World Happiness Report

Analysis by Claudio P.

Setting the environment and explanation [point 1]

I decided to use the data provided by the World Happiness Report; here's a brief explanation of the columns and their meaning:

- **Score** - the percentage of Happiness of a Country: people where asked to rate their happiness from 0 to 100, this number it's the average.
- **GDP** - indicator of wealth, it's the logarithm of the GDP indicator and it represents how much money the average person earns per year.
- **Health** - this is the life expectancy, so its scale is a number of years.
- **Sociality** - social support of a country, this is the average of the answers to the question "Do you have someone to rely on?". It's a percentage (example: sociality = 50 means that 50% of the population has at least a friend or a loved one)
- **Corruption** - percentage, people were asked how corrupted their country is from 0 to 100.
- **Generosity** - it goes from -1 to 1 and it represents selfishness or generosity of a country based on money given to charity.
- **Country** - qualitative variable, they simply state the name of the country and the geographical region.

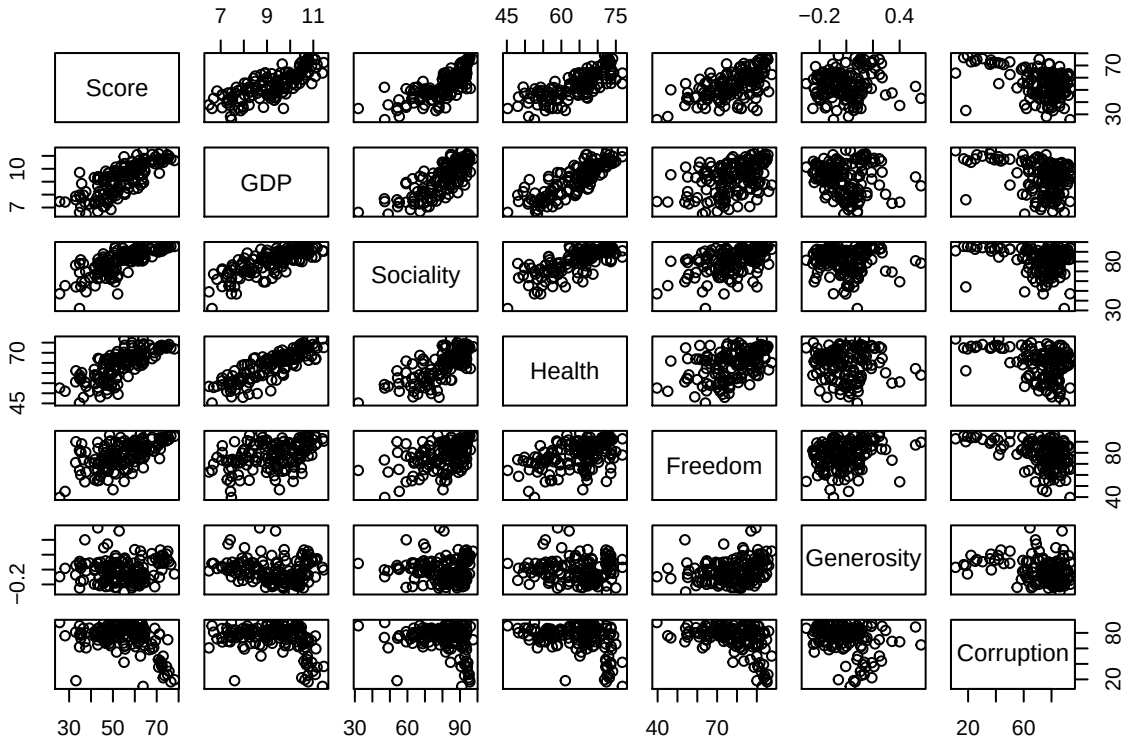
Applied Goals and data discussion [point 2]

The goal here is to investigate the data in order to create a reliable model to make predictions. Since the Score is gathered using answers from users, it is not always an available information, so having a linear model that could give an estimate using some more general data could be useful. Also, it is possible that the Score column is not always reliable: during surveys, people could be prone to overestimate their happiness or give a false answer, therefore compromising the results. The model, on the other hand, relies on more objective numbers (or, at least, not as prone to false results as the Happiness Score): GDP, Generosity and Health won't depend on the user, Corruption of a Country and Social Support are less personal than one person's own satisfaction.

Scatterplot matrix [point 3]

Let's plot the matrix, excluding the Country column (since they're purely descriptive and make the plot unreadable):

```
plot(happy[,2:8])
```



As we can see, Score has a neat, positive, linear relationship with GDP, Health, Sociality and Freedom, while the relationships with Corruption and generosity seem to be weaker (the former being negative). GDP has a linear relationship with Health, Sociality and Freedom as well; Sociality and Health seem to be related as well.

