    Estimate    Std. Error    z value     Pr(>|z|)
-----------------------------------------------------------------------
  (Intercept)       1     -5.1347        0.3728    -13.7729      0.0000
      age           1      0.0096        0.0067      1.4299      0.1528
    duration        1      0.0042         2e-04     20.7853      0.0000
    previous        1     -0.0357        0.0392     -0.9089      0.3634
   housingno        1      0.7894        0.1232      6.4098      0.0000
   defaultyes       1     -0.8691        0.6919     -1.2562      0.2091
     loanno         1      0.6598        0.1945      3.3925       7e-04
poutcomefailure     1      0.6085        0.2012      3.0248      0.0025
 poutcomeother      1      1.1354        0.2700      4.2057      0.0000
poutcomesuccess     1      3.2481        0.2462     13.1913      0.0000
 jobtechnician      1     -0.2713        0.1806     -1.5019      0.1331
jobentrepreneur     1     -0.7041        0.3809     -1.8486      0.0645
 jobblue-collar     1     -0.6132        0.1867     -3.2851      0.0010
   jobunknown       1     -0.9932        0.8226     -1.2073      0.2273
   jobretired       1      0.3197        0.2729      1.1713      0.2415
   jobadmin.        1      0.1120        0.2001      0.5599      0.5755
  jobservices       1     -0.1750        0.2265     -0.7728      0.4397
jobself-employed    1     -0.1408        0.3009     -0.4680      0.6398
 jobunemployed      1     -0.6581        0.3432     -1.9174      0.0552
  jobhousemaid      1     -0.7456        0.3932     -1.8963      0.0579
   jobstudent       1      0.1927        0.3433      0.5613      0.5746
 maritalsingle      1      0.5451        0.1387      3.9299       1e-04
maritaldivorced     1     -0.1989        0.1986     -1.0012      0.3167
-----------------------------------------------------------------------

 Association of Predicted Probabilities and Observed Responses
```

```
---------------------------------------------------------------
% Concordant          0.8886          Somers' D        0.7773
% Discordant          0.1114          Gamma            0.7773
% Tied                0.0000          Tau-a            0.1575
Pairs                2070068          c                0.8886
---------------------------------------------------------------
```

Model Fit Statistics

Model fit statistics are available to assess how well the model fits the data and to compare two different models.The output includes likelihood ratio test, AIC, BIC and a host of pseudo r-squared measures. You can read more about pseudo r-squared at https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/ (https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/).

Single Model

Hide

```
blr_model_fit_stats(model)
```

```
                               Model Fit Statistics
---------------------------------------------------------------------------------
Log-Lik Intercept Only:     -1607.330    Log-Lik Full Model:             -1123.340
Deviance(4498):              2246.679    LR(22):                           967.980
                                         Prob > LR:                          0.000
MCFadden's R2                   0.301    McFadden's Adj R2:                  0.287
ML (Cox-Snell) R2:              0.193    Cragg-Uhler(Nagelkerke) R2:         0.379
McKelvey & Zavoina's R2:        0.388    Efron's R2:                         0.278
Count R2:                       0.904    Adj Count R2:                       0.157
BIC:                         2440.259    AIC:                             2292.679
---------------------------------------------------------------------------------
```

Model Validation

Confusion Matrix

In the event of deciding a cut-off point on the probability scores of a logistic regression model, a confusion matrix is created corresponding to a particular cut-off. The observations with probability scores above the cut-off score are predicted to be events and those below the cut-off score, as non-events. The confusion matrix, a 2X2 table, then calculates the number of correctly classified and miss-classified observations.

Hide

```
blr_confusion_matrix(model, cutoff = 0.5)
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3920  352
         1   84  165

                Accuracy : 0.9036
     No Information Rate : 0.8856

                   Kappa : 0.3851

McNemars's Test P-Value  : 0.0000

             Sensitivity : 0.3191
             Specificity : 0.9790
          Pos Pred Value : 0.6627
          Neg Pred Value : 0.9176
              Prevalence : 0.1144
          Detection Rate : 0.0365
    Detection Prevalence : 0.0551
       Balanced Accuracy : 0.6491
               Precision : 0.6627
                  Recall : 0.3191

        'Positive' Class : 1
```

The validity of a cut-off is measured using sensitivity, specificity and accuracy.

- **Sensitivity**: The % of correctly classified events out of all events = TP / (TP + FN)
- **Specificity**: The % of correctly classified non-events out of all non-events = TN / (TN + FP)
- **Accuracy**: The % of correctly classified observation over all observations = (TP + TN) / (TP + FP + TN + FN)
- *True Positive (TP)* : Events correctly classified as events.
- *True Negative (TN)* : Non-Events correctly classified as non-events.
- *False Positive (FP)*: Non-events miss-classified as events.
- *False Negative (FN)*: Events miss-classified as non-events.

For a standard logistic model, the higher is the cut-off, the lower will be the sensitivity and the higher would be the specificity. As the cut-off is decreased, sensitivity will go up, as then more events would be captured. Also, specificity will go down, as more non-events would miss-classified as events. Hence a trade-off is done based on the requirements. For example, if we are looking to capture as many events as possible, and we can afford to have miss-classified non-events, then a low cut-off is taken.

Hosmer Lemeshow Test

Hosmer and Lemeshow developed a goodness-of-fit test for logistic regression models with binary responses. The test involves dividing the data into approximately ten groups of roughly equal size based on the percentiles of the estimated probabilities. The observations are sorted in increasing order of their estimated probability of having an even outcome. The discrepancies between the observed and expected number of observations in these groups are summarized by the Pearson chi-square statistic, which is then compared to chi-square distribution with t degrees of freedom, where t is the number of groups minus 2. Lower values of Goodness-of-fit are preferred.

Hide

