STAT 4355 Life Expectancy Project

Muhammad Munir, Vismaya Joseph, Caleb Captain

5/12/2021

Loading Required Libraries

```
shhh <- suppressPackageStartupMessages
shhh(library(ggplot2))
shhh(library(plotly))
shhh(library(tidyverse))
shhh(library(stringr))
shhh(library(mltools))
shhh(library(data.table))
shhh(library(car))
shhh(library(dplyr))
shhh(library(Hmisc))
shhh(library(corrplot))
shhh(library(MASS))
shhh(library(faraway))
shhh(library(viridis))
shhh(library(stargazer))
shhh(library(gtsummary))
shhh(library(sensemakr))
```

Data Exploration and Cleaning

```
data <- read.csv("Life Expectancy Data.csv")
describe(data)
```

```
## data
##
##  22  Variables      2938  Observations
## ---------------------------------------------------------------------------
## Country
##        n  missing distinct
##     2938        0      193
##
## lowest : Afghanistan                         Albania                           Algeria
## highest: Venezuela (Bolivarian Republic of) Viet Nam                          Yemen
## ---------------------------------------------------------------------------
## Year
##        n  missing distinct     Info     Mean      Gmd      .05      .10
```

```
##     2938        0       16    0.996     2008    5.318     2000     2001
##      .25      .50      .75      .90      .95
##     2004     2008     2012     2014     2015
##
## lowest : 2000 2001 2002 2003 2004, highest: 2011 2012 2013 2014 2015
##
## Value       2000  2001  2002  2003  2004  2005  2006  2007  2008  2009  2010
## Frequency    183   183   183   183   183   183   183   183   183   183   183
## Proportion 0.062 0.062 0.062 0.062 0.062 0.062 0.062 0.062 0.062 0.062 0.062
##
## Value       2011  2012  2013  2014  2015
## Frequency    183   183   193   183   183
## Proportion 0.062 0.062 0.066 0.062 0.062
## ------------------------------------------------------------------------------
## Status
##        n  missing distinct
##     2938        0        2
##
## Value       Developed Developing
## Frequency         512       2426
## Proportion      0.174      0.826
## ------------------------------------------------------------------------------
## Life.expectancy
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2928       10      362        1    69.22    10.62     51.4     54.8
##      .25      .50      .75      .90      .95
##     63.1     72.1     75.7     79.7     82.0
##
## lowest : 36.3 39.0 41.0 41.5 42.3, highest: 85.0 86.0 87.0 88.0 89.0
## ------------------------------------------------------------------------------
## Adult.Mortality
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2928       10      425        1    164.8    134.1     13.0     19.0
##      .25      .50      .75      .90      .95
##     74.0    144.0    228.0    336.0    398.3
##
## lowest :   1   2   3   4   5, highest: 693 699 715 717 723
## ------------------------------------------------------------------------------
## infant.deaths
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2938        0      209    0.974     30.3    51.15     0.00     0.00
##      .25      .50      .75      .90      .95
##     0.00     3.00    22.00    58.00    94.15
##
## lowest :    0    1    2    3    4, highest: 1400 1500 1600 1700 1800
## ------------------------------------------------------------------------------
## Alcohol
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2744      194     1076    0.999    4.603    4.555   0.0100   0.0100
##      .25      .50      .75      .90      .95
##   0.8775   3.7550   7.7025  10.7570  11.9600
##
## lowest :  0.01  0.02  0.03  0.04  0.05, highest: 16.35 16.58 16.99 17.31 17.87
## ------------------------------------------------------------------------------
```

```
## percentage.expenditure
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2938        0     2328    0.991    738.3     1238    0.000    0.000
##      .25      .50      .75      .90      .95
##    4.685   64.913  441.534 1852.948 4506.638
##
## lowest : 0.000000e+00 9.987219e-02 1.080560e-01 2.756483e-01 3.284181e-01
## highest: 1.837933e+04 1.882287e+04 1.896135e+04 1.909905e+04 1.947991e+04
## ------------------------------------------------------------------------------
## Hepatitis.B
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2385      553       87    0.997    80.94    23.25        9       44
##      .25      .50      .75      .90      .95
##       77       92       97       99       99
##
## lowest :  1  2  4  5  6, highest: 95 96 97 98 99
## ------------------------------------------------------------------------------
## Measles
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2938        0      958    0.962     2420     4492      0.0      0.0
##      .25      .50      .75      .90      .95
##      0.0     17.0    360.2   3580.1   9985.6
##
## lowest :      0      1      2      3      4, highest: 133802 141258 168107 182485 212183
## ------------------------------------------------------------------------------
## BMI
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2904       34      608        1    38.32    22.87     5.20    11.90
##      .25      .50      .75      .90      .95
##    19.30    43.50    56.20    61.80    64.78
##
## lowest :  1.0  1.4  1.8  1.9  2.0, highest: 79.3 81.6 82.8 83.3 87.3
## ------------------------------------------------------------------------------
## under.five.deaths
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2938        0      252    0.979    42.04    71.19        0        0
##      .25      .50      .75      .90      .95
##        0        4       28       87      138
##
## lowest :    0    1    2    3    4, highest: 2100 2200 2300 2400 2500
## ------------------------------------------------------------------------------
## Polio
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2919       19       73    0.996    82.55    21.45        9       52
##      .25      .50      .75      .90      .95
##       78       93       97       99       99
##
## lowest :  3  4  5  6  7, highest: 95 96 97 98 99
## ------------------------------------------------------------------------------
## Total.expenditure
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2712      226      818        1    5.938    2.771    1.930    2.840
##      .25      .50      .75      .90      .95
##    4.260    5.755    7.492    9.120    9.760
```

```
##
## lowest :  0.37  0.65  0.74  0.76  0.92, highest: 17.00 17.14 17.20 17.24 17.60
## ---------------------------------------------------------------------------
## Diphtheria
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2919       19       81    0.996    82.32    21.69        9       49
##      .25      .50      .75      .90      .95
##       78       93       97       99       99
##
## lowest :  2  3  4  5  6, highest: 95 96 97 98 99
## ---------------------------------------------------------------------------
## HIV.AIDS
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2938        0      200    0.777    1.742    2.892    0.100    0.100
##      .25      .50      .75      .90      .95
##    0.100    0.100    0.800    4.400    8.515
##
## lowest :  0.1  0.2  0.3  0.4  0.5, highest: 48.8 49.1 49.9 50.3 50.6
## ---------------------------------------------------------------------------
## GDP
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2490      448     2490        1     7483    11067    68.05   161.46
##      .25      .50      .75      .90      .95
##   463.94  1766.95  5910.81 23726.14 41606.85
##
## lowest : 1.681350e+00 3.685949e+00 4.613575e+00 5.668726e+00 8.376432e+00
## highest: 8.973971e+04 1.137519e+05 1.142938e+05 1.157616e+05 1.191727e+05
## ---------------------------------------------------------------------------
## Population
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2286      652     2278        1 12753375 21575245     9618    29383
##      .25      .50      .75      .90      .95
##   195793  1386542  7420359 25787136 47554416
##
## lowest :         34         36         41         43        123
## highest: 1126135777 1144118674 1161977719 1179681239 1293859294
## ---------------------------------------------------------------------------
## thinness..1.19.years
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2904       34      200        1     4.84    4.527      0.6      0.8
##      .25      .50      .75      .90      .95
##      1.6      3.3      7.2      9.8     13.8
##
## lowest :  0.1  0.2  0.3  0.4  0.5, highest: 27.2 27.3 27.4 27.5 27.7
## ---------------------------------------------------------------------------
## thinness.5.9.years
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2904       34      207        1     4.87    4.596      0.5      0.8
##      .25      .50      .75      .90      .95
##      1.5      3.3      7.2      9.7     13.8
##
## lowest :  0.1  0.2  0.3  0.4  0.5, highest: 28.2 28.3 28.4 28.5 28.6
## ---------------------------------------------------------------------------
## Income.composition.of.resources
```

```
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2771      167      625        1   0.6276   0.2271    0.277    0.387
##      .25      .50      .75      .90      .95
##    0.493    0.677    0.779    0.864    0.892
##
## lowest : 0.000 0.253 0.255 0.261 0.266, highest: 0.939 0.941 0.942 0.945 0.948
## ---------------------------------------------------------------------------
## Schooling
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2775      163      173        1    11.99    3.713      5.8      7.7
##      .25      .50      .75      .90      .95
##     10.1     12.3     14.3     15.9     16.8
##
## lowest :  0.0  2.8  2.9  3.0  3.1, highest: 20.3 20.4 20.5 20.6 20.7
## ---------------------------------------------------------------------------
```

These are the features of the data. Also, the number of unique entries for each feature. Some of these we will want to drop. Country is a good example: we would not like the model to learn life expectancy from a country name. It needs to learn from the values, we can use the names to as an outlier check.

```
colnames(data)
```

```
##  [1] "Country"                        "Year"
##  [3] "Status"                         "Life.expectancy"
##  [5] "Adult.Mortality"                "infant.deaths"
##  [7] "Alcohol"                        "percentage.expenditure"
##  [9] "Hepatitis.B"                    "Measles"
## [11] "BMI"                            "under.five.deaths"
## [13] "Polio"                          "Total.expenditure"
## [15] "Diphtheria"                     "HIV.AIDS"
## [17] "GDP"                            "Population"
## [19] "thinness..1.19.years"           "thinness.5.9.years"
## [21] "Income.composition.of.resources" "Schooling"
```

```
unique(data$Status)
```

```
## [1] "Developing" "Developed"
```

```
unique(data$Country)
```

```
##   [1] "Afghanistan"
##   [2] "Albania"
##   [3] "Algeria"
##   [4] "Angola"
##   [5] "Antigua and Barbuda"
##   [6] "Argentina"
##   [7] "Armenia"
##   [8] "Australia"
##   [9] "Austria"
##  [10] "Azerbaijan"
##  [11] "Bahamas"
##  [12] "Bahrain"
```

```
##  [13] "Bangladesh"
##  [14] "Barbados"
##  [15] "Belarus"
##  [16] "Belgium"
##  [17] "Belize"
##  [18] "Benin"
##  [19] "Bhutan"
##  [20] "Bolivia (Plurinational State of)"
##  [21] "Bosnia and Herzegovina"
##  [22] "Botswana"
##  [23] "Brazil"
##  [24] "Brunei Darussalam"
##  [25] "Bulgaria"
##  [26] "Burkina Faso"
##  [27] "Burundi"
##  [28] "CÃ´te d'Ivoire"
##  [29] "Cabo Verde"
##  [30] "Cambodia"
##  [31] "Cameroon"
##  [32] "Canada"
##  [33] "Central African Republic"
##  [34] "Chad"
##  [35] "Chile"
##  [36] "China"
##  [37] "Colombia"
##  [38] "Comoros"
##  [39] "Congo"
##  [40] "Cook Islands"
##  [41] "Costa Rica"
##  [42] "Croatia"
##  [43] "Cuba"
##  [44] "Cyprus"
##  [45] "Czechia"
##  [46] "Democratic People's Republic of Korea"
##  [47] "Democratic Republic of the Congo"
##  [48] "Denmark"
##  [49] "Djibouti"
##  [50] "Dominica"
##  [51] "Dominican Republic"
##  [52] "Ecuador"
##  [53] "Egypt"
##  [54] "El Salvador"
##  [55] "Equatorial Guinea"
##  [56] "Eritrea"
##  [57] "Estonia"
##  [58] "Ethiopia"
##  [59] "Fiji"
##  [60] "Finland"
##  [61] "France"
##  [62] "Gabon"
##  [63] "Gambia"
##  [64] "Georgia"
##  [65] "Germany"
##  [66] "Ghana"
```

```
##  [67] "Greece"
##  [68] "Grenada"
##  [69] "Guatemala"
##  [70] "Guinea"
##  [71] "Guinea-Bissau"
##  [72] "Guyana"
##  [73] "Haiti"
##  [74] "Honduras"
##  [75] "Hungary"
##  [76] "Iceland"
##  [77] "India"
##  [78] "Indonesia"
##  [79] "Iran (Islamic Republic of)"
##  [80] "Iraq"
##  [81] "Ireland"
##  [82] "Israel"
##  [83] "Italy"
##  [84] "Jamaica"
##  [85] "Japan"
##  [86] "Jordan"
##  [87] "Kazakhstan"
##  [88] "Kenya"
##  [89] "Kiribati"
##  [90] "Kuwait"
##  [91] "Kyrgyzstan"
##  [92] "Lao People's Democratic Republic"
##  [93] "Latvia"
##  [94] "Lebanon"
##  [95] "Lesotho"
##  [96] "Liberia"
##  [97] "Libya"
##  [98] "Lithuania"
##  [99] "Luxembourg"
## [100] "Madagascar"
## [101] "Malawi"
## [102] "Malaysia"
## [103] "Maldives"
## [104] "Mali"
## [105] "Malta"
## [106] "Marshall Islands"
## [107] "Mauritania"
## [108] "Mauritius"
## [109] "Mexico"
## [110] "Micronesia (Federated States of)"
## [111] "Monaco"
## [112] "Mongolia"
## [113] "Montenegro"
## [114] "Morocco"
## [115] "Mozambique"
## [116] "Myanmar"
## [117] "Namibia"
## [118] "Nauru"
## [119] "Nepal"
## [120] "Netherlands"
```

```
## [121] "New Zealand"
## [122] "Nicaragua"
## [123] "Niger"
## [124] "Nigeria"
## [125] "Niue"
## [126] "Norway"
## [127] "Oman"
## [128] "Pakistan"
## [129] "Palau"
## [130] "Panama"
## [131] "Papua New Guinea"
## [132] "Paraguay"
## [133] "Peru"
## [134] "Philippines"
## [135] "Poland"
## [136] "Portugal"
## [137] "Qatar"
## [138] "Republic of Korea"
## [139] "Republic of Moldova"
## [140] "Romania"
## [141] "Russian Federation"
## [142] "Rwanda"
## [143] "Saint Kitts and Nevis"
## [144] "Saint Lucia"
## [145] "Saint Vincent and the Grenadines"
## [146] "Samoa"
## [147] "San Marino"
## [148] "Sao Tome and Principe"
## [149] "Saudi Arabia"
## [150] "Senegal"
## [151] "Serbia"
## [152] "Seychelles"
## [153] "Sierra Leone"
## [154] "Singapore"
## [155] "Slovakia"
## [156] "Slovenia"
## [157] "Solomon Islands"
## [158] "Somalia"
## [159] "South Africa"
## [160] "South Sudan"
## [161] "Spain"
## [162] "Sri Lanka"
## [163] "Sudan"
## [164] "Suriname"
## [165] "Swaziland"
## [166] "Sweden"
## [167] "Switzerland"
## [168] "Syrian Arab Republic"
## [169] "Tajikistan"
## [170] "Thailand"
## [171] "The former Yugoslav republic of Macedonia"
## [172] "Timor-Leste"
## [173] "Togo"
## [174] "Tonga"
```

```
## [175] "Trinidad and Tobago"
## [176] "Tunisia"
## [177] "Turkey"
## [178] "Turkmenistan"
## [179] "Tuvalu"
## [180] "Uganda"
## [181] "Ukraine"
## [182] "United Arab Emirates"
## [183] "United Kingdom of Great Britain and Northern Ireland"
## [184] "United Republic of Tanzania"
## [185] "United States of America"
## [186] "Uruguay"
## [187] "Uzbekistan"
## [188] "Vanuatu"
## [189] "Venezuela (Bolivarian Republic of)"
## [190] "Viet Nam"
## [191] "Yemen"
## [192] "Zambia"
## [193] "Zimbabwe"
```

```
Developed <- unique(data$Country[which(data$Status == "Developed")])
Developing <- unique(data$Country[which(data$Status == "Developing")])
```

```
table(Developed)
```

```
## Developed
##                                          Australia
##                                                  1
##                                            Austria
##                                                  1
##                                            Belgium
##                                                  1
##                                           Bulgaria
##                                                  1
##                                            Croatia
##                                                  1
##                                             Cyprus
##                                                  1
##                                            Czechia
##                                                  1
##                                            Denmark
##                                                  1
##                                            Germany
##                                                  1
##                                            Hungary
##                                                  1
##                                            Iceland
##                                                  1
##                                            Ireland
##                                                  1
##                                              Italy
##                                                  1
##                                              Japan
```

```
##                                                 1
##                                            Latvia
##                                                 1
##                                         Lithuania
##                                                 1
##                                        Luxembourg
##                                                 1
##                                             Malta
##                                                 1
##                                       Netherlands
##                                                 1
##                                       New Zealand
##                                                 1
##                                            Norway
##                                                 1
##                                            Poland
##                                                 1
##                                          Portugal
##                                                 1
##                                           Romania
##                                                 1
##                                         Singapore
##                                                 1
##                                          Slovakia
##                                                 1
##                                          Slovenia
##                                                 1
##                                             Spain
##                                                 1
##                                            Sweden
##                                                 1
##                                       Switzerland
##                                                 1
## United Kingdom of Great Britain and Northern Ireland
##                                                 1
##                           United States of America
##                                                 1
```

```
table(Developing)
```

```
## Developing
##                             Afghanistan
##                                       1
##                                 Albania
##                                       1
##                                 Algeria
##                                       1
##                                  Angola
##                                       1
##                      Antigua and Barbuda
##                                       1
##                               Argentina
##                                       1
##                                 Armenia
```

```
##                                     1
##                                Azerbaijan
##                                     1
##                                   Bahamas
##                                     1
##                                   Bahrain
##                                     1
##                                Bangladesh
##                                     1
##                                  Barbados
##                                     1
##                                   Belarus
##                                     1
##                                    Belize
##                                     1
##                                     Benin
##                                     1
##                                    Bhutan
##                                     1
##          Bolivia (Plurinational State of)
##                                     1
##                    Bosnia and Herzegovina
##                                     1
##                                  Botswana
##                                     1
##                                    Brazil
##                                     1
##                          Brunei Darussalam
##                                     1
##                               Burkina Faso
##                                     1
##                                   Burundi
##                                     1
##                             CÃ´te d'Ivoire
##                                     1
##                               Cabo Verde
##                                     1
##                                  Cambodia
##                                     1
##                                  Cameroon
##                                     1
##                                    Canada
##                                     1
##                  Central African Republic
##                                     1
##                                      Chad
##                                     1
##                                     Chile
##                                     1
##                                     China
##                                     1
##                                  Colombia
##                                     1
##                                   Comoros
```

```
##                                     1
##                                 Congo
##                                     1
##                           Cook Islands
##                                     1
##                            Costa Rica
##                                     1
##                                  Cuba
##                                     1
##    Democratic People's Republic of Korea
##                                     1
##         Democratic Republic of the Congo
##                                     1
##                                Djibouti
##                                     1
##                                Dominica
##                                     1
##                      Dominican Republic
##                                     1
##                                 Ecuador
##                                     1
##                                   Egypt
##                                     1
##                             El Salvador
##                                     1
##                       Equatorial Guinea
##                                     1
##                                 Eritrea
##                                     1
##                                 Estonia
##                                     1
##                                Ethiopia
##                                     1
##                                    Fiji
##                                     1
##                                Finland
##                                     1
##                                 France
##                                     1
##                                  Gabon
##                                     1
##                                 Gambia
##                                     1
##                                Georgia
##                                     1
##                                  Ghana
##                                     1
##                                 Greece
##                                     1
##                                Grenada
##                                     1
##                              Guatemala
##                                     1
##                                 Guinea
```

```
##                                   1
##                          Guinea-Bissau
##                                   1
##                                Guyana
##                                   1
##                                 Haiti
##                                   1
##                              Honduras
##                                   1
##                                 India
##                                   1
##                             Indonesia
##                                   1
##           Iran (Islamic Republic of)
##                                   1
##                                  Iraq
##                                   1
##                                Israel
##                                   1
##                               Jamaica
##                                   1
##                                Jordan
##                                   1
##                            Kazakhstan
##                                   1
##                                 Kenya
##                                   1
##                              Kiribati
##                                   1
##                                Kuwait
##                                   1
##                            Kyrgyzstan
##                                   1
##     Lao People's Democratic Republic
##                                   1
##                                Lebanon
##                                   1
##                                Lesotho
##                                   1
##                                Liberia
##                                   1
##                                  Libya
##                                   1
##                             Madagascar
##                                   1
##                                 Malawi
##                                   1
##                               Malaysia
##                                   1
##                               Maldives
##                                   1
##                                   Mali
##                                   1
##                        Marshall Islands
```

```
##                                     1
##                               Mauritania
##                                     1
##                                Mauritius
##                                     1
##                                   Mexico
##                                     1
##          Micronesia (Federated States of)
##                                     1
##                                   Monaco
##                                     1
##                                 Mongolia
##                                     1
##                               Montenegro
##                                     1
##                                  Morocco
##                                     1
##                               Mozambique
##                                     1
##                                  Myanmar
##                                     1
##                                  Namibia
##                                     1
##                                    Nauru
##                                     1
##                                    Nepal
##                                     1
##                                Nicaragua
##                                     1
##                                    Niger
##                                     1
##                                  Nigeria
##                                     1
##                                     Niue
##                                     1
##                                     Oman
##                                     1
##                                 Pakistan
##                                     1
##                                    Palau
##                                     1
##                                   Panama
##                                     1
##                           Papua New Guinea
##                                     1
##                                 Paraguay
##                                     1
##                                     Peru
##                                     1
##                                Philippines
##                                     1
##                                    Qatar
##                                     1
##                           Republic of Korea
```

```
##                                     1
##                       Republic of Moldova
##                                     1
##                        Russian Federation
##                                     1
##                                  Rwanda
##                                     1
##                     Saint Kitts and Nevis
##                                     1
##                               Saint Lucia
##                                     1
##          Saint Vincent and the Grenadines
##                                     1
##                                   Samoa
##                                     1
##                                San Marino
##                                     1
##                     Sao Tome and Principe
##                                     1
##                               Saudi Arabia
##                                     1
##                                  Senegal
##                                     1
##                                   Serbia
##                                     1
##                                Seychelles
##                                     1
##                               Sierra Leone
##                                     1
##                           Solomon Islands
##                                     1
##                                  Somalia
##                                     1
##                               South Africa
##                                     1
##                                South Sudan
##                                     1
##                                  Sri Lanka
##                                     1
##                                    Sudan
##                                     1
##                                 Suriname
##                                     1
##                                Swaziland
##                                     1
##                       Syrian Arab Republic
##                                     1
##                                Tajikistan
##                                     1
##                                  Thailand
##                                     1
## The former Yugoslav republic of Macedonia
##                                     1
##                                Timor-Leste
```

```
##                                     1
##                                  Togo
##                                     1
##                                 Tonga
##                                     1
##                     Trinidad and Tobago
##                                     1
##                                Tunisia
##                                     1
##                                Turkey
##                                     1
##                            Turkmenistan
##                                     1
##                                 Tuvalu
##                                     1
##                                 Uganda
##                                     1
##                                Ukraine
##                                     1
##                     United Arab Emirates
##                                     1
##             United Republic of Tanzania
##                                     1
##                                Uruguay
##                                     1
##                              Uzbekistan
##                                     1
##                                 Vanuatu
##                                     1
##      Venezuela (Bolivarian Republic of)
##                                     1
##                                Viet Nam
##                                     1
##                                   Yemen
##                                     1
##                                 Zambia
##                                     1
##                                Zimbabwe
##                                     1
```

This section of code shows us which features/columns have null values.

```
colnames(data)[sapply(data, anyNA)]
```

```
##  [1] "Life.expectancy"                "Adult.Mortality"
##  [3] "Alcohol"                        "Hepatitis.B"
##  [5] "BMI"                            "Polio"
##  [7] "Total.expenditure"              "Diphtheria"
##  [9] "GDP"                            "Population"
## [11] "thinness..1.19.years"           "thinness.5.9.years"
## [13] "Income.composition.of.resources" "Schooling"
```

```
table(is.na(data))
```

```
##
## FALSE  TRUE
## 62073  2563
```

There are 2,563 null values in the data. But they are confined to 1,293 rows of null values.

```
table(is.na.data.frame(data$Hepatitis.B))
```

```
##
## FALSE  TRUE
##  2385   553
```

```
table(is.na.data.frame(data$Life.expectancy))
```

```
##
## FALSE  TRUE
##  2928    10
```

```
table(is.na.data.frame(data$Polio))
```

```
##
## FALSE  TRUE
##  2919    19
```

```
table(is.na.data.frame(data$BMI))
```

```
##
## FALSE  TRUE
##  2904    34
```

```
table(is.na.data.frame(data$Diphtheria))
```

```
##
## FALSE  TRUE
##  2919    19
```

```
table(is.na.data.frame(data$thinness..1.19.years))
```

```
##
## FALSE  TRUE
##  2904    34
```

```
table(is.na.data.frame(data$thinness.5.9.years))
```

```
##
## FALSE  TRUE
##  2904    34
```

```
table(is.na.data.frame(data$Total.expenditure))
```

```
##
## FALSE  TRUE
##  2712   226
```

```
table(is.na.data.frame(data$Population))
```

```
##
## FALSE  TRUE
##  2286   652
```

```
table(is.na.data.frame(data$Income.composition.of.resources))
```

```
##
## FALSE  TRUE
##  2771   167
```

```
table(is.na.data.frame(data$Adult.Mortality))
```

```
##
## FALSE  TRUE
##  2928    10
```

```
table(is.na.data.frame(data$Schooling))
```

```
##
## FALSE  TRUE
##  2775   163
```

```
table(is.na.data.frame(data$Alcohol))
```

```
##
## FALSE  TRUE
##  2744   194
```

```
table(is.na.data.frame(data$GDP))
```

```
##
## FALSE  TRUE
##  2490   448
```

These are the features/columns with null values.

```
colnames(data)[sapply(data, anyNA)]
```

```
##  [1] "Life.expectancy"                 "Adult.Mortality"
##  [3] "Alcohol"                         "Hepatitis.B"
##  [5] "BMI"                             "Polio"
##  [7] "Total.expenditure"               "Diphtheria"
##  [9] "GDP"                             "Population"
## [11] "thinness..1.19.years"            "thinness.5.9.years"
## [13] "Income.composition.of.resources" "Schooling"
```

```
summary(data$Life.expectancy)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   36.30   63.10   72.10   69.22   75.70   89.00      10
```

```
summary(data$Hepatitis.B)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   77.00   92.00   80.94   97.00   99.00     553
```

```
summary(data$Total.expenditure)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.370   4.260   5.755   5.938   7.492  17.600     226
```

```
summary(data$Population)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## 3.400e+01 1.958e+05 1.387e+06 1.275e+07 7.420e+06 1.294e+09       652
```

```
summary(data$Income.composition.of.resources)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.4930  0.6770  0.6276  0.7790  0.9480     167
```

```
summary(data$Adult.Mortality)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     1.0    74.0   144.0   164.8   228.0   723.0      10
```

```
summary(data$BMI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   19.30   43.50   38.32   56.20   87.30      34
```

```
summary(data$Diphtheria)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2.00   78.00   93.00   82.32   97.00   99.00      19
```

```
summary(data$thinness..1.19.years)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.10    1.60    3.30    4.84    7.20   27.70      34
```

```
summary(data$Schooling)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   10.10   12.30   11.99   14.30   20.70     163
```

```
summary(data$Alcohol)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0100  0.8775  3.7550  4.6029  7.7025 17.8700     194
```

```
summary(data$Polio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    3.00   78.00   93.00   82.55   97.00   99.00      19
```

```
summary(data$GDP)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##      1.68    463.94   1766.95   7483.16   5910.81 119172.74       448
```

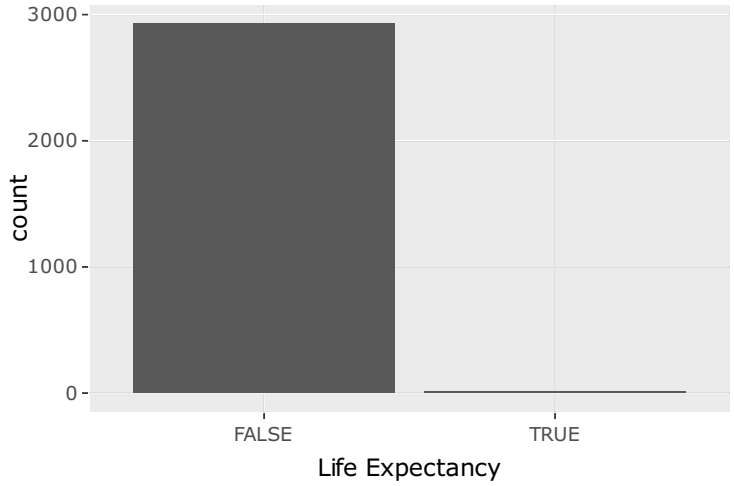
```
summary(data$thinness.5.9.years)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.10    1.50    3.30    4.87    7.20   28.60      34
```

Boxplots and Histigrams of each of the null columns/features is shown below.