Subset selection [point 4]

In order to make a subset selection, I based on the results of the previous point: Health, GDP, Generosity, Freedom, Corruption and Sociality are obviously included. I added three more interaction terms, the most significant ones according to the plot: Health-GDP, Sociality-GDP, Sociality-Health. So we have a total of 9 possible predictors.

```
p <- 9 #number of maximum predictors
ols1<-regsubsets(Score ~ Health + GDP + Generosity + Freedom +
                   Corruption + Sociality + Health*GDP + Sociality*GDP + Sociality*Health,
                 data=happy, nvmax=p)
summ<- summary(ols1)
summ
```

```
## Subset selection object
## Call: regsubsets.formula(Score ~ Health + GDP + Generosity + Freedom +
##     Corruption + Sociality + Health * GDP + Sociality * GDP +
##     Sociality * Health, data = happy, nvmax = p)
## 9 Variables  (and intercept)
```

```
##                   Forced in Forced out
## Health                FALSE      FALSE
## GDP                   FALSE      FALSE
## Generosity            FALSE      FALSE
## Freedom               FALSE      FALSE
## Corruption            FALSE      FALSE
## Sociality             FALSE      FALSE
## Health:GDP            FALSE      FALSE
## GDP:Sociality         FALSE      FALSE
## Health:Sociality      FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: exhaustive
##          Health GDP Generosity Freedom Corruption Sociality Health:GDP
## 1  ( 1 ) " "    " " " "        " "     " "        " "       " "
## 2  ( 1 ) " "    " " " "        "*"     " "        " "       " "
## 3  ( 1 ) " "    " " " "        "*"     " "        " "       "*"
## 4  ( 1 ) "*"    " " " "        "*"     " "        "*"       " "
## 5  ( 1 ) "*"    " " " "        "*"     " "        "*"       "*"
## 6  ( 1 ) "*"    " " " "        "*"     "*"        "*"       "*"
## 7  ( 1 ) "*"    " " "*"        "*"     "*"        "*"       "*"
## 8  ( 1 ) "*"    "*" "*"        "*"     "*"        "*"       "*"
## 9  ( 1 ) "*"    "*" "*"        "*"     "*"        "*"       "*"
##          GDP:Sociality Health:Sociality
## 1  ( 1 ) " "           "*"
## 2  ( 1 ) " "           "*"
## 3  ( 1 ) " "           "*"
## 4  ( 1 ) " "           "*"
## 5  ( 1 ) " "           "*"
## 6  ( 1 ) " "           "*"
## 7  ( 1 ) " "           "*"
## 8  ( 1 ) " "           "*"
## 9  ( 1 ) "*"           "*"
```

Of all the predictors, the interaction between Health and Sociality is the most relevant one, since it's present in every model, while the interaction between GDP and Sociality is the least useful (being present only in the complete model). Freedom comes in from the model with 2 parameters. Surprisingly, GDP is not very relevant according to the subset selection, but this can be easily explained by the fact that there are several interaction terms and GDP is present through the Health-GDP interaction anyway.

BIC, R-Squared, CP and Cross-Validation [point 5]