```
blr_test_hosmer_lemeshow(model)
```

```
           Partition for the Hosmer & Lemeshow Test
--------------------------------------------------------------
                        def = 1                 def = 0
Group    Total    Observed    Expected    Observed    Expected
--------------------------------------------------------------
  1       453        2          5.14        451        447.86
  2       452        3          8.63        449        443.37
  3       452        4         11.88        448        440.12
  4       452        7         15.29        445        436.71
  5       452        14        19.39        438        432.61
  6       452        10        24.97        442        427.03
  7       452        31        33.65        421        418.35
  8       452        62        49.74        390        402.26
  9       452       128        88.10        324        363.90
 10       452       256        260.21       196        191.79
--------------------------------------------------------------

     Goodness of Fit Test
------------------------------
Chi-Square    DF    Pr > ChiSq
------------------------------
 52.9942      8       0.0000
------------------------------
```

Gains Table & Lift Chart

A lift curve is a graphical representation of the % of cumulative events captured at a specific cut-off. The cut-off can be a particular decile or a percentile. Similar, to rank ordering procedure, the data is in descending order of the scores and is then grouped into deciles/percentiles. The cumulative number of observations and events are then computed for each decile/percentile. The lift curve is the created using the cumulative % population as the x-axis and the cumulative percentage of events as the y-axis.

Hide

```
blr_gains_table(model)
```

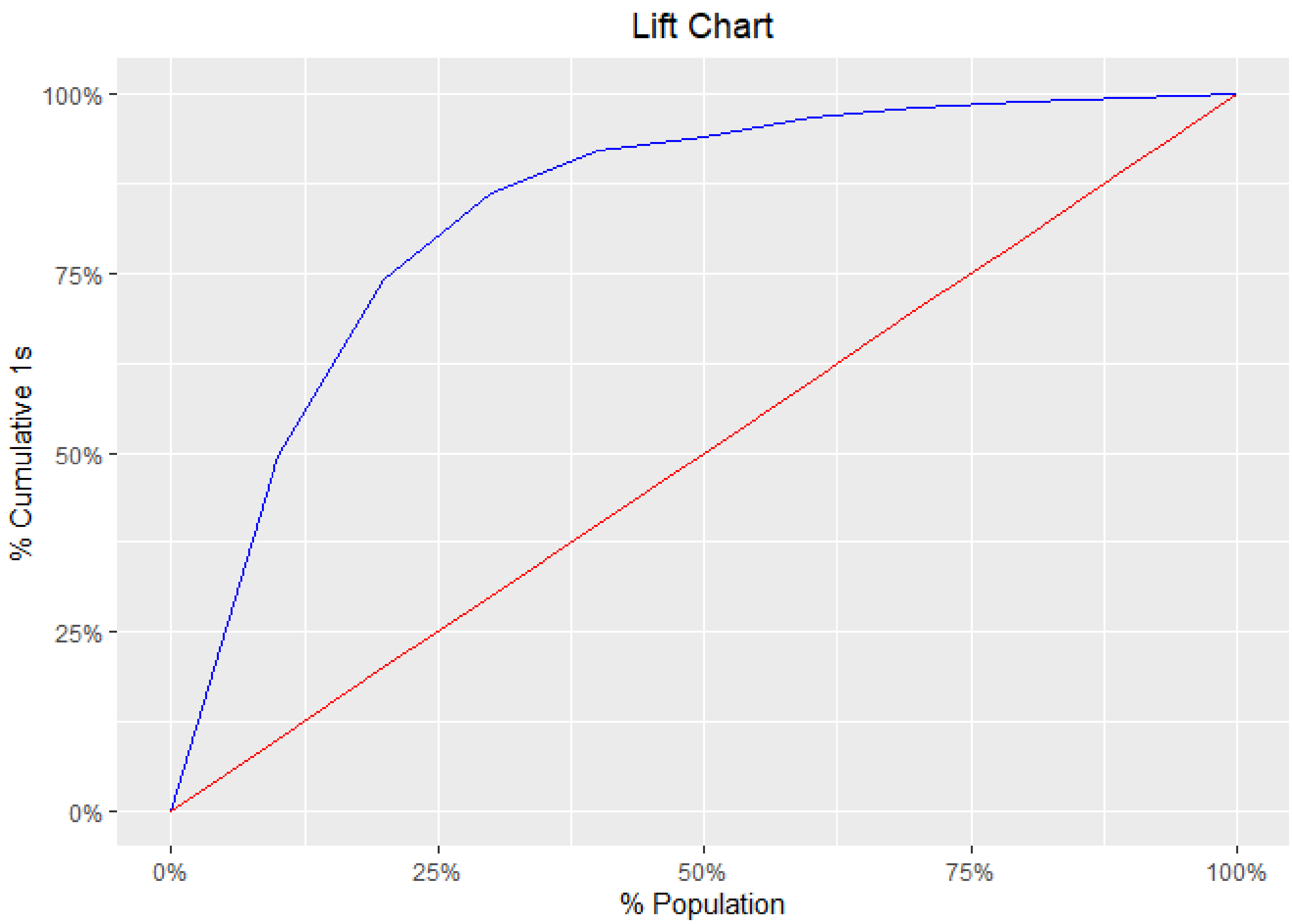
decile <dbl>	total <int>	1 <int>	0 <int>	ks <dbl>	tp <int>	tn <int>	fp <int>	fn <int>	sensitivity <dbl>	▸
1	452	256	196	44.62134	256	3808	196	261	49.51644	
2	452	128	324	61.28765	384	3484	520	133	74.27466	
3	452	62	390	63.53965	446	3094	910	71	86.26692	
4	452	31	421	59.02130	477	2673	1331	40	92.26306	
5	452	10	442	49.91657	487	2231	1773	30	94.19729	
6	452	14	438	41.68544	501	1793	2211	16	96.90522	
7	452	7	445	31.92552	508	1348	2656	9	98.25919	
8	452	4	448	21.51040	512	900	3104	5	99.03288	
9	452	3	449	10.87689	515	451	3553	2	99.61315	
10	453	2	451	0.00000	517	0	4004	0	100.00000	

1-10 of 10 rows | 1-10 of 12 columns

Lift Chart

Hide

