Obesity

EDA (Exploratory Data Analysis)

```r
df <- read.csv("C:/Users/hecto/OneDrive/Documents/School/Semesters/Spring 2025/Project/Data/

head(df)
```

```
      Country Year     Status Life.expectancy Adult.Mortality infant.deaths
1 Afghanistan 2015 Developing            65.0             263            62
2 Afghanistan 2014 Developing            59.9             271            64
3 Afghanistan 2013 Developing            59.9             268            66
4 Afghanistan 2012 Developing            59.5             272            69
5 Afghanistan 2011 Developing            59.2             275            71
6 Afghanistan 2010 Developing            58.8             279            74
  Alcohol percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths
1    0.01              71.279624          65    1154 19.1                83
2    0.01              73.523582          62     492 18.6                86
3    0.01              73.219243          64     430 18.1                89
4    0.01              78.184215          67    2787 17.6                93
5    0.01               7.097109          68    3013 17.2                97
6    0.01              79.679367          66    1989 16.7               102
  Polio Total.expenditure Diphtheria HIV.AIDS       GDP Population
1     6              8.16         65      0.1 584.25921   33736494
2    58              8.18         62      0.1 612.69651     327582
3    62              8.13         64      0.1 631.74498   31731688
4    67              8.52         67      0.1 669.95900    3696958
5    68              7.87         68      0.1  63.53723    2978599
6    66              9.20         66      0.1 553.32894    2883167
  thinness..1.19.years thinness.5.9.years Income.composition.of.resources
1                 17.2               17.3                           0.479
2                 17.5               17.5                           0.476
```

```
3                 17.7                17.7                            0.470
4                 17.9                18.0                            0.463
5                 18.2                18.2                            0.454
6                 18.4                18.4                            0.448
  Schooling
1      10.1
2      10.0
3       9.9
4       9.8
5       9.5
6       9.2
```

```r
str(df)
```

```
'data.frame':   2938 obs. of  22 variables:
 $ Country                        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanis
 $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 .
 $ Status                         : chr  "Developing" "Developing" "Developing" "Developing"
 $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
 $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
 $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
 $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 .
 $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
 $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
 $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
 $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 .
 $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 116 ...
 $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
 $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ..
 $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
 $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
 $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
 $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
 $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
 $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
 $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415
 $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```

```r
df$Status <- as.factor(df$Status)
```

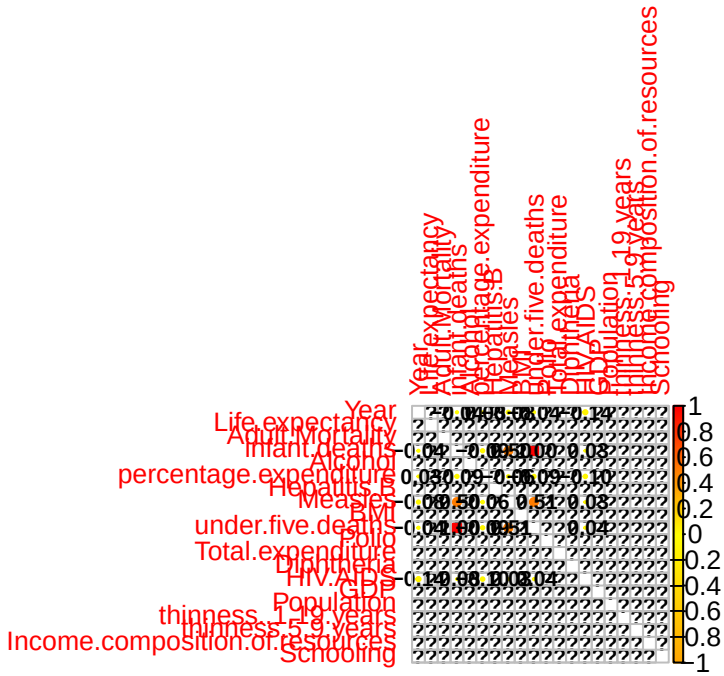
```r
dim(df)
```

```
[1] 2938   22
```

Multicolinearity

```r
# Create a new dataframe excluding column 9 (response variable)


# Optional: Double check that column 9 was removed
# names(grad2)

# Recreate the cool color palette
cool_colors <- colorRampPalette(c("orange", "yellow", "red"))(200)

# Run corrplot only on numeric columns (predictors only)
corrplot(cor(df[, sapply(df, is.numeric)]),
         method = "circle",
         col = cool_colors,
         tl.cex = 0.8,
         number.cex = 0.7,
         addCoef.col = "black",
         diag = FALSE)
```

```
Warning in ind1:ind2: numerical expression has 2 elements: only the first used
```




```r
df <- df[ , -1]
```

```r
library(car)
fit <- lm(Life.expectancy ~  ., data = df)  # drop Serial.No.
summary(fit)
```

```
Call:
lm(formula = Life.expectancy ~ ., data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.7681  -2.1427   0.0273   2.1776  12.4242 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      3.090e+02  4.621e+01   6.687 3.12e-11 ***
Year                            -1.272e-01  2.308e-02  -5.510 4.18e-08 ***
StatusDeveloping                -8.865e-01  3.353e-01  -2.644  0.00827 ** 
Adult.Mortality                 -1.621e-02  9.441e-04 -17.171  < 2e-16 ***
infant.deaths                    8.873e-02  1.059e-02   8.376  < 2e-16 ***
Alcohol                         -1.313e-01  3.366e-02  -3.901 9.95e-05 ***
```

```
percentage.expenditure           3.026e-04  1.789e-04   1.691  0.09096 .
Hepatitis.B                     -3.258e-03  4.449e-03  -0.732  0.46413
Measles                         -1.033e-05  1.070e-05  -0.966  0.33439
BMI                              3.183e-02  5.955e-03   5.345 1.03e-07 ***
under.five.deaths               -6.662e-02  7.673e-03  -8.682  < 2e-16 ***
Polio                            5.797e-03  5.121e-03   1.132  0.25776
Total.expenditure                9.220e-02  4.042e-02   2.281  0.02268 *
Diphtheria                       1.403e-02  5.877e-03   2.387  0.01712 *
HIV.AIDS                        -4.481e-01  1.780e-02 -25.174  < 2e-16 ***
GDP                              2.451e-05  2.826e-05   0.867  0.38594
Population                      -6.085e-10  1.733e-09  -0.351  0.72558
thinness..1.19.years            -5.815e-03  5.254e-02  -0.111  0.91189
thinness.5.9.years              -5.010e-02  5.185e-02  -0.966  0.33412
Income.composition.of.resources  1.045e+01  8.327e-01  12.549  < 2e-16 ***
Schooling                        8.949e-01  5.910e-02  15.142  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.556 on 1628 degrees of freedom
  (1289 observations deleted due to missingness)
Multiple R-squared:  0.8386,    Adjusted R-squared:  0.8366
F-statistic: 422.9 on 20 and 1628 DF,  p-value: < 2.2e-16
```

```r
sum(is.na(df))
```

```
[1] 2563
```

```r
df2 <- na.omit(df)
df2 <- df2[ , -1]
```

```r
sum(is.na(df2))
```

```
[1] 0
```

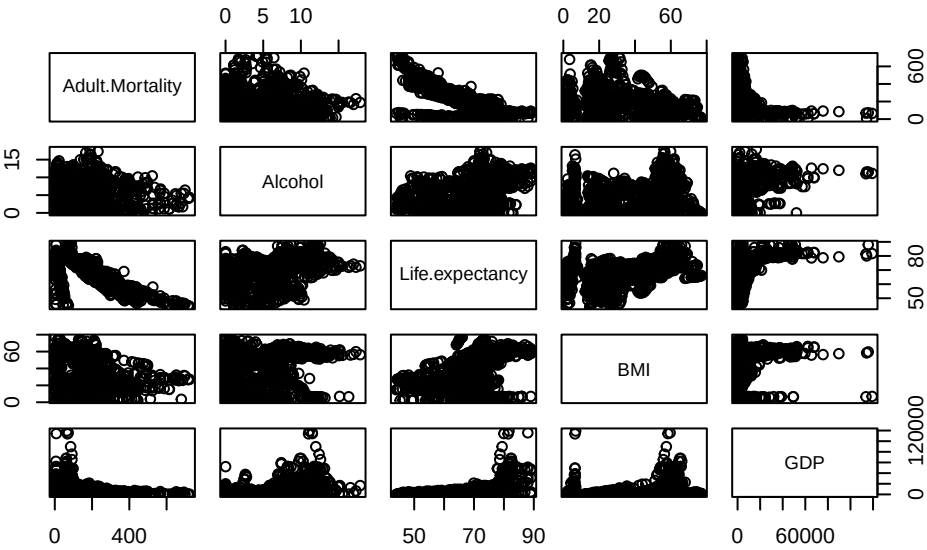
```r
dim(df2)
```

```
[1] 1649   20
```

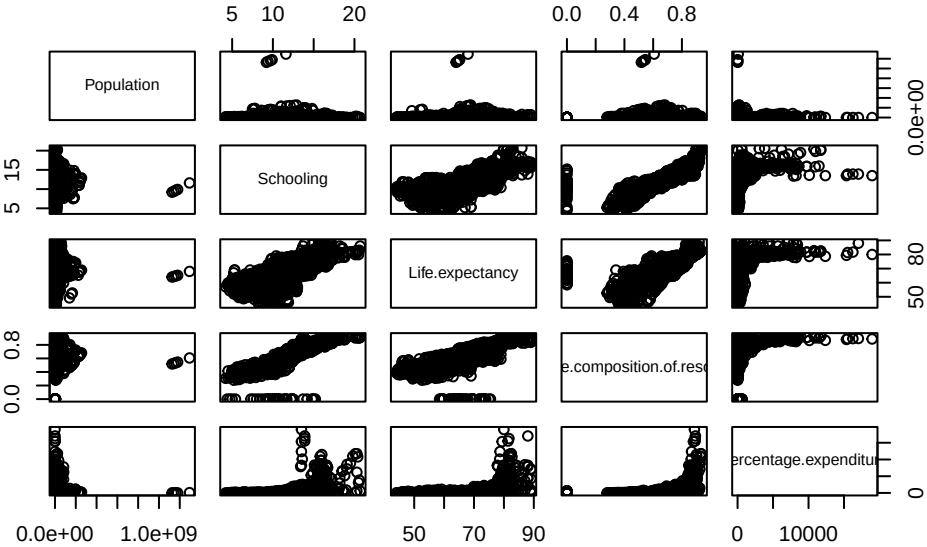
```r
str(df2)
```

```
'data.frame':   1649 obs. of  20 variables:
 $ Status                         : Factor w/ 2 levels "Developed","Developing": 2 2 2 2 2 2
 $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
 $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
 $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
 $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 .
 $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
 $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
 $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
 $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 .
 $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 116 ...
 $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
 $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ..
 $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
 $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
 $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
 $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
 $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
 $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
 $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415
 $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```

```r
pairs(subset(df2, select = c(Adult.Mortality, Alcohol ,Life.expectancy, BMI, GDP)))
```




```
pairs(subset(df2, select = c(Population, Schooling,Life.expectancy, Income.composition.of.res
```