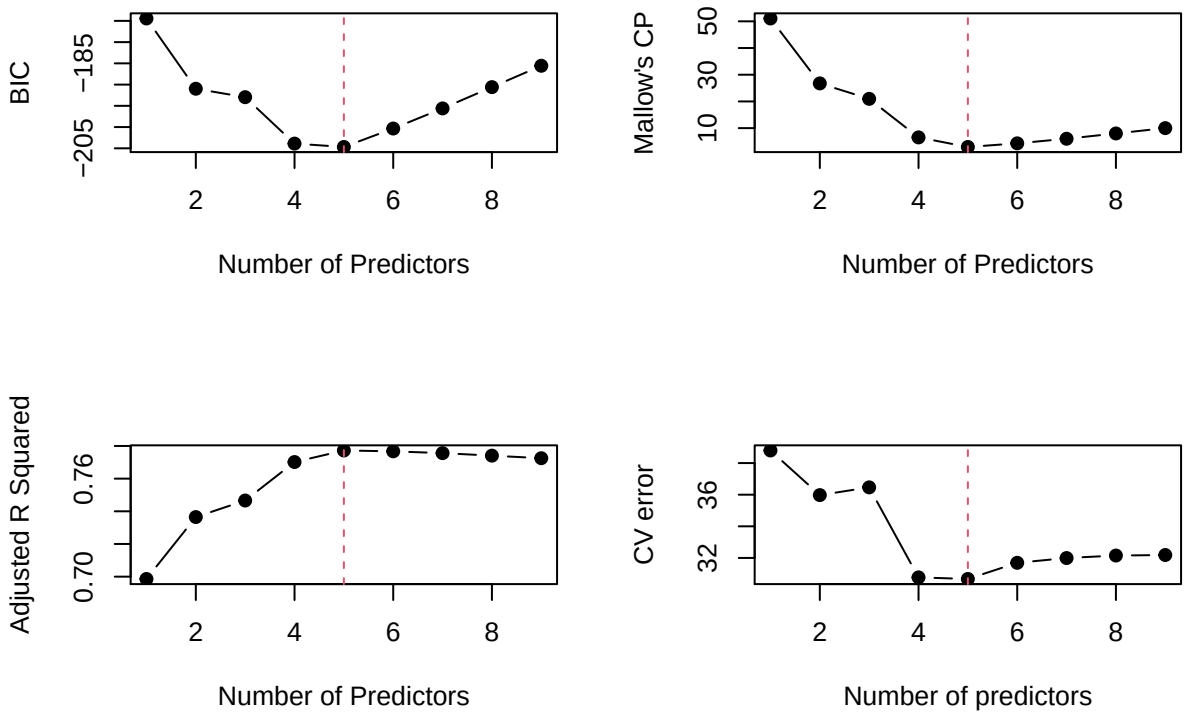
BIC, R-Squared and CP have already been computed and are stored inside the **summ** variable. About the Cross-Validation, I'll use a fold selection of 10 and repeat the process 100 times, changing the seed every time. I'll use the average to compute the cross-validation error per number of predictors:

```r
k <- 10 #fold selection
cv.mean<-c() #initializing the matrix that will contain the errors
for(seed in 1:100) {
  set.seed (seed)
  folds <- sample (1:k,nrow(happy),replace =TRUE) #creating the fold
  cv.errors <- matrix (NA ,k, p, dimnames =list(NULL, paste (1:p))) #initializing
```

```
  for(j in 1:k){
    best.fit <- regsubsets (Score ~ Health + GDP + Generosity + Freedom +
                              Corruption + Sociality + Health*GDP +
                              Sociality*GDP + Sociality*Health,
                            data=happy[folds!=j,], nvmax=p)
    for(i in 1:p) {
      mat <- model.matrix(as.formula(best.fit$call[[2]]), happy[folds==j,])
      coefi <- coef(best.fit,id = i)
      xvars <- names(coefi)
      pred <- mat[,xvars ]%*% coefi
      cv.errors[j,i] <- mean( (happy$Score[folds==j] - pred)^2) #computing the error with this seed
    }
  }
  cv.mean <- cbind(cv.mean, colMeans(cv.errors)) #storing the newly computed errors

}
cv.mean<- rowMeans(cv.mean) #average of errors per number of predictors
```

Now let's see the plots. I'm adding a vertical red line to highlight the best selected number of predictors. Obviously, for CV error, BIC and CP it will be the minimum (since they represent an error), while the R Squared needs to be as high as possible, since it represents the variability explained by the model.



All 4 methods agree that the best model is the one with 5 predictors. With Cross validation, it's not always the case: sometimes, depending on the seed, some other numbers can occur (example, with seed=1, 5 is not the optimal number). Since the average of the seeds and the other three methods indicates 5, I'll pick 5 as the best number.

Combining the results with the previous point the best model is composed by: Health, Freedom, Sociality abd the interaction terms between Sociality-Health and GDP-Health.

```
ols1<-lm(Score ~ Health + Freedom + Sociality + Health*GDP + Sociality*Health, data=happy)
```

Collinearity Issues [point 6]

Let's check the collinearity of the model:

```
vif(ols1)
```

```
##           Health          Freedom        Sociality              GDP
##        73.185244         1.338861       180.527154       242.297719
##       Health:GDP Health:Sociality
##       737.455113       471.787461
```

There are a lot of issues, since the VIFs are very high. Let's start by dropping the highest one, interaction between Health and GDP:

```
ols1<-lm(Score ~ Health + Freedom + Sociality + GDP + Sociality*Health, data=happy)
vif(ols1)
```

```
##           Health          Freedom        Sociality              GDP
##        44.255963         1.336424        69.579381         4.482356
## Health:Sociality
##       194.008999
```

Health, Sociality and interaction term between them are still too high, so I'll drop the interaction term:

```
ols1<-lm(Score ~ Health + Freedom + Sociality + GDP, data=happy)
vif(ols1)
```

```
##    Health   Freedom Sociality       GDP
##  3.864075  1.331919  2.884978  4.373218
```

Now every VIF is below the acceptable threshold, so the new model will only have Health, Freedom, Sociality and GDP as predictors.

```
ols<-lm(Score ~ Health + Freedom + Sociality + GDP, data=happy)
summ<- summary(ols)
```