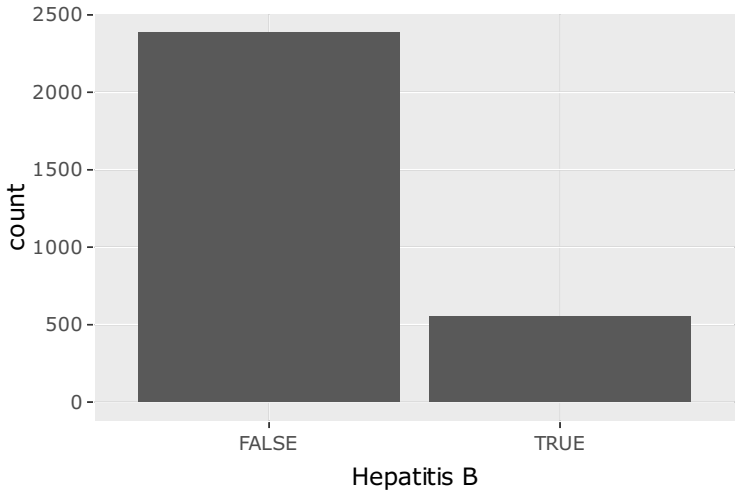
```
ggplotly(ggplot(data) +
           geom_histogram(stat="count", aes(x=is.na(Life.expectancy))) +
           labs(x="Life Expectancy"))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



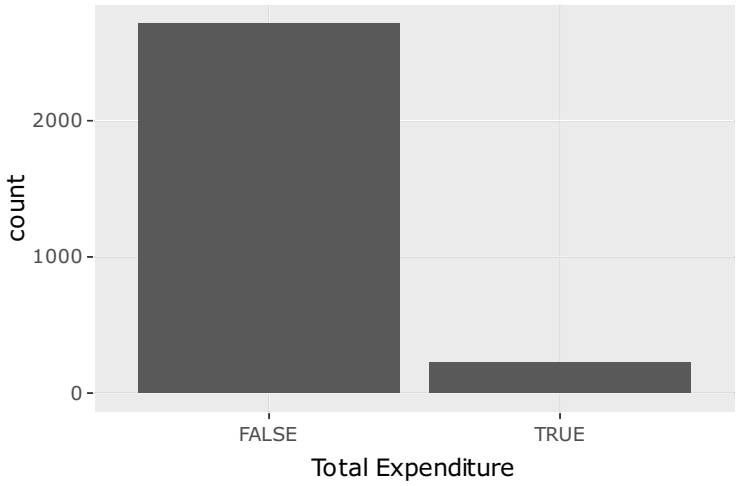

```
ggplotly(ggplot(data) +
           geom_histogram(stat="count", aes(x=is.na(Hepatitis.B))) +
           labs(x="Hepatitis B"))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



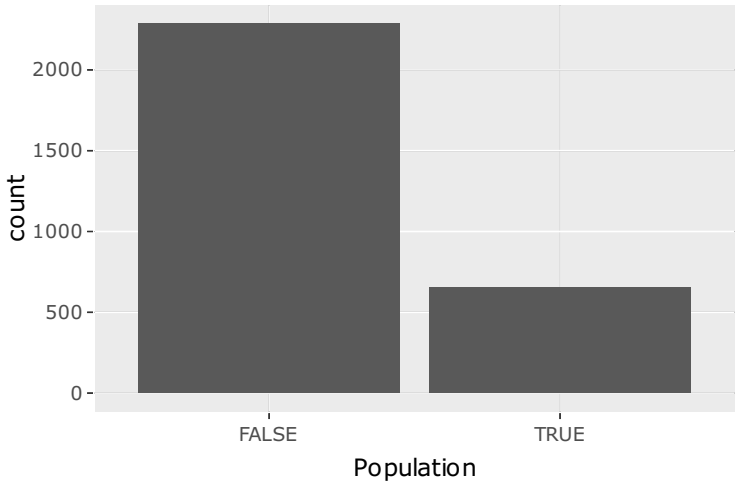

```
ggplotly(ggplot(data) +
           geom_histogram(stat="count", aes(x=is.na(Total.expenditure))) +
           labs(x="Total Expenditure"))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



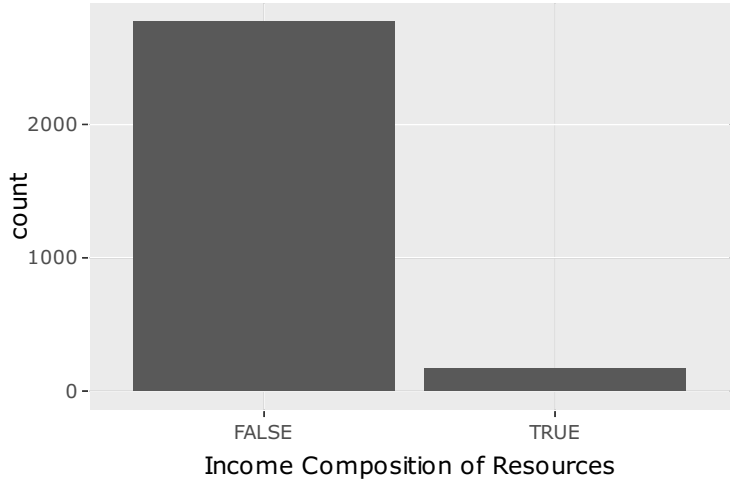

```
ggplotly(ggplot(data) +
           geom_histogram(stat="count", aes(x=is.na(Population))) +
           labs(x="Population"))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```




```
ggplotly(ggplot(data) +
           geom_histogram(stat="count", aes(x=is.na(Income.composition.of.resources))) +
           labs(x="Income Composition of Resources"))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

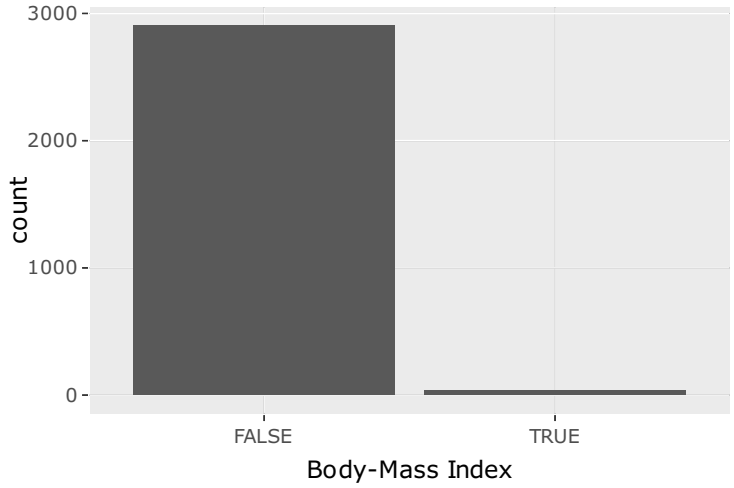



```
ggplotly(ggplot(data) +
           geom_histogram(stat="count", aes(x=is.na(Adult.Mortality))) +
           labs(x="Adult Mortality"))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

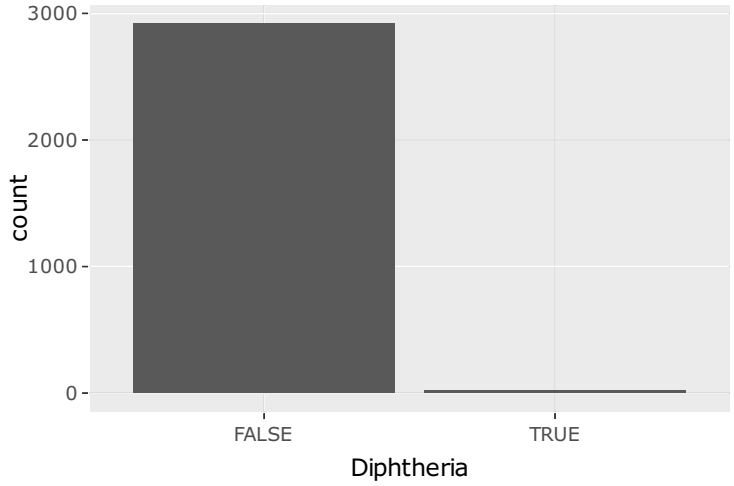

```
ggplotly(ggplot(data) +
           geom_histogram(stat="count", aes(x=is.na(BMI))) +
           labs(x="Body-Mass Index"))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



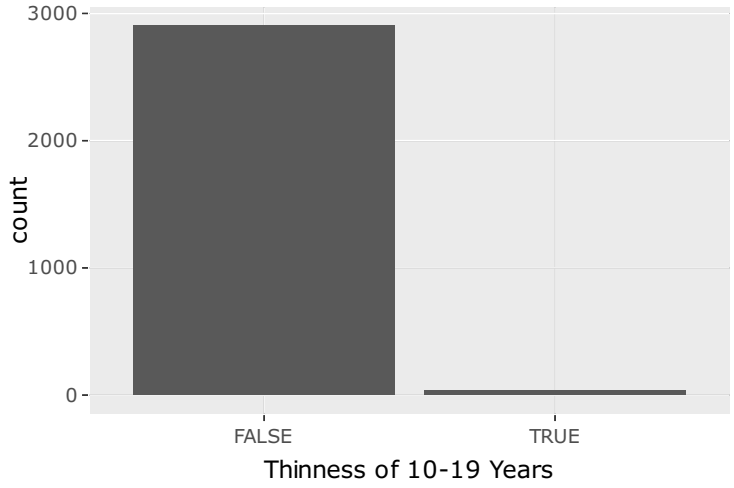

```
ggplotly(ggplot(data) +
           geom_histogram(stat="count", aes(x=is.na(Diphtheria))) +
           labs(x="Diphtheria"))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



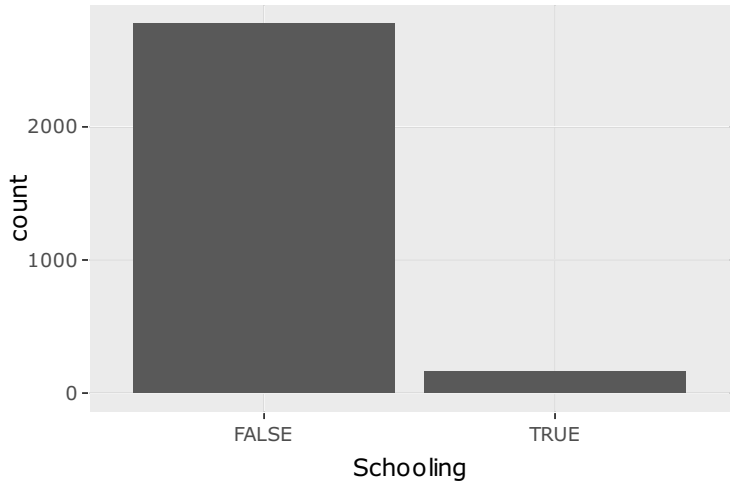

```
ggplotly(ggplot(data) +
           geom_histogram(stat="count", aes(x=is.na(thinness..1.19.years))) +
           labs(x="Thinness of 10-19 Years"))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



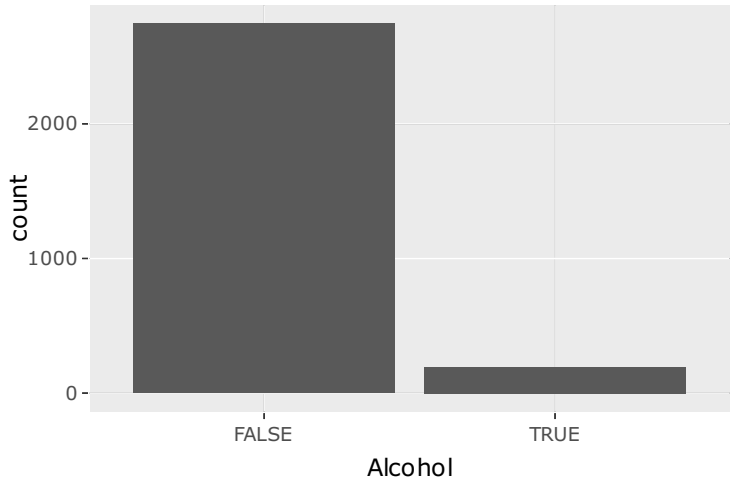

```
ggplotly(ggplot(data) +
           geom_histogram(stat="count", aes(x=is.na(Schooling))) +
           labs(x="Schooling"))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```




```
ggplotly(ggplot(data) +
           geom_histogram(stat="count", aes(x=is.na(Alcohol))) +
           labs(x="Alcohol"))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

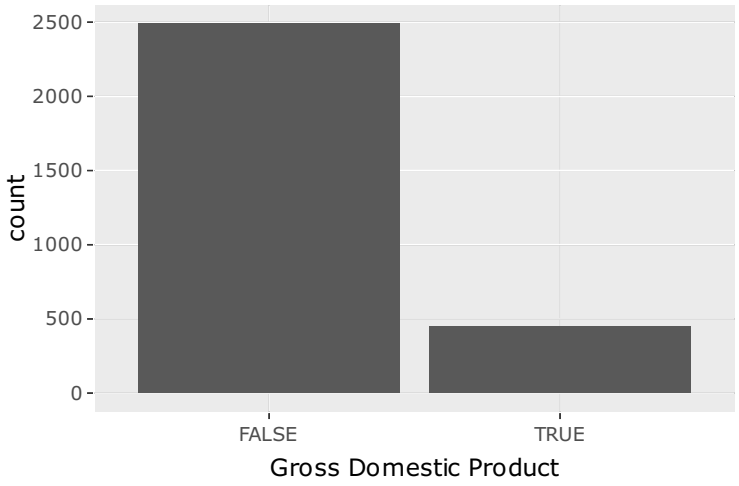



```
ggplotly(ggplot(data) +
           geom_histogram(stat="count", aes(x=is.na(Polio))) +
           labs(x="Polio"))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```


```
ggplotly(ggplot(data) +
           geom_histogram(stat="count", aes(x=is.na(GDP))) +
           labs(x="Gross Domestic Product"))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



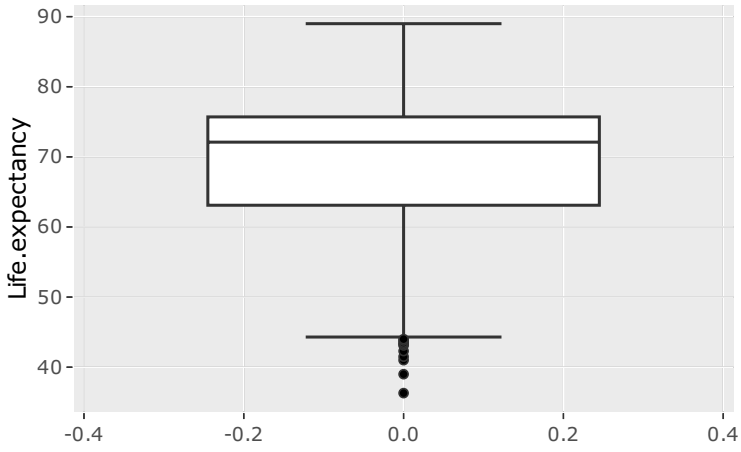

```
ggplotly(ggplot(data) +
           geom_histogram(stat="count", aes(x=is.na(thinness.5.9.years))) +
           labs(x="Thinness of 5-9 Years"))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

Note that population is flat. Essentially the population is so sparse that it does not show a range of values.

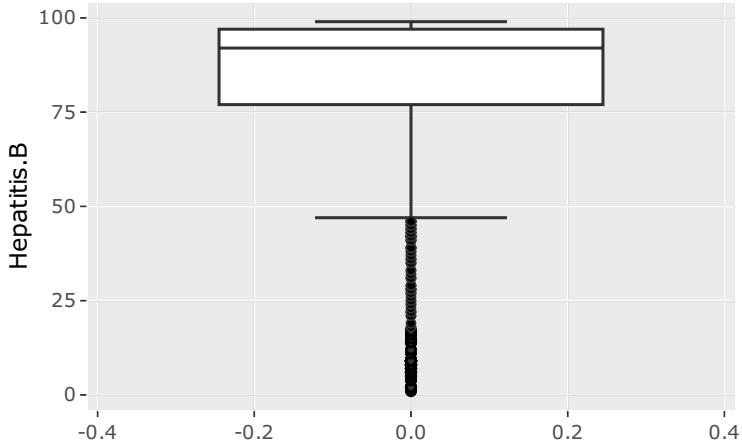
```
ggplotly(ggplot(data, aes(y=Life.expectancy)) + geom_boxplot())
```

```
## Warning: Removed 10 rows containing non-finite values (stat_boxplot).
```



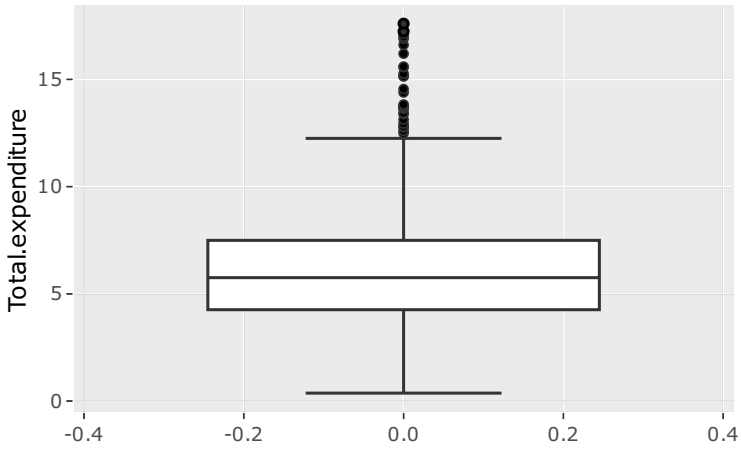

```
ggplotly(ggplot(data, aes(y=Hepatitis.B)) + geom_boxplot())
```

```
## Warning: Removed 553 rows containing non-finite values (stat_boxplot).
```



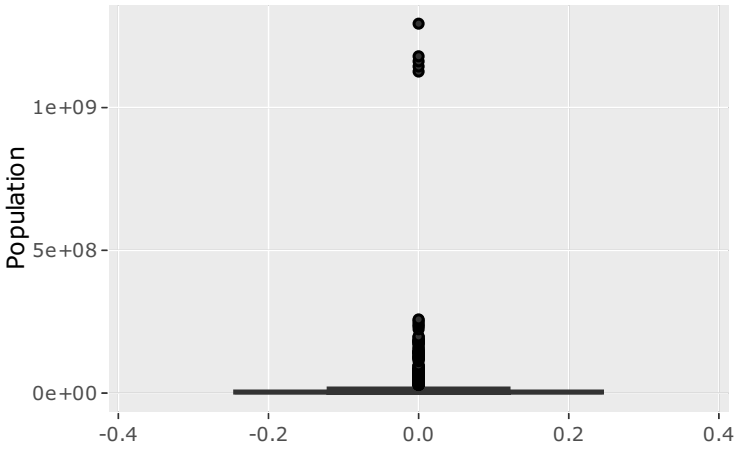

```
ggplotly(ggplot(data, aes(y=Total.expenditure)) + geom_boxplot())
```

```
## Warning: Removed 226 rows containing non-finite values (stat_boxplot).
```



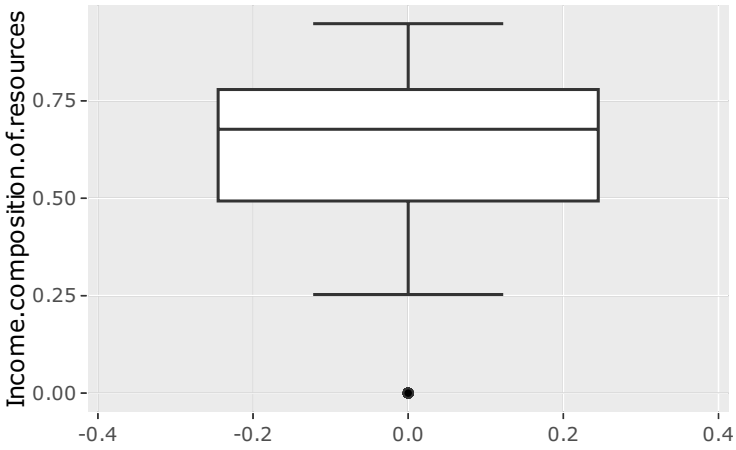

```
ggplotly(ggplot(data, aes(y=Population)) + geom_boxplot())
```

```
## Warning: Removed 652 rows containing non-finite values (stat_boxplot).
```



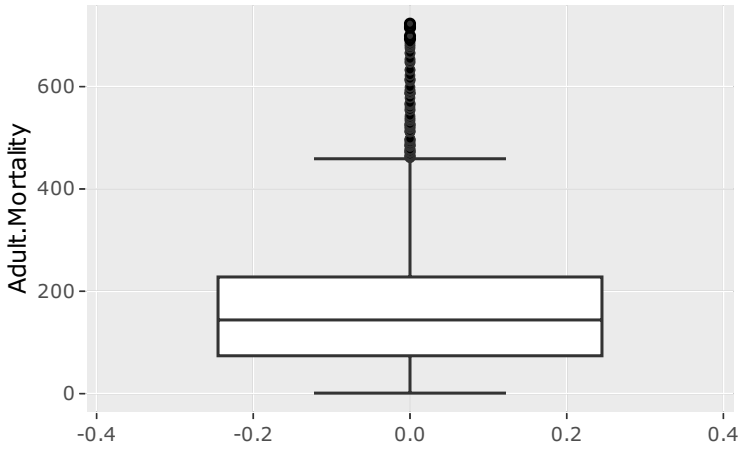

```
ggplotly(ggplot(data, aes(y=Income.composition.of.resources)) + geom_boxplot())
```

```
## Warning: Removed 167 rows containing non-finite values (stat_boxplot).
```



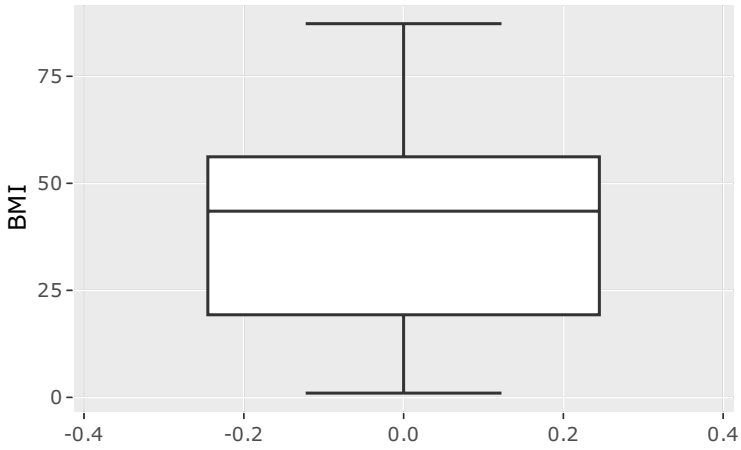

```
ggplotly(ggplot(data, aes(y=Adult.Mortality)) + geom_boxplot())
```

```
## Warning: Removed 10 rows containing non-finite values (stat_boxplot).
```



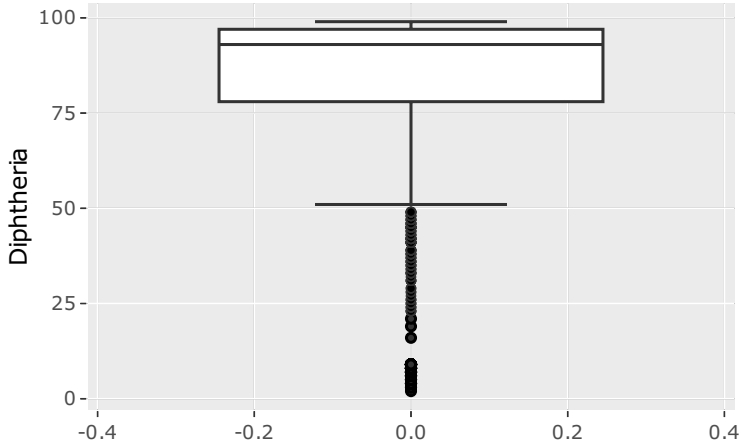

```
ggplotly(ggplot(data, aes(y=BMI)) + geom_boxplot())
```

```
## Warning: Removed 34 rows containing non-finite values (stat_boxplot).
```



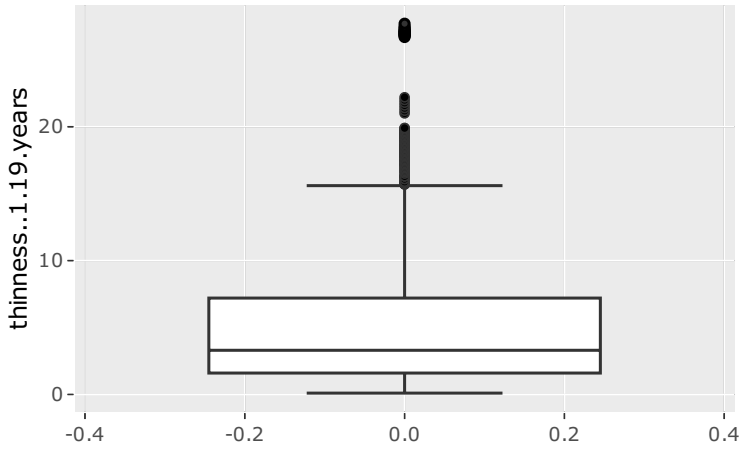

```
ggplotly(ggplot(data, aes(y=Diphtheria)) + geom_boxplot())
```

```
## Warning: Removed 19 rows containing non-finite values (stat_boxplot).
```



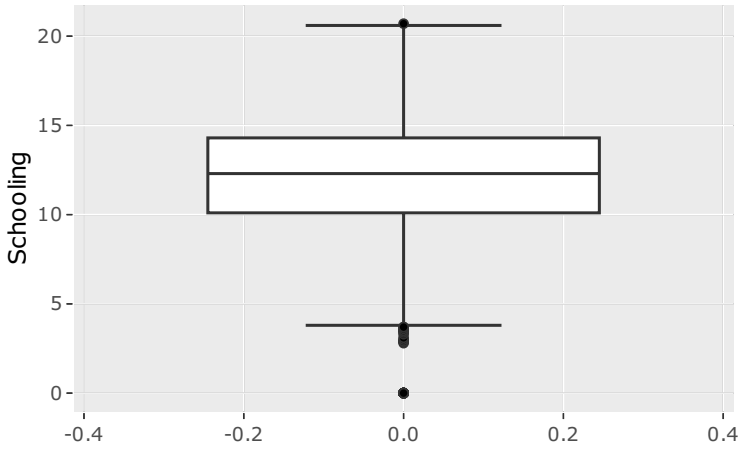

```
ggplotly(ggplot(data, aes(y=thinness..1.19.years)) + geom_boxplot())
```

```
## Warning: Removed 34 rows containing non-finite values (stat_boxplot).
```



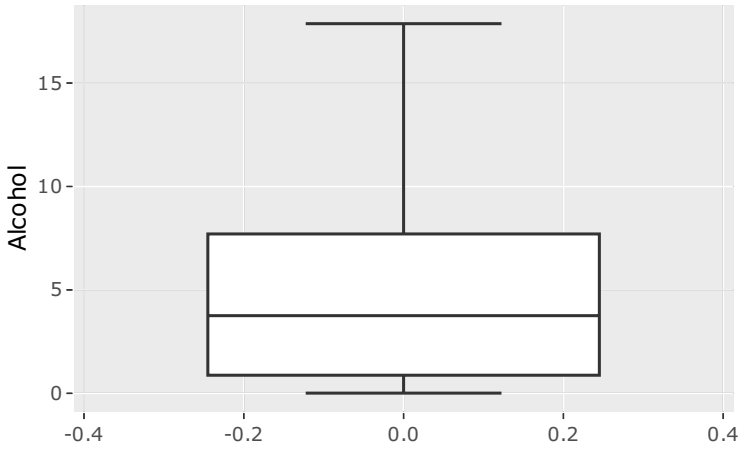

```
ggplotly(ggplot(data, aes(y=Schooling)) + geom_boxplot())
```

```
## Warning: Removed 163 rows containing non-finite values (stat_boxplot).
```