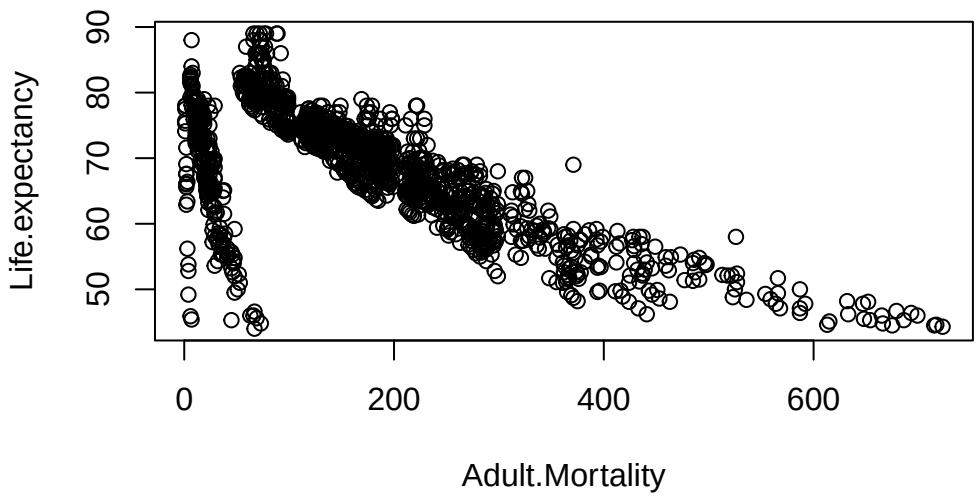


```r
attach(df2)
```

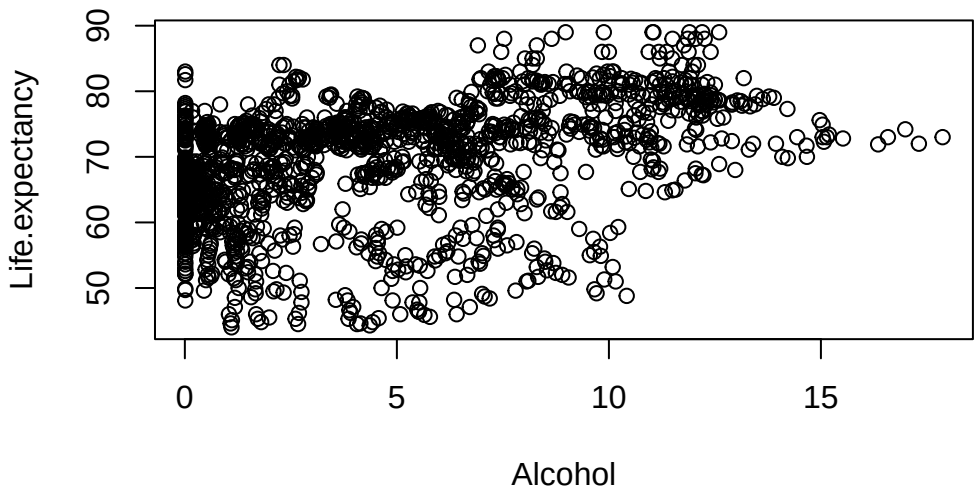
```r
summary(df2$Life.expectancy)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   44.0    64.4    71.7    69.3    75.0    89.0 
```

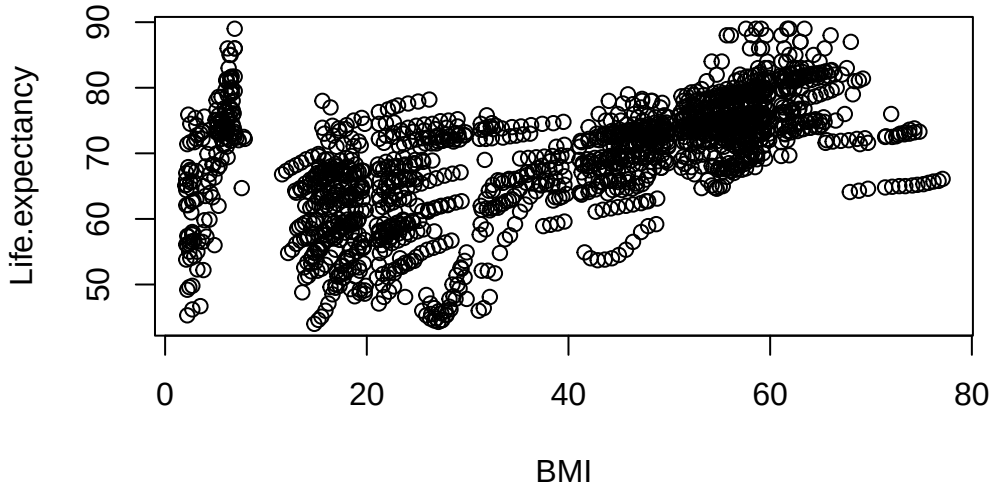
```r
plot(Life.expectancy ~ Adult.Mortality)
```



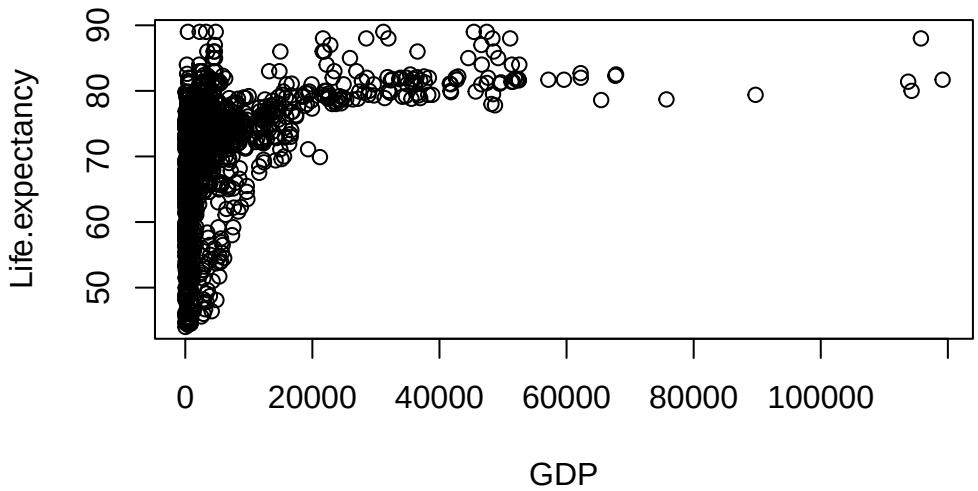

```r
plot(Life.expectancy ~ Alcohol)
```




```r
plot(Life.expectancy ~ BMI)
```

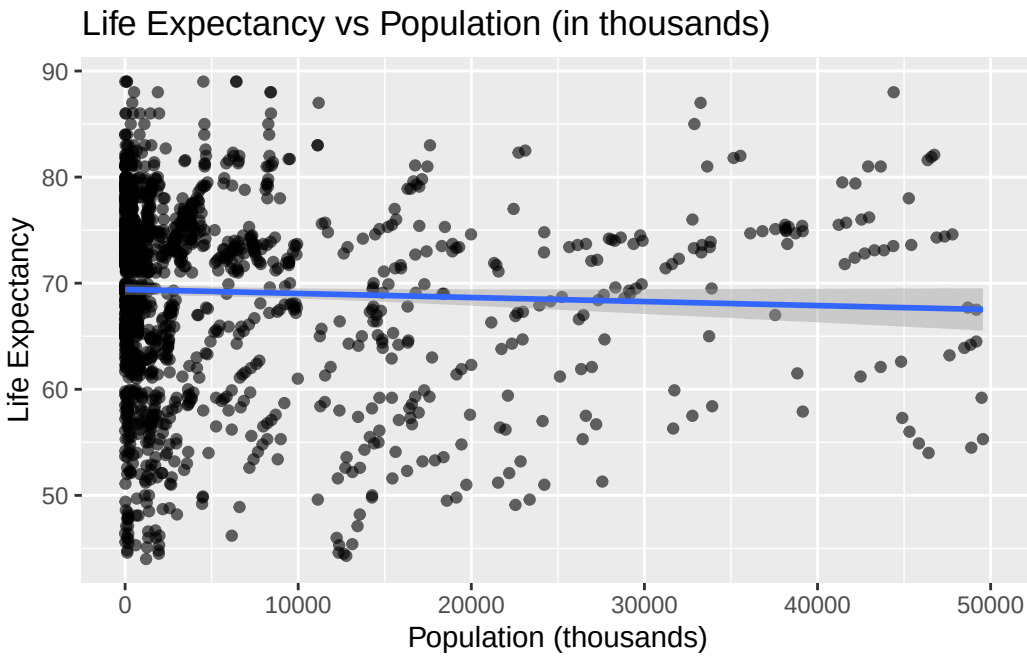


```r
plot(Life.expectancy ~ GDP)
```

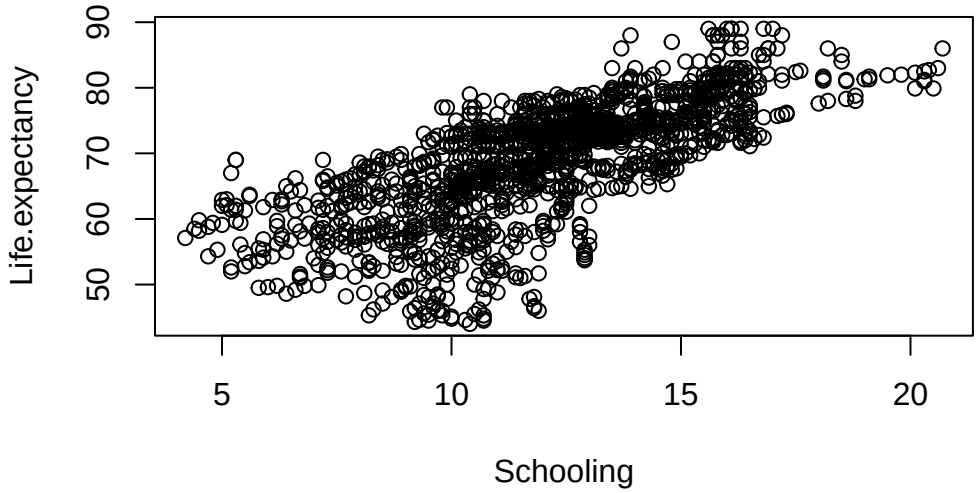



```r
library(ggplot2)

ggplot(df2, aes(x = Population / 1000, y = Life.expectancy)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm") +
  xlim(0, 50000) +
  labs(
    title = "Life Expectancy vs Population (in thousands)",
    x = "Population (thousands)",
    y = "Life Expectancy"
  )
```

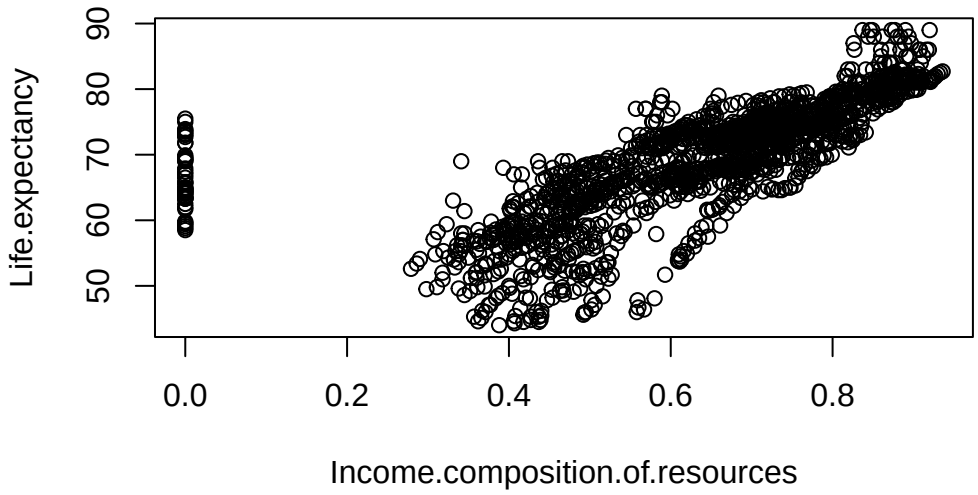
```
`geom_smooth()` using formula = 'y ~ x'

Warning: Removed 89 rows containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 89 rows containing missing values or values outside the scale range
(`geom_point()`).
```



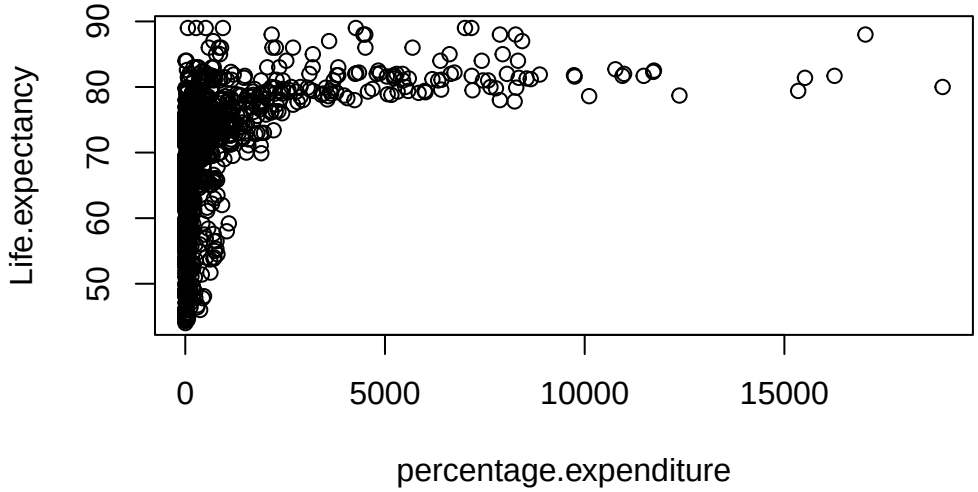

```r
plot(Life.expectancy ~ Schooling)
```



```r
plot(Life.expectancy ~ Income.composition.of.resources)
```