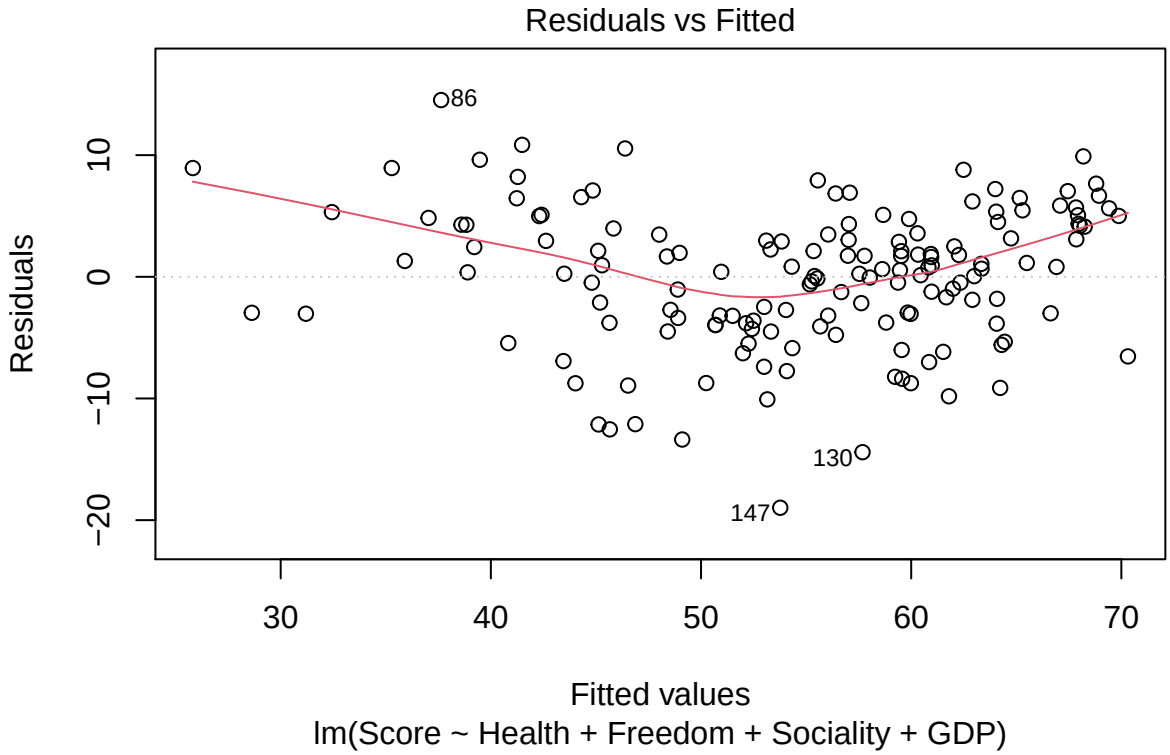
Diagnostics [point 7]

Constant Variance assumption [a]

Let's check the Residuals vs Fitted plot:

```r
#I'll store variables I'm about to use in the next lines of code
res<- residuals(ols) #residuals
fit <- fitted(ols) #fitted values
rsta<- rstandard(ols) #standardized residuals

plot(ols, which=1)
```



At first sight, it would seem that the variance changes when fitted values are around 50. let's run a test:

```r
var.test(res[fitted(ols)>50],res[fitted(ols)<50])
```

```
##
##  F test to compare two variances
##
## data:  res[fitted(ols) > 50] and res[fitted(ols) < 50]
## F = 0.58195, num df = 108, denom df = 43, p-value = 0.02607
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3415853 0.9380538
## sample estimates:
## ratio of variances
##          0.5819523
```

Indeed, the p-value allow us to reject the null hypothesis of constant variance. Ratio=1 is not between the 95% confidence interval.

Relationship between predictors and response [b]