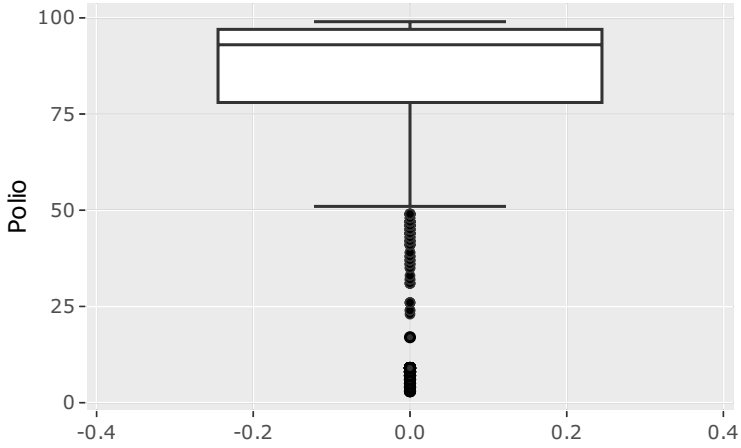

```
ggplotly(ggplot(data, aes(y=Alcohol)) + geom_boxplot())
```

```
## Warning: Removed 194 rows containing non-finite values (stat_boxplot).
```



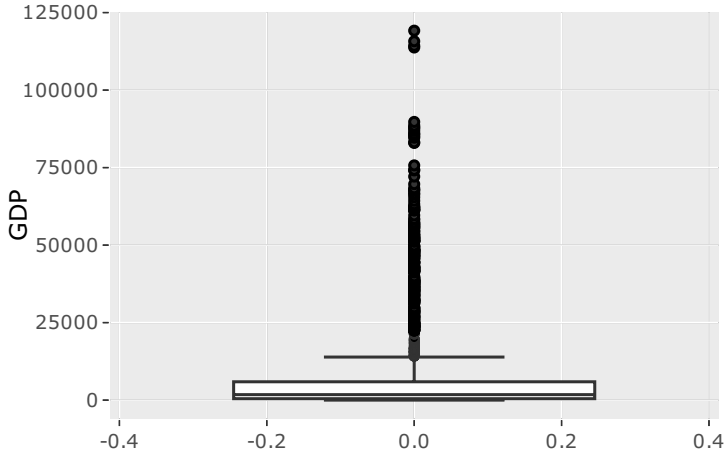

```
ggplotly(ggplot(data, aes(y=Polio)) + geom_boxplot())
```

```
## Warning: Removed 19 rows containing non-finite values (stat_boxplot).
```



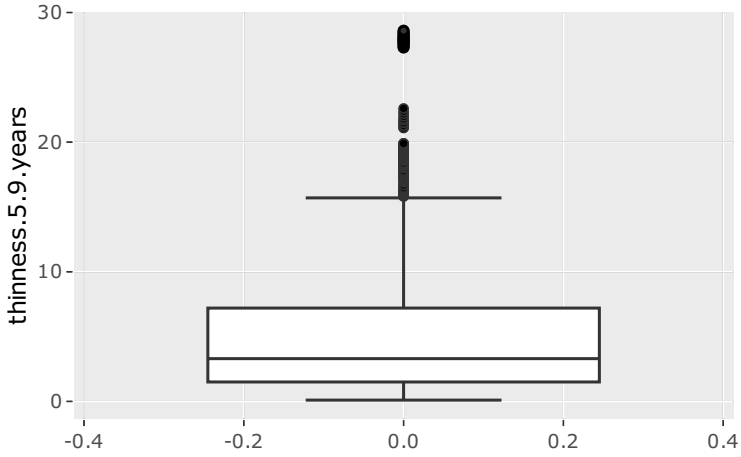

```
ggplotly(ggplot(data, aes(y=GDP)) + geom_boxplot())
```

```
## Warning: Removed 448 rows containing non-finite values (stat_boxplot).
```




```
ggplotly(ggplot(data, aes(y=thinness.5.9.years)) + geom_boxplot())
```

```
## Warning: Removed 34 rows containing non-finite values (stat_boxplot).
```




```
data <- as.data.table(data)

data1 <- data %>% filter(
  !is.na(data$Country),
  !is.na(data$Status),
  !is.na(data$Adult.Mortality),
  !is.na(data$Life.expectancy),
  !is.na(data$infant.deaths),
  !is.na(data$Alcohol),
  !is.na(data$percentage.expenditure),
  !is.na(data$Hepatitis.B),
  !is.na(data$Measles),
  !is.na(data$BMI),
  !is.na(data$under.five.deaths),
  !is.na(data$Polio),
  !is.na(data$Total.expenditure),
  !is.na(data$Diphtheria),
  !is.na(data$HIV.AIDS),
  !is.na(data$GDP),
  !is.na(data$Population),
  !is.na(data$thinness..1.19.years),
  !is.na(data$thinness.5.9.years),
  !is.na(data$Income.composition.of.resources),
  !is.na(data$Schooling)
)
```

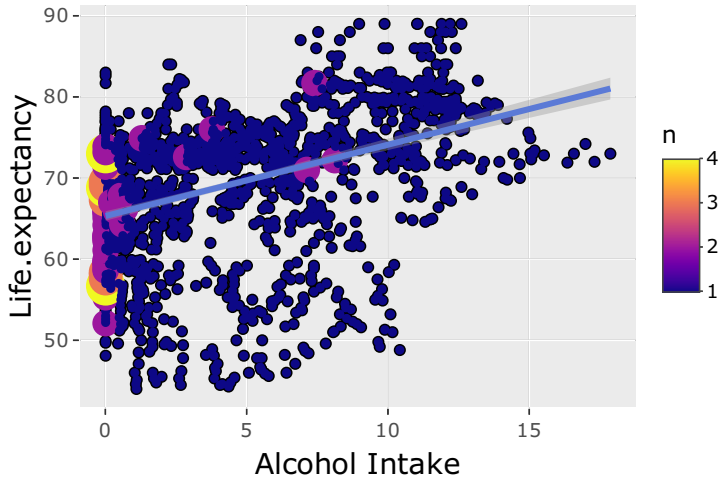
```
data1$Status <- factor(x=data1$Status, levels=c("Developing", "Developed"), labels= c(0,1))
data1 <- one_hot(as.data.table(data1), cols = "Status")
data1$Status <- data1$Status_1
```

Graphing

These graphs show the relations between Life Expectancy and various features of interest.

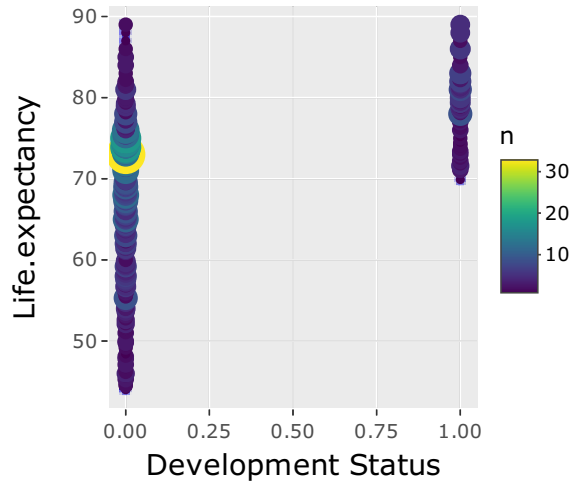
```
ggplotly(ggplot(data1, aes(x=Alcohol, y=Life.expectancy)) +
           geom_point() +
           geom_count(aes(colour= ..n.., size=..n..)) +
           scale_color_viridis(option = "C") +
           theme(axis.title.x=element_text(angle=0,size = rel(1.25),
           margin = margin(1, unit = "cm"),vjust = 1)) +
           theme(axis.title.y=element_text(angle=90,size = rel(1.25))) +
           geom_smooth(method="lm") +
           labs(x="Alcohol Intake"))
```

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



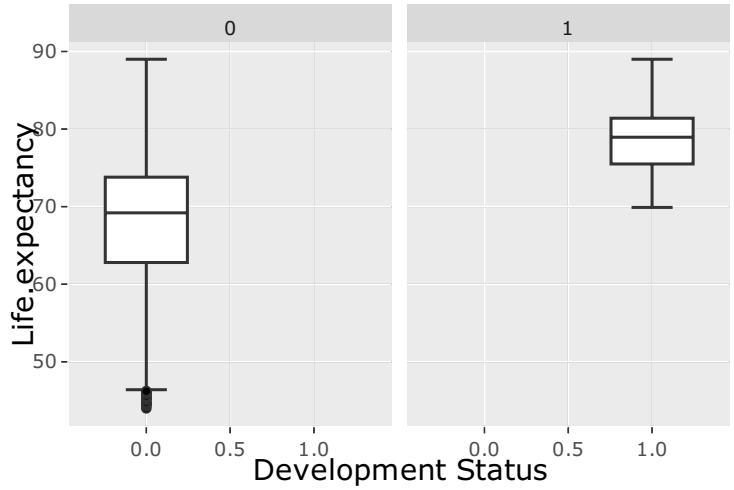

```
ggplotly(ggplot(data1, aes(x=Status, y=Life.expectancy)) +
           geom_point(alpha=.5, colour="blue") +
```

```
            geom_count(aes(colour= ..n.., size=..n..)) +
            scale_color_viridis(option="D") +
            theme(axis.title.x=element_text(angle=0,size = rel(1.25),
            margin = margin(1, unit = "cm"),vjust = 1)) +
            theme(axis.title.y=element_text(angle=0,size = rel(1.25))) +
            labs(x="Development Status"))
```



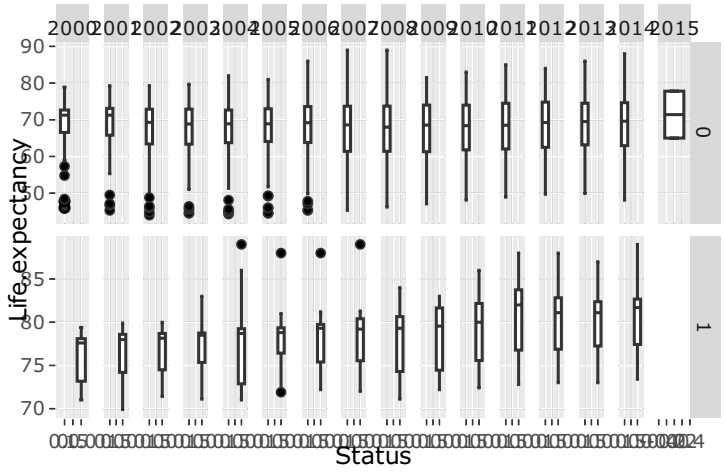

```
ggplotly(ggplot(data1, aes(x=Status, y=Life.expectancy)) +
            geom_boxplot() +
            theme(axis.title.x=element_text(angle=0,size = rel(1.25),
            margin = margin(1, unit = "cm"),vjust = 1)) +
            theme(axis.title.y=element_text(angle=90,size = rel(1.25))) +
            facet_wrap(data1$Status) +
            labs(x="Development Status"))
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



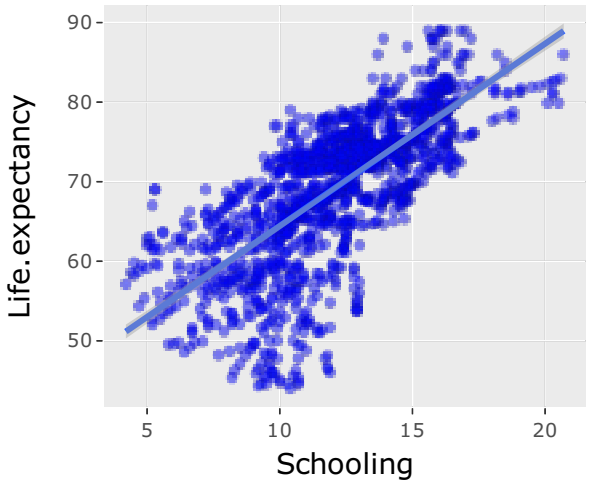

```
ggplotly(ggplot(data1, aes(x=Status, y=Life.expectancy)) +
           theme(axis.title.x=element_text(angle=0,size = rel(1))) +
           theme(axis.title.y=element_text(angle=90, size = rel(1))) +
           geom_boxplot() +
           facet_grid(rows = vars(data1$Status), cols=vars(data1$Year), scales = "free"))
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```




```
ggplotly(ggplot(data1, aes(x=Schooling, y=Life.expectancy)) +
           geom_point(alpha=.5, colour="blue") +
           geom_smooth(method="lm") + theme(axis.title.x=element_text(angle=0,size = rel(1.25),
           margin = margin(1, unit = "cm"),vjust =1)) +
           theme(axis.title.y=element_text(angle=0,size = rel(1.25))))
```

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




```
mean(data1$Life.expectancy[which(data1$Year == 2000)])
```

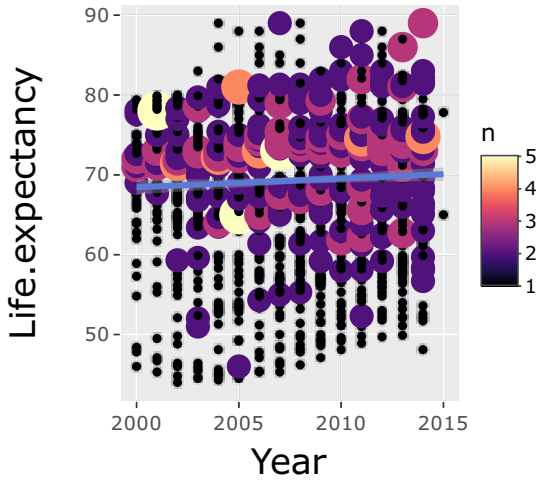
```
## [1] 70.2
```

```
mean(data1$Life.expectancy[which(data1$Year == 2015)])
```

```
## [1] 71.4
```

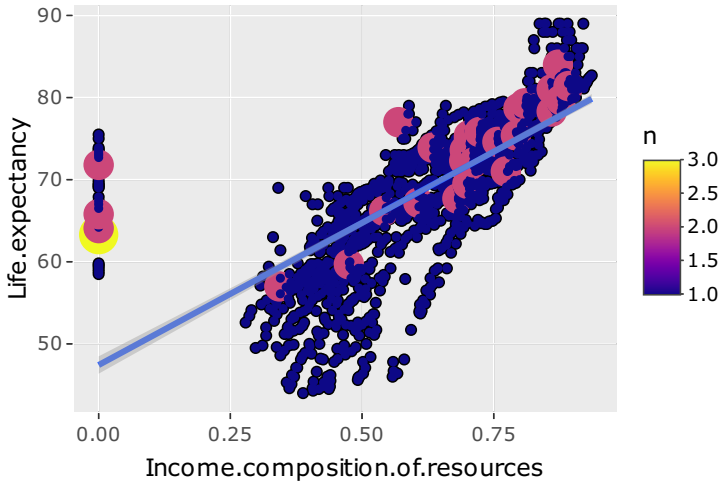
```
ggplotly(ggplot(data1, aes(x=Year, y=Life.expectancy)) + geom_point(alpha=.25) +
  geom_count(aes(colour= ..n.., size=..n..)) +
  scale_color_viridis(option = "A") +
  geom_smooth(method="lm") +
    theme(axis.title.x=element_text(angle=0,size = rel(1.5),
    margin = margin(1, unit = "cm"),vjust =1)) +
    theme(axis.title.y=element_text(angle=0,size = rel(1.5))))
```

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

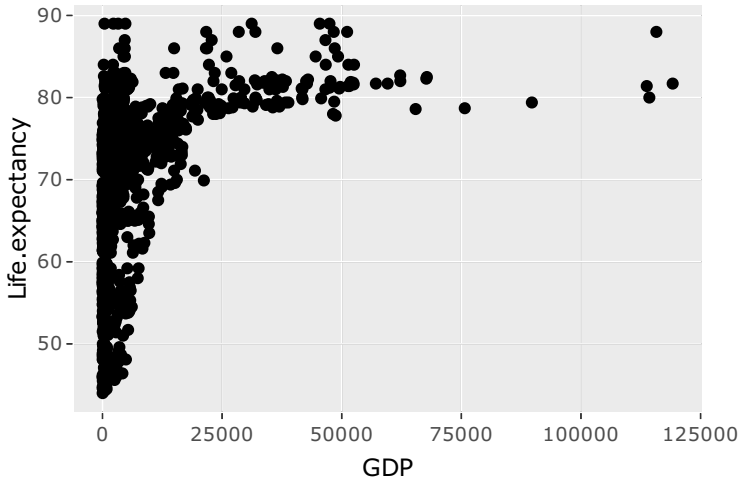



```
ggplotly(ggplot(data1, aes(x= Income.composition.of.resources,y= Life.expectancy)) +
  geom_point() +
  geom_count(aes(colour= ..n.., size=..n..)) +
  scale_color_viridis(option = "C") +
  geom_smooth(method="lm"))
```

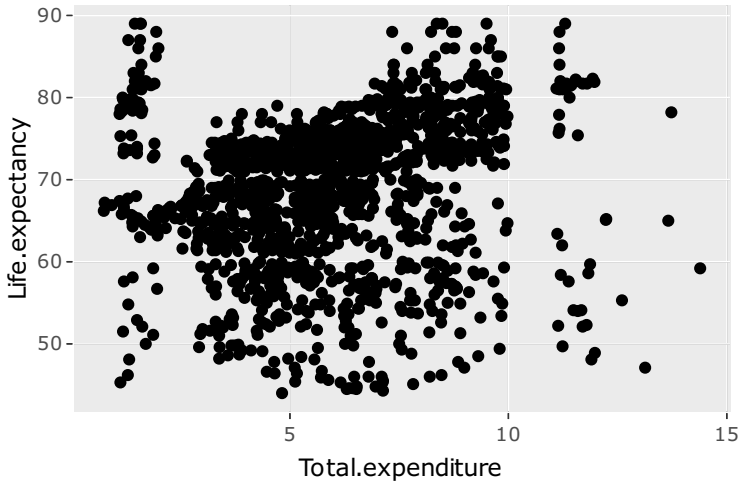
```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplotly(ggplot(data1, aes(x= GDP,y= Life.expectancy)) +
  geom_point())
```

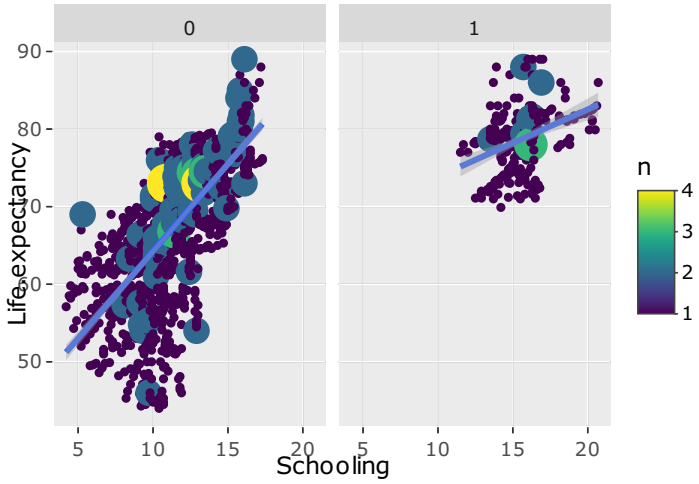



```
ggplotly(ggplot(data1, aes(x= Total.expenditure,y= Life.expectancy)) +
  geom_point())
```

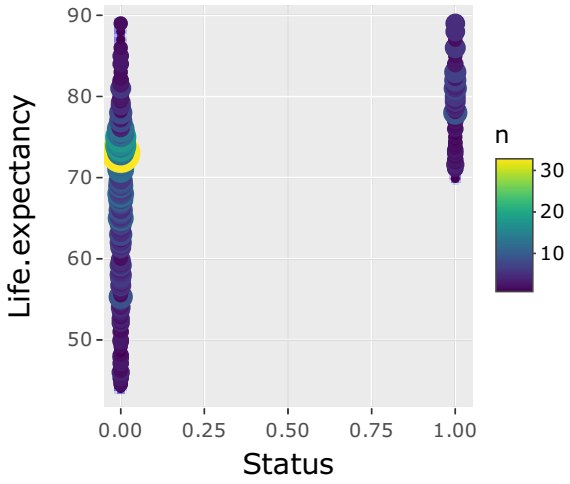



```
ggplotly(ggplot(data1, aes(x=Schooling,y=Life.expectancy)) +
           geom_count(aes(colour=..n..,size=..n..)) +
           facet_wrap(data1$Status) +
           scale_color_viridis(option="D") +
           geom_smooth(method="lm"))
```

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

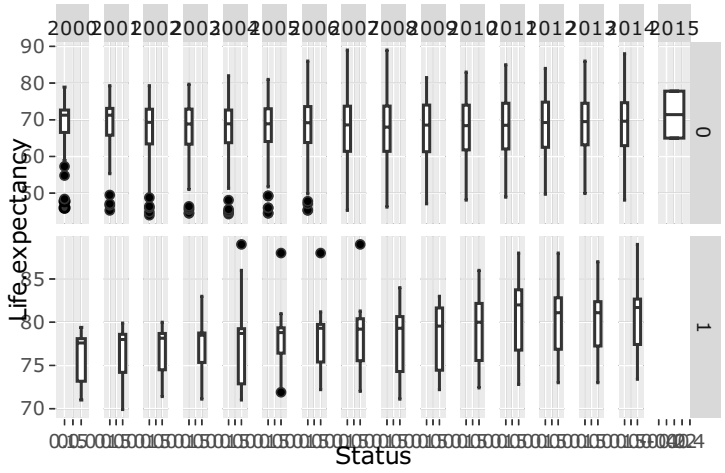



```
ggplotly(ggplot(data1, aes(x=Status, y=Life.expectancy)) +
           geom_point(alpha=.5, colour="blue") +
           geom_count(aes(colour= ..n.., size=..n..)) +
           scale_color_viridis(option="D") +
           theme(axis.title.x=element_text(angle=0,size = rel(1.25),
           margin = margin(1, unit = "cm"),vjust = 1)) +
           theme(axis.title.y=element_text(angle=0,size = rel(1.25))))
```



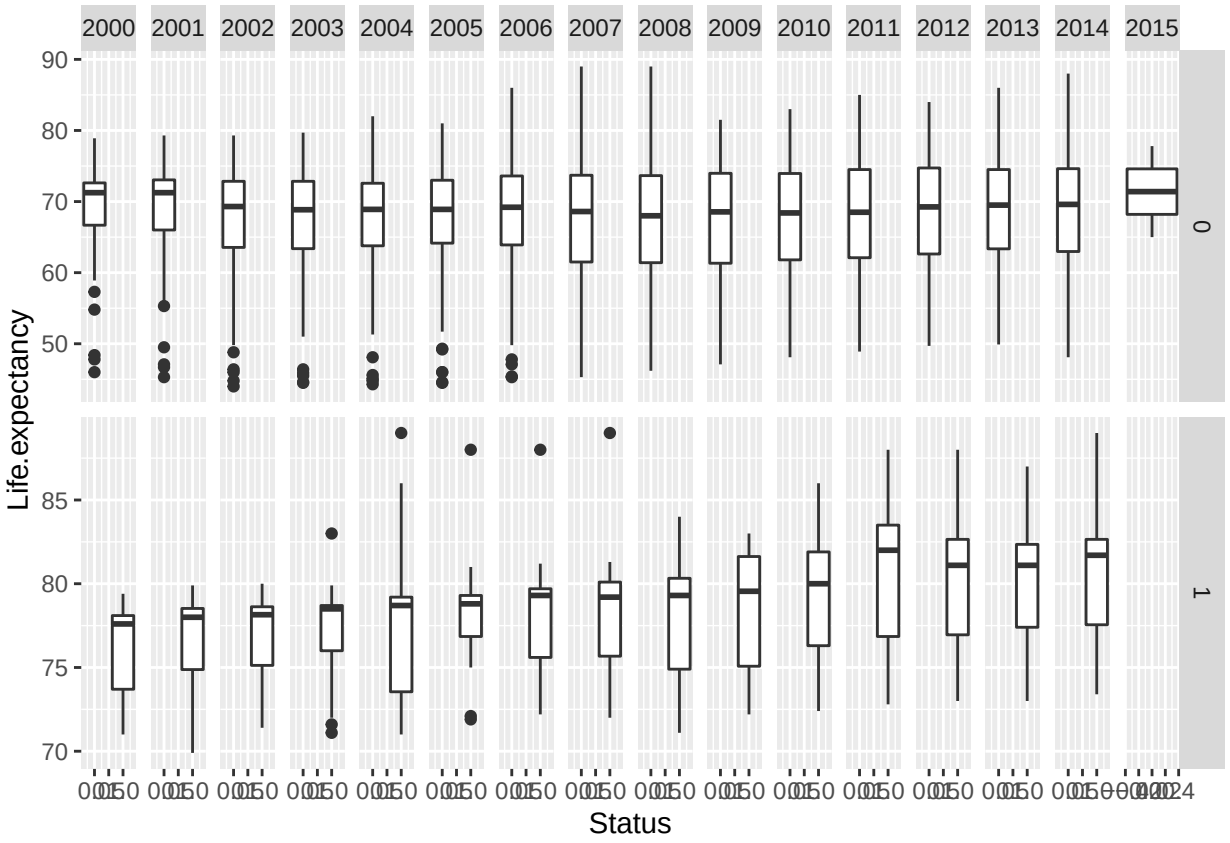

```
ggplotly(ggplot(data1, aes(x=Status, y=Life.expectancy)) +
           theme(axis.title.x=element_text(angle=0,size = rel(1))) +
           theme(axis.title.y=element_text(angle=90, size = rel(1))) +
           geom_boxplot() +
           facet_grid(rows = vars(data1$Status), cols=vars(data1$Year), scales = "free"))
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```




```
ggplot(data1, aes(x=Status, y=Life.expectancy)) +
           theme(axis.title.x=element_text(angle=0,size = rel(1))) +
           theme(axis.title.y=element_text(angle=90, size = rel(1))) +
           geom_boxplot() +
           facet_grid(rows = vars(data1$Status), cols=vars(data1$Year), scales = "free")
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