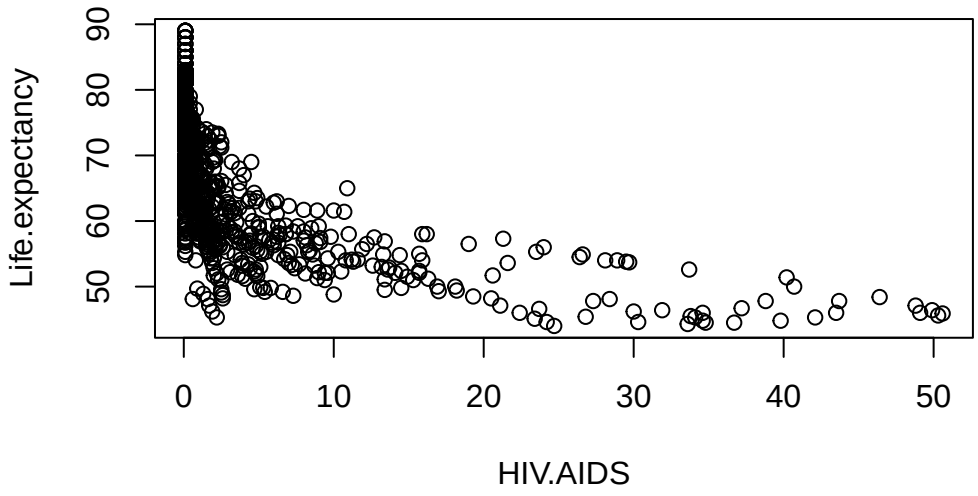



```r
plot(Life.expectancy ~  percentage.expenditure)
```

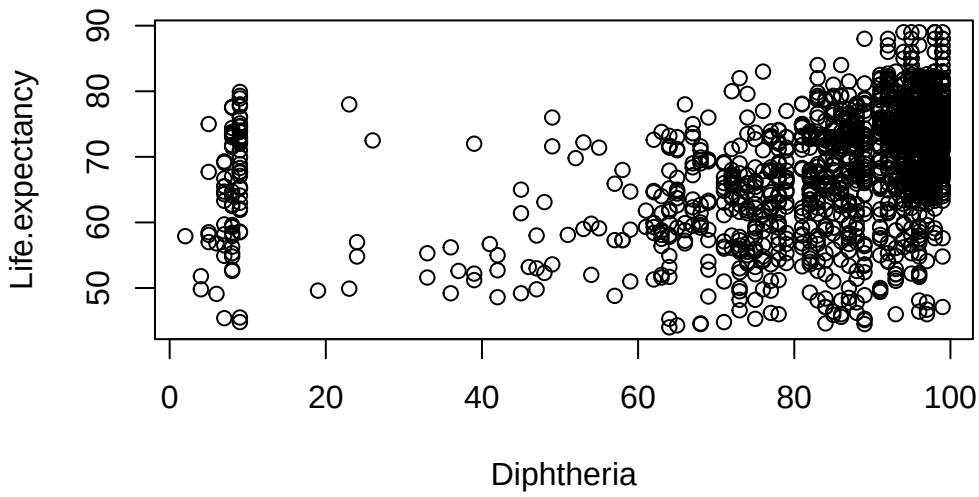


Potentially Problematic Predictors

```
plot(Life.expectancy ~ HIV.AIDS)
```

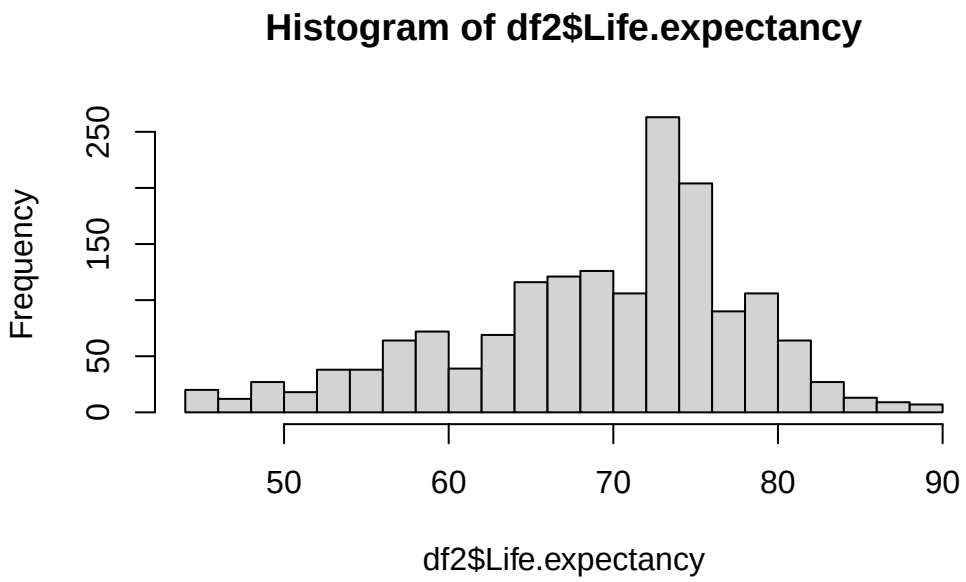



```
plot(Life.expectancy ~ Diphtheria)
```

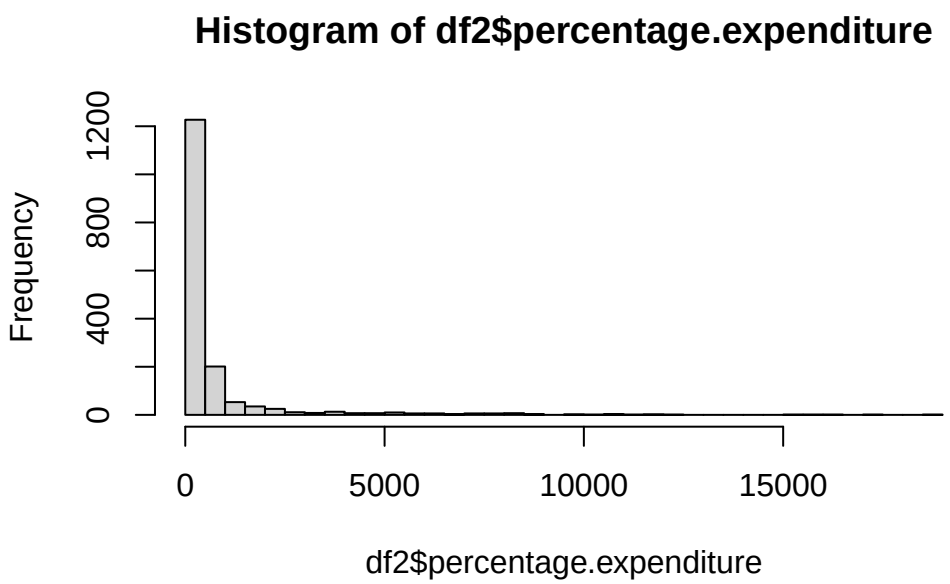


Distribution Checks

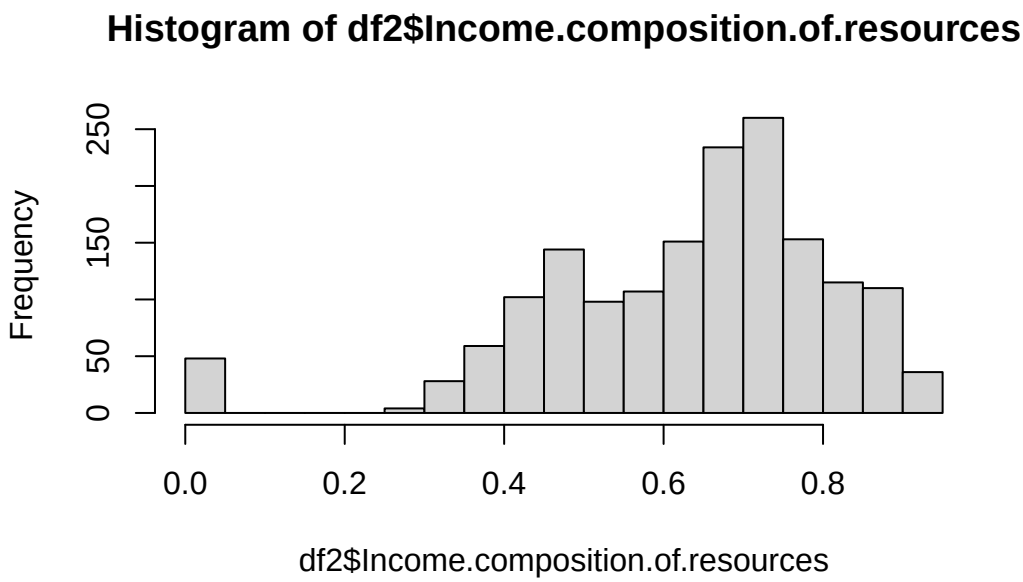
```
hist(df2$Life.expectancy, breaks=30)
```



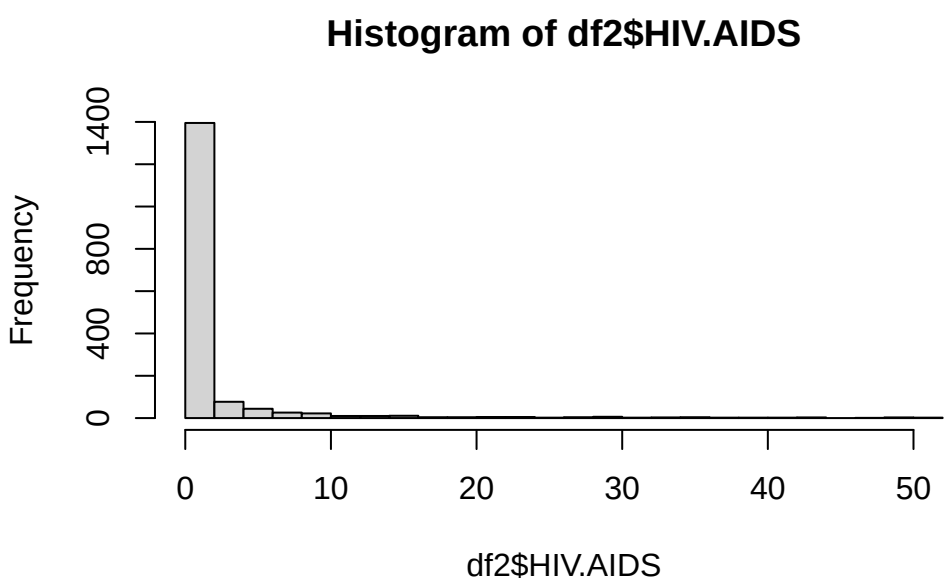

```r
hist(df2$percentage.expenditure, breaks=30)
```