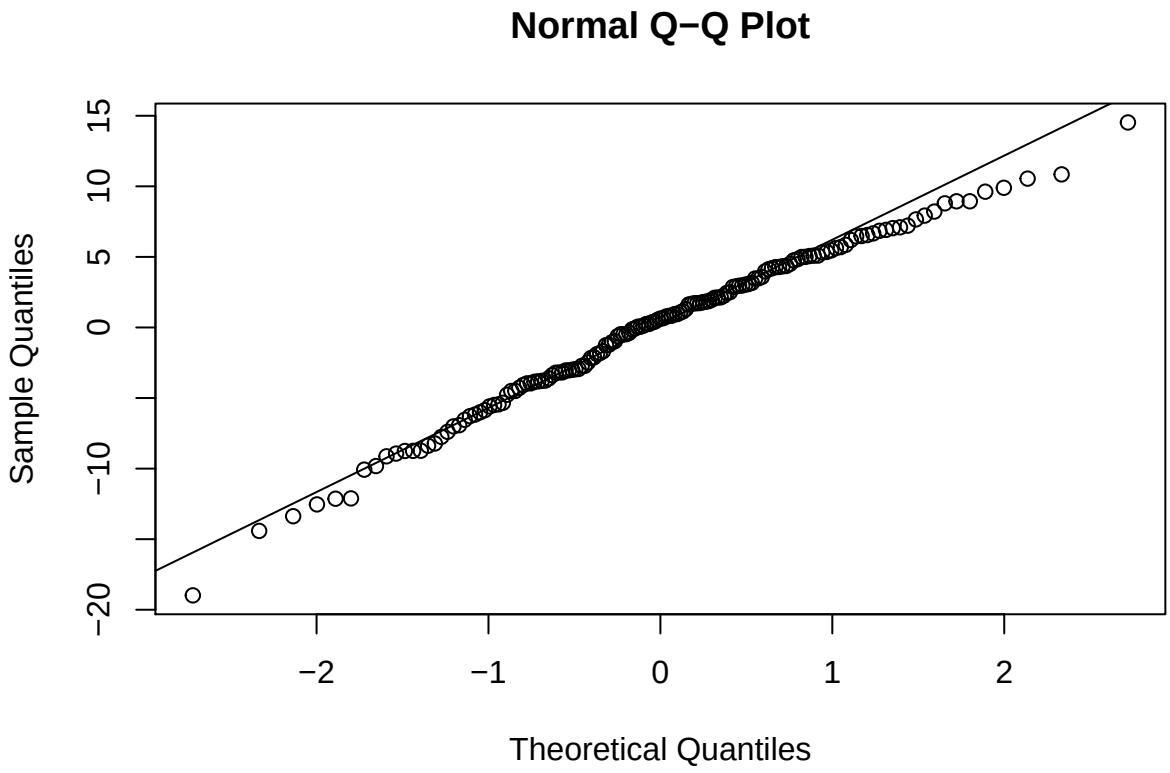
From the previous plot, we can see that the relationship doesn't seem to be linear. The red line is parabolic, suggesting some kind on non-linear relationship. On the other hand, it's worth noticing that on the left side, which seems more parabolic, there are very few points, so I'm assuming that those few points influence a lot the red line. Indeed, even if I try to apply some transformation to the Score (for example the square root), the line doesn't change much.

In any case, the line suggests **the relationship is not linear.**

Normality Assumption [c]

Let's verify if the residuals are normally distributed.

```
qqnorm(res)
qqline(res)
```



It would appear that they are normally distributed, but let's run a test:

```
shapiro.test(res)
```

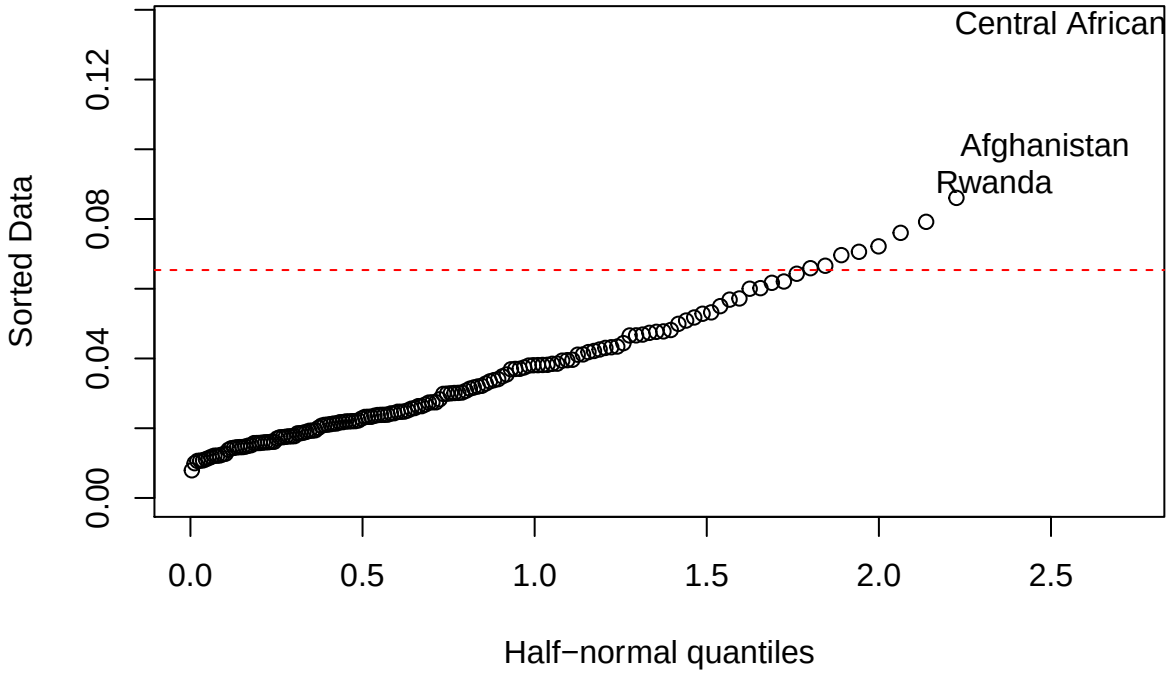
```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.98786, p-value = 0.2056
```