Model Fitting

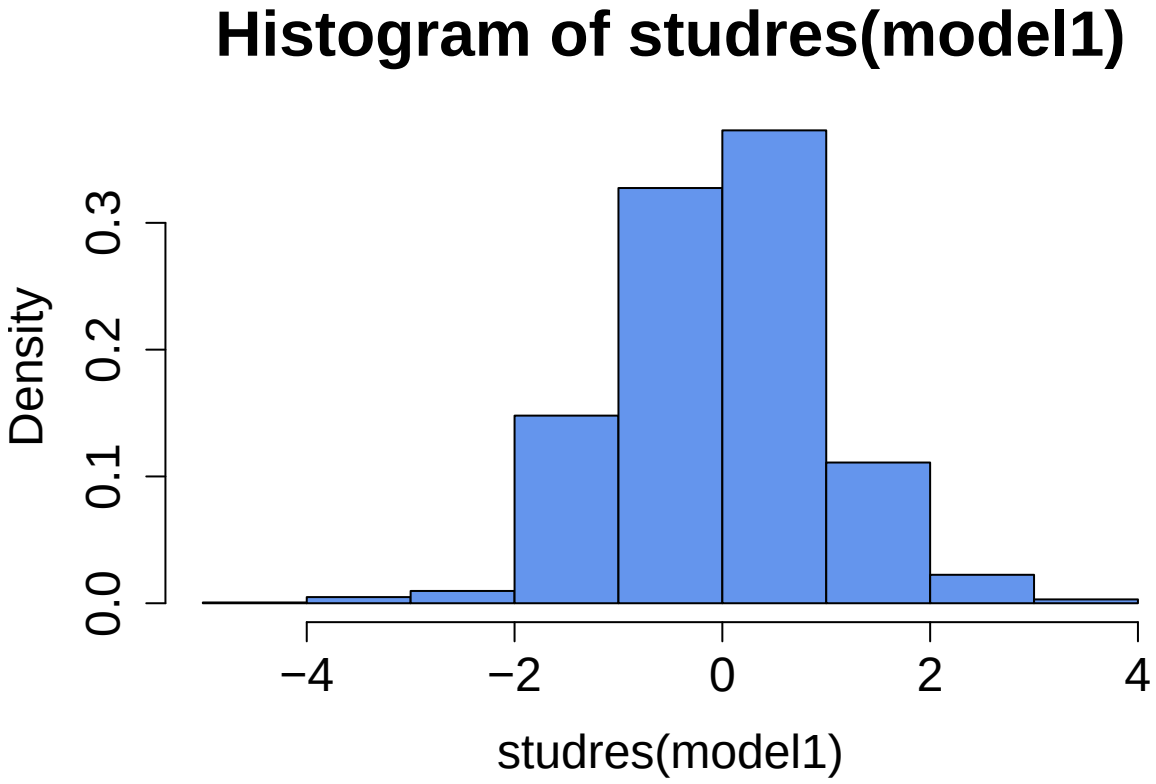
The first model for variable selection

```
data1 <- within(data1, rm(Year, Country, Adult.Mortality, infant.deaths, Status_0, Status_1))

model1 <- lm(Life.expectancy ~ Status+Alcohol+percentage.expenditure+
               Hepatitis.B+Measles+BMI+under.five.deaths+Polio+Total.expenditure+
               Diphtheria+HIV.AIDS+GDP+Population+thinness..1.19.years+
               thinness.5.9.years+Income.composition.of.resources+Schooling, data1)

summary(model1)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Status + Alcohol + percentage.expenditure +
##     Hepatitis.B + Measles + BMI + under.five.deaths + Polio +
##     Total.expenditure + Diphtheria + HIV.AIDS + GDP + Population +
##     thinness..1.19.years + thinness.5.9.years + Income.composition.of.resources +
##     Schooling, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.9334  -2.5678   0.1136   2.6868  13.4855
```

```
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.661e+01  7.384e-01  63.128  < 2e-16 ***
## Status                          1.593e+00  3.777e-01   4.219 2.59e-05 ***
## Alcohol                        -2.263e-01  3.636e-02  -6.224 6.14e-10 ***
## percentage.expenditure          4.475e-04  2.021e-04   2.214 0.026982 *
## Hepatitis.B                    -1.053e-02  4.990e-03  -2.110 0.034967 *
## Measles                         2.607e-05  1.164e-05   2.240 0.025229 *
## BMI                             4.557e-02  6.704e-03   6.798 1.48e-11 ***
## under.five.deaths              -2.611e-03  1.003e-03  -2.604 0.009309 **
## Polio                           1.484e-02  5.771e-03   2.571 0.010229 *
## Total.expenditure               9.098e-02  4.564e-02   1.993 0.046403 *
## Diphtheria                      2.206e-02  6.621e-03   3.332 0.000882 ***
## HIV.AIDS                       -5.987e-01  1.759e-02 -34.036  < 2e-16 ***
## GDP                             4.159e-06  3.185e-05   0.131 0.896121
## Population                      2.786e-09  1.904e-09   1.463 0.143668
## thinness..1.19.years            5.388e-03  5.947e-02   0.091 0.927819
## thinness.5.9.years             -5.850e-02  5.849e-02  -1.000 0.317426
## Income.composition.of.resources 1.242e+01  9.234e-01  13.452  < 2e-16 ***
## Schooling                       1.009e+00  6.624e-02  15.227  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.028 on 1631 degrees of freedom
## Multiple R-squared:  0.7925, Adjusted R-squared:  0.7903
## F-statistic: 366.3 on 17 and 1631 DF,  p-value: < 2.2e-16
```

```
hist(studres(model1),
     breaks=10, freq=F, col="cornflowerblue", cex.axis=1.5, cex.lab=1.5, cex.main=2)
```

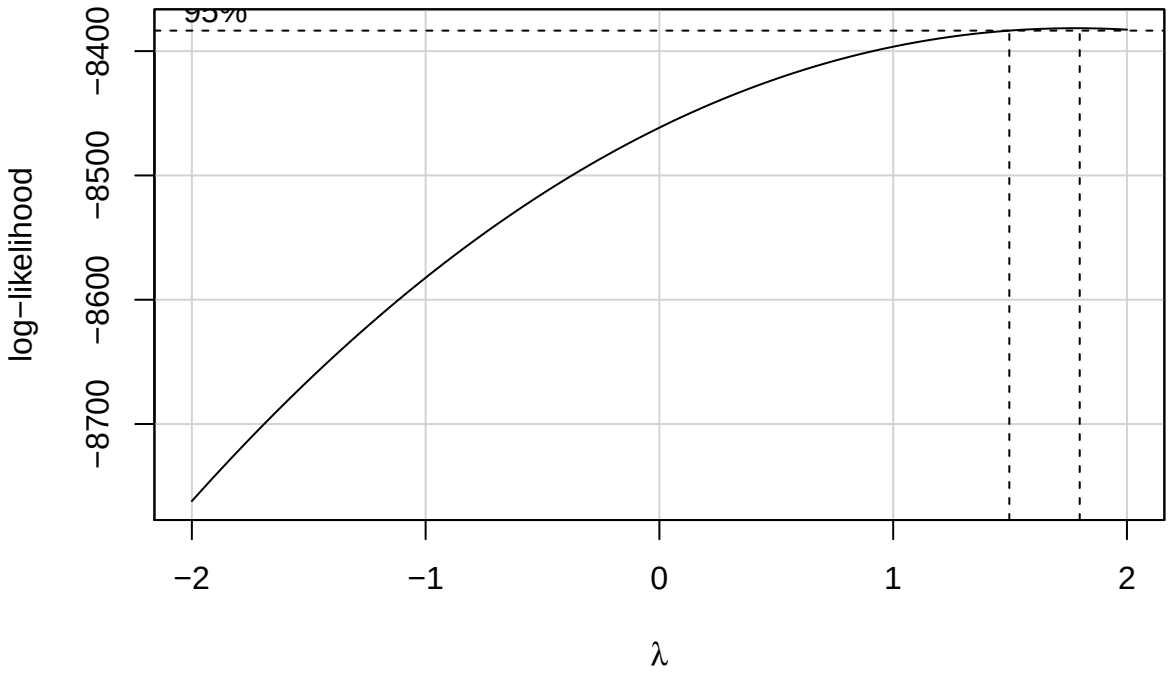



```
vif(model1)
```

```
##                          Status                        Alcohol
##                        1.814944                        2.179064
##          percentage.expenditure                    Hepatitis.B
##                       12.841644                        1.657777
##                        Measles                            BMI
##                        1.399114                        1.780987
##                under.five.deaths                          Polio
##                        2.709214                        1.704478
##               Total.expenditure                     Diphtheria
##                        1.118563                        2.073027
##                        HIV.AIDS                             GDP
##                        1.143260                       13.568641
##                      Population            thinness..1.19.years
##                        1.828447                        7.596814
##             thinness.5.9.years Income.composition.of.resources
##                        7.525020                        2.902866
##                      Schooling
##                        3.482467
```

```
bc1 <- boxCox(data1$Life.expectancy ~ data1$Status+data1$Alcohol+
         data1$percentage.expenditure+data1$Hepatitis.B+data1$Measles+
         data1$BMI+data1$under.five.deaths+data1$Polio+data1$Total.expenditure+
         data1$Diphtheria+data1$HIV.AIDS+data1$GDP+data1$Population+
```

```
        data1$thinness..1.19.years+data1$thinness.5.9.years+
        data1$Income.composition.of.resources+data1$Schooling)
```



Notice the histogram: It is an obvious normal distribution, so we won't need transformations.

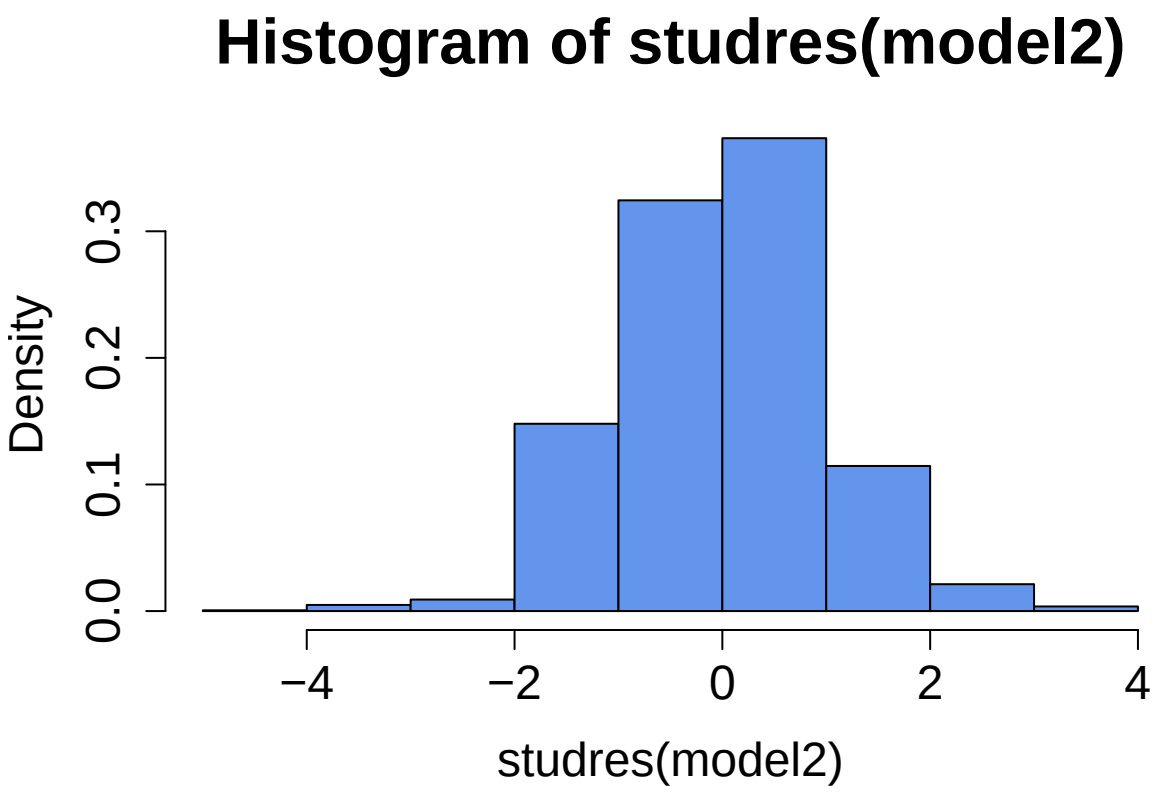
The second model is one with the optimized variables selected

```
model2 <- lm(Life.expectancy ~ Status+Alcohol+percentage.expenditure+
               Hepatitis.B+Measles+BMI+under.five.deaths+Polio+Total.expenditure+
               Diphtheria+HIV.AIDS+Income.composition.of.resources+Schooling, data1)

summary(model2)
```

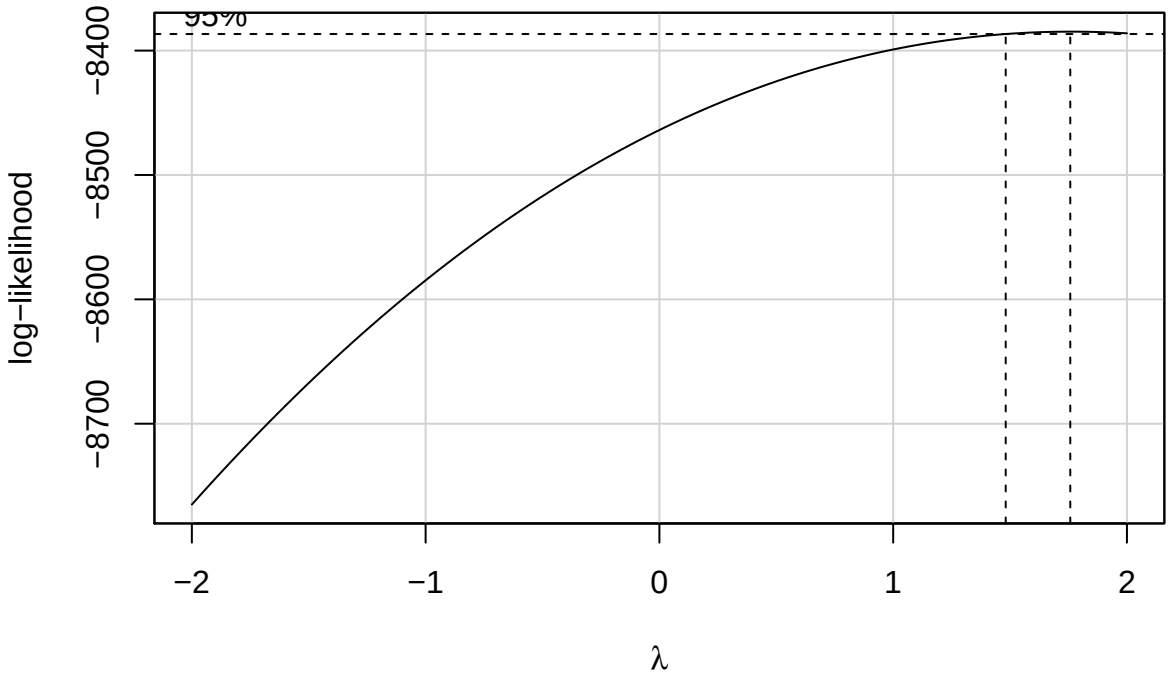
```
##
## Call:
## lm(formula = Life.expectancy ~ Status + Alcohol + percentage.expenditure +
##     Hepatitis.B + Measles + BMI + under.five.deaths + Polio +
##     Total.expenditure + Diphtheria + HIV.AIDS + Income.composition.of.resources +
##     Schooling, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.9880  -2.5468   0.1077   2.7279  13.7823
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                      4.596e+01  6.617e-01  69.454  < 2e-16 ***
## Status                           1.596e+00  3.769e-01   4.235 2.42e-05 ***
## Alcohol                         -2.191e-01  3.583e-02  -6.115 1.20e-09 ***
## percentage.expenditure           4.761e-04  6.671e-05   7.138 1.42e-12 ***
## Hepatitis.B                     -1.111e-02  4.982e-03  -2.231 0.025806 *
## Measles                          2.801e-05  1.155e-05   2.425 0.015419 *
## BMI                              5.019e-02  6.248e-03   8.033 1.80e-15 ***
## under.five.deaths               -2.368e-03  7.449e-04  -3.179 0.001508 **
## Polio                            1.483e-02  5.757e-03   2.576 0.010078 *
## Total.expenditure                9.657e-02  4.548e-02   2.123 0.033880 *
## Diphtheria                       2.254e-02  6.609e-03   3.410 0.000666 ***
## HIV.AIDS                        -6.022e-01  1.750e-02 -34.421  < 2e-16 ***
## Income.composition.of.resources  1.254e+01  9.187e-01  13.654  < 2e-16 ***
## Schooling                        1.018e+00  6.571e-02  15.494  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.03 on 1635 degrees of freedom
## Multiple R-squared:  0.7918, Adjusted R-squared:  0.7901
## F-statistic: 478.3 on 13 and 1635 DF,  p-value: < 2.2e-16
```

```
hist(studres(model2),
     breaks=10, freq=F, col="cornflowerblue", cex.axis=1.5, cex.lab=1.5, cex.main=2)
```



```
bc2 <- boxCox(data1$Life.expectancy~data1$Status+data1$Alcohol+
         data1$percentage.expenditure+data1$Hepatitis.B+data1$Measles+
         data1$BMI+data1$under.five.deaths+data1$Polio+data1$Total.expenditure+
         data1$Diphtheria+data1$HIV.AIDS+
         data1$Income.composition.of.resources+data1$Schooling)
```



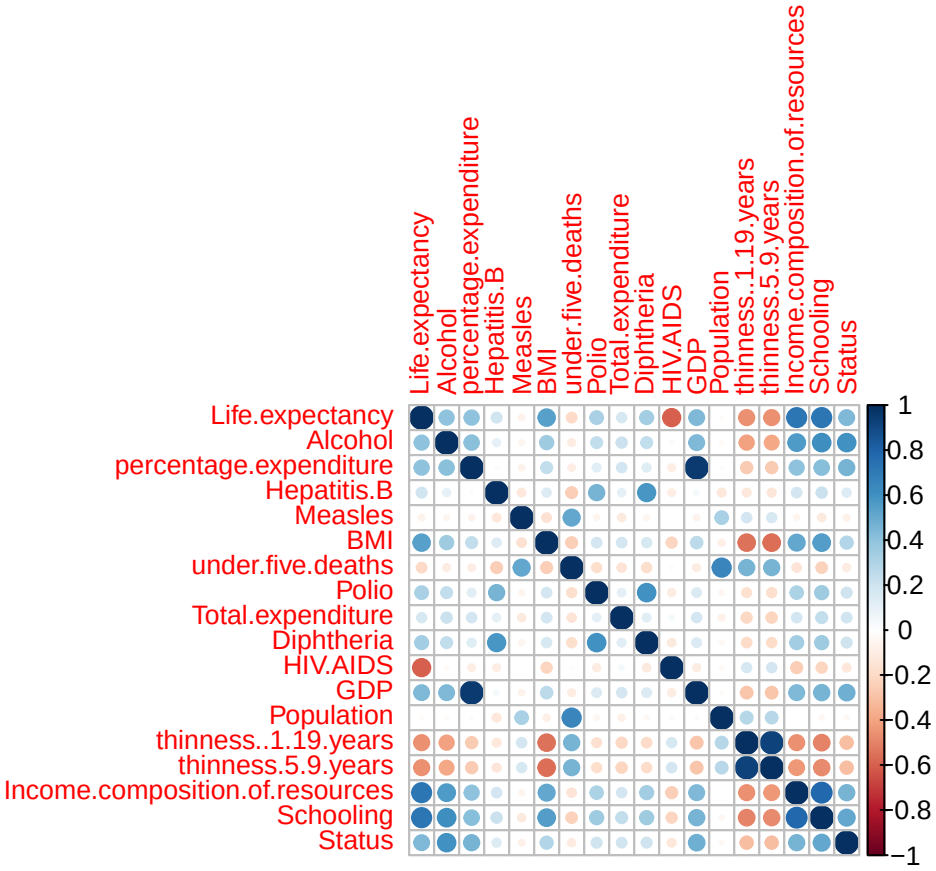
Another normal distribution. No transformation needed.

#Correlation Diagram

```
typeof(data1)
```

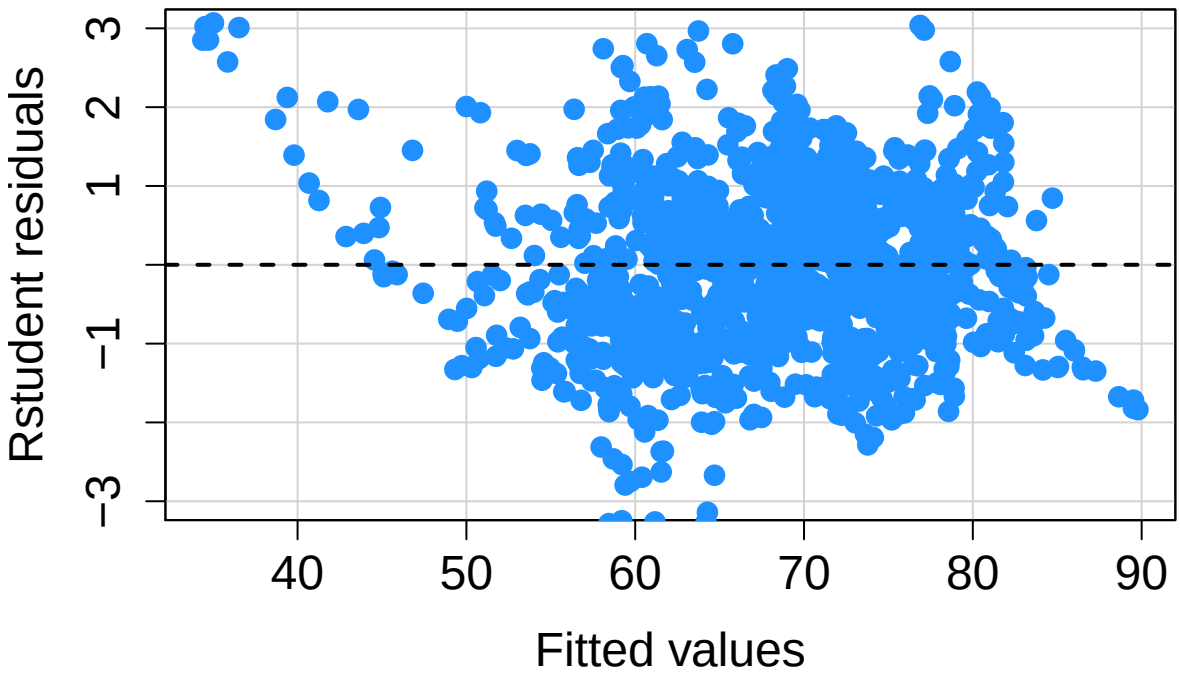
```
## [1] "list"
```

```
data1.cor <- cor(data1)

corrplot(data1.cor, tl.cex=.8)
```



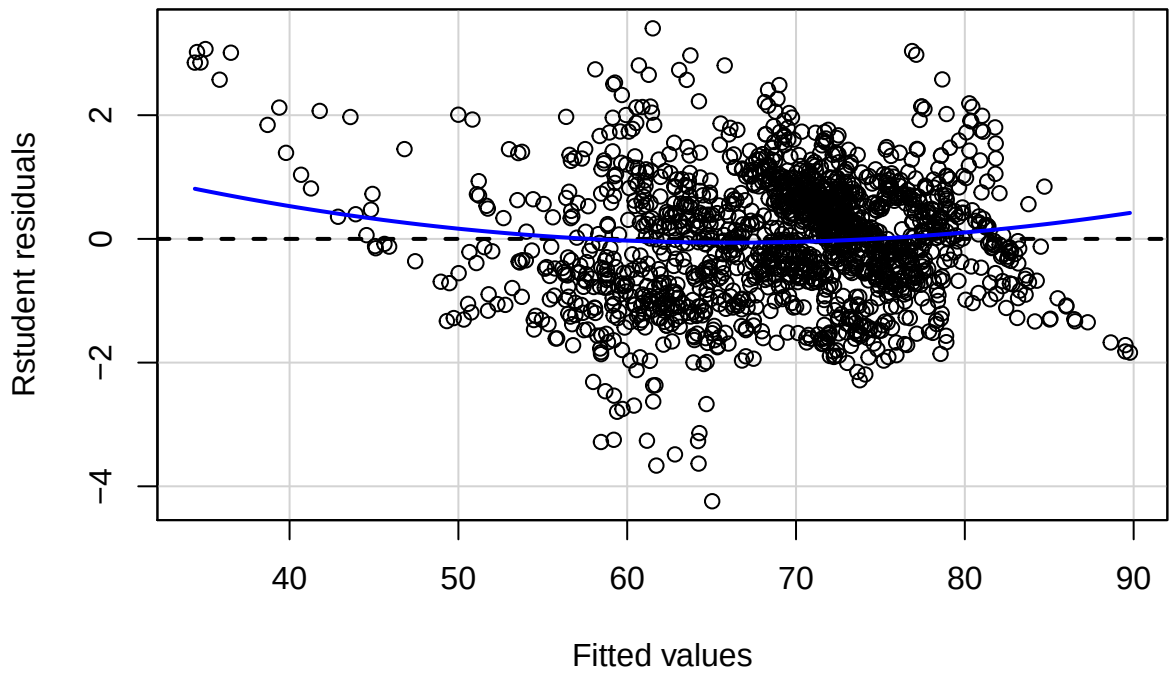
#Residual Analysis

```
summary(residualPlot(model1, type="rstudent", quadratic=F,
              col = "dodgerblue", pch=16, cex=1.5,
              cex.axis=1.5, cex.lab=1.5, ylim=c(-3,3)))
```

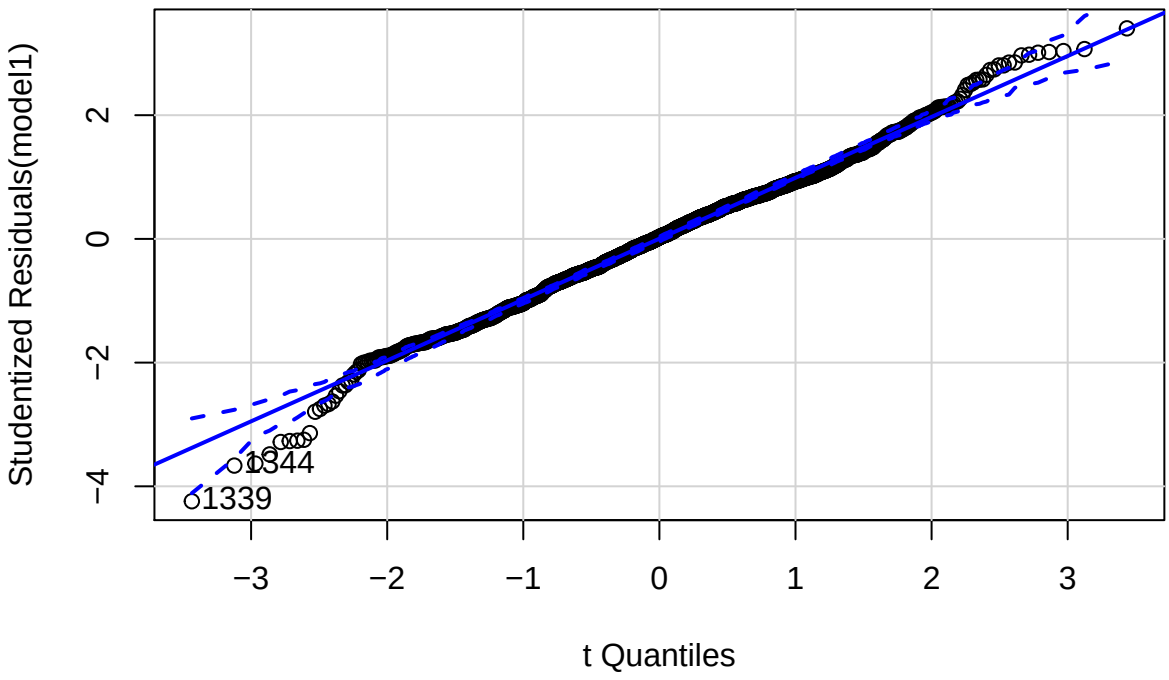



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.853   3.707   3.707   5.560   7.414
```

```
residualPlot(model1, variable='fitted', type='rstudent')
```

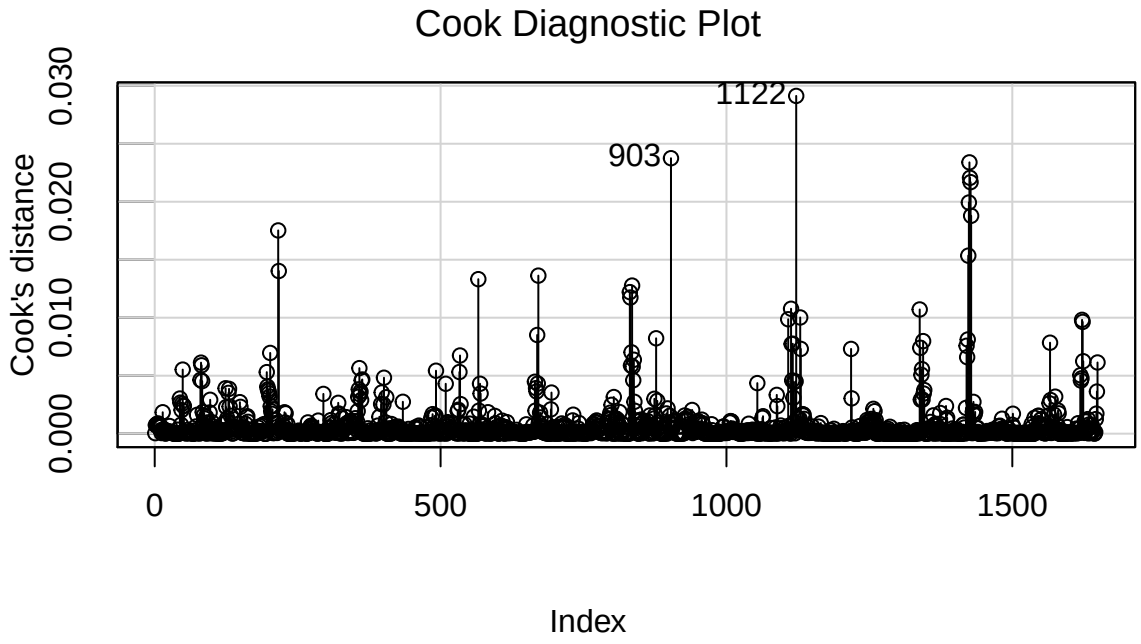