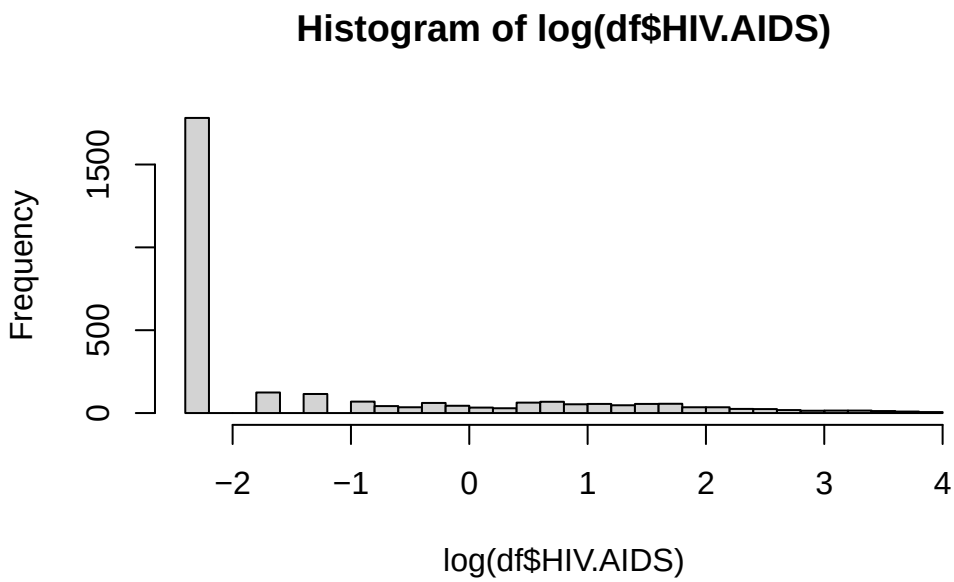
```r
hist(df2$Income.composition.of.resources, breaks=30)
```




```r
hist(df2$HIV.AIDS, breaks=30)
```



```r
hist(log(df$HIV.AIDS), breaks = 30)
```



Forward Stepwise AIC

```r
mod.full <- lm(Life.expectancy ~ ., data = df2)
mod.none <- lm(Life.expectancy ~ 1, data = df2)

aic.fw <- step(mod.none, scope = list(lower = mod.none, upper = mod.full), direction = "forwa
aic.fw
```

```
Call:
lm(formula = Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality +
    Income.composition.of.resources + percentage.expenditure +
    BMI + Diphtheria + under.five.deaths + infant.deaths + Status +
    Alcohol + thinness.5.9.years + Total.expenditure, data = df2)

Coefficients:
                     (Intercept)                        Schooling
                      54.4368068                        0.8742890
```

```
                      HIV.AIDS                  Adult.Mortality
                    -0.4358946                       -0.0167246
Income.composition.of.resources           percentage.expenditure
                     9.8706691                        0.0004259
                           BMI                        Diphtheria
                     0.0336846                        0.0148825
              under.five.deaths                     infant.deaths
                    -0.0695149                        0.0918587
               StatusDeveloping                          Alcohol
                    -0.9461072                       -0.0874716
             thinness.5.9.years                Total.expenditure
                    -0.0569469                        0.0748772
```

Backwards Stepwise AIC

```r
mod.full <- lm(Life.expectancy ~ ., data = df2)

aic.bkw <- step(mod.full, direction = "backward", trace = 0)
aic.bkw
```

```
Call:
lm(formula = Life.expectancy ~ Status + Adult.Mortality + infant.deaths +
    Alcohol + percentage.expenditure + Hepatitis.B + BMI + under.five.deaths +
    Polio + Total.expenditure + Diphtheria + HIV.AIDS + thinness.5.9.years +
    Income.composition.of.resources + Schooling, data = df2)

Coefficients:
                     (Intercept)                 StatusDeveloping
                      54.3875461                       -0.9798311
                 Adult.Mortality                    infant.deaths
                      -0.0166288                        0.0907175
                         Alcohol           percentage.expenditure
                      -0.0917812                        0.0004199
                     Hepatitis.B                               BMI
                      -0.0064329                        0.0341115
              under.five.deaths                             Polio
                      -0.0687641                        0.0078606
              Total.expenditure                       Diphtheria
                       0.0765870                        0.0148547
                        HIV.AIDS                thinness.5.9.years
```

```
                     -0.4372398                       -0.0560117
Income.composition.of.resources                        Schooling
                      9.8478217                        0.8685707
```

Forward Stepwise BIC

```r
n <- nrow(df2)
mod.full <- lm(Life.expectancy ~ ., data = df2)
mod.none <- lm(Life.expectancy ~ 1, data = df2)

mod.fw <- step(mod.none,
     scope = list(lower = mod.none, upper = mod.full),
     direction = "forward",
     k = log(n),     # BIC penalty
     trace = 0)

mod.fw
```

```
Call:
lm(formula = Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality +
    Income.composition.of.resources + percentage.expenditure +
    BMI + Diphtheria + under.five.deaths + infant.deaths, data = df2)

Coefficients:
                     (Intercept)                       Schooling
                       53.394564                        0.873643
                        HIV.AIDS                 Adult.Mortality
                       -0.438326                       -0.017266
Income.composition.of.resources           percentage.expenditure
                        9.778268                        0.000462
                             BMI                       Diphtheria
                        0.038435                        0.014797
               under.five.deaths                    infant.deaths
                       -0.070036                        0.091575
```

Backward Stepwise BIC

```r
n <- nrow(df2)
mod.full <- lm(Life.expectancy ~ ., data = df2)

mod.bkw <- step(mod.full,
     direction = "backward",
     k = log(n),     # BIC penalty
     trace = 0)
mod.bkw
```

```
Call:
lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +
    percentage.expenditure + BMI + under.five.deaths + Diphtheria +
    HIV.AIDS + Income.composition.of.resources + Schooling, data = df2)

Coefficients:
                     (Intercept)                  Adult.Mortality
                       53.394564                        -0.017266
                   infant.deaths           percentage.expenditure
                        0.091575                         0.000462
                             BMI                 under.five.deaths
                        0.038435                        -0.070036
                      Diphtheria                          HIV.AIDS
                        0.014797                        -0.438326
Income.composition.of.resources                         Schooling
                        9.778268                         0.873643
```

```r
vif(mod.bkw)
```

```
                 Adult.Mortality                    infant.deaths
                        1.766417                       176.934458
          percentage.expenditure                              BMI
                        1.248657                         1.549990
               under.five.deaths                      Diphtheria
                      178.045561                         1.209337
                        HIV.AIDS Income.composition.of.resources
                        1.444584                         2.849230
                       Schooling
                        3.066677
```

Variables of concern: `under.five.deaths` & `infant.deaths`

`gdp` and `percentage.expenditure` = (Health expenditure / GDP) × 100

If GDP increases and health spending stays constant, `percentage.expenditure` drops.

`thinness.1.19.years` and `thinness.5.9.years`

Variables Worth Dropping:

- `under.five.deaths` & `infant.deaths` : removed due to explaining similar things as `adult.mortality`
- `Measles` / `Polio` / `Hepatitis.B` / `Diptheria` : Highly correlated and mostly useful for explaining lower income countries
- `thinness 1-19` / `thinness 5-9`: not useful for explaining `adult.mortality`
- `GDP`: Better explained by `percentage.expenditure`
- `Population` : Not measured in per capita, so may mess up measurments due to not being normalized

Fitting First Model

```r
fit1 <- lm(Life.expectancy ~  Status +  Adult.Mortality + Alcohol +
             percentage.expenditure + BMI + HIV.AIDS
           + Income.composition.of.resources + Schooling,
           data = df2)

summary(fit1)
```

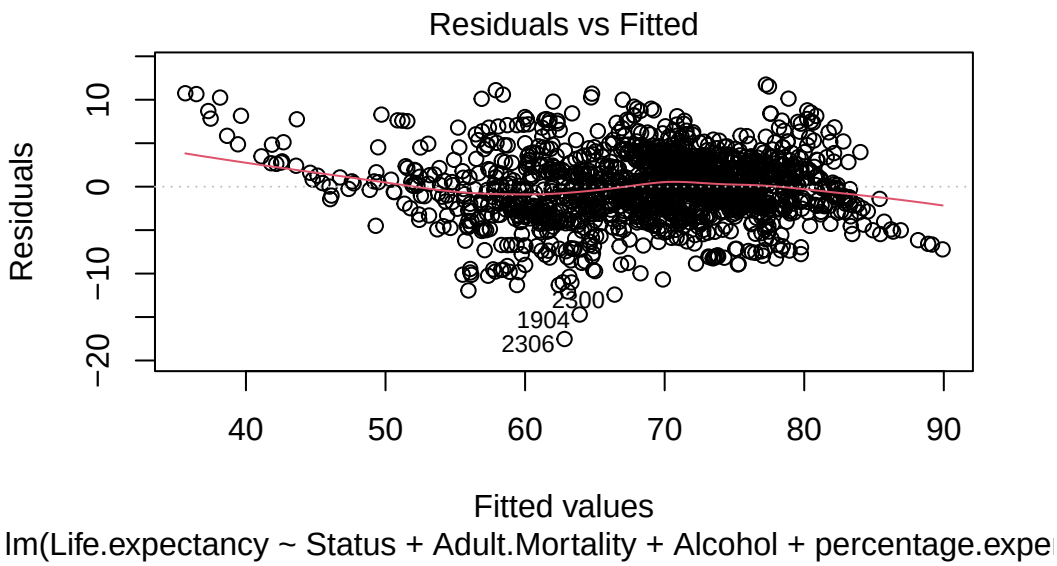
```
Call:
lm(formula = Life.expectancy ~ Status + Adult.Mortality + Alcohol +
    percentage.expenditure + BMI + HIV.AIDS + Income.composition.of.resources +
    Schooling, data = df2)

Residuals:
     Min       1Q   Median       3Q      Max
-17.5266  -2.1474   0.0724   2.3554  11.7443

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)                      5.410e+01  7.128e-01  75.896  < 2e-16 ***
StatusDeveloping                -9.944e-01  3.482e-01  -2.856 0.004346 **
Adult.Mortality                 -1.788e-02  9.715e-04 -18.410  < 2e-16 ***
Alcohol                         -1.293e-01  3.323e-02  -3.891 0.000104 ***
percentage.expenditure           4.156e-04  6.118e-05   6.792 1.54e-11 ***
BMI                              4.043e-02  5.693e-03   7.102 1.83e-12 ***
HIV.AIDS                        -4.368e-01  1.833e-02 -23.828  < 2e-16 ***
Income.composition.of.resources  1.080e+01  8.480e-01  12.734  < 2e-16 ***
Schooling                        9.790e-01  5.968e-02  16.405  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.721 on 1640 degrees of freedom
Multiple R-squared:  0.8219,    Adjusted R-squared:  0.821
F-statistic: 946.1 on 8 and 1640 DF,  p-value: < 2.2e-16
```

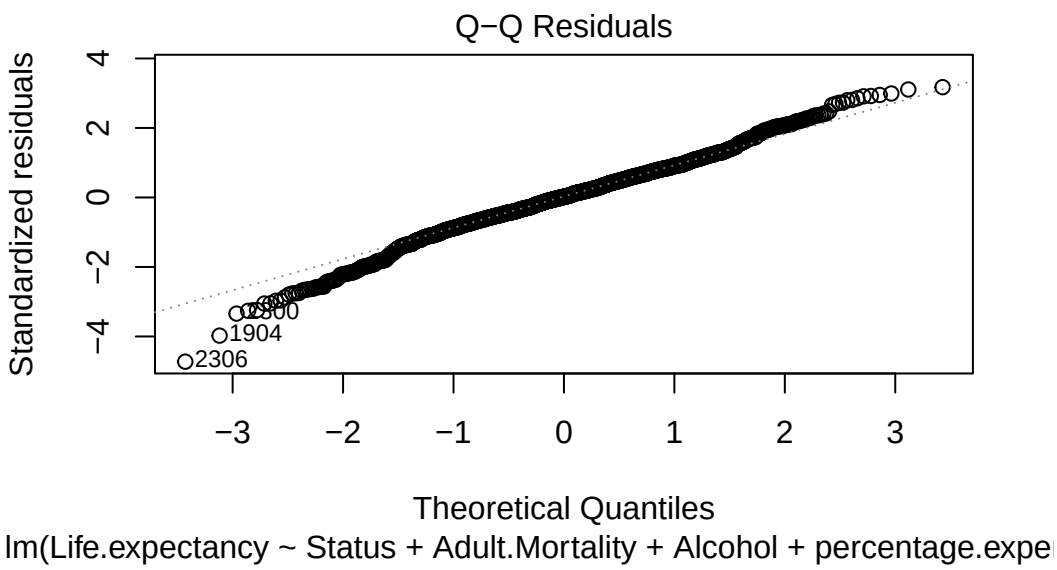
```r
vif(fit1)
```

```
                         Status                  Adult.Mortality
                       1.807872                         1.763606
                        Alcohol           percentage.expenditure
                       2.133401                         1.378734
                            BMI                         HIV.AIDS
                       1.504862                         1.455272
Income.composition.of.resources                        Schooling
                       2.868679                         3.311640
```

```r
plot(fit1, 1)
```