Indeed, the p-value is very high and therefore there's no evidence against the null hypothesis of normality.

Large Leverage points [d]

Let's visualize the halfnorm plot. The red line represents the threshold, above which points are considered of large leverage.

```
infl <- influence(ols)
hat <- infl$hat
halfnorm(hat,3,labs=happy$Country)
abline(h=2*sum(hat)/nrow(happy), lty=2, col="red")
```

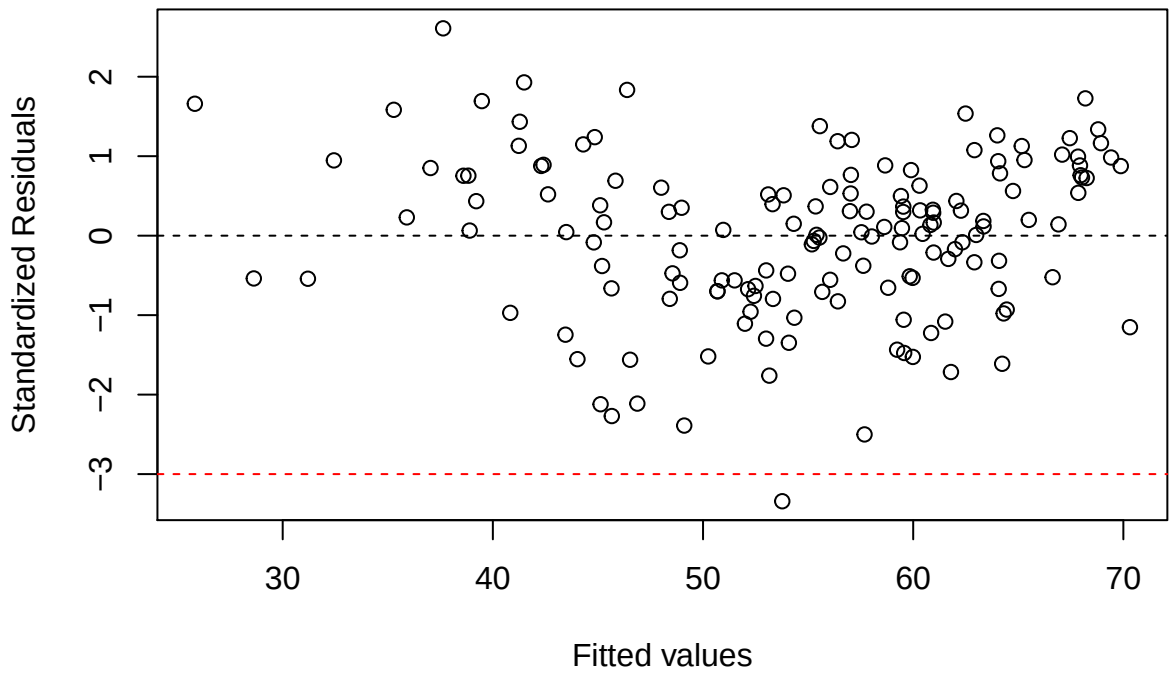


Central African Republic, Rwanda and Afghanistan are large leverage points, but there are several large leverage points above the line.

Outliers [e]

First, let's plot the standardized residuals vs fitted values.

```
plot(fitted(ols), rsta, xlab="Fitted values", ylab="Standardized Residuals")
abline(h=0, lty=2)
abline(h=-3, lty=2, col="red")
```



There are no points above 3, but there is one below -3 which is:

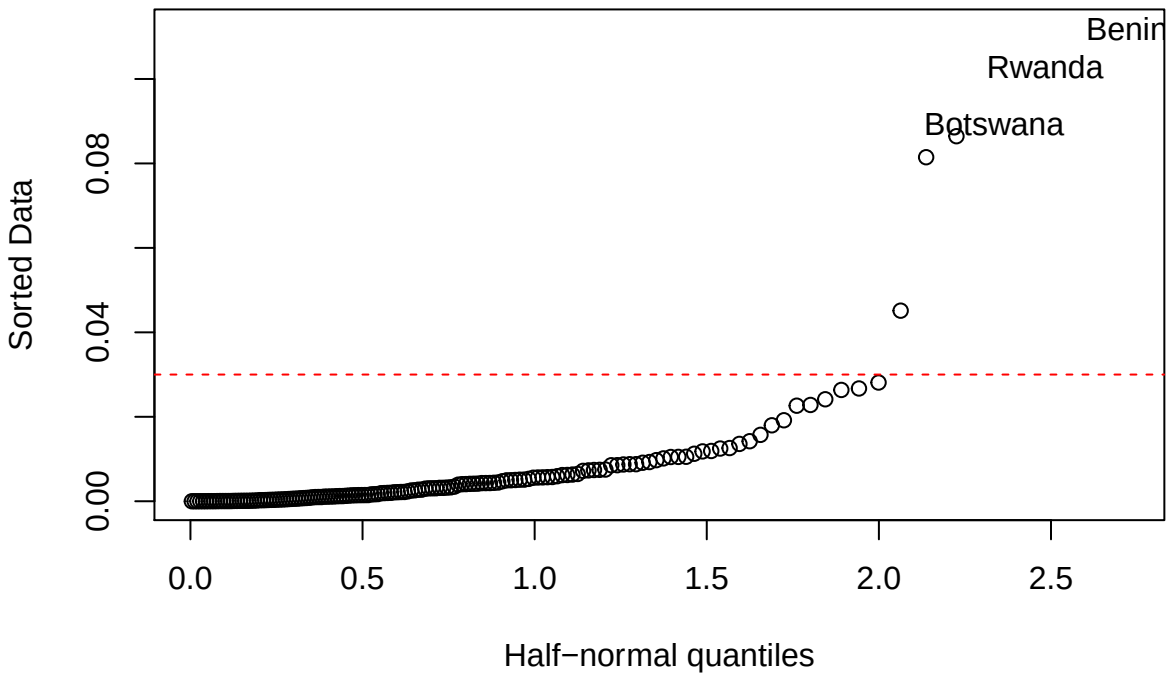
```
happy$Country[as.numeric(names(rsta[which(abs(rsta)>3)]))]
```

```
## [1] "Botswana"
```

Influential points [f]

Let's compute and plot the cook's distance on a halfnorm plot. I used 0.03 as threshold because after that number points tend to have a bigger distance.

```
cook <- cooks.distance(ols)
halfnorm(cook, 3, labs=happy$Country)
abline(h=0.03,lty=2, col='red')
```



So all the influential points are:

```
happy$Country[as.numeric(names(cook[which(cook>0.03)]))]
```

```
## [1] "Ivory Coast"              "Benin"
## [3] "India"                    "Botswana"
## [5] "Central African Republic" "Rwanda"
```

Improving the model [point 8]

As I said before, I tried to apply a transformation without obtaining much benefits, so in order to obtain a better model, I'll just exclude points with a Cook's distance greater than 0.03. In this specific case, it's not a big issue to drop a few countries: Happiness is influenced by a few things that cannot really be measured, so it totally makes sense that there are states very different from the expectation. An example: two countries with the same GDP, Health, Sociality and Freedom but very different cultures can have different happiness scores because their respective population value their own happiness in different ways.

```
ols<- lm(Score ~ Health + GDP + Freedom + Sociality, data=happy, subset=(cook<0.03))
summ<-summary(ols)
summ
```

```
##
## Call:
## lm(formula = Score ~ Health + GDP + Freedom + Sociality, data = happy,
```

```
##     subset = (cook < 0.03))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6829  -3.6241   0.5199   3.6948  10.7470
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.31726    4.26732  -7.808 1.16e-12 ***
## Health        0.41480    0.12305   3.371 0.000965 ***
## GDP           2.33476    0.77114   3.028 0.002927 **
## Freedom       0.25655    0.04316   5.944 2.05e-08 ***
## Sociality     0.24210    0.06839   3.540 0.000541 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.2 on 142 degrees of freedom
## Multiple R-squared:  0.7757, Adjusted R-squared:  0.7694
## F-statistic: 122.7 on 4 and 142 DF,  p-value: < 2.2e-16
```

Compared to the model before the improvements, R squared increased by 4 (from 73 to 77) and the Residual Standard Error decreased of 0.5.