```
qqPlot(model1)
```

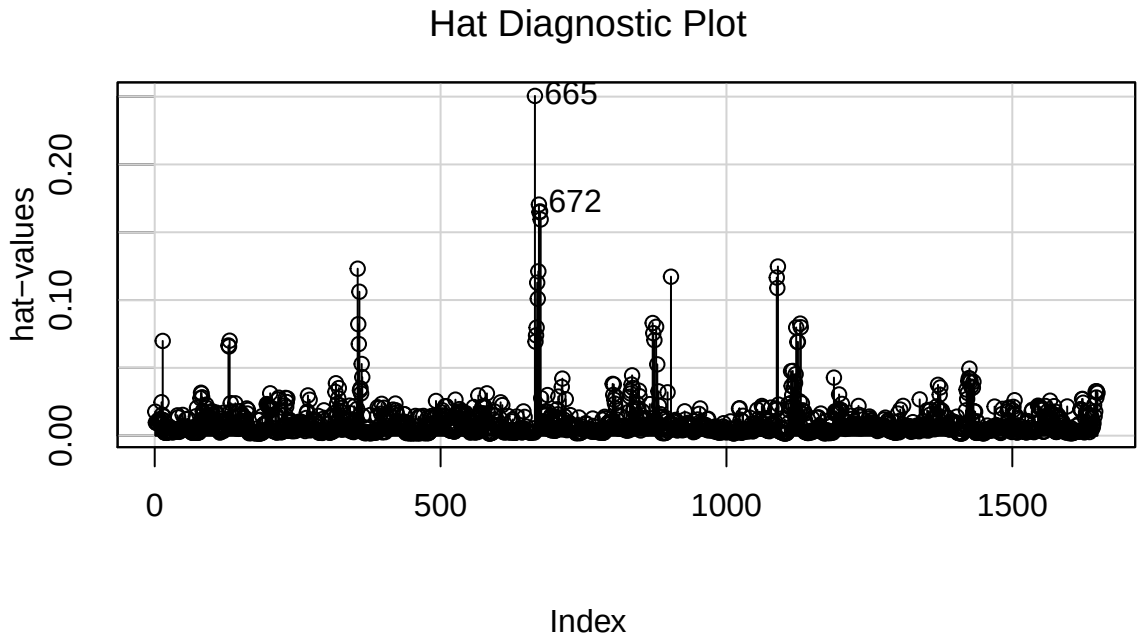



```
## [1] 1339 1344
```

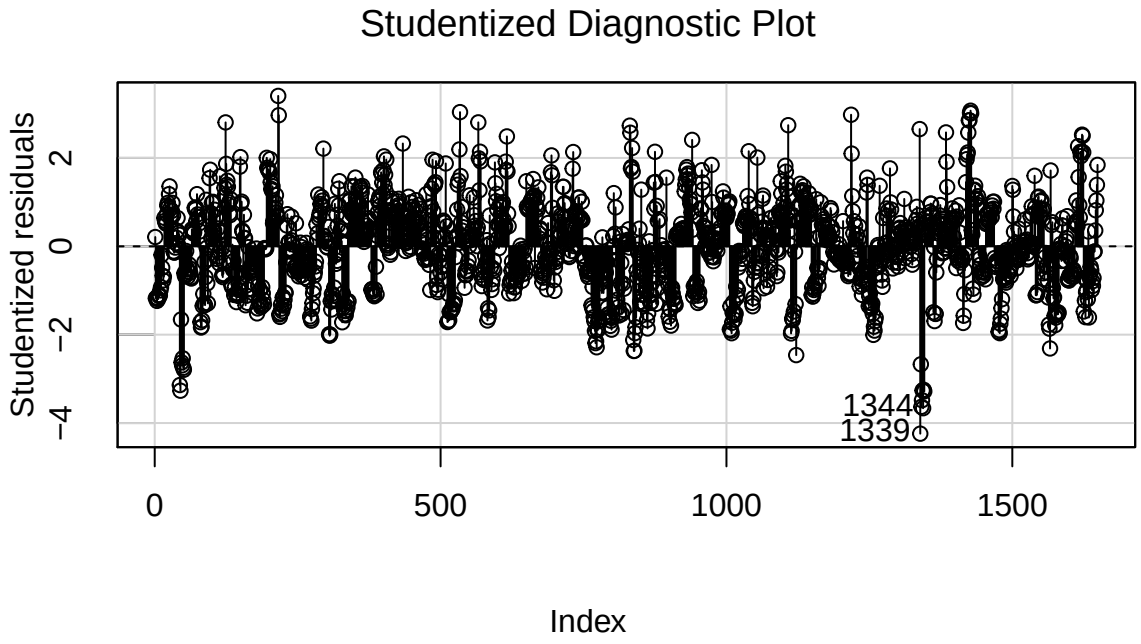
```
influenceIndexPlot(model1, vars=c('Cook'), data1, main = "Cook Diagnostic Plot")
```



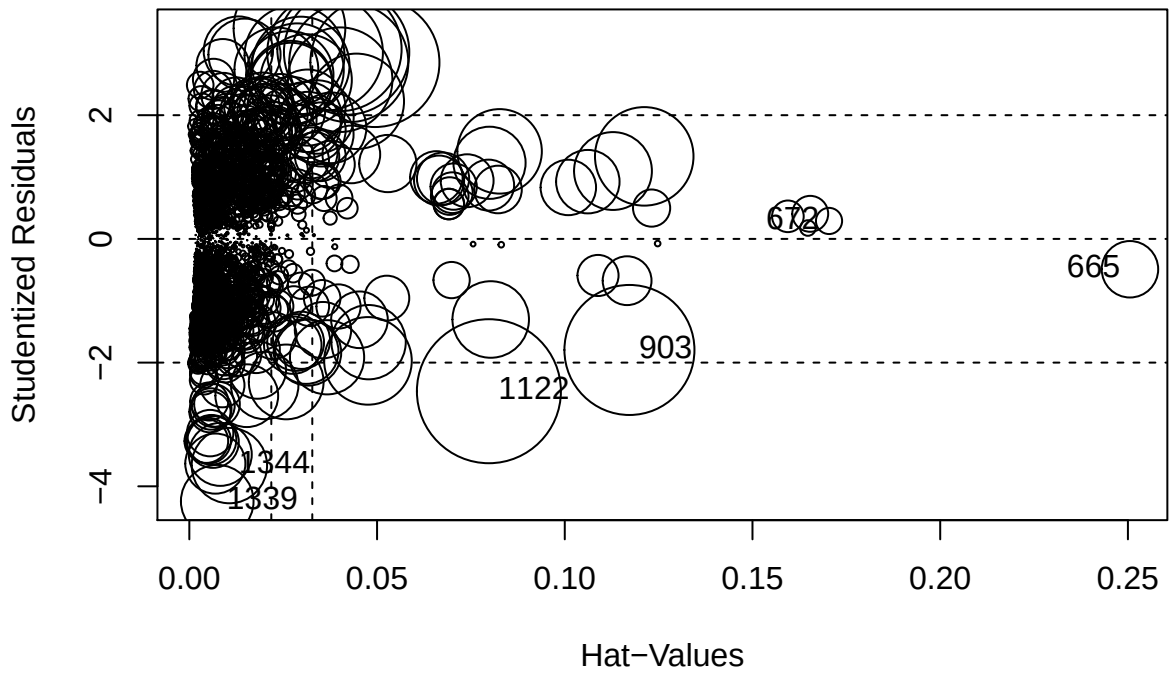

```
influenceIndexPlot(model1, vars=c('hat'), data1, main = "Hat Diagnostic Plot")
```




```
influenceIndexPlot(model1, vars=c('Studentized'), data1, main = "Studentized Diagnostic Plot")
```

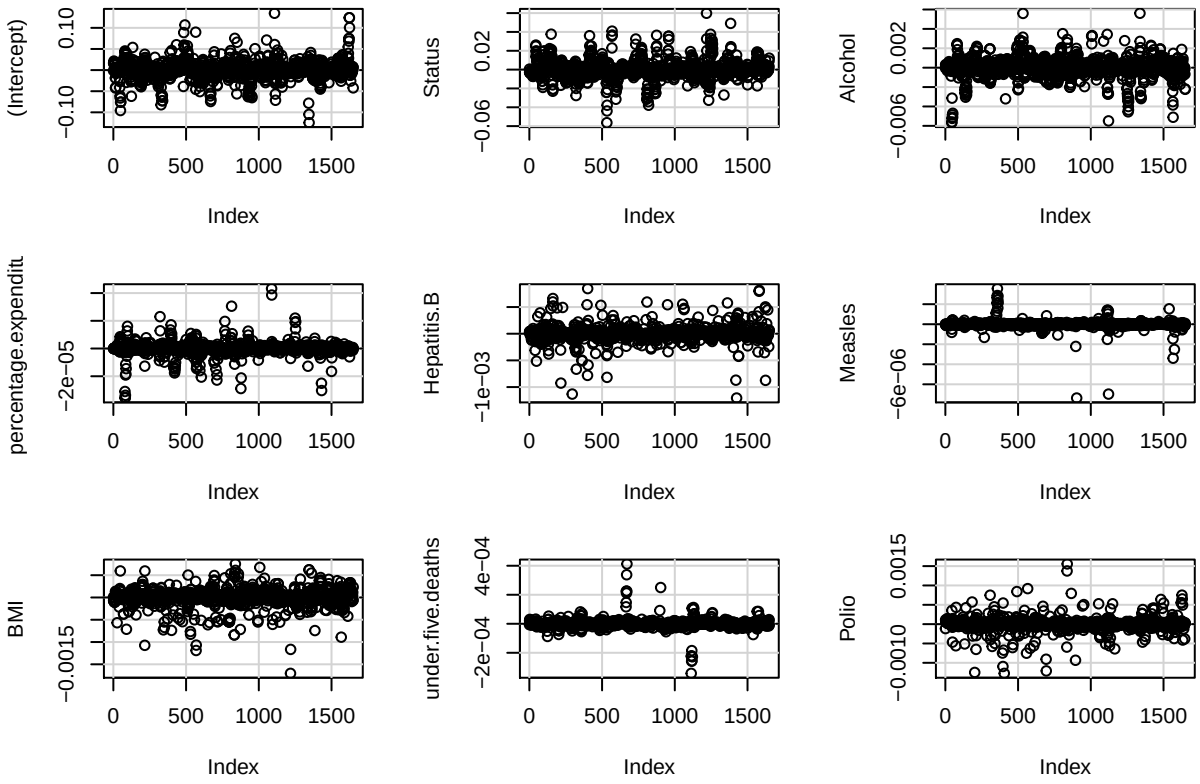


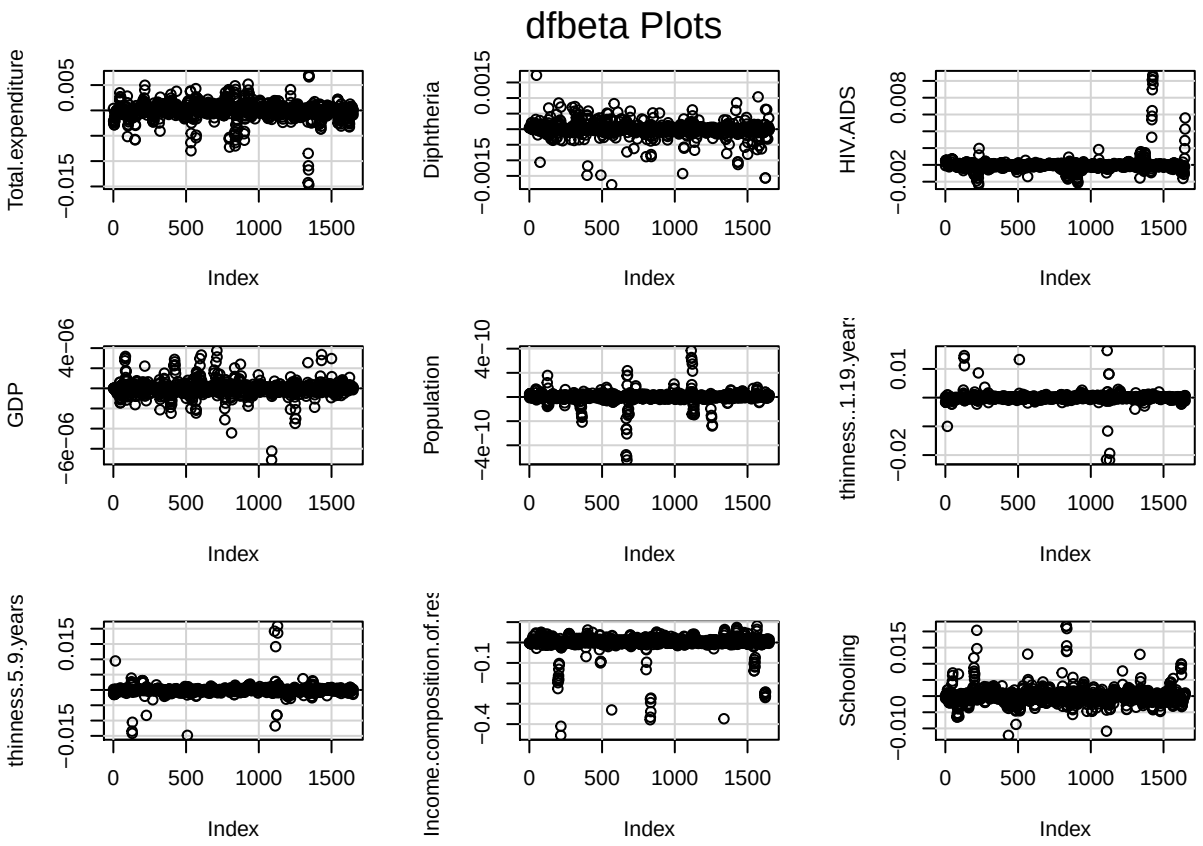

```
influencePlot(model1)
```



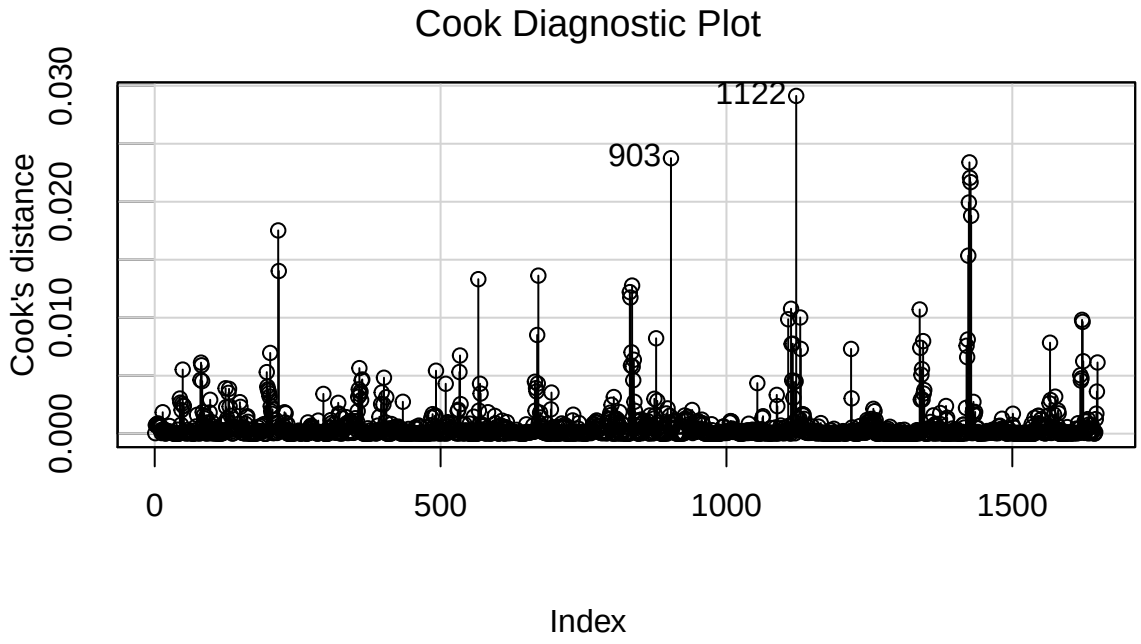

```
##          StudRes        Hat        CookD
## 665  -0.4908460 0.25054118 0.0044766325
## 672   0.2902808 0.17038757 0.0009619905
## 903  -1.7955325 0.11722488 0.0237515497
## 1122 -2.4626656 0.07978945 0.0291239396
## 1339 -4.2411744 0.00743226 0.0074056049
## 1344 -3.6646995 0.01066266 0.0079804630
```

```
dfbetaPlots(model1, intercept=T)
```

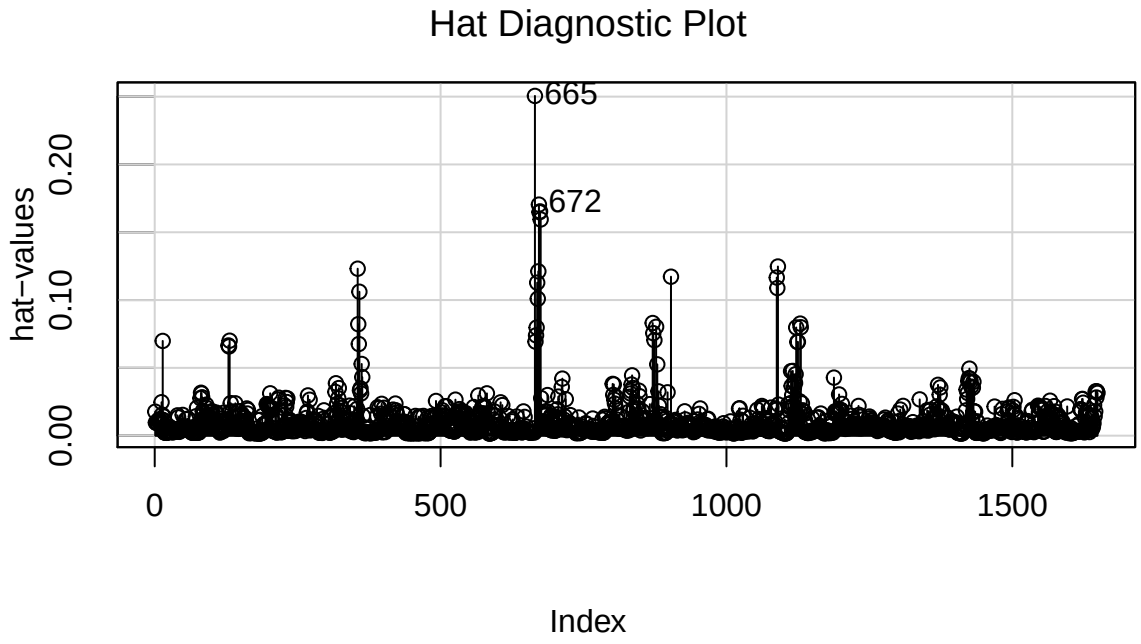

(Intercept)
−0.10
0.10
Status
−0.06
0.02
Alcohol
−0.006
0.002
percentage.expenditu
−2e−05
Hepatitis.B
−1e−03
Measles
−6e−06
BMI
−0.0015
under.five.deaths
−2e−04
4e−04
Polio
−0.0010
0.0015
0
500
1000
1500
Index



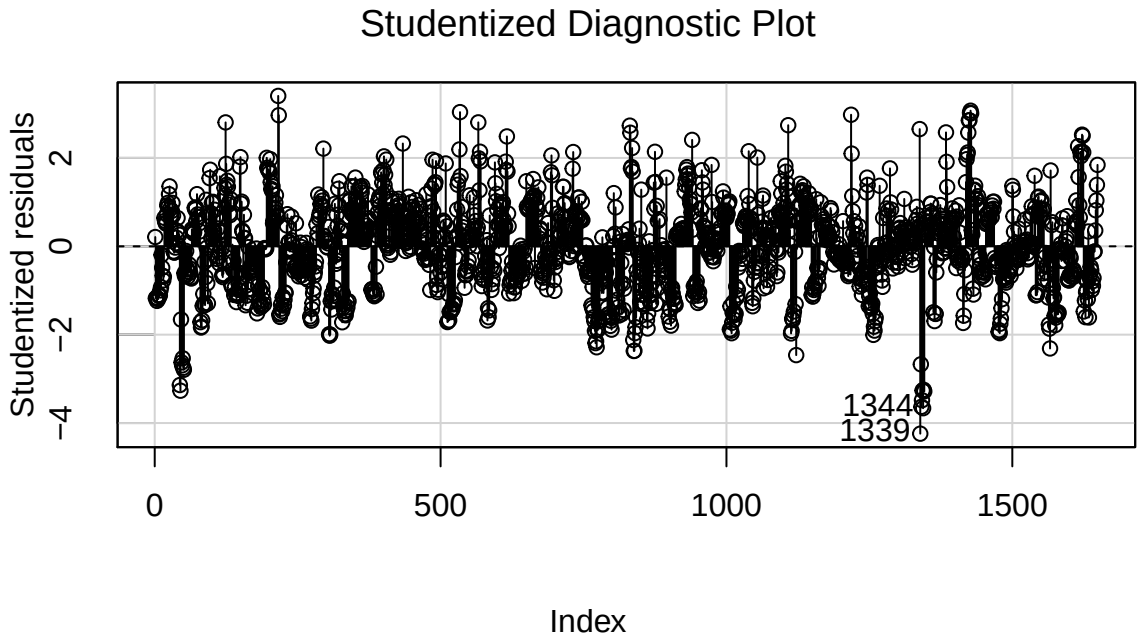

```
influenceIndexPlot(model1, vars=c('Cook'), data1, main = "Cook Diagnostic Plot")
```



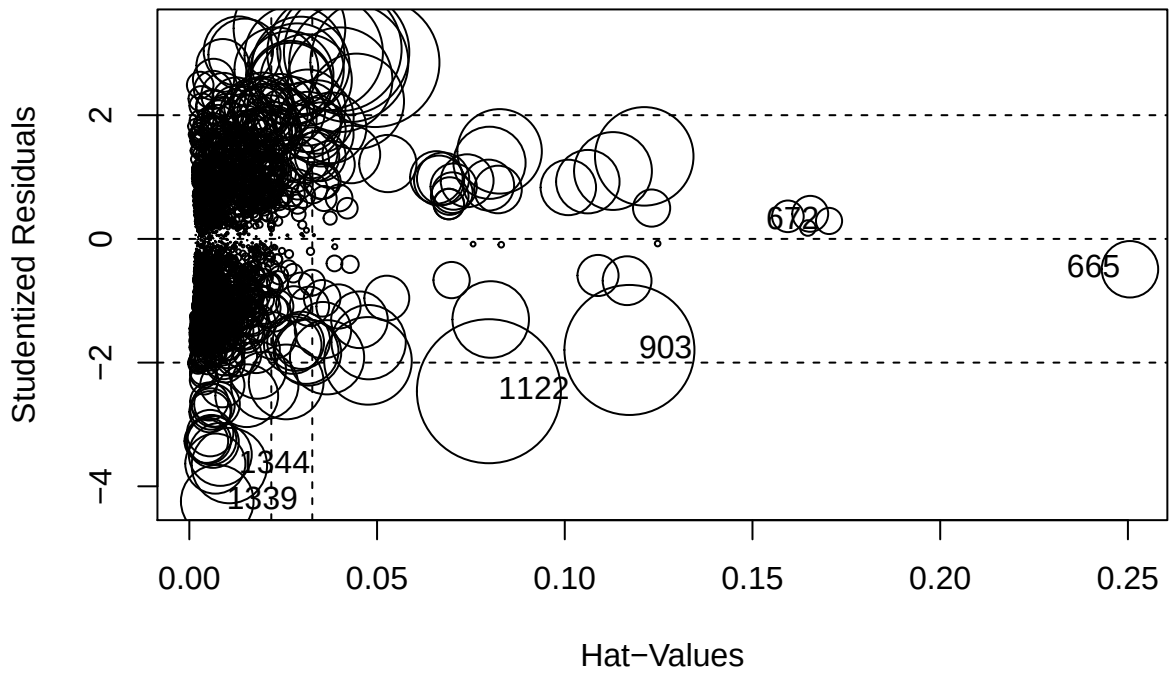

```
influenceIndexPlot(model1, vars=c('hat'), data1, main = "Hat Diagnostic Plot")
```