Residuals vs Fitted
Residuals
10
0
−10
−20
40
50
60
70
80
90
2300
1904
2306
Fitted values
lm(Life.expectancy ~ Status + Adult.Mortality + Alcohol + percentage.expe

```r
plot(fit1, 2)
```


Q−Q Residuals
Standardized residuals
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
2300
1904
2306
Theoretical Quantiles
lm(Life.expectancy ~ Status + Adult.Mortality + Alcohol + percentage.expe

```r
shapiro.test(fit1$residuals)
```

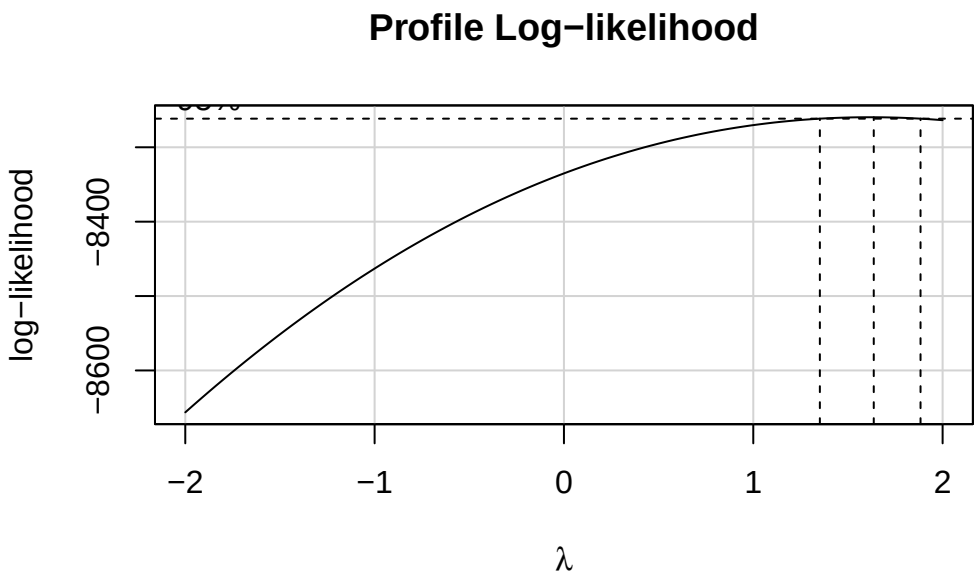
```
    Shapiro-Wilk normality test

data:  fit1$residuals
W = 0.99031, p-value = 5.204e-09
```

Power Transformations for X & Y

```r
pt1 <- powerTransform(df2$Life.expectancy)
summary(pt1)
```

```
bcPower Transformation to Normality
                    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
df2$Life.expectancy    2.7739        2.77       2.4416       3.1062

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                           LRT df       pval
LR test, lambda = (0) 288.2944  1 < 2.22e-16

Likelihood ratio test that no transformation is needed
                          LRT df       pval
LR test, lambda = (1) 114.902  1 < 2.22e-16
```

```r
boxCox(fit1)
```




```
df2$percentage.expenditure_adj <- df2$percentage.expenditure + 0.01
df2$Income.composition.adj <- df2$Income.composition.of.resources + 0.01
```

```
pt2 <- powerTransform(cbind(Status ,Adult.Mortality ,
    Alcohol , BMI, HIV.AIDS,
    Schooling, percentage.expenditure_adj,Income.composition.adj) ~ 1, data = df2)
summary(pt2)
```

```
bcPower Transformations to Multinormality
                            Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Status                         9.7526        9.75       9.2689      10.2364
Adult.Mortality                0.5770        0.58       0.5345       0.6195
Alcohol                        0.4799        0.50       0.4513       0.5085
BMI                            1.1036        1.10       1.0284       1.1788
HIV.AIDS                      -0.5741       -0.57      -0.6187      -0.5296
Schooling                      1.4503        1.45       1.3425       1.5581
percentage.expenditure_adj     0.1008        0.10       0.0836       0.1180
Income.composition.adj         2.4192        2.42       2.3245       2.5139

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)
                                          LRT df       pval
```

```
LR test, lambda = (0 0 0 0 0 0 0 0) 13283.22  8 < 2.22e-16

Likelihood ratio test that no transformations are needed
                                          LRT df       pval
LR test, lambda = (1 1 1 1 1 1 1 1) 21365.74  8 < 2.22e-16
```

We see a major issue with Status ($\lambda = 9.75$)

Model Fit 2

```r
fit2 <- lm(Life.expectancy ~ Adult.Mortality + Alcohol +
             percentage.expenditure + BMI + log(HIV.AIDS)
           + Income.composition.of.resources + Schooling)
summary(fit2)
```

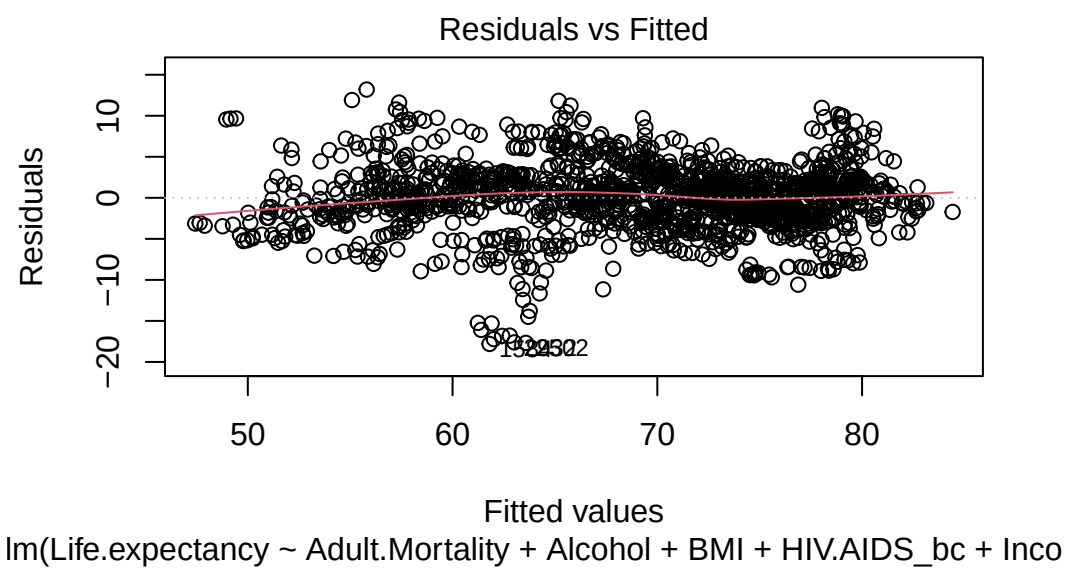
```
Call:
lm(formula = Life.expectancy ~ Adult.Mortality + Alcohol + percentage.expenditure +
    BMI + log(HIV.AIDS) + Income.composition.of.resources + Schooling)

Residuals:
    Min      1Q  Median      3Q     Max
-15.215  -1.846   0.011   1.902  13.281

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.424e+01  5.481e-01  98.962  < 2e-16 ***
Adult.Mortality                 -1.543e-02  9.290e-04 -16.613  < 2e-16 ***
Alcohol                         -3.114e-03  2.909e-02  -0.107 0.914758
percentage.expenditure           4.588e-04  5.603e-05   8.188 5.26e-16 ***
BMI                              2.080e-02  5.430e-03   3.831 0.000132 ***
log(HIV.AIDS)                   -2.244e+00  7.676e-02 -29.239  < 2e-16 ***
Income.composition.of.resources  9.956e+00  8.013e-01  12.424  < 2e-16 ***
Schooling                        6.261e-01  5.730e-02  10.927  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.51 on 1641 degrees of freedom
Multiple R-squared:  0.8415,    Adjusted R-squared:  0.8408
F-statistic:  1245 on 7 and 1641 DF,  p-value: < 2.2e-16
```

We take `log(HIV.AIDS)` due to stand out powerTransform suggestion (-0.5), we take log, and -0.5 and compare with AIC.

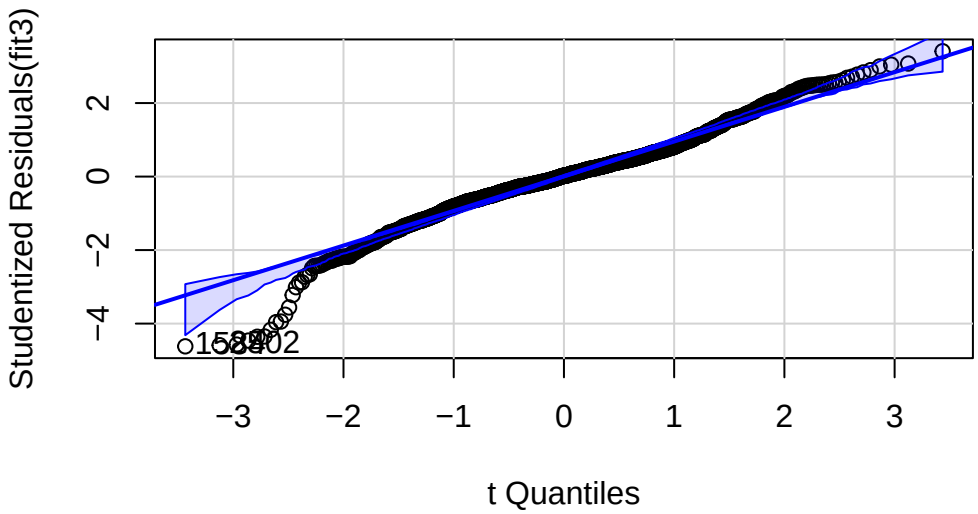
Model Fit 3 (λ = -0.5)

```
# Apply Box-Cox transformation with   = -0.5
df2$HIV.AIDS_bc <- ((df2$HIV.AIDS)^(-0.5) - 1)/(-0.5)

# Now fit the model
fit3 <- lm(Life.expectancy ~
             Adult.Mortality + Alcohol +
             BMI + HIV.AIDS_bc +
             Income.composition.of.resources + Schooling,
           data = df2)
```

```
plot(fit3,1)
```




```
qqPlot(fit3)
```




```
1584 2502
 911 1425
```

```
AIC(fit2, fit3)
```

```
     df      AIC
fit2  9 8830.216
fit3  8 9172.232
```

AIC shows that the `log(HIV.AIDS)` model performs better, as both satisfy constant variance.

Model Fit 4 (Moving Forward with Model 2)

```
fit4 <- lm(Life.expectancy ~ Adult.Mortality + Alcohol +
             percentage.expenditure + BMI + log(HIV.AIDS)
           + Income.composition.of.resources + Schooling)
summary(fit4)
```

```
Call:
lm(formula = Life.expectancy ~ Adult.Mortality + Alcohol + percentage.expenditure +
    BMI + log(HIV.AIDS) + Income.composition.of.resources + Schooling)

Residuals:
    Min      1Q  Median      3Q     Max
-15.215  -1.846   0.011   1.902  13.281

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.424e+01  5.481e-01  98.962  < 2e-16 ***
Adult.Mortality                 -1.543e-02  9.290e-04 -16.613  < 2e-16 ***
Alcohol                         -3.114e-03  2.909e-02  -0.107 0.914758
percentage.expenditure           4.588e-04  5.603e-05   8.188 5.26e-16 ***
BMI                              2.080e-02  5.430e-03   3.831 0.000132 ***
log(HIV.AIDS)                   -2.244e+00  7.676e-02 -29.239  < 2e-16 ***
Income.composition.of.resources  9.956e+00  8.013e-01  12.424  < 2e-16 ***
Schooling                        6.261e-01  5.730e-02  10.927  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.51 on 1641 degrees of freedom
Multiple R-squared:  0.8415,    Adjusted R-squared:  0.8408
F-statistic:  1245 on 7 and 1641 DF,  p-value: < 2.2e-16
```

We now see that `Alcohol` is statistically insignificant (p-value = 0.9147) so we remove it from the model.