Coefficients and uncertainty [point 9]

Here's the coefficients, as in the previous point, with standard errors and their confidence intervals:

```
summ$coefficients[,1:2]
```

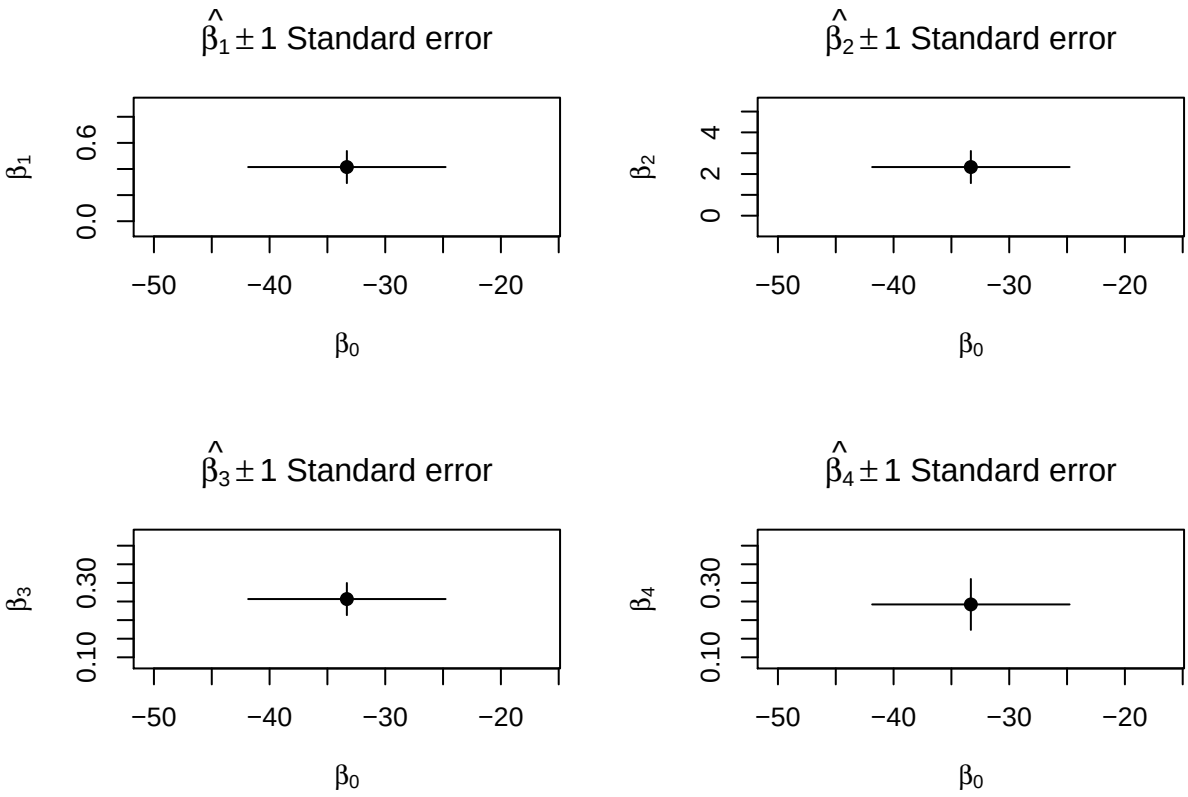
```
##                Estimate Std. Error
## (Intercept) -33.3172629 4.26731821
## Health        0.4148026 0.12305344
## GDP           2.3347649 0.77113595
## Freedom       0.2565547 0.04316220
## Sociality     0.2420976 0.06838976
```

```
confint(ols)
```

```
##                   2.5 %      97.5 %
## (Intercept) -41.7529444 -24.8815813
## Health        0.1715492   0.6580560
## GDP           0.8103748   3.8591549
## Freedom       0.1712312   0.3418782
## Sociality     0.1069040   0.3772912
```

None of the confidence intervals include 0, so it's safe to assume that all the parameters are relevant. The estimates are good, since the standard errors are small compared to them. The intercept could, technically, cause the Score (a percentage) to go below zero: this cannot happen since it's impossible to have every predictor equal to zero (it makes no sense to have a country with 0 life expectancy or 0 GDP).

Here's some plots, showing the various betas vs the intercept. The lines represent their standard error.



Sigma and R Squared [point 10]

Here's the Residual Standard Error (Sigma) and R squared:

```
sigma(ols)
```

```
## [1] 5.200256
```

```
summ$r.squared
```

```
## [1] 0.7756698
```

Sigma represents the error on the Happiness Score: since Score is a percentage, so is sigma. It means that the variability of the score is around 5.2%

R squared represents the variability explained by the model, which is around 77%.

P-values [point 11]

Here's the p-values of the beta:

```r
n<- nrow(happy) - 6 #n without influential points
p <- 4 #predictors

test <- c()
p_value <- c()
p_names<-c("Health","GDP","Freedom","Sociality")

test[1:4]<- summ$coefficients[2