```
model %>%
    blr_gains_table() %>%
    plot()
```

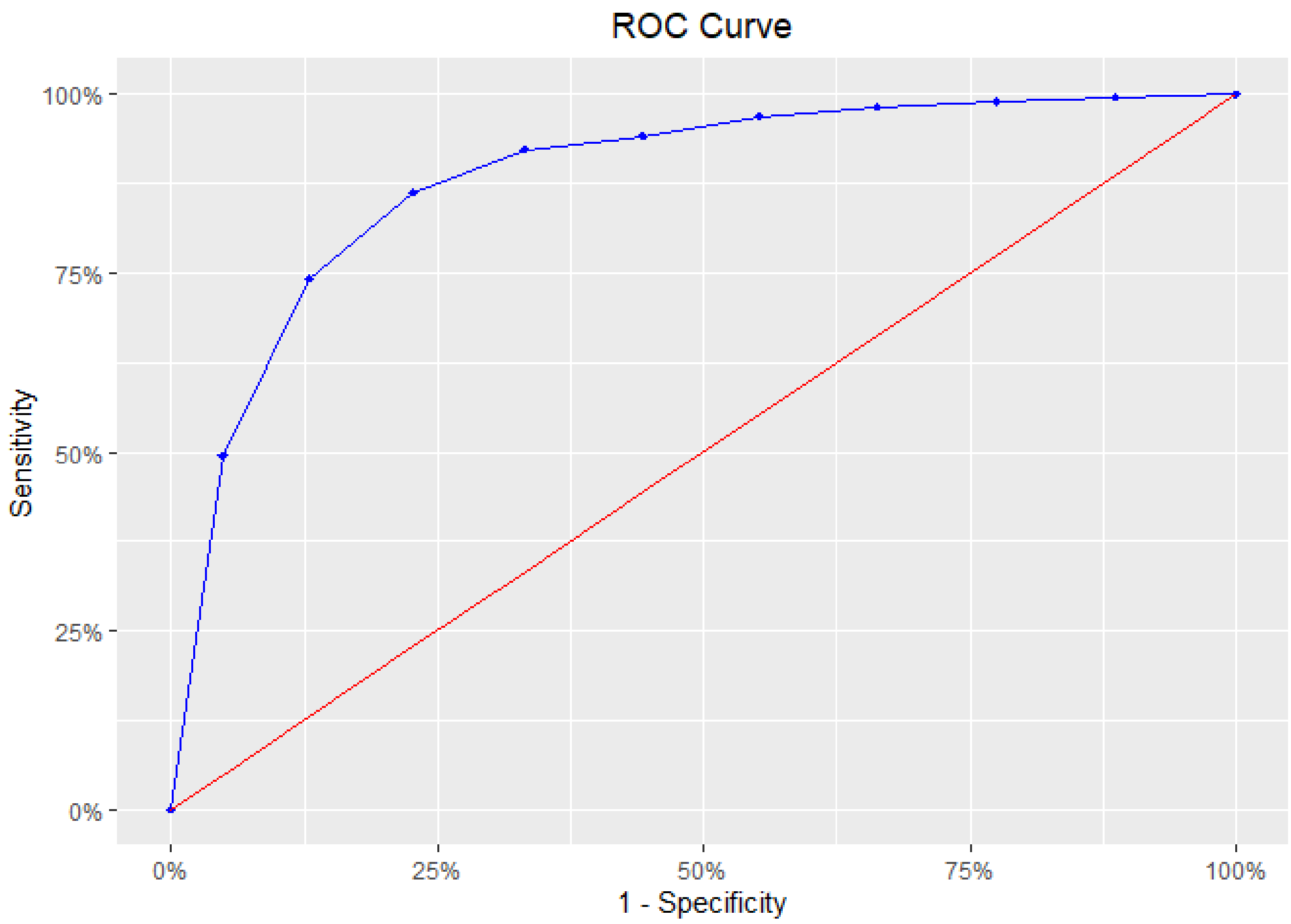


ROC Curve

ROC curve is a graphical representation of the validity of cut-offs for a logistic regression model. The ROC curve is plotted using the sensitivity and specificity for all possible cut-offs, i.e., all the probability scores. The graph is plotted using sensitivity on the y-axis and 1-specificity on the x-axis. Any point on the ROC curve represents a sensitivity X (1-specificity) measure corresponding to a cut-off. The area under the ROC curve is used as a validation measure for the model – the bigger the area the better is the model.

Hide