Model Fit 5

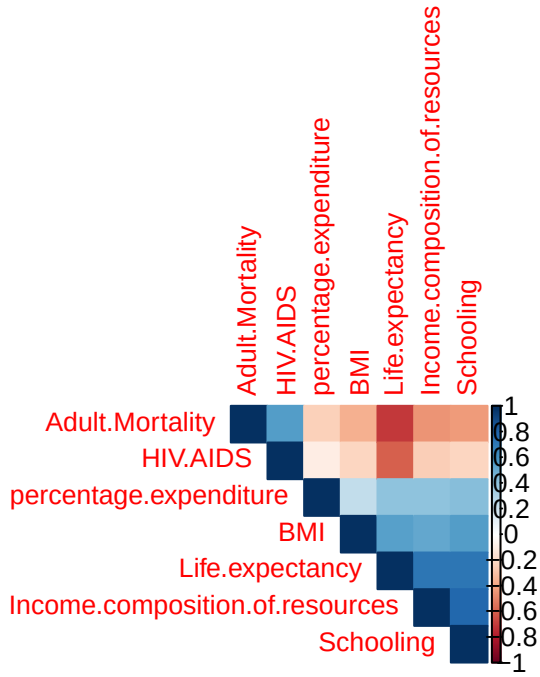
```r
fit5 <- lm(Life.expectancy ~ Adult.Mortality +
             percentage.expenditure + BMI + log(HIV.AIDS)
           + Income.composition.of.resources + Schooling, data = df2)
summary(fit5)
```

```
Call:
lm(formula = Life.expectancy ~ Adult.Mortality + percentage.expenditure +
```

```
    BMI + log(HIV.AIDS) + Income.composition.of.resources + Schooling,
    data = df2)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.225  -1.844   0.011   1.900  13.280 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      5.426e+01  5.093e-01 106.545  < 2e-16 ***
Adult.Mortality                 -1.544e-02  9.276e-04 -16.643  < 2e-16 ***
percentage.expenditure           4.575e-04  5.479e-05   8.350  < 2e-16 ***
BMI                              2.078e-02  5.422e-03   3.832 0.000132 ***
log(HIV.AIDS)                   -2.246e+00  7.548e-02 -29.754  < 2e-16 ***
Income.composition.of.resources  9.941e+00  7.889e-01  12.601  < 2e-16 ***
Schooling                        6.240e-01  5.383e-02  11.593  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.508 on 1642 degrees of freedom
Multiple R-squared:  0.8415,	Adjusted R-squared:  0.8409 
F-statistic:  1453 on 6 and 1642 DF,  p-value: < 2.2e-16
```

Compared to the full model we now see that Adult.Mortality, percentage.expenditure, BMI, log(HIV.AIDS), Income.composition.of.resources, and Schooling are all highly significant after removing some variables from the stepwise selection process, as well as seeing as Alcohol after seeing it become insignificant after taking log(HIV.AIDS). About 84.09% of the adj. R^2 variability in total is explained by a multiple linear regression model with all the predictors collectively.

```r
vars_to_check <- c("Life.expectancy", "Adult.Mortality",
                   "percentage.expenditure", "BMI",
                   "HIV.AIDS", "Income.composition.of.resources", "Schooling")
model_vars_df <- df2[, vars_to_check]
model_vars_df <- model_vars_df[sapply(model_vars_df, is.numeric)]
cor_matrix <- cor(model_vars_df, use = "complete.obs")
corrplot(cor_matrix, method = "color", type = "upper", order = "hclust", tl.cex = 0.8)
```



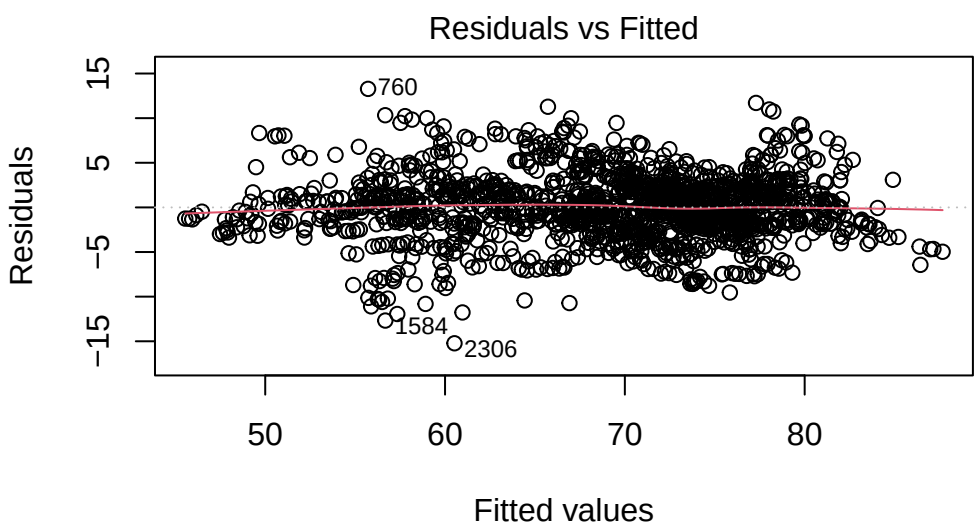

```r
# Calculate correlation matrix (only numbers)
cor_matrix <- cor(model_vars_df, use = "complete.obs")

# Print correlation matrix nicely
round(cor_matrix, 2)
```

```
                                Life.expectancy Adult.Mortality
Life.expectancy                            1.00           -0.70
Adult.Mortality                           -0.70            1.00
percentage.expenditure                     0.41           -0.24
BMI                                        0.54           -0.35
HIV.AIDS                                  -0.59            0.55
Income.composition.of.resources            0.72           -0.44
Schooling                                  0.73           -0.42
                                percentage.expenditure   BMI HIV.AIDS
Life.expectancy                                   0.41  0.54    -0.59
Adult.Mortality                                  -0.24 -0.35     0.55
percentage.expenditure                            1.00  0.24    -0.10
BMI                                               0.24  1.00    -0.21
HIV.AIDS                                         -0.10 -0.21     1.00
Income.composition.of.resources                   0.40  0.51    -0.25
Schooling                                         0.42  0.55    -0.21
```

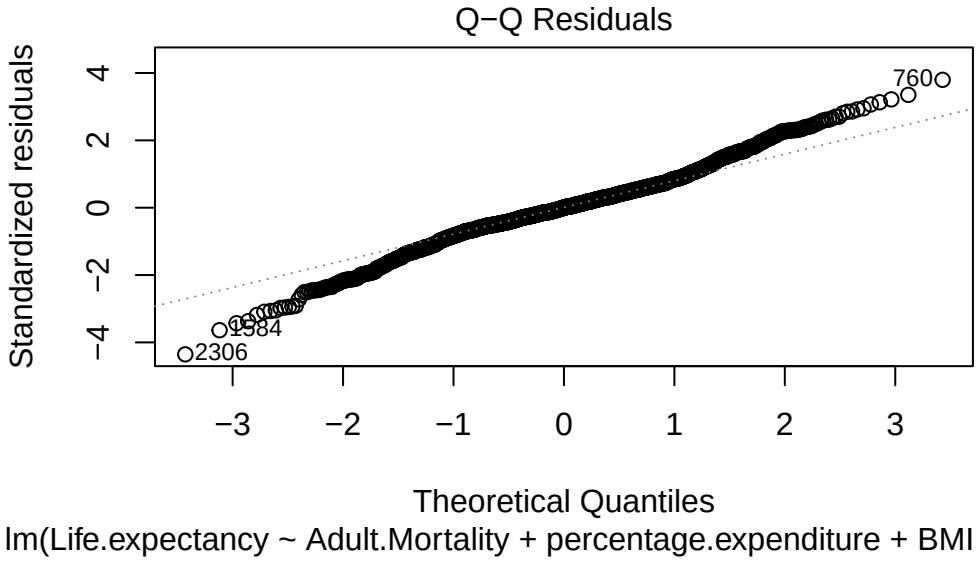
```
                                   Income.composition.of.resources Schooling
Life.expectancy                                               0.72      0.73
Adult.Mortality                                              -0.44     -0.42
percentage.expenditure                                        0.40      0.42
BMI                                                           0.51      0.55
HIV.AIDS                                                     -0.25     -0.21
Income.composition.of.resources                               1.00      0.78
Schooling                                                     0.78      1.00
```

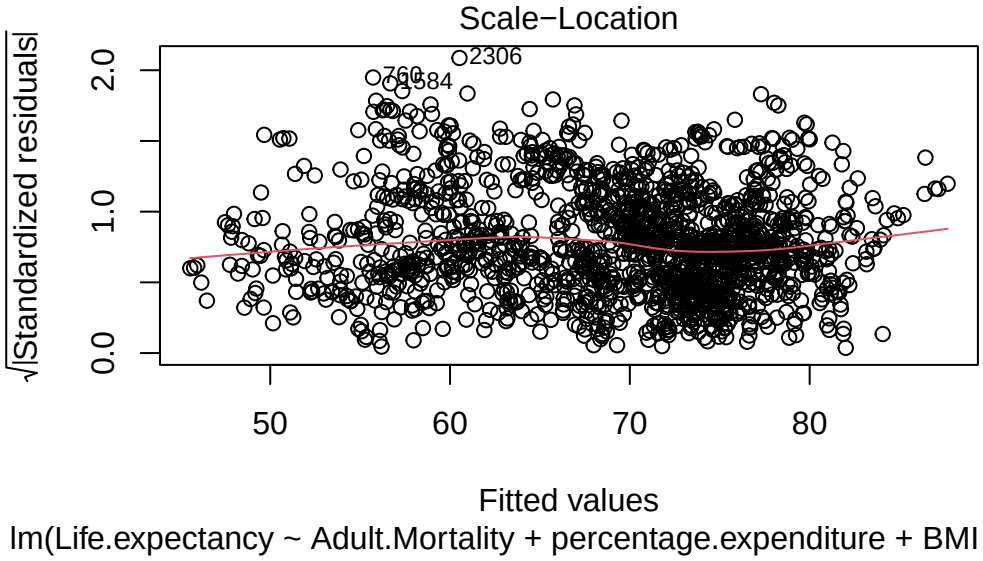
Model 5 Diagnostics

```r
plot(fit5, 1:3)
```