```
influenceIndexPlot(model1, vars=c('Studentized'), data1, main = "Studentized Diagnostic Plot")
```

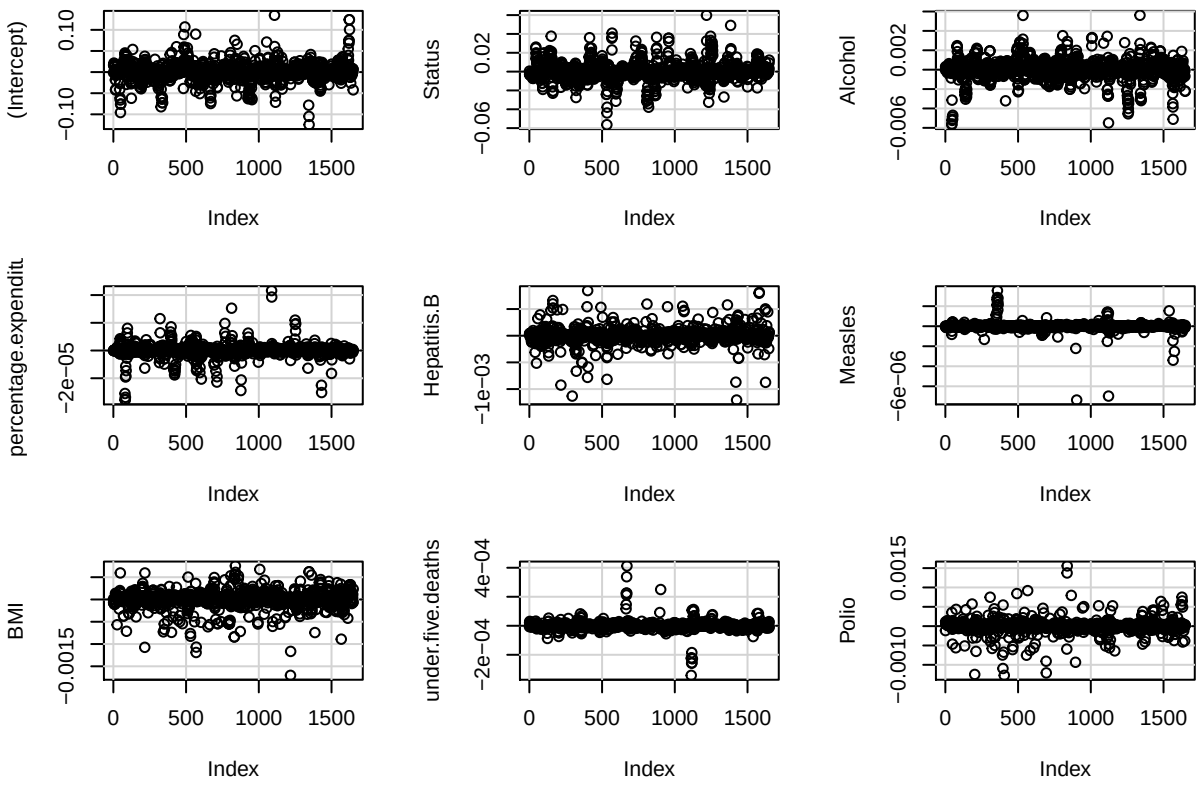


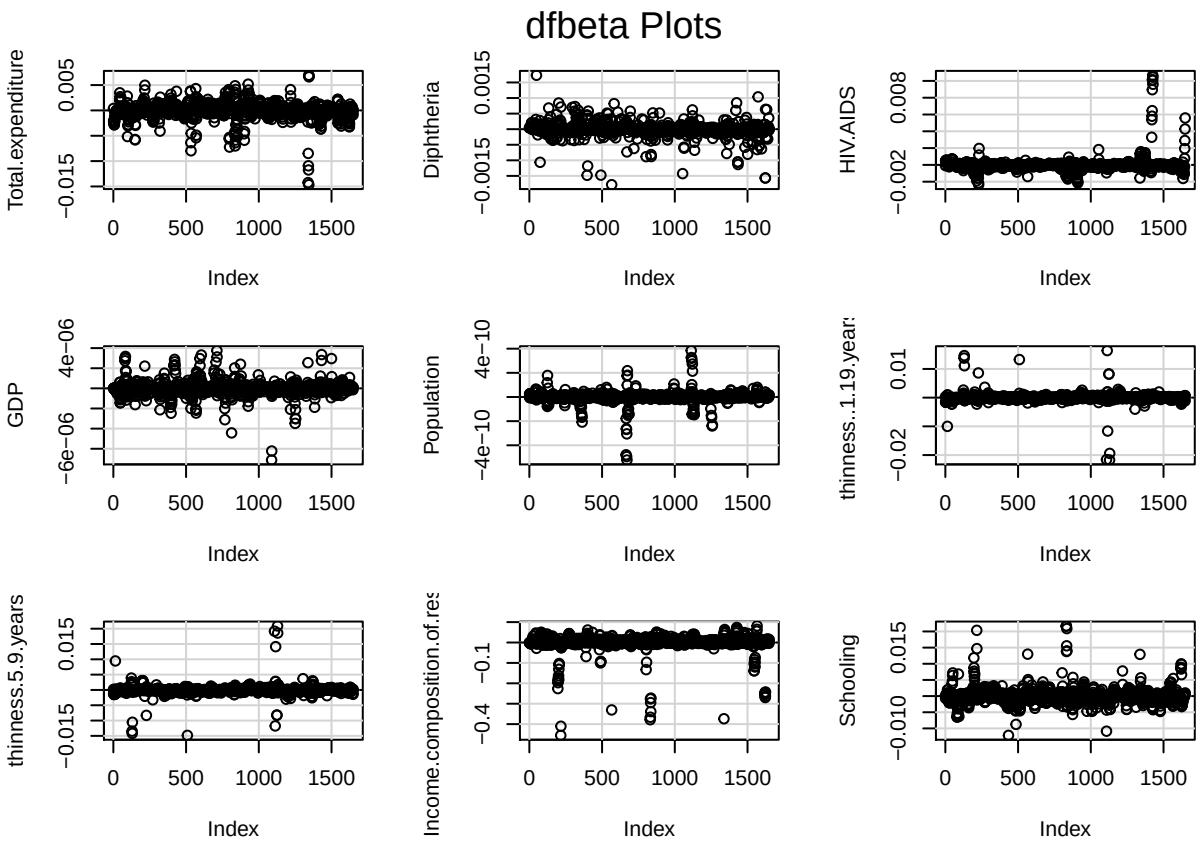

```
influencePlot(model1)
```




```
##          StudRes        Hat        CookD
## 665  -0.4908460 0.25054118 0.0044766325
## 672   0.2902808 0.17038757 0.0009619905
## 903  -1.7955325 0.11722488 0.0237515497
## 1122 -2.4626656 0.07978945 0.0291239396
## 1339 -4.2411744 0.00743226 0.0074056049
## 1344 -3.6646995 0.01066266 0.0079804630
```

```
dfbetaPlots(model1, intercept=T)
```


(Intercept)
−0.10
0.10
Status
−0.06
0.02
Alcohol
−0.006
0.002
percentage.expenditu
−2e−05
Hepatitis.B
−1e−03
Measles
−6e−06
BMI
−0.0015
under.five.deaths
−2e−04
4e−04
Polio
−0.0010
0.0015
0
500
1000
1500
Index



From the future: Outliers/Problem indexes are 903, 1122, 665, 672, 1344, 1339.

```
data1[c(903,1122,665,672,1344,1339,355),]
```

```
##    Life.expectancy Alcohol percentage.expenditure Hepatitis.B Measles  BMI
## 1:            52.9    1.08               9.728005          93  118712 17.6
## 2:            49.2    9.71               6.416253          18  110927 19.3
## 3:            68.0    3.07              86.521539          79   79563 18.1
## 4:            65.2    1.59               5.234770           6   41144 14.4
## 5:            47.1    3.97              49.837127          84      31 21.2
## 6:            48.1    0.01               1.443286          83    1006 23.8
## 7:            74.5    4.27              39.225097          95  131441 26.5
##    under.five.deaths Polio Total.expenditure Diphtheria HIV.AIDS        GDP
## 1:                54    86               1.50         93     13.7  458.86817
## 2:               893    45               4.11         36      5.4   87.89387
## 3:              1200    84               4.69         85      0.2 1573.11889
## 4:              1900    67               4.23         64      0.3  118.16637
## 5:                42    81              13.13         84      1.7  394.59324
## 6:                32    83              11.90         83      0.6   78.43948
## 7:               308    99               4.59         97      0.1 3471.24755
##    Population thinness..1.19.years thinness.5.9.years
## 1:    1516795                  6.8                6.7
## 2:  138939478                 12.9               12.9
## 3: 1293859294                 26.8               27.4
## 4: 1179681239                 27.1               28.0
## 5:      63126                  8.5                8.4
```

```
## 6:     779162                  7.5               7.4
## 7:    1324655                  4.5               4.0
##    Income.composition.of.resources Schooling Status
## 1:                           0.430      10.2      0
## 2:                           0.463       8.9      0
## 3:                           0.607      11.6      0
## 4:                           0.546       9.9      0
## 5:                           0.375       8.5      0
## 6:                           0.426       9.5      0
## 7:                           0.672      11.9      0
```

Lets Take them out and try it again.

```
data1 <- data1[-c(903,1122,665,672,1344,1339),]
data1
```

```
##       Life.expectancy Alcohol percentage.expenditure Hepatitis.B Measles  BMI
##    1:            65.0    0.01              71.279624          65    1154 19.1
##    2:            59.9    0.01              73.523582          62     492 18.6
##    3:            59.9    0.01              73.219243          64     430 18.1
##    4:            59.5    0.01              78.184215          67    2787 17.6
##    5:            59.2    0.01               7.097109          68    3013 17.2
##   ---
## 1639:            44.3    4.36               0.000000          68      31 27.1
## 1640:            44.5    4.06               0.000000           7     998 26.7
## 1641:            44.8    4.43               0.000000          73     304 26.3
## 1642:            45.3    1.72               0.000000          76     529 25.9
## 1643:            46.0    1.68               0.000000          79    1483 25.5
##       under.five.deaths Polio Total.expenditure Diphtheria HIV.AIDS       GDP
##    1:                83     6              8.16         65      0.1 584.25921
##    2:                86    58              8.18         62      0.1 612.69651
##    3:                89    62              8.13         64      0.1 631.74498
##    4:                93    67              8.52         67      0.1 669.95900
##    5:                97    68              7.87         68      0.1  63.53723
##   ---
## 1639:                42    67              7.13         65     33.6 454.36665
## 1640:                41     7              6.52         68     36.7 453.35116
## 1641:                40    73              6.53         71     39.8  57.34834
## 1642:                39    76              6.16         75     42.1 548.58731
## 1643:                39    78              7.10         78     43.5 547.35888
##       Population thinness..1.19.years thinness.5.9.years
##    1:   33736494                 17.2               17.3
##    2:     327582                 17.5               17.5
##    3:   31731688                 17.7               17.7
##    4:    3696958                 17.9               18.0
##    5:    2978599                 18.2               18.2
##   ---
## 1639:   12777511                  9.4                9.4
## 1640:   12633897                  9.8                9.9
## 1641:     125525                  1.2                1.3
## 1642:   12366165                  1.6                1.7
## 1643:   12222251                 11.0               11.2
##       Income.composition.of.resources Schooling Status
##    1:                           0.479      10.1      0
```

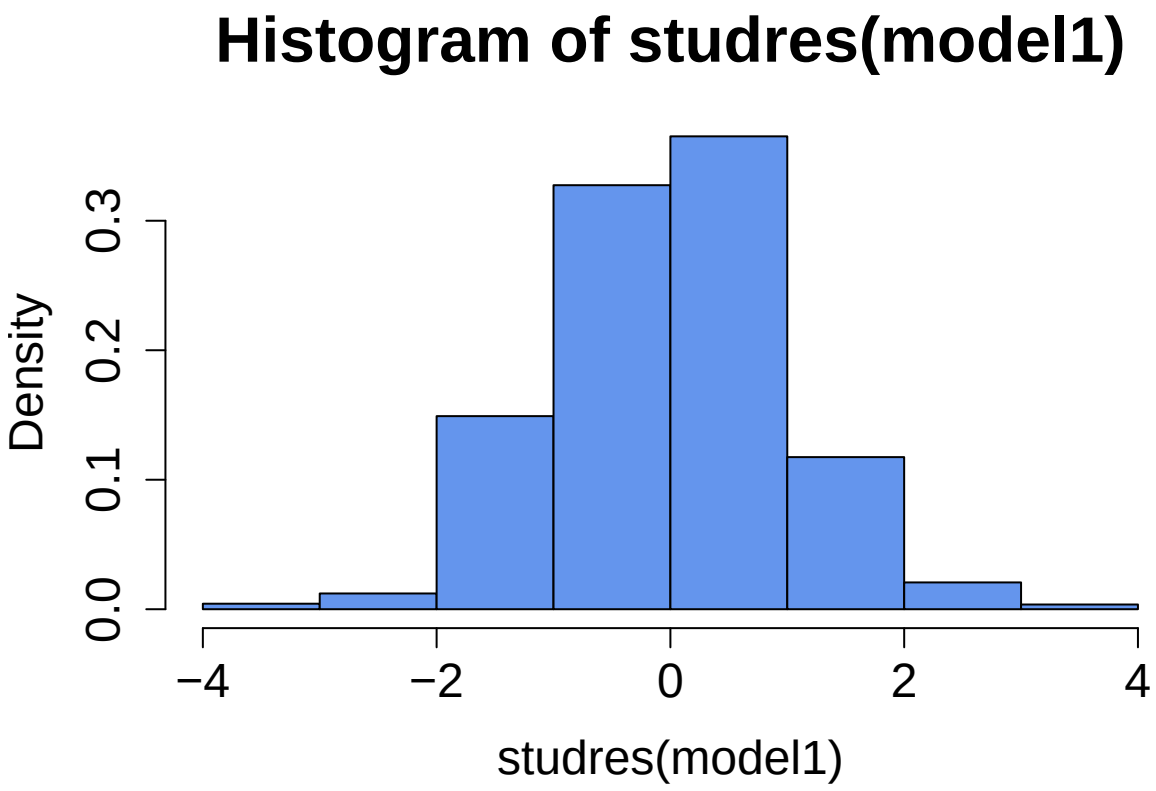
```
##    2:                          0.476      10.0      0
##    3:                          0.470       9.9      0
##    4:                          0.463       9.8      0
##    5:                          0.454       9.5      0
##   ---
## 1639:                          0.407       9.2      0
## 1640:                          0.418       9.5      0
## 1641:                          0.427      10.0      0
## 1642:                          0.427       9.8      0
## 1643:                          0.434       9.8      0
```

```
model1 <- lm(Life.expectancy ~ Status+Alcohol+percentage.expenditure+
               Hepatitis.B+Measles+BMI+under.five.deaths+Polio+Total.expenditure+
               Diphtheria+HIV.AIDS+GDP+Population+thinness..1.19.years+
               thinness.5.9.years+Income.composition.of.resources+Schooling, data1)

summary(model1)
```

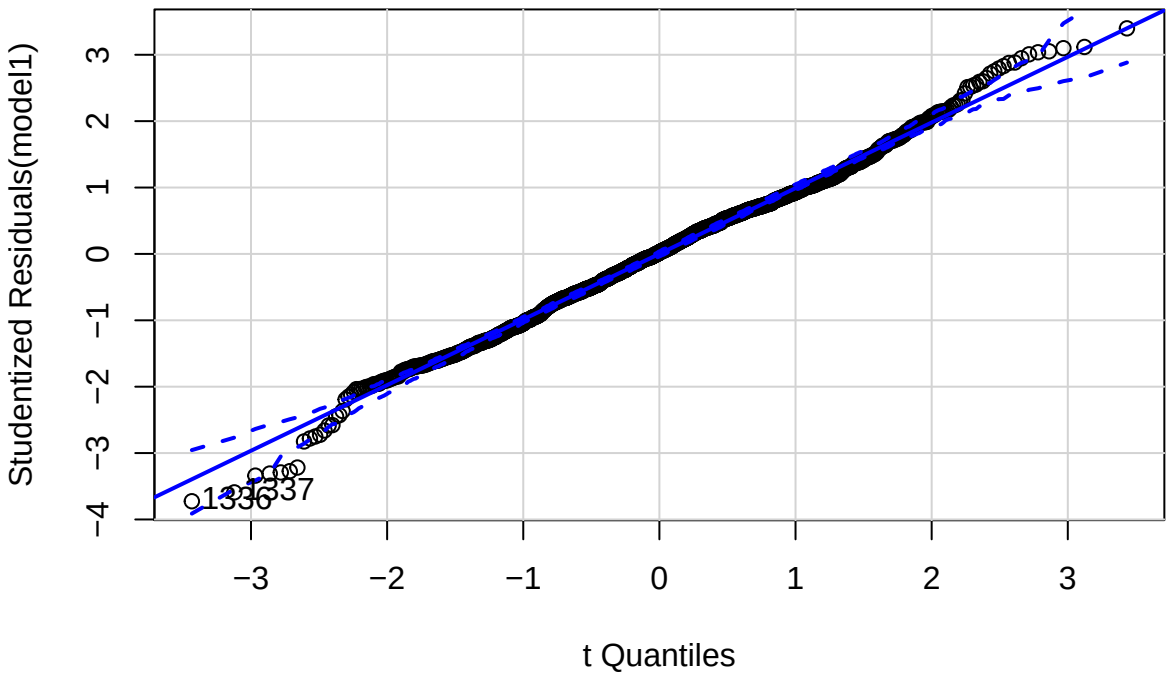
```
##
## Call:
## lm(formula = Life.expectancy ~ Status + Alcohol + percentage.expenditure +
##     Hepatitis.B + Measles + BMI + under.five.deaths + Polio +
##     Total.expenditure + Diphtheria + HIV.AIDS + GDP + Population +
##     thinness..1.19.years + thinness.5.9.years + Income.composition.of.resources +
##     Schooling, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7306  -2.5913   0.0868   2.6657  13.3105
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.660e+01  7.305e-01  63.791  < 2e-16 ***
## Status                          1.592e+00  3.736e-01   4.261 2.16e-05 ***
## Alcohol                        -2.208e-01  3.605e-02  -6.126 1.13e-09 ***
## percentage.expenditure          4.479e-04  1.999e-04   2.241 0.025181 *
## Hepatitis.B                    -1.019e-02  4.942e-03  -2.062 0.039404 *
## Measles                         4.334e-05  1.280e-05   3.387 0.000725 ***
## BMI                             4.600e-02  6.634e-03   6.934 5.88e-12 ***
## under.five.deaths              -3.061e-03  1.030e-03  -2.971 0.003014 **
## Polio                           1.465e-02  5.708e-03   2.567 0.010358 *
## Total.expenditure               1.179e-01  4.549e-02   2.593 0.009606 **
## Diphtheria                      2.164e-02  6.553e-03   3.302 0.000981 ***
## HIV.AIDS                       -5.996e-01  1.743e-02 -34.406  < 2e-16 ***
## GDP                             4.126e-06  3.150e-05   0.131 0.895789
## Population                      3.103e-09  2.273e-09   1.365 0.172325
## thinness..1.19.years            7.837e-03  5.882e-02   0.133 0.894016
## thinness.5.9.years             -5.353e-02  5.790e-02  -0.925 0.355280
## Income.composition.of.resources  1.222e+01  9.140e-01  13.373  < 2e-16 ***
## Schooling                       1.003e+00  6.554e-02  15.308  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.984 on 1625 degrees of freedom
```

```
## Multiple R-squared:  0.7952, Adjusted R-squared:  0.793
## F-statistic: 371.1 on 17 and 1625 DF,  p-value: < 2.2e-16
```

```
hist(studres(model1),
     breaks=10, freq=F, col="cornflowerblue", cex.axis=1.5, cex.lab=1.5, cex.main=2)
```

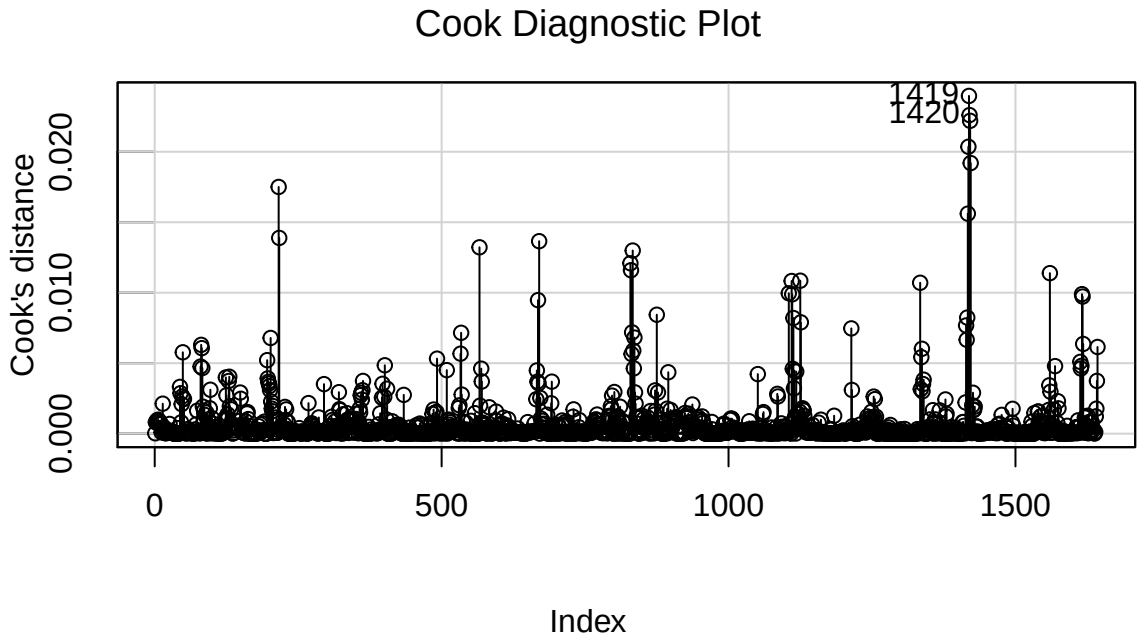



```
qqPlot(model1)
```

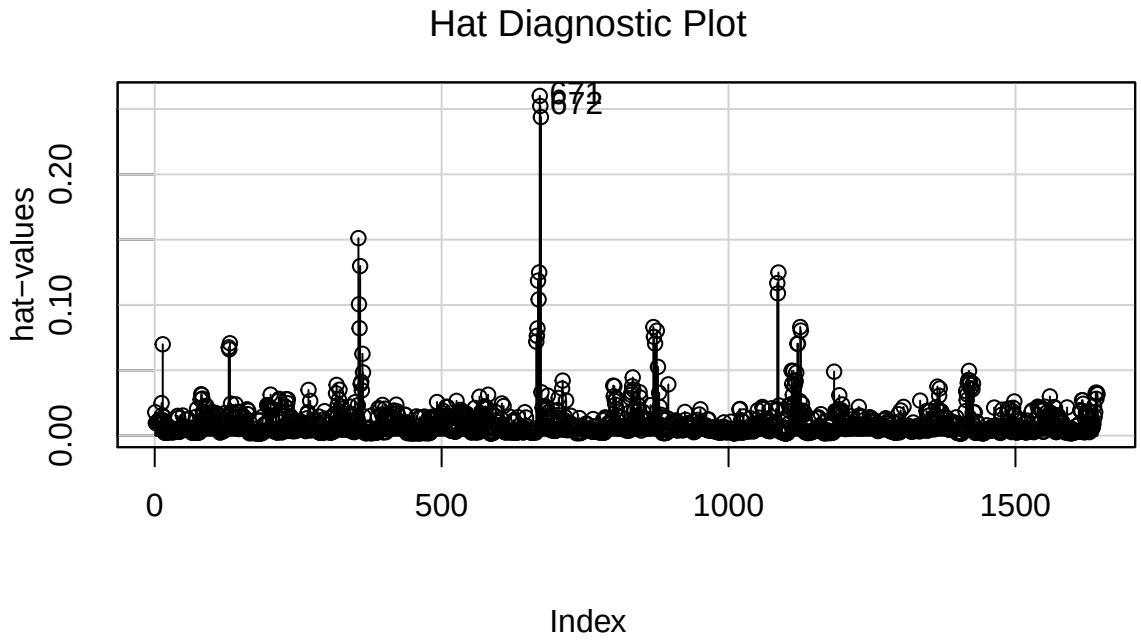



```
## [1] 1336 1337
```

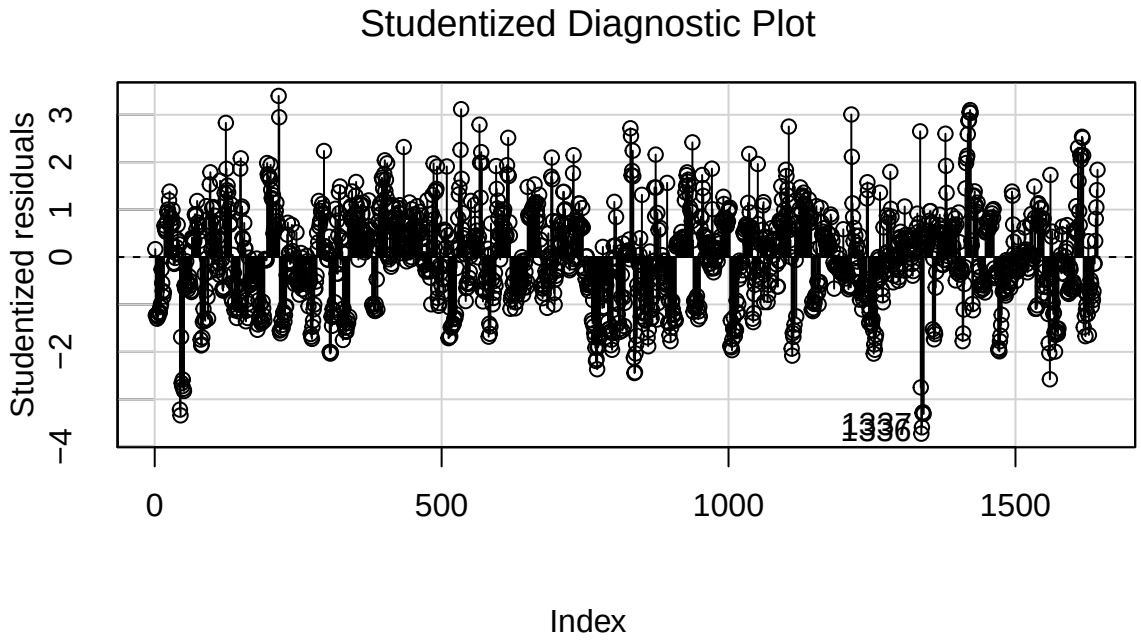
```
influenceIndexPlot(model1, vars=c('Cook'), data1, main = "Cook Diagnostic Plot")
```



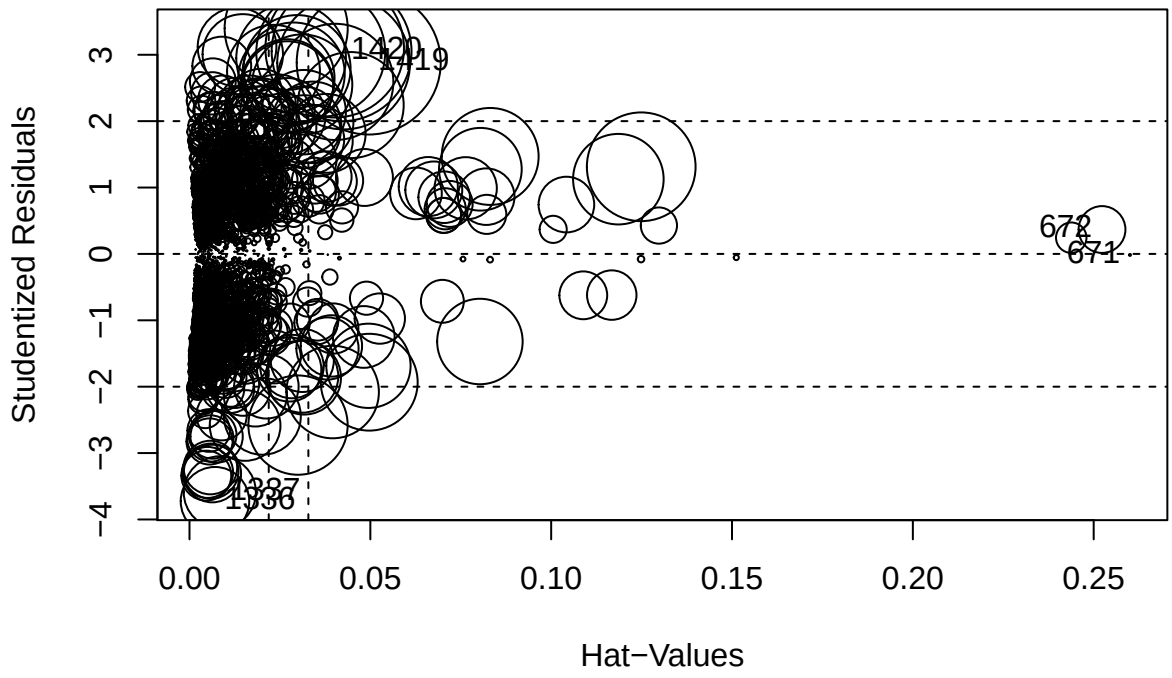

```
influenceIndexPlot(model1, vars=c('hat'), data1, main = "Hat Diagnostic Plot")
```