```
model %>%
    blr_gains_table() %>%
  blr_roc_curve()
```

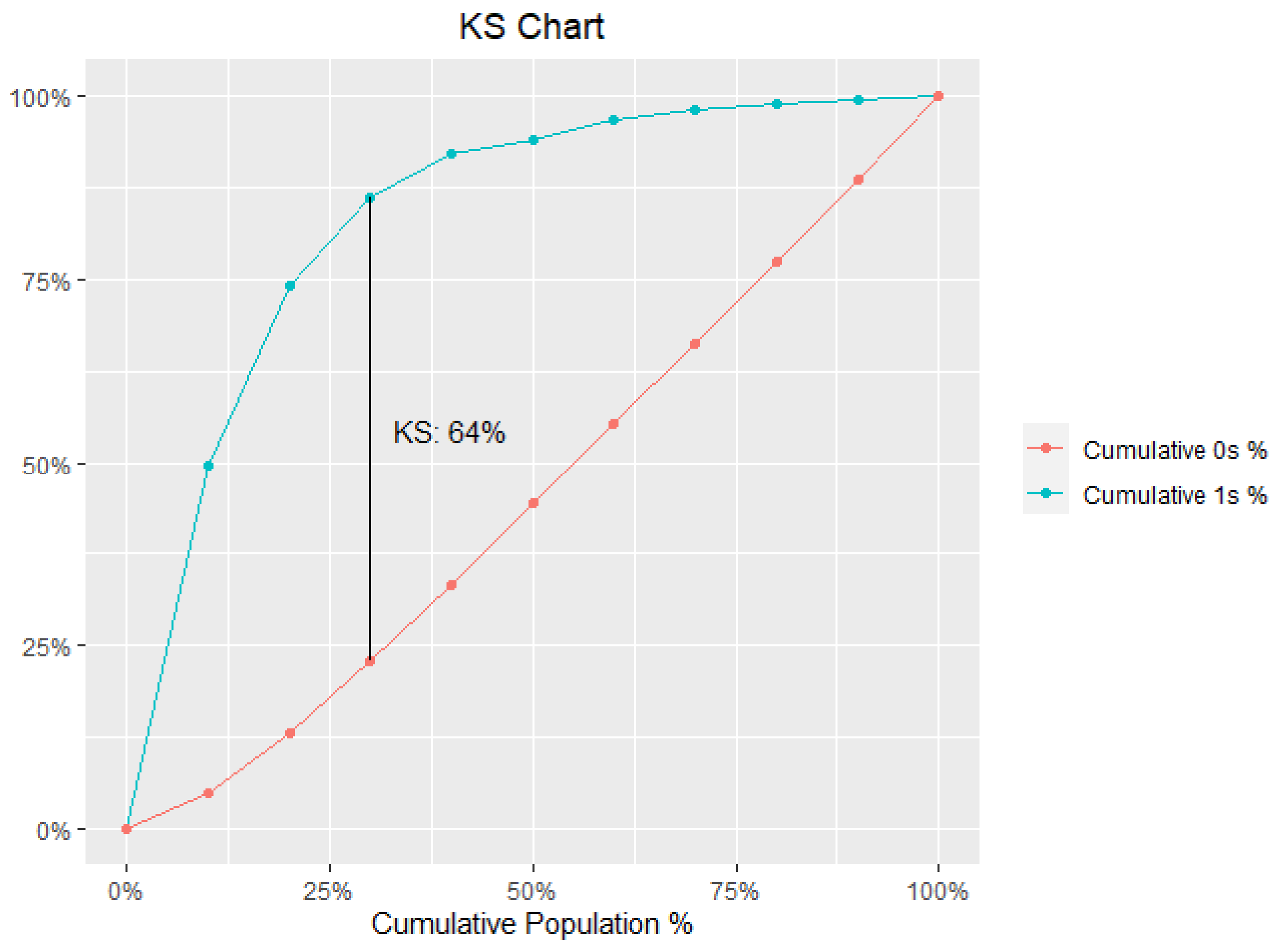


KS Chart

The KS Statistic is again a measure of model efficiency, and it is created using the lift curve. The lift curve is created to plot % events. If we also plot % non-events on the same scale, with % population at x-axis, we would get another curve. The maximum distance between the lift curve for events and that for non-events is termed as KS. For a good model, KS should be big (>=0.3) and should occur as close to the event rate as possible.

Hide

```
model %>%
    blr_gains_table() %>%
  blr_ks_chart()
```

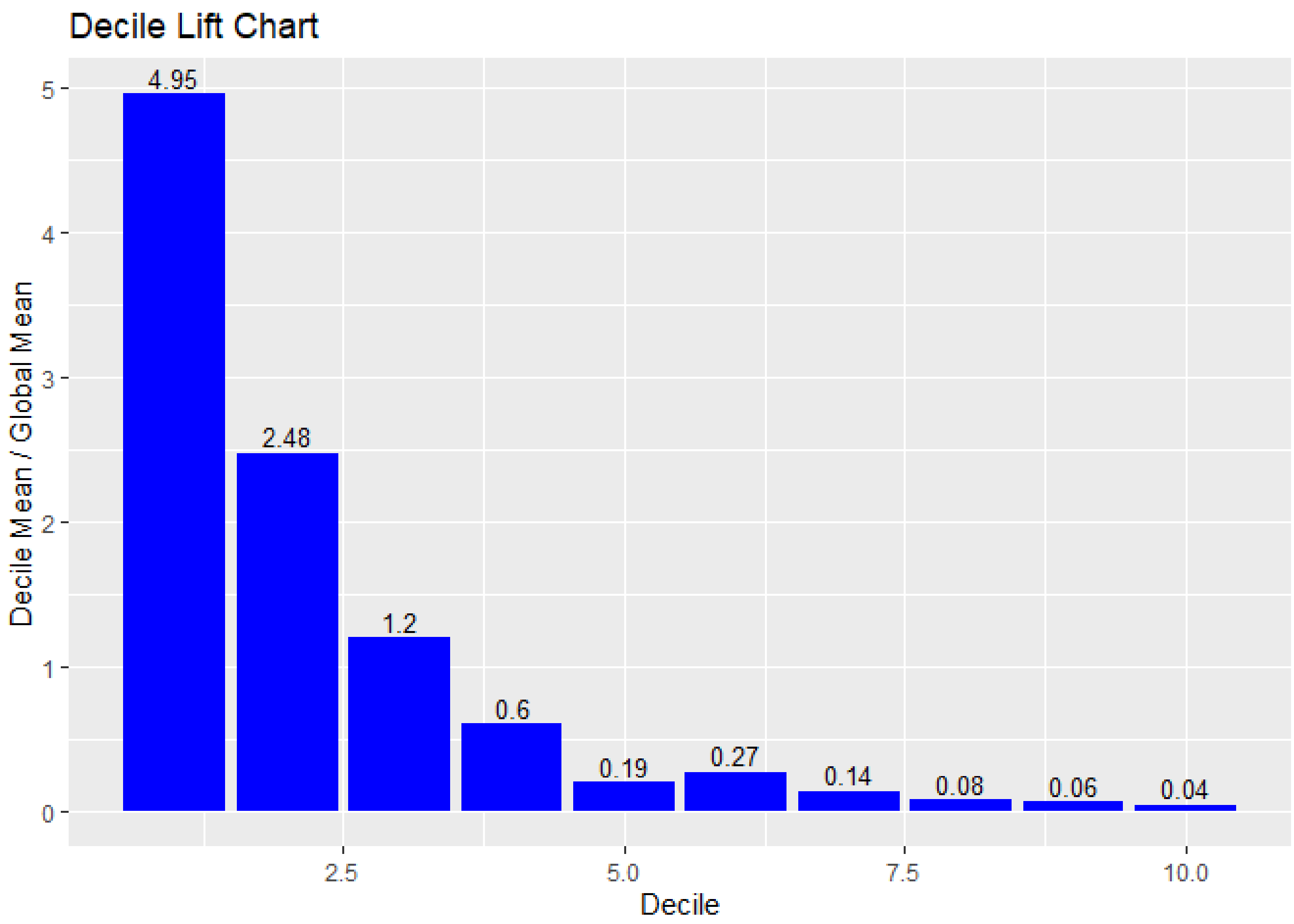


Decile Lift Chart

The decile lift chart displays the lift over the global mean event rate for each decile. For a model with good discriminatory power, the top deciles should have an event/conversion rate greater than the global mean.

Hide