lm(Life.expectancy ~ Adult.Mortality + percentage.expenditure + BMI + log

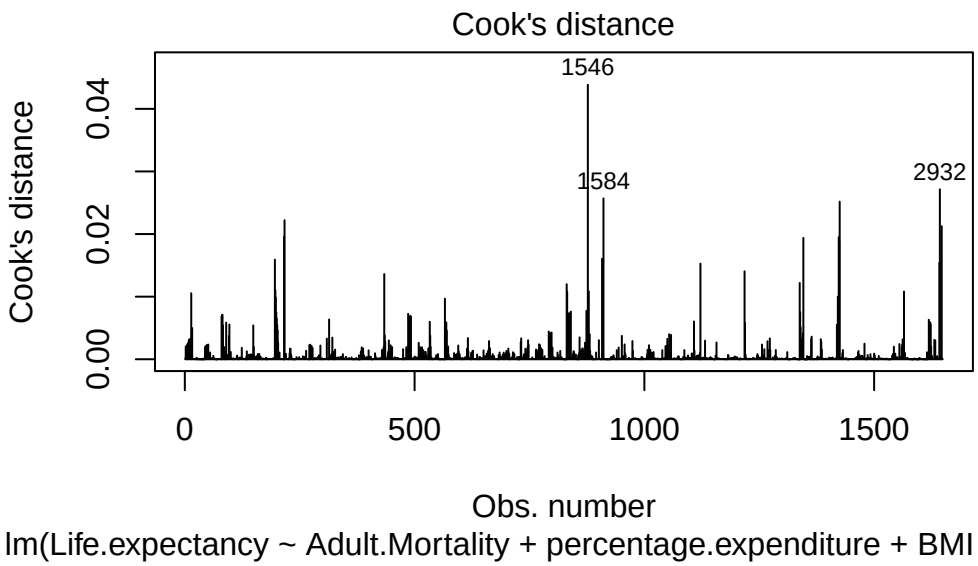





```
shapiro.test(fit5$residuals)
```

```
    Shapiro-Wilk normality test

data:  fit5$residuals
W = 0.98518, p-value = 5.347e-12
```

```r
plot(fit5, 4)
```




```r
df2[1346, ]
```

```
         Status Life.expectancy Adult.Mortality infant.deaths Alcohol
2306 Developing            45.3               45            29    3.86
     percentage.expenditure Hepatitis.B Measles BMI under.five.deaths Polio
2306               45.57109          63       0 2.2                45    63
     Total.expenditure Diphtheria HIV.AIDS      GDP Population
2306              1.12         64      2.2 358.8275     615417
     thinness..1.19.years thinness.5.9.years Income.composition.of.resources
2306                  8.9                8.9                           0.357
     Schooling percentage.expenditure_adj Income.composition.adj HIV.AIDS_bc
2306       8.2                    45.58109                  0.367   0.6516003
```

```r
df2[1340:1348,]
```

```
         Status Life.expectancy Adult.Mortality infant.deaths Alcohol
2300 Developing            54.0              47            23    0.01
2301 Developing            49.7             411            25    0.01
2302 Developing            48.9             418            26    3.78
2303 Developing            48.1             424            27    3.84
2304 Developing            47.1             433            28    3.97
2305 Developing            46.2             441            29    3.91
2306 Developing            45.3              45            29    3.86
2363 Developing            68.8             183             0    0.01
2364 Developing            68.8             182             0    0.01
     percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths Polio
2300               1.321464          92      15 23.3                33    92
2301              54.560337          91     678 22.7                35    91
2302              54.665917          89    1865 22.2                38    88
2303               5.347718          86    1089 21.7                40    84
2304              49.837127          84      31 21.2                42    81
2305               5.379606          77      44  2.7                44    75
2306              45.571089          63       0  2.2                45    63
2363               3.718439          88       0 49.7                 0    94
2364              24.733286          94       0 48.9                 0    99
     Total.expenditure Diphtheria HIV.AIDS       GDP Population
2300             11.59         92      0.8  71.81870     692279
2301             11.24         91      0.9 561.89842     676613
2302             11.98         89      1.3 445.52500    6611692
2303              1.32         86      1.6  45.12842     645872
2304             13.13         84      1.7 394.59324      63126
2305              1.29         77      1.9  46.37592    6165372
2306              1.12         64      2.2 358.82747     615417
2363              5.50         88      0.1  29.65262      57554
2364              5.42         94      0.1 188.51590     563513
     thinness..1.19.years thinness.5.9.years Income.composition.of.resources
2300                  7.7                7.6                           0.413
2301                  7.9                7.8                           0.401
2302                  8.1                8.0                           0.392
2303                  8.3                8.2                           0.384
2304                  8.5                8.4                           0.375
2305                  8.7                8.7                           0.367
2306                  8.9                8.9                           0.357
2363                  1.1                1.2                           0.512
2364                  1.1                1.2                           0.509
```

```
     Schooling percentage.expenditure_adj Income.composition.adj HIV.AIDS_bc
2300       9.3                   1.331464                  0.423  -0.2360680
2301       9.1                  54.570337                  0.411  -0.1081851
2302       8.9                  54.675917                  0.402   0.2458840
2303       8.7                   5.357718                  0.394   0.4188612
2304       8.5                  49.847127                  0.385   0.4660700
2305       8.3                   5.389606                  0.377   0.5490475
2306       8.2                  45.581089                  0.367   0.6516003
2363       9.6                   3.728439                  0.522  -4.3245553
2364       9.5                  24.743286                  0.519  -4.3245553
```

We can see that HIV.AIDS rates were at an an unusual high in Sierra Leone in 2008

```
df2[877, ]
```

```
        Status Life.expectancy Adult.Mortality infant.deaths Alcohol
1546 Developed              80               8             0   11.53
     percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths Polio
1546               18961.35          94       1 58.4                 0    96
     Total.expenditure Diphtheria HIV.AIDS      GDP Population
1546              7.34         99      0.1 114293.8      48865
     thinness..1.19.years thinness.5.9.years Income.composition.of.resources
1546                  0.9                0.9                           0.887
     Schooling percentage.expenditure_adj Income.composition.adj HIV.AIDS_bc
1546      13.5                   18961.36                  0.897   -4.324555
```

```
df2[874:880,]
```

```
        Status Life.expectancy Adult.Mortality infant.deaths Alcohol
1543 Developed            88.0              72             0   11.50
1544 Developed            86.0              73             0   11.36
1545 Developed            83.0              76             0   11.42
1546 Developed            80.0               8             0   11.53
1547 Developed            79.7              82             0   11.75
1548 Developed            79.4              85             0   11.98
1549 Developed            78.8               9             0   11.84
     percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths Polio
1543             17028.5280          95       6 59.6                 0    96
1544              2267.2522          94       0 59.2                 0    96
1545              2052.3931          95       0 58.8                 0    96
1546             18961.3486          94       1 58.4                 0    96
```

```
1547               267.0513          87       0 57.9                 0    96
1548             15345.4907          95       8 57.5                 0    99
1549              1346.2467          95       0 57.0                 0    99
     Total.expenditure Diphtheria HIV.AIDS        GDP Population
1543              7.34         99      0.1 115761.577     518347
1544              7.68         99      0.1  14965.361      56953
1545              8.11         99      0.1  13198.669     497783
1546              7.34         99      0.1 114293.843      48865
1547              6.80         99      0.1   1618.493     479993
1548              7.75         99      0.1  89739.712     472637
1549              7.95         99      0.1   8289.696     465158
     thinness..1.19.years thinness.5.9.years Income.composition.of.resources
1543                  0.9                0.9                           0.894
1544                  0.9                0.9                           0.884
1545                  0.9                0.9                           0.888
1546                  0.9                0.9                           0.887
1547                  0.9                0.9                           0.877
1548                  0.9                0.9                           0.880
1549                  1.0                0.9                           0.874
     Schooling percentage.expenditure_adj Income.composition.adj HIV.AIDS_bc
1543      13.9                 17028.5380                  0.904   -4.324555
1544      13.7                  2267.2622                  0.894   -4.324555
1545      13.5                  2052.4031                  0.898   -4.324555
1546      13.5                 18961.3586                  0.897   -4.324555
1547      13.5                   267.0613                  0.887   -4.324555
1548      13.5                 15345.5007                  0.890   -4.324555
1549      13.5                  1346.2567                  0.884   -4.324555
```

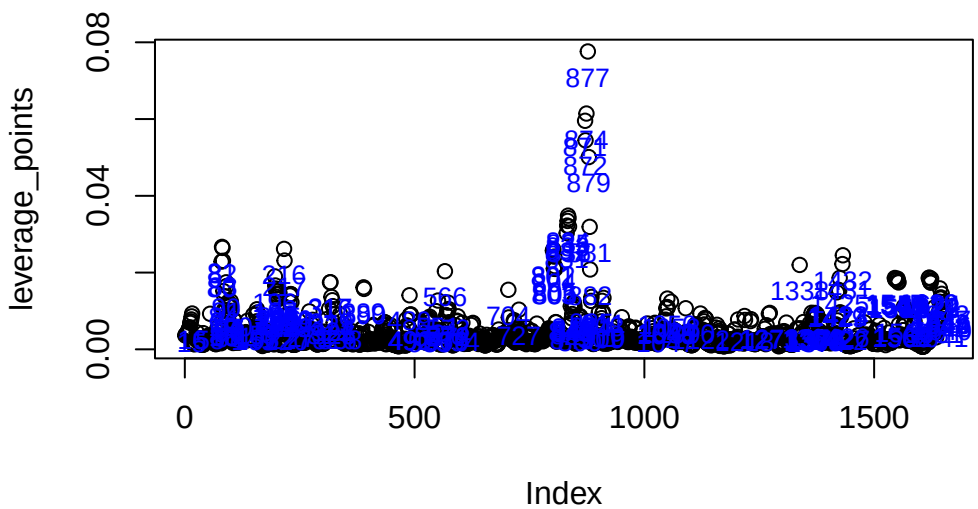
Percentage.expenditure is an outlier here in Luxembourg, explained by economic and historical factors

```
#n <- nrow(df2)
#p <- 6
#ind_h <- which(hatvalues(fit5) > 2*(p+1) / n)
#df2[ind_h, ]
```

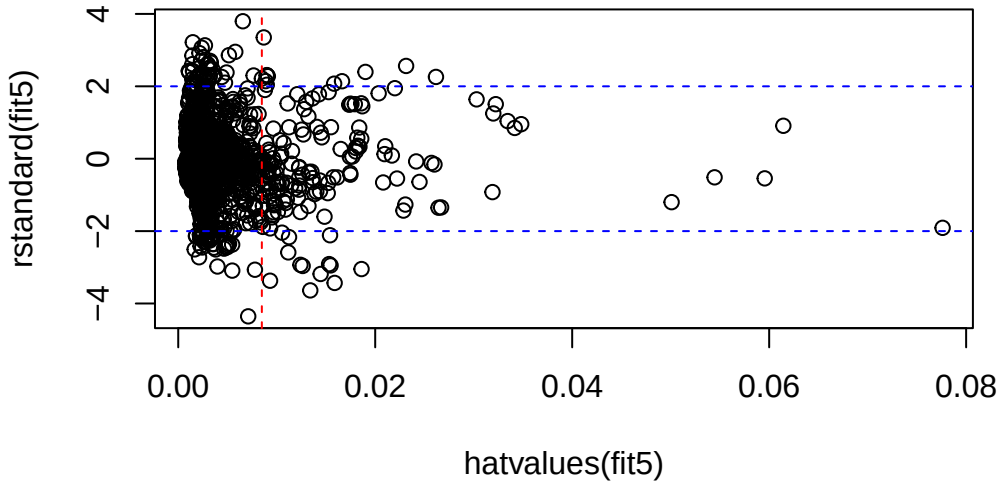
```
#ind <- which(abs(rstandard(fit5)) > 2)
#df2[ind, ]
```

```
#intersect(ind, ind_h)
```

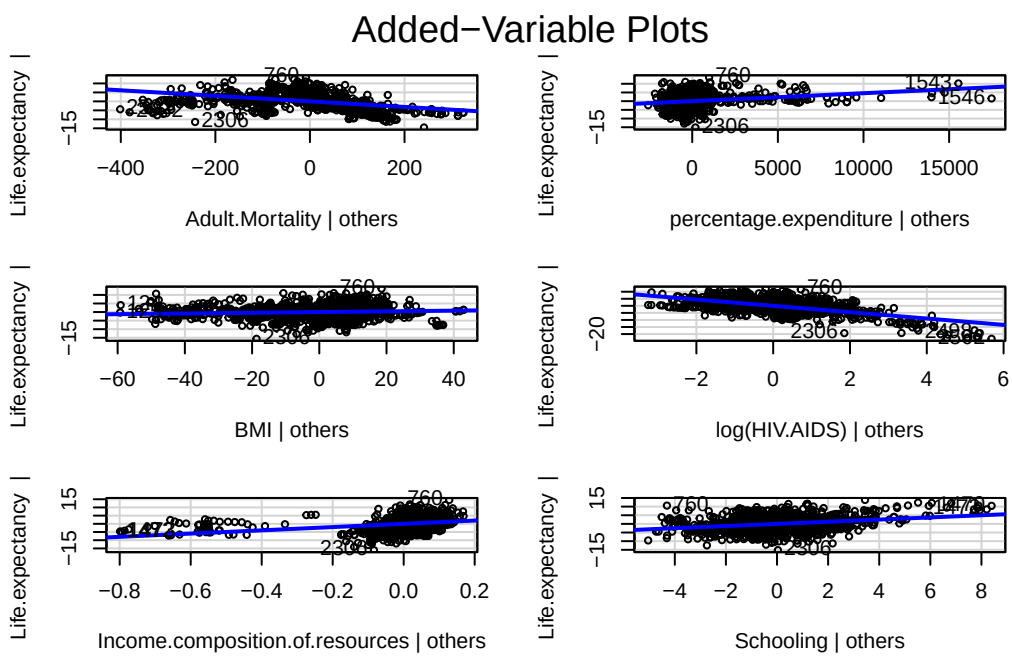
```r
leverage_points <- hatvalues(fit5)
plot(leverage_points)

high_lev <- which(leverage_points > 2 * mean(leverage_points))
text(high_lev, leverage_points[high_lev], labels=high_lev, pos=1, cex=0.8, col="blue")
```




```r
p <- 6
n <- nrow(df2)
plot(hatvalues(fit5), rstandard(fit5))
abline(h = c(-2,2), lty =2, col = "blue")
abline(v = 2 * (p +1) / n, col = "red", lty = 2)
```