```
influenceIndexPlot(model1, vars=c('Studentized'), data1, main = "Studentized Diagnostic Plot")
```

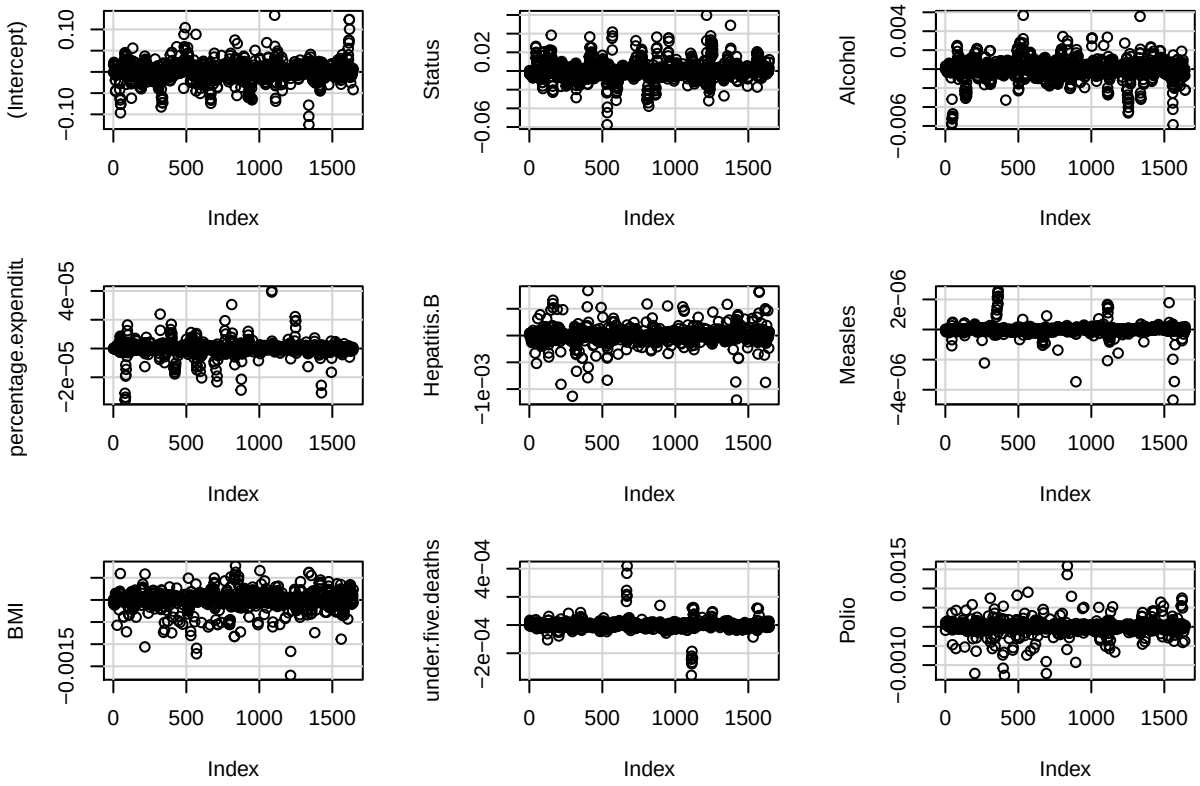


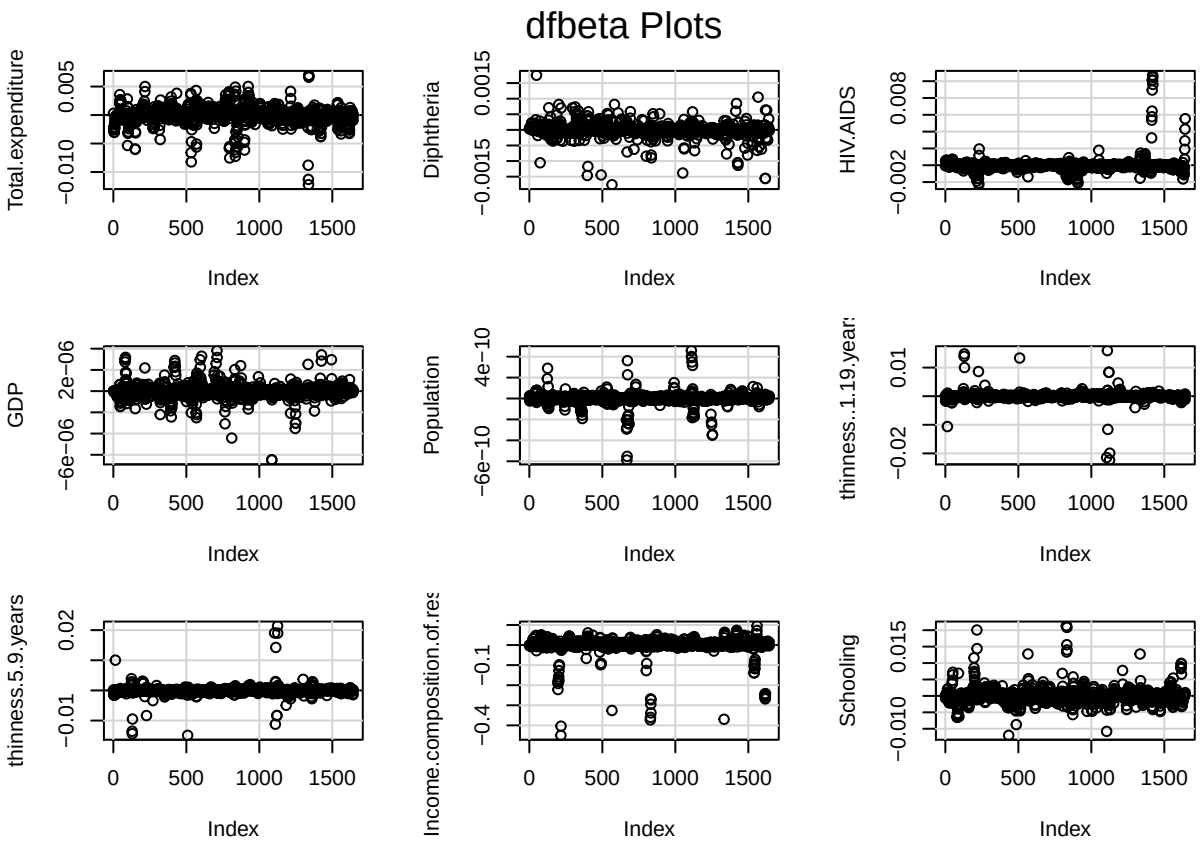

```
influencePlot(model1)
```



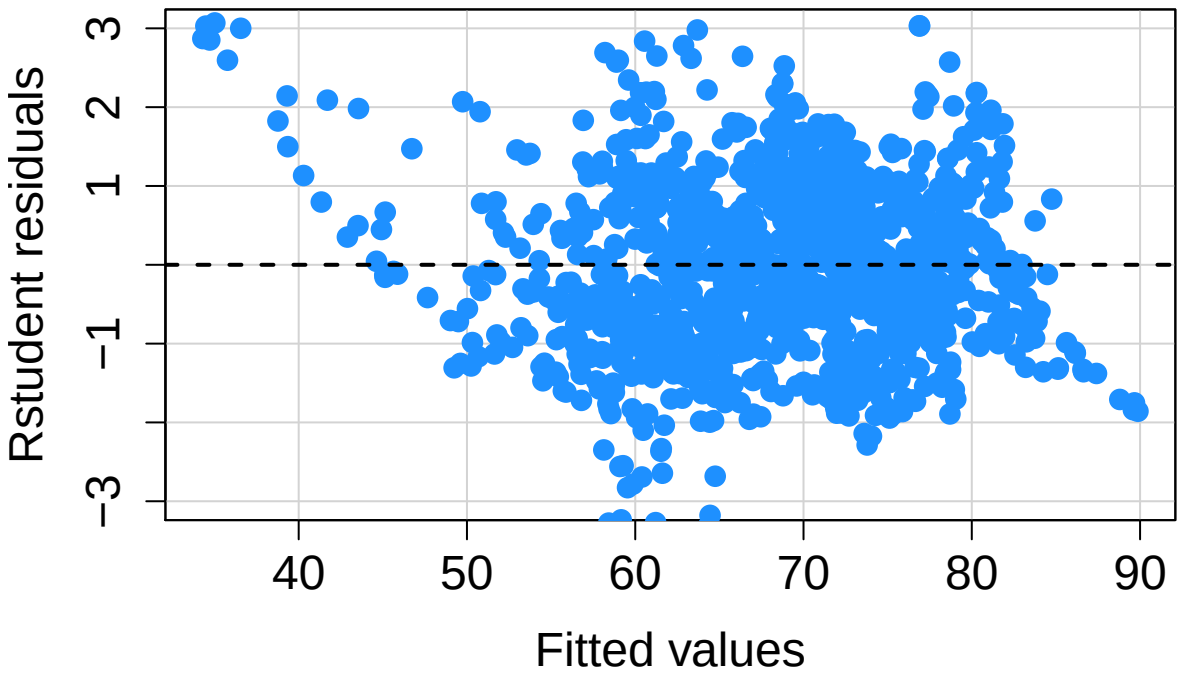

```
##          StudRes         Hat        CookD
## 671  -0.01531099 0.260029582 4.579412e-06
## 672   0.36747834 0.252306124 2.532947e-03
## 1336 -3.72554658 0.007043799 5.426949e-03
## 1337 -3.59173451 0.008393577 6.022472e-03
## 1419  2.88284725 0.049549663 2.396252e-02
## 1420  3.05214672 0.042050095 2.260194e-02
```

```
dfbetaPlots(model1, intercept=T)
```


(Intercept)
−0.10
0.10
Status
−0.06
0.02
Alcohol
−0.006
0.004
percentage.expenditu
−2e−05
4e−05
Hepatitis.B
−1e−03
Measles
−4e−06
2e−06
BMI
−0.0015
under.five.deaths
−2e−04
4e−04
Polio
−0.0010
0.0015
0
500
1000
1500
Index

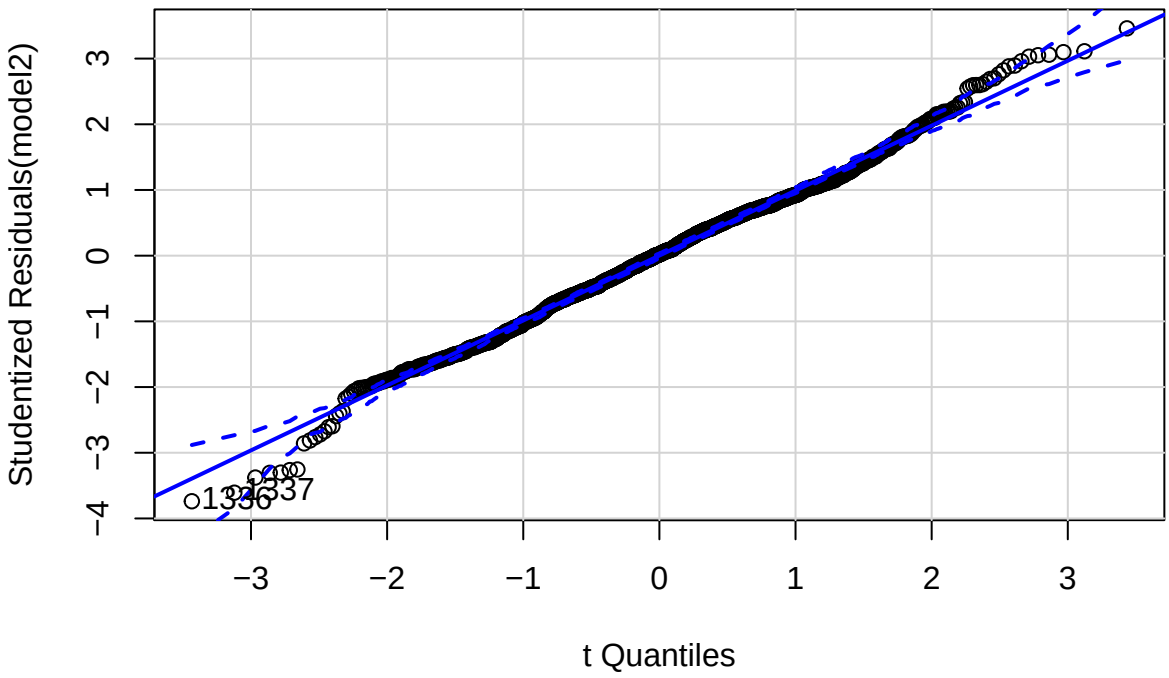



```
summary(residualPlot(model2, type="rstudent", quadratic=F,
             col = "dodgerblue", pch=16, cex=1.5,
             cex.axis=1.5, cex.lab=1.5, ylim=c(-3,3)))
```



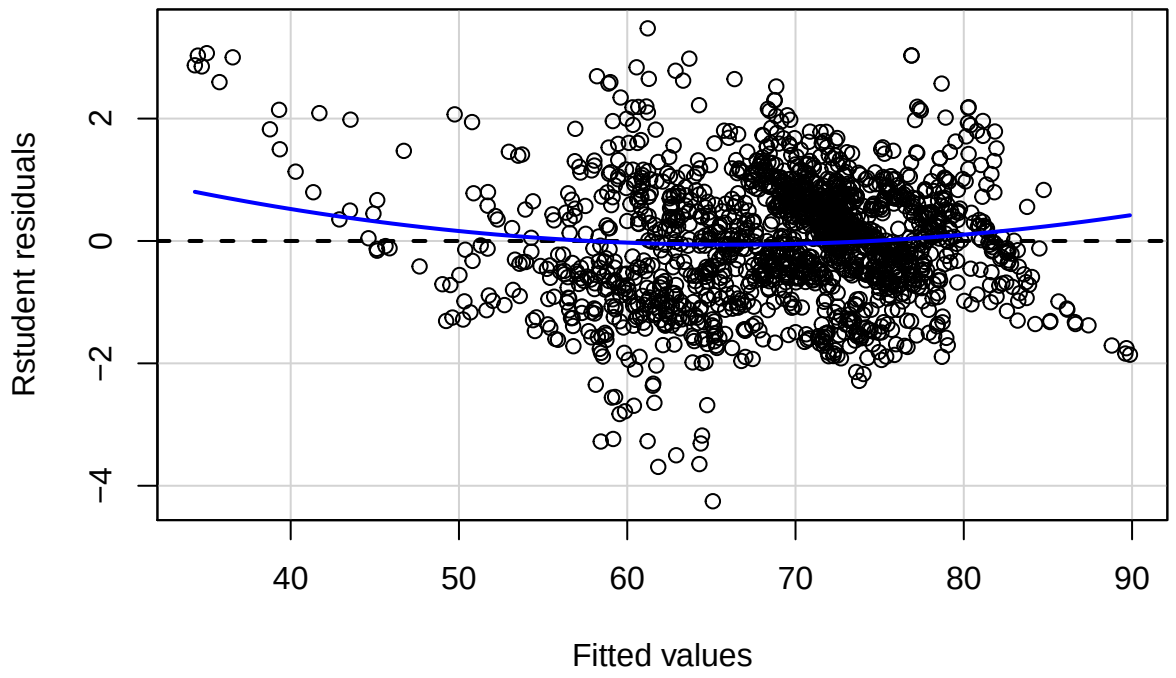

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.818   3.637   3.637   5.455   7.274
```

```
qqPlot(model2)
```

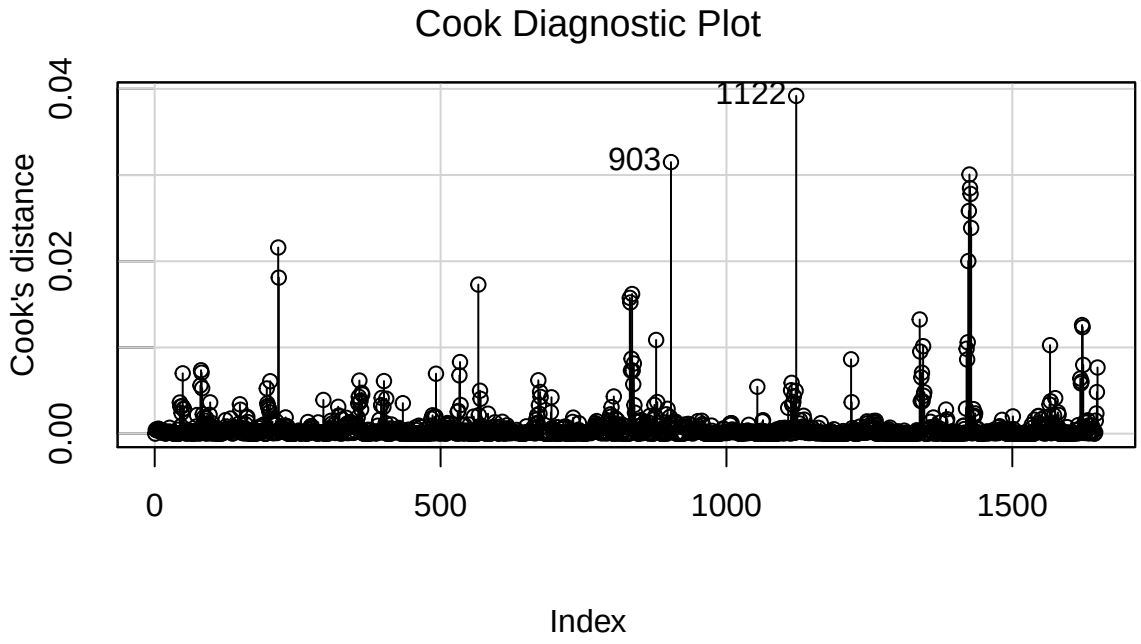



```
## [1] 1336 1337
```

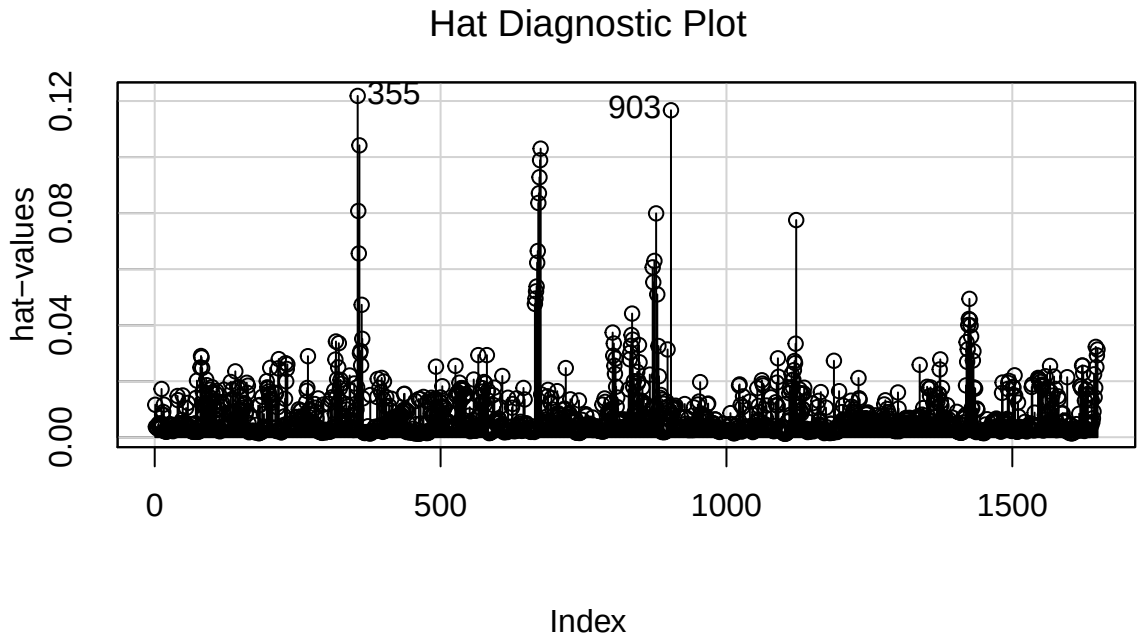
```
residualPlot(model2, variable='fitted', type='rstudent')
```



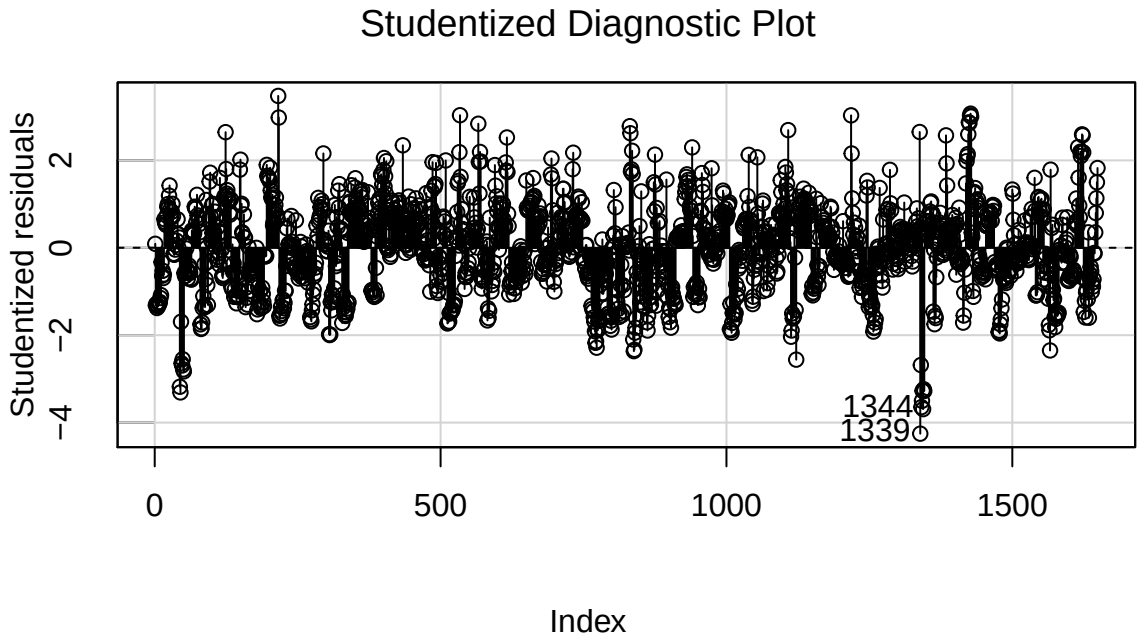

```
influenceIndexPlot(model2, vars=c('Cook'), data1, main = "Cook Diagnostic Plot")
```



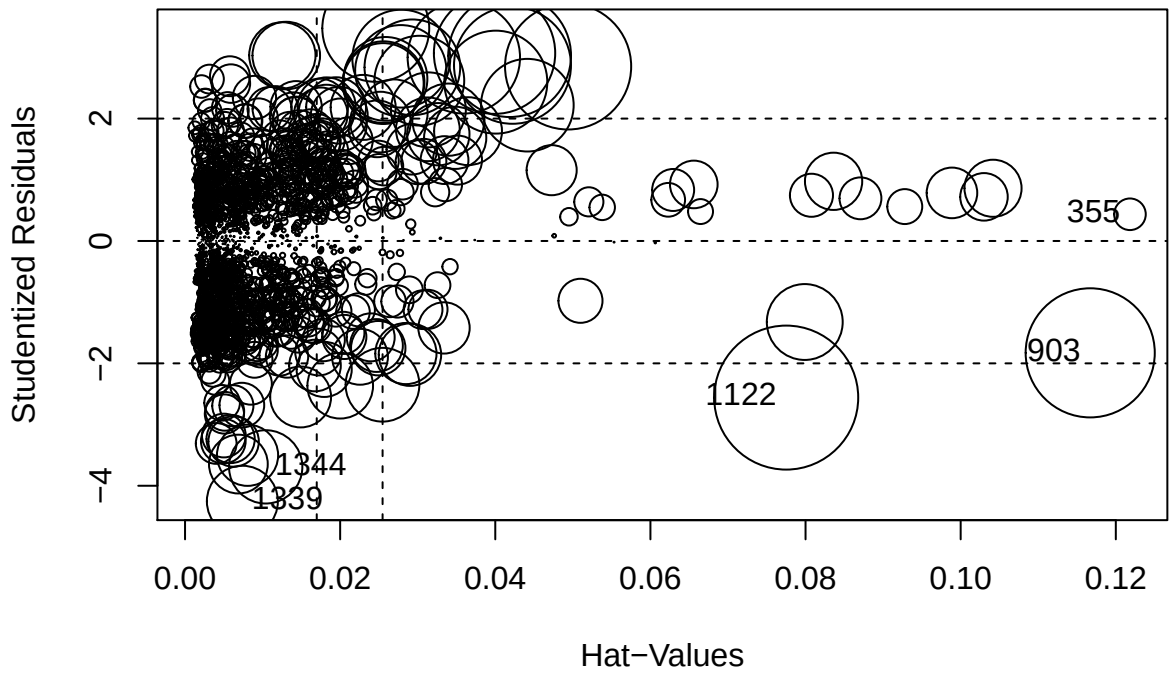

```
influenceIndexPlot(model2, vars=c('hat'), data1, main = "Hat Diagnostic Plot")
```




```
influenceIndexPlot(model2, vars=c('Studentized'), data1, main = "Studentized Diagnostic Plot")
```

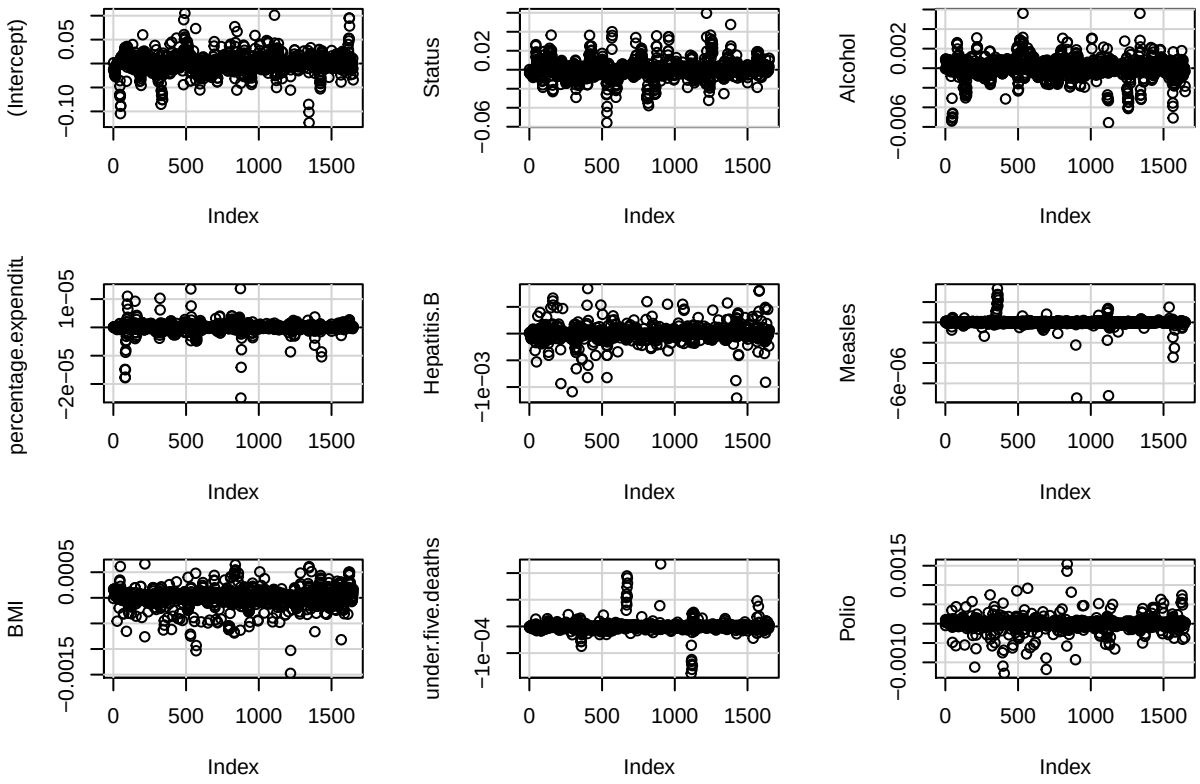


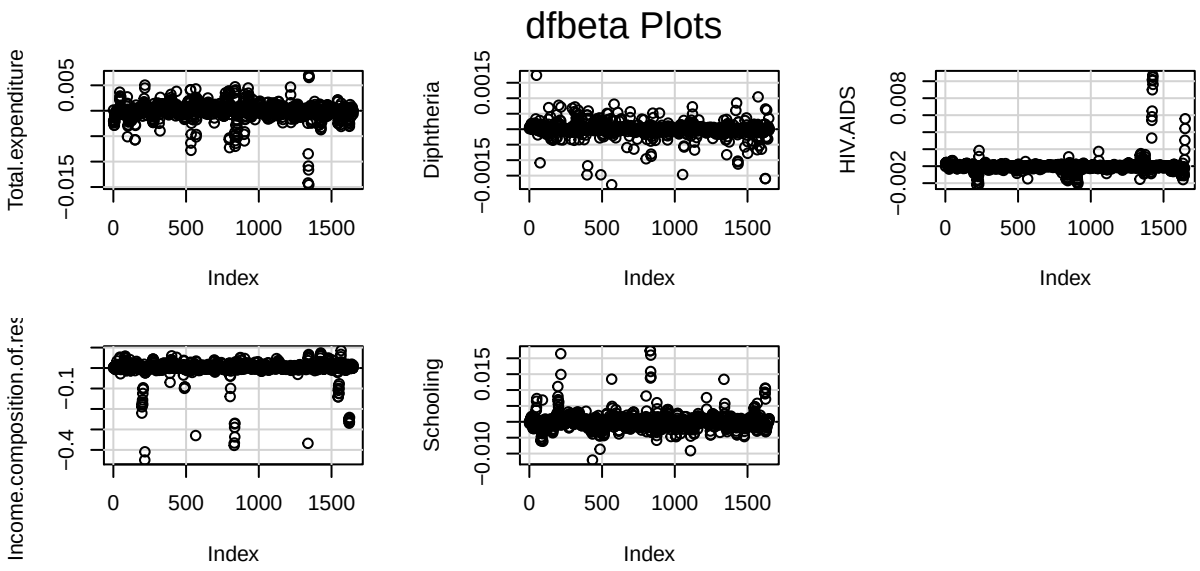

```
influencePlot(model2)
```



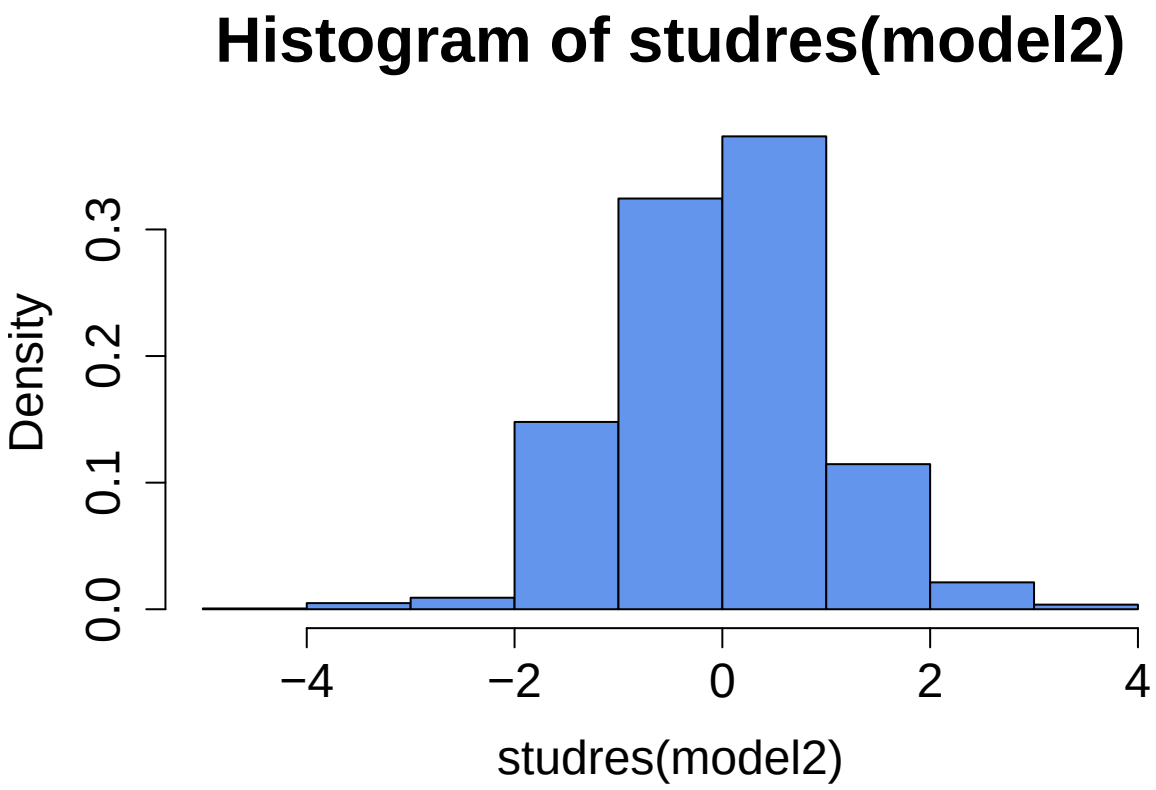

```
##          StudRes         Hat       CookD
## 355    0.4361693 0.121824292 0.001886035
## 903   -1.8282502 0.116699973 0.031498071
## 1122 -2.5593299 0.077507494 0.039177219
## 1339 -4.2531218 0.007391309 0.009521716
## 1344 -3.6914116 0.010399170 0.010149730
```

```
dfbetaPlots(model2, intercept=T)
```