```
model %>%
  blr_gains_table() %>%
  blr_decile_lift_chart()
```

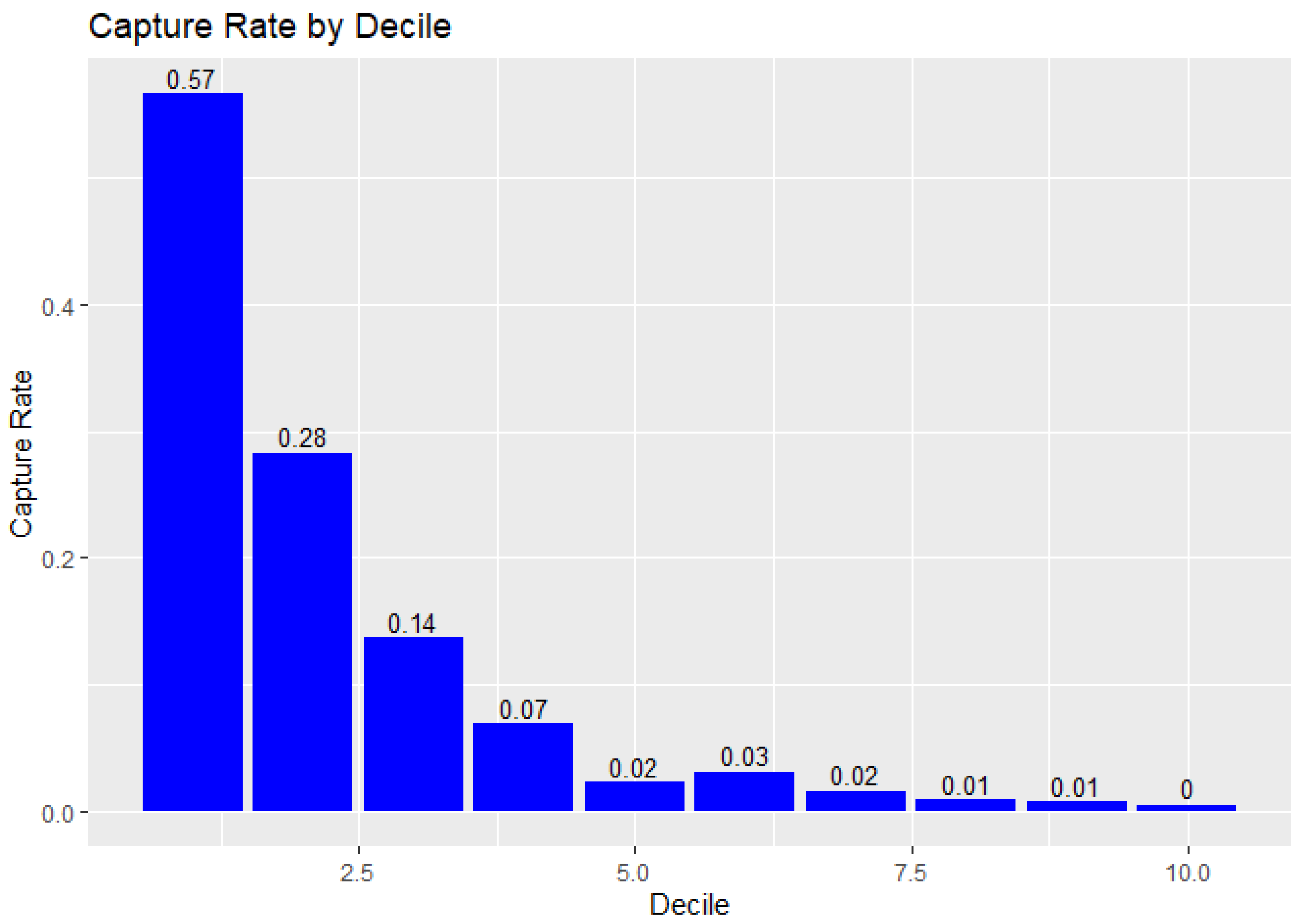


Capture Rate by Decile

If the model has good discriminatory power, the top deciles should have a higher event/conversion rate compared to the bottom deciles.

Hide

```
model %>%
  blr_gains_table() %>%
  blr_decile_capture_rate()
```

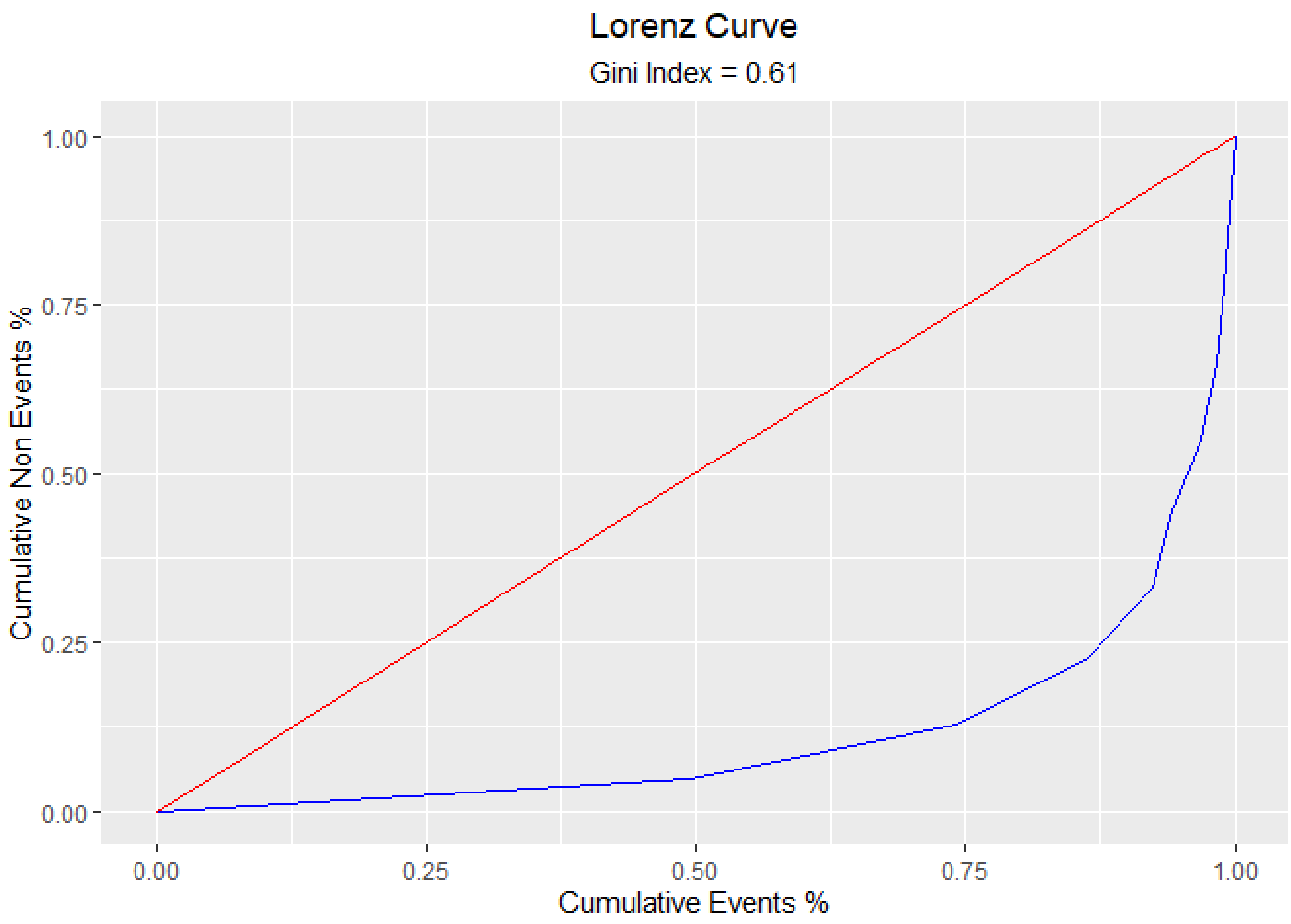


Lorenz Curve

The Lorenz curve is a simple graphic device which illustrates the degree of inequality in the distribution of thevariable concerned. It is a visual representation of inequality used to measure the discriminatory power of the predictive model.

Hide

```
blr_lorenz_curve(model)
```



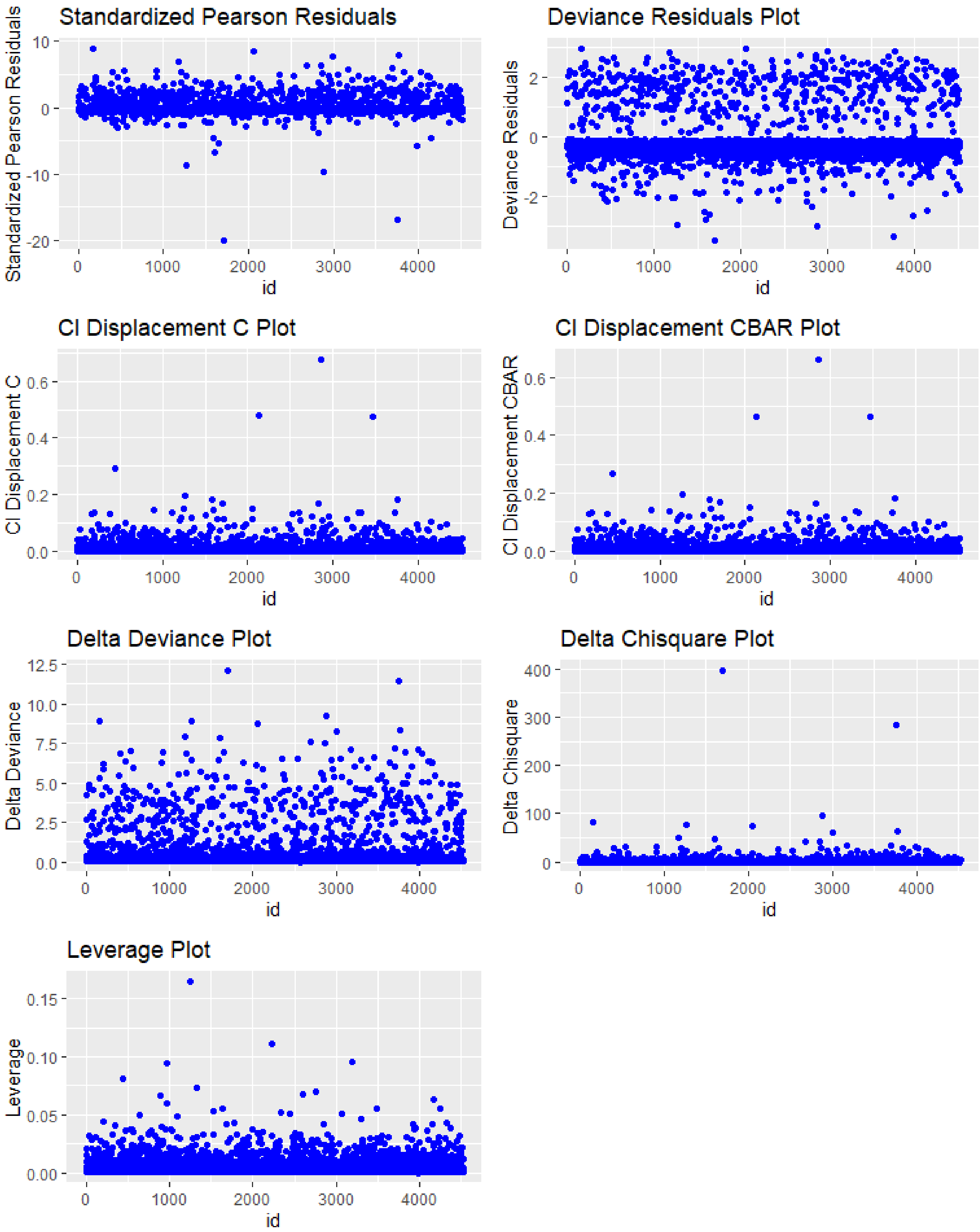
Residual & Influence Diagnostics

blorr can generate 22 plots for residual, influence and leverage diagnostics.

Influence Diagnostics

Hide

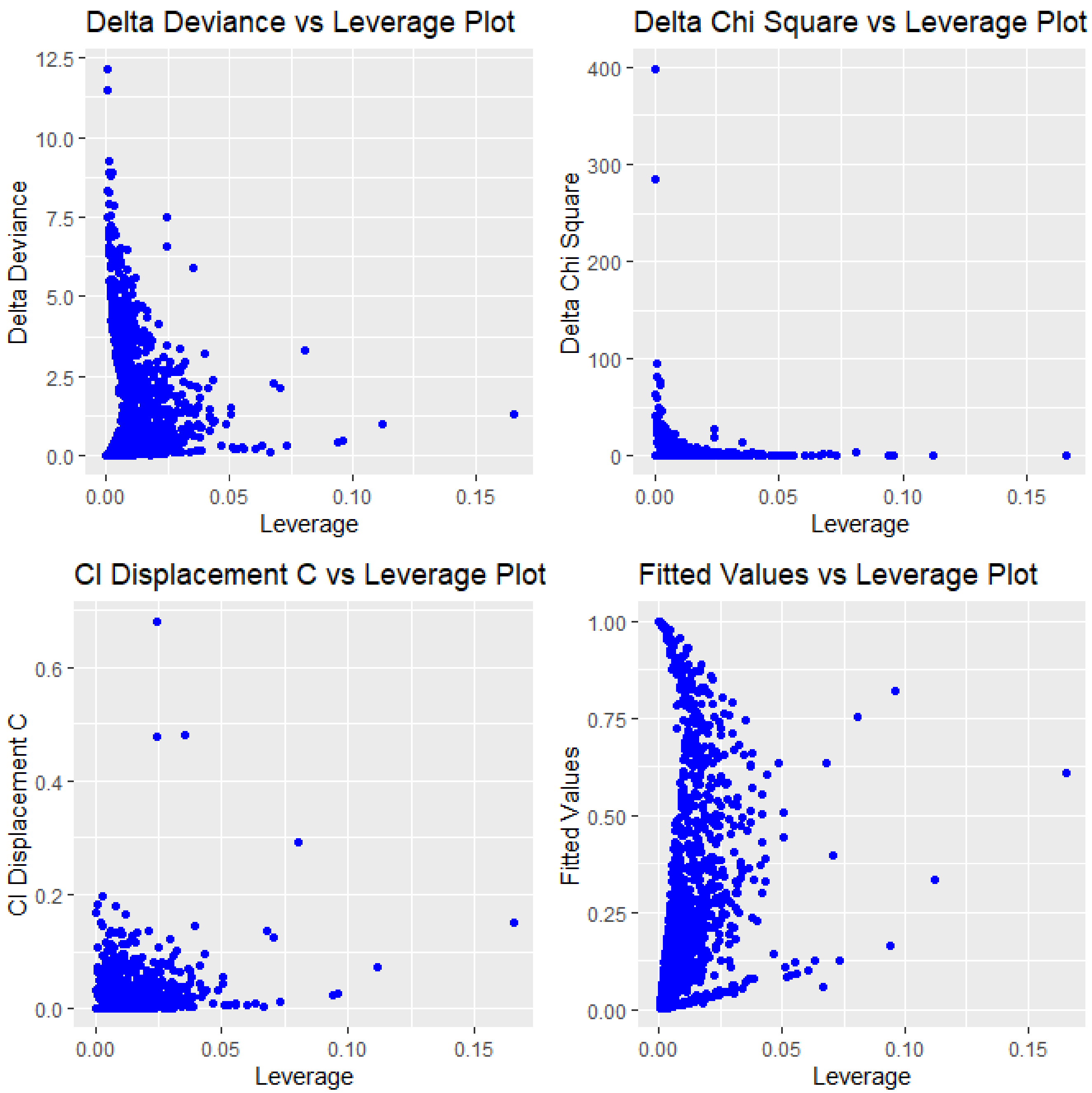
```
blr_plot_diag_influence(model)
```



Leverage Diagnostics

Hide

```
blr_plot_diag_leverage(model)
```



Delta Deviance vs Leverage Plot

Delta Chi Square vs Leverage Plot

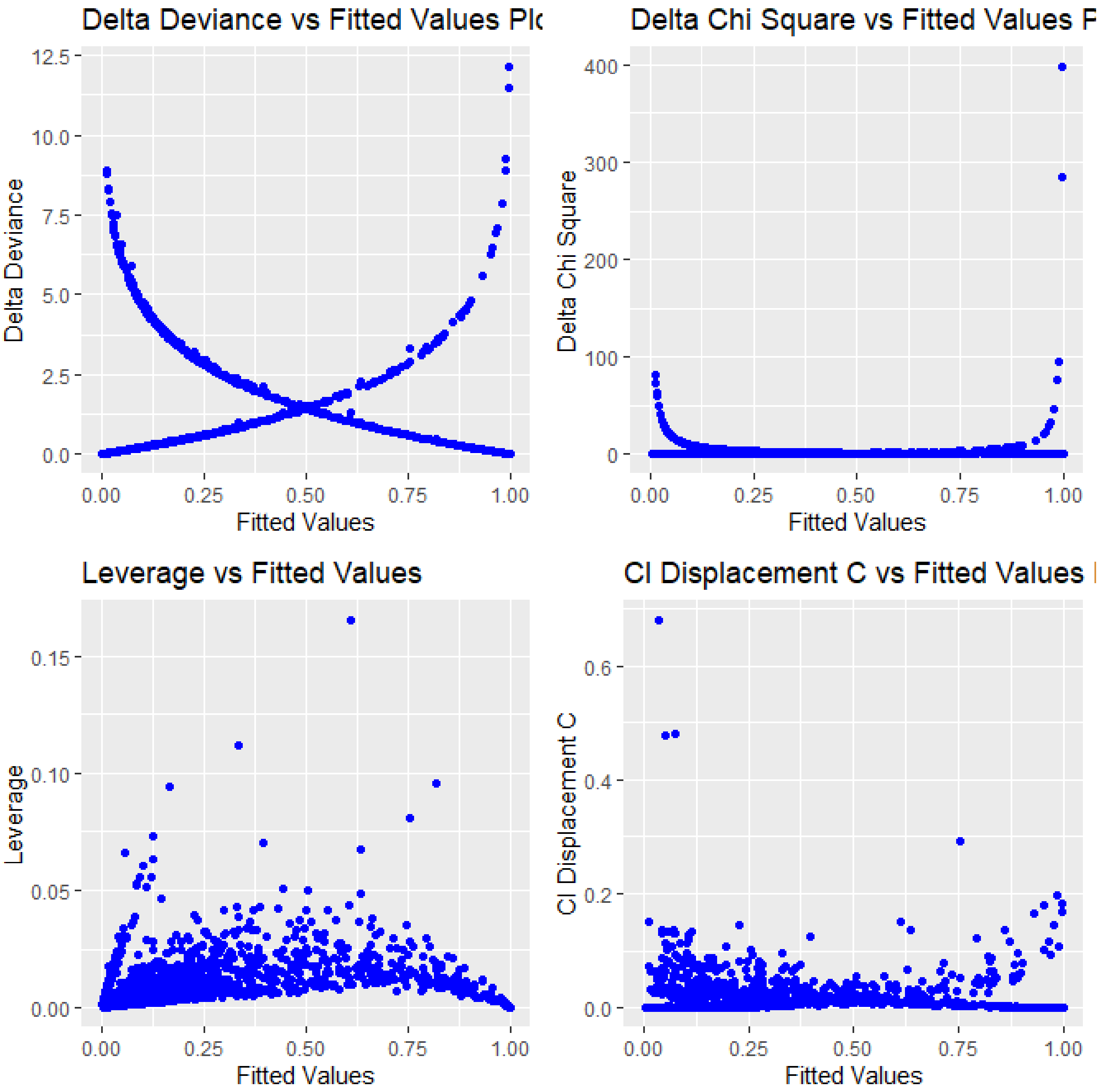
CI Displacement C vs Leverage Plot

Fitted Values vs Leverage Plot

Fitted Values Diagnostics

Hide

```
blr_plot_diag_fit(model)
```


Delta Deviance vs Fitted Values Pl
Delta Deviance
12.5
10.0
7.5
5.0
2.5
0.0
0.00
0.25
0.50
0.75
1.00
Fitted Values
Delta Chi Square vs Fitted Values F
Delta Chi Square
400
300
200
100
0
0.00
0.25
0.50
0.75
1.00
Fitted Values
Leverage vs Fitted Values
Leverage
0.15
0.10
0.05
0.00
0.00
0.25
0.50
0.75
1.00
Fitted Values
CI Displacement C vs Fitted Values
CI Displacement C
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
1.00
Fitted Values