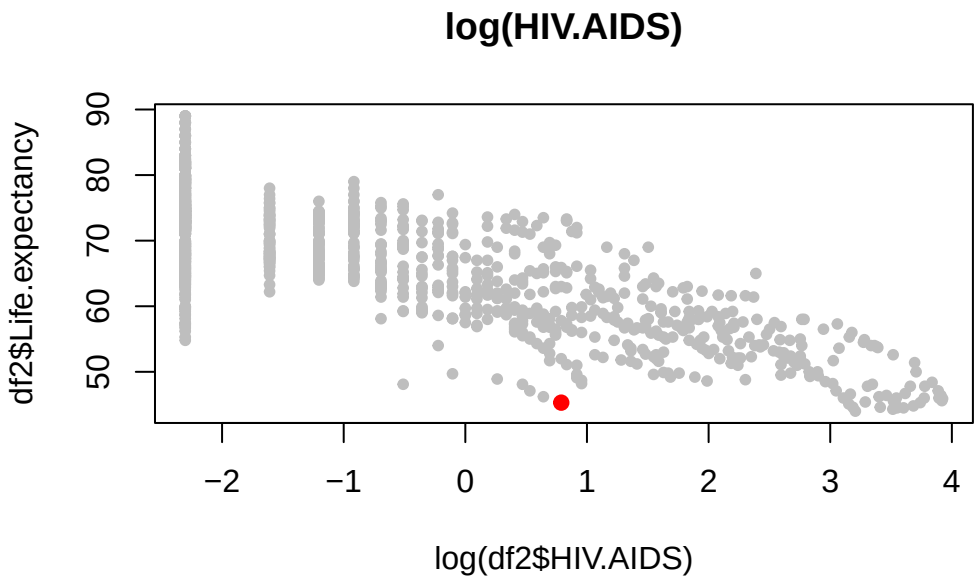

```r
avPlots(fit5)
```



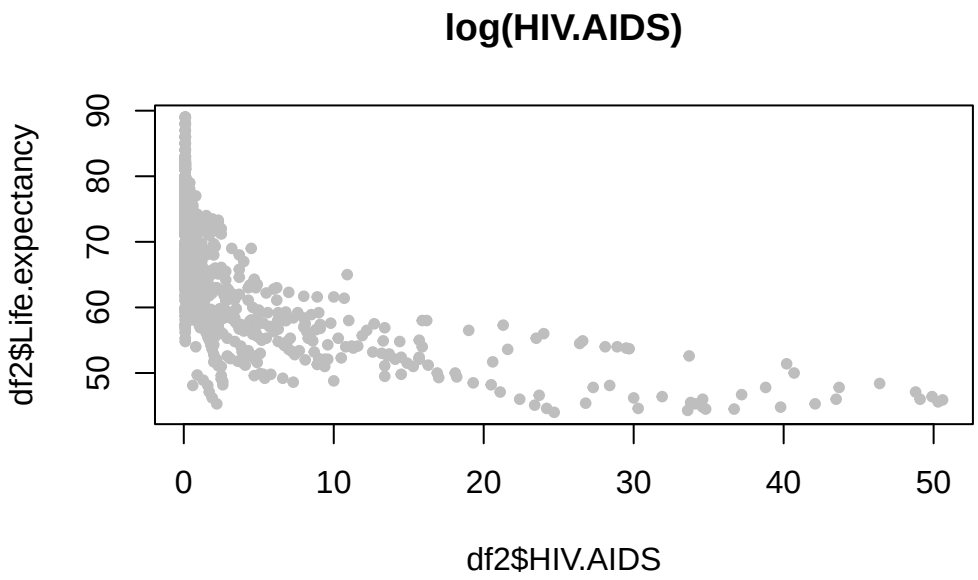
From the added variable plots we can see the predictors all have significant predicting power for Life.expectancy when controlling for other predictors like those in this model

Checking for [1346, 0] outliers

```
plot(log(df2$HIV.AIDS), df2$Life.expectancy,
     col = "gray", pch = 20, main = "log(HIV.AIDS)")
points(log(df2$HIV.AIDS[1346]), df2$Life.expectancy[1346],
       col = "red", pch = 19)
```



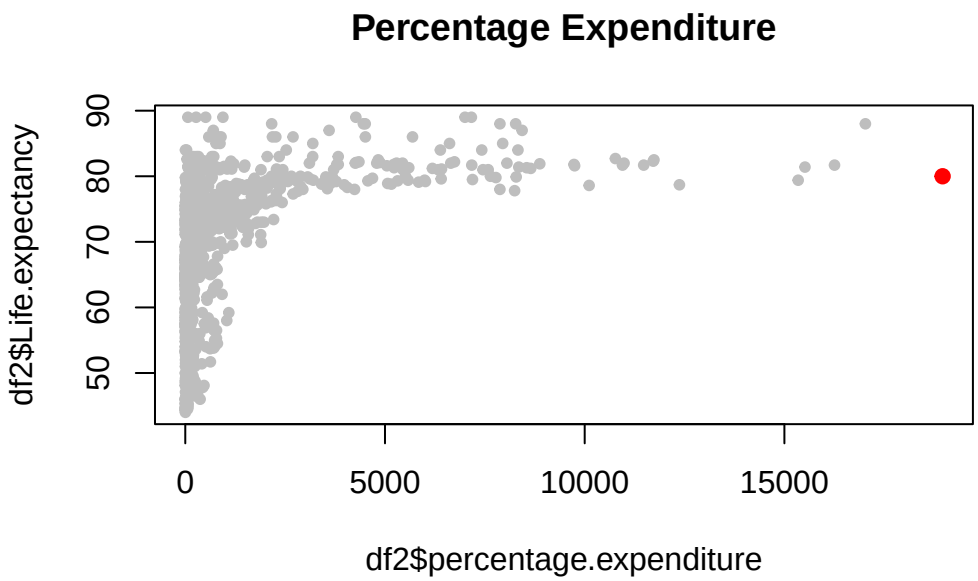

```
# Life.expectancy vs log(HIV.AIDS)
plot(df2$HIV.AIDS, df2$Life.expectancy, col = "gray", pch = 20, main = "log(HIV.AIDS)")
```



High HIV.AIDS rate in Sierra Leone in 2008

Checking for [877, 0] outliers

```
# Life.expectancy vs percentage.expenditure
plot(df2$percentage.expenditure, df2$Life.expectancy, col = "gray", pch = 20, main = "Percent
points(df2$percentage.expenditure[877], df2$Life.expectancy[877], col = "red", pch = 19)
```



Row 877 - Luxemburg high spending on GPD % for healthcare in 2009

Model Fit 6 (Removing Outlier)

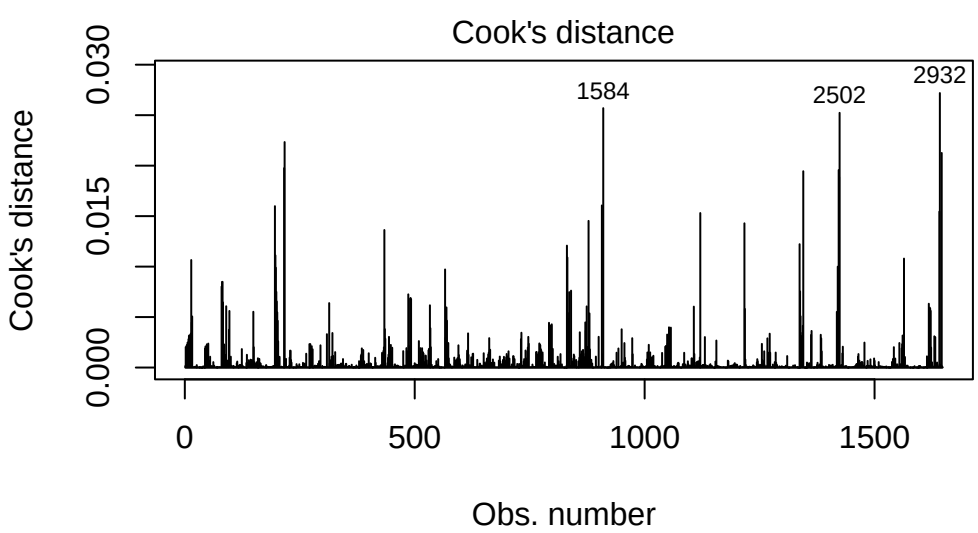
```
fit6<- lm(Life.expectancy ~ Adult.Mortality +
                      percentage.expenditure + BMI + log(HIV.AIDS) +
                      Income.composition.of.resources + Schooling,
                    data = df2[-877,])
summary(fit6)
```

```
Call:
lm(formula = Life.expectancy ~ Adult.Mortality + percentage.expenditure +
    BMI + log(HIV.AIDS) + Income.composition.of.resources + Schooling,
    data = df2[-877, ])

Residuals:
     Min       1Q   Median       3Q      Max
-15.2335  -1.8344  -0.0034   1.9086  13.2391
```

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.433e+01  5.104e-01 106.462  < 2e-16 ***
Adult.Mortality                 -1.544e-02  9.268e-04 -16.659  < 2e-16 ***
percentage.expenditure           4.873e-04  5.692e-05   8.561  < 2e-16 ***
BMI                              2.095e-02  5.419e-03   3.867 0.000115 ***
log(HIV.AIDS)                   -2.247e+00  7.542e-02 -29.791  < 2e-16 ***
Income.composition.of.resources  9.942e+00  7.883e-01  12.613  < 2e-16 ***
Schooling                        6.157e-01  5.396e-02  11.411  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.506 on 1641 degrees of freedom
Multiple R-squared:  0.8417,    Adjusted R-squared:  0.8411
F-statistic:  1454 on 6 and 1641 DF,  p-value: < 2.2e-16
```

Model 6 Diagnostics

```r
plot(fit6, 4)
```



```r
AIC(fit5, fit6)
```

```
Warning in AIC.default(fit5, fit6): models are not all fitted to the same
number of observations

     df      AIC
fit5  8 8828.228
fit6  8 8820.218
```

AIC improvement is marginal and not significant in its predicting power

Exploring Fit 5 Further

```r
confint(fit5)
```

```
                                        2.5 %       97.5 %
(Intercept)                     53.2593093315 55.257012605
Adult.Mortality                 -0.0172568896 -0.013618194
percentage.expenditure           0.0003500706  0.000565017
BMI                              0.0101405151  0.031411525
log(HIV.AIDS)                   -2.3937893940 -2.097705097
Income.composition.of.resources  8.3933428467 11.488045183
Schooling                        0.5184104542  0.729560271
```

```r
library(broom)
tidy(fit5, conf.int = TRUE)
```

```
# A tibble: 7 x 7
  term                 estimate std.error statistic   p.value conf.low conf.high
  <chr>                   <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept)           5.43e+1 0.509        107.   0          5.33e+1 55.3
2 Adult.Mortality      -1.54e-2 0.000928     -16.6  1.35e- 57 -1.73e-2 -0.0136
3 percentage.expendit~  4.58e-4 0.0000548      8.35 1.43e- 16  3.50e-4  0.000565
4 BMI                   2.08e-2 0.00542        3.83 1.32e-  4  1.01e-2  0.0314
5 log(HIV.AIDS)        -2.25e+0 0.0755       -29.8  5.79e-156 -2.39e+0 -2.10
6 Income.composition.~  9.94e+0 0.789         12.6  8.08e- 35  8.39e+0 11.5
7 Schooling             6.24e-1 0.0538        11.6  6.34e- 30  5.18e-1  0.730
```

```r
detach(df2)
```