(Intercept)
−0.10
0.05
Status
−0.06
0.02
Alcohol
−0.006
0.002
percentage.expenditu
−2e−05
1e−05
Hepatitis.B
−1e−03
Measles
−6e−06
BMI
−0.0015
0.0005
under.five.deaths
−1e−04
Polio
−0.0010
0.0015
0
500
1000
1500
Index




```
hist(studres(model2),
     breaks=10, freq=F, col="cornflowerblue", cex.axis=1.5, cex.lab=1.5, cex.main=2)
```



Problem indices: 903, 355, 1122, 1344, 1339. We took out 903, 1122, 1344, and 1339, so now we only should remove 355

```
data1 <- data1[-c(355),]
data1
```

```
##       Life.expectancy Alcohol percentage.expenditure Hepatitis.B Measles  BMI
##    1:            65.0    0.01              71.279624          65     1154 19.1
##    2:            59.9    0.01              73.523582          62      492 18.6
##    3:            59.9    0.01              73.219243          64      430 18.1
##    4:            59.5    0.01              78.184215          67     2787 17.6
##    5:            59.2    0.01               7.097109          68     3013 17.2
##   ---
## 1638:            44.3    4.36               0.000000          68       31 27.1
## 1639:            44.5    4.06               0.000000           7      998 26.7
## 1640:            44.8    4.43               0.000000          73      304 26.3
## 1641:            45.3    1.72               0.000000          76      529 25.9
## 1642:            46.0    1.68               0.000000          79     1483 25.5
##       under.five.deaths Polio Total.expenditure Diphtheria HIV.AIDS       GDP
##    1:                83     6              8.16         65      0.1 584.25921
##    2:                86    58              8.18         62      0.1 612.69651
##    3:                89    62              8.13         64      0.1 631.74498
##    4:                93    67              8.52         67      0.1 669.95900
##    5:                97    68              7.87         68      0.1  63.53723
##   ---
## 1638:                42    67              7.13         65     33.6 454.36665
```

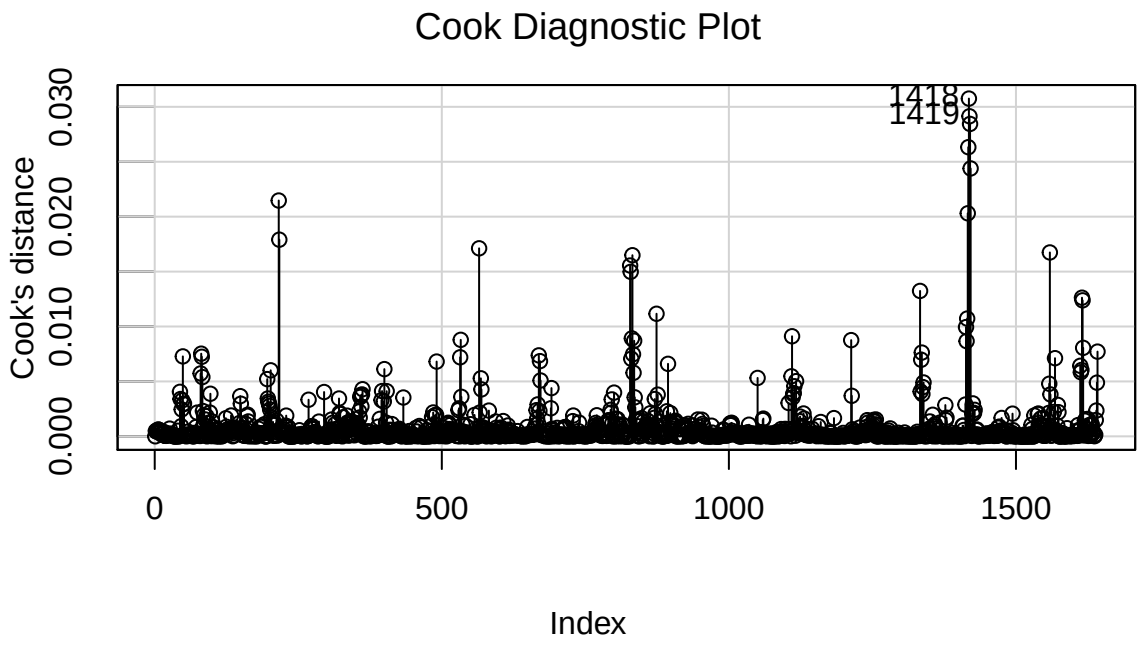
```
## 1639:                41     7              6.52         68     36.7 453.35116
## 1640:                40    73              6.53         71     39.8  57.34834
## 1641:                39    76              6.16         75     42.1 548.58731
## 1642:                39    78              7.10         78     43.5 547.35888
##       Population thinness..1.19.years thinness.5.9.years
##    1:   33736494                 17.2               17.3
##    2:     327582                 17.5               17.5
##    3:   31731688                 17.7               17.7
##    4:    3696958                 17.9               18.0
##    5:    2978599                 18.2               18.2
##   ---
## 1638:   12777511                  9.4                9.4
## 1639:   12633897                  9.8                9.9
## 1640:     125525                  1.2                1.3
## 1641:   12366165                  1.6                1.7
## 1642:   12222251                 11.0               11.2
##       Income.composition.of.resources Schooling Status
##    1:                           0.479      10.1      0
##    2:                           0.476      10.0      0
##    3:                           0.470       9.9      0
##    4:                           0.463       9.8      0
##    5:                           0.454       9.5      0
##   ---
## 1638:                           0.407       9.2      0
## 1639:                           0.418       9.5      0
## 1640:                           0.427      10.0      0
## 1641:                           0.427       9.8      0
## 1642:                           0.434       9.8      0
```

```
model2 <- lm(Life.expectancy ~ Status+Alcohol+percentage.expenditure+
               Hepatitis.B+Measles+BMI+under.five.deaths+Polio+Total.expenditure+
               Diphtheria+HIV.AIDS+Income.composition.of.resources+Schooling, data1)

summary(model2)
```

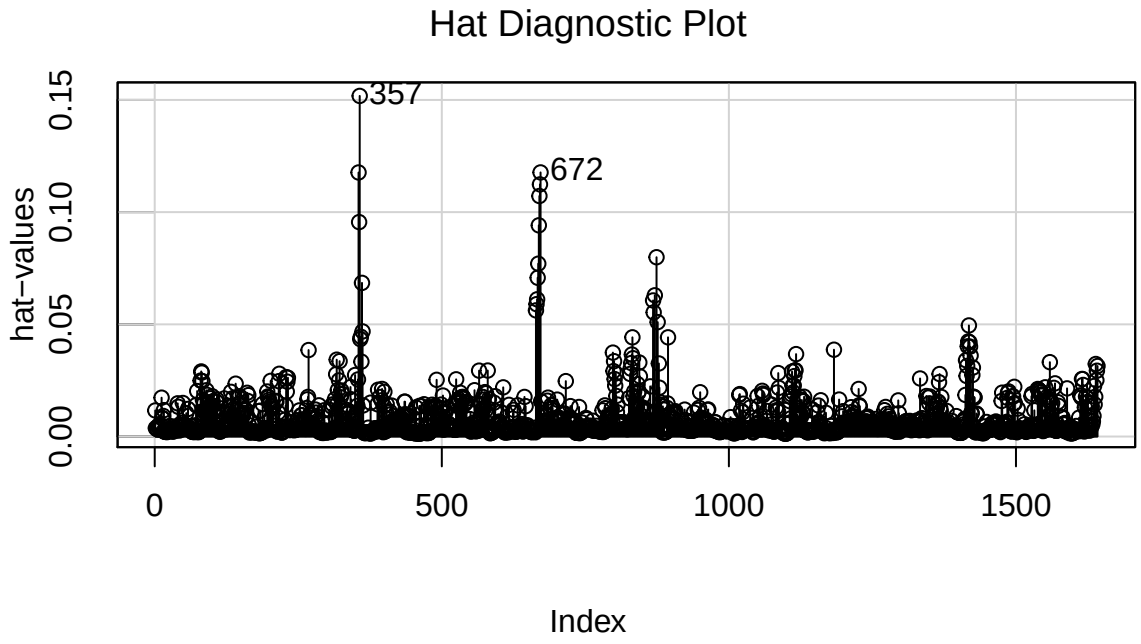
```
##
## Call:
## lm(formula = Life.expectancy ~ Status + Alcohol + percentage.expenditure +
##     Hepatitis.B + Measles + BMI + under.five.deaths + Polio +
##     Total.expenditure + Diphtheria + HIV.AIDS + Income.composition.of.resources +
##     Schooling, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7821  -2.5371   0.0822   2.7177  13.5691
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.605e+01  6.567e-01  70.122  < 2e-16 ***
## Status                          1.593e+00  3.729e-01   4.271 2.06e-05 ***
## Alcohol                        -2.145e-01  3.554e-02  -6.035 1.97e-09 ***
## percentage.expenditure          4.762e-04  6.598e-05   7.218 8.08e-13 ***
## Hepatitis.B                    -1.075e-02  4.935e-03  -2.178 0.029567 *
## Measles                         4.513e-05  1.370e-05   3.295 0.001006 **
```

```
## BMI                              5.001e-02  6.183e-03   8.089 1.16e-15 ***
## under.five.deaths               -2.806e-03  7.905e-04  -3.550 0.000396 ***
## Polio                            1.465e-02  5.695e-03   2.572 0.010201 *
## Total.expenditure                1.222e-01  4.531e-02   2.697 0.007062 **
## Diphtheria                       2.209e-02  6.541e-03   3.377 0.000750 ***
## HIV.AIDS                        -6.028e-01  1.734e-02 -34.763  < 2e-16 ***
## Income.composition.of.resources  1.233e+01  9.095e-01  13.563  < 2e-16 ***
## Schooling                        1.011e+00  6.508e-02  15.537  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.985 on 1628 degrees of freedom
## Multiple R-squared:  0.7945, Adjusted R-squared:  0.7929
## F-statistic: 484.3 on 13 and 1628 DF,  p-value: < 2.2e-16
```

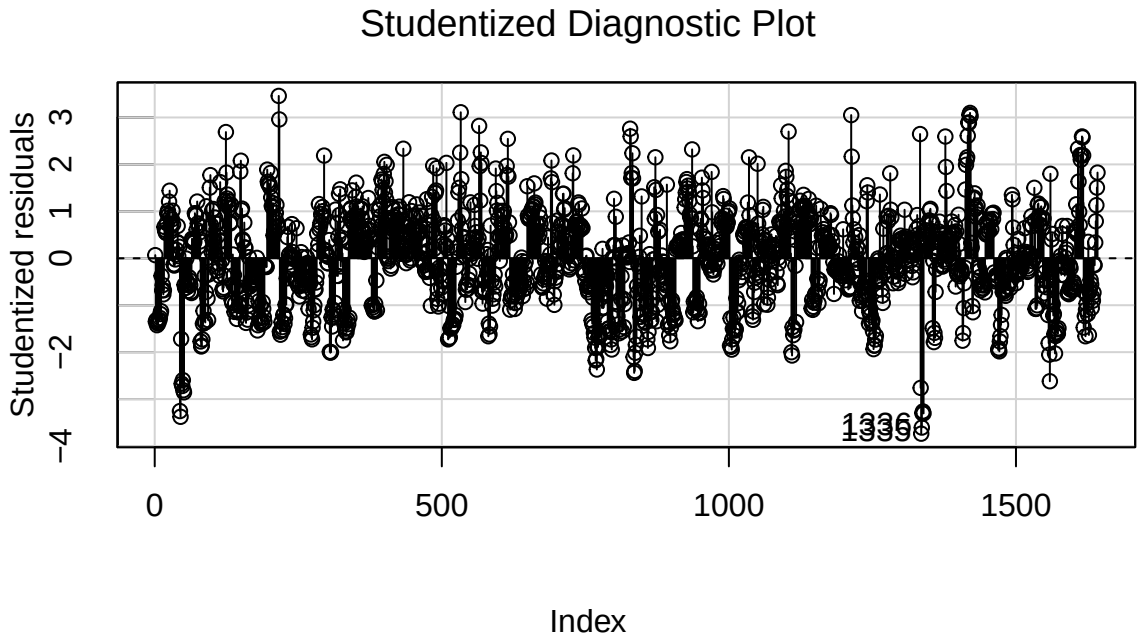
```
influenceIndexPlot(model2, vars=c('Cook'), main = "Cook Diagnostic Plot")
```



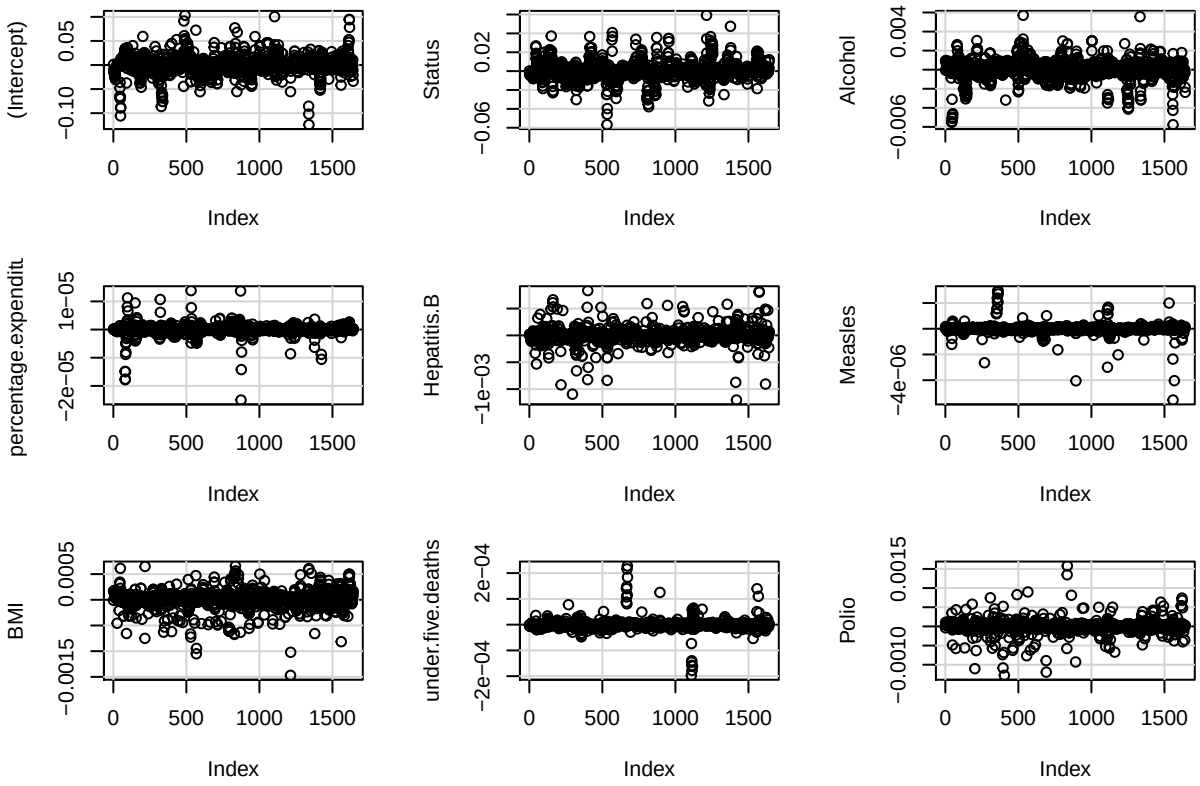

```
influenceIndexPlot(model2, vars=c('hat'), main = "Hat Diagnostic Plot")
```

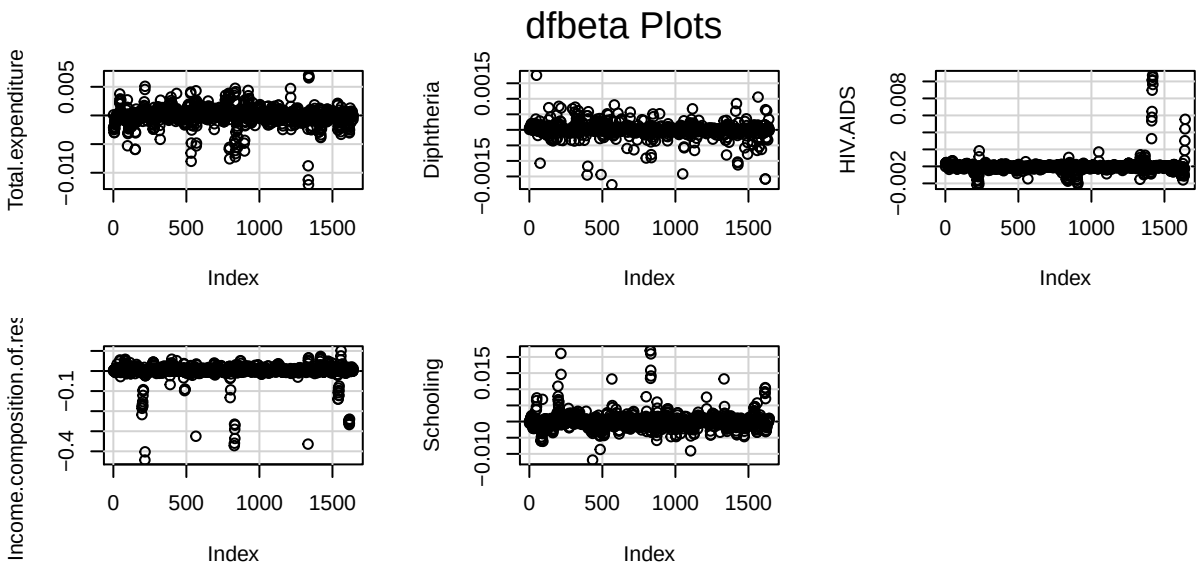



```
influenceIndexPlot(model2, vars=c('Studentized'), main = "Studentized Diagnostic Plot")
```

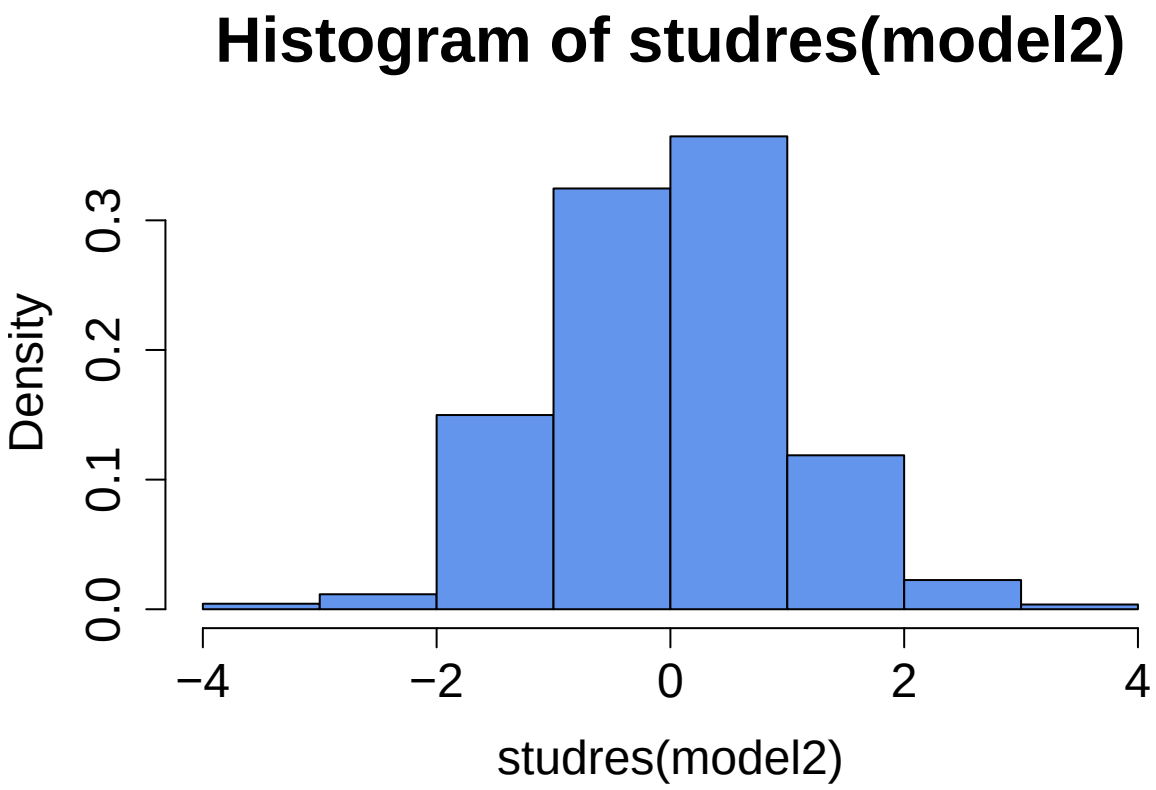



```
dfbetaPlots(model2, intercept=T)
```

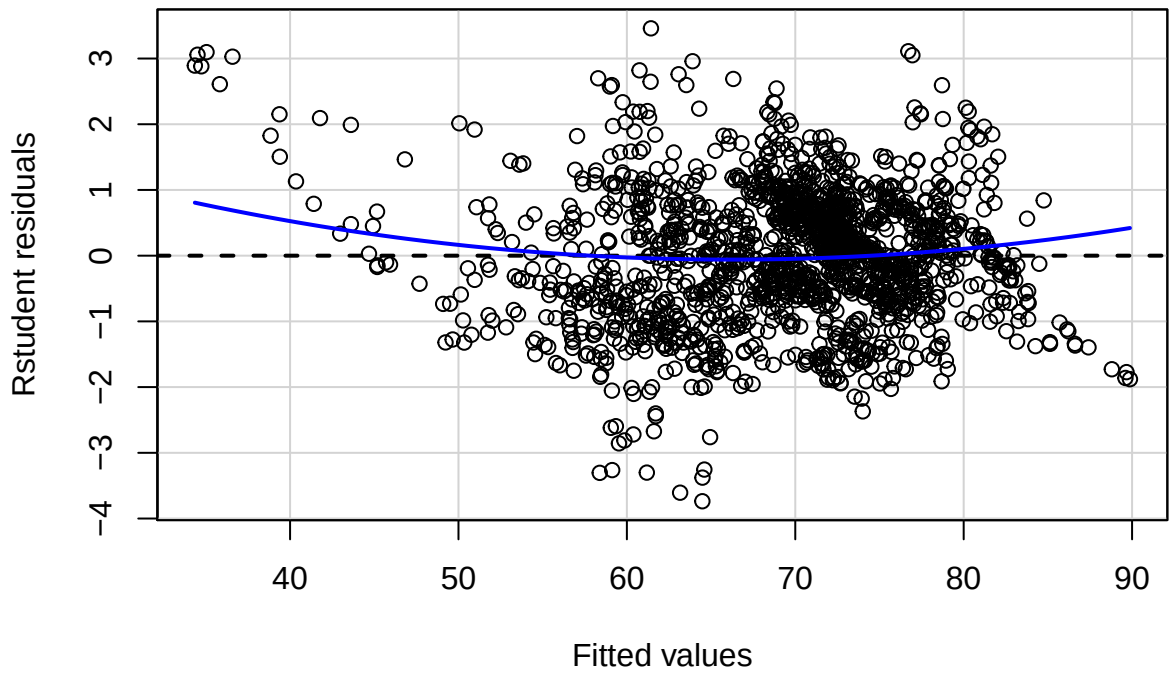

(Intercept)
−0.10
0.05
Status
−0.06
0.02
Alcohol
−0.006
0.004
percentage.expenditu
−2e−05
1e−05
Hepatitis.B
−1e−03
Measles
−4e−06
BMI
−0.0015
0.0005
under.five.deaths
−2e−04
2e−04
Polio
−0.0010
0.0015
0
500
1000
1500
Index




```
hist(studres(model2),
     breaks=10, freq=F, col="cornflowerblue", cex.axis=1.5, cex.lab=1.5, cex.main=2)
```




```
residualPlot(model2, variable='fitted', type='rstudent')
```




```
vif(model2)
```

```
##                          Status                         Alcohol
##                        1.806138                        2.121342
##          percentage.expenditure                      Hepatitis.B
##                        1.396866                        1.641887
##                         Measles                             BMI
##                        1.392989                        1.541872
##                under.five.deaths                           Polio
##                        1.500720                        1.691693
##               Total.expenditure                       Diphtheria
##                        1.111328                        2.059209
##                         HIV.AIDS Income.composition.of.resources
##                        1.132194                        2.866982
##                       Schooling
##                        3.423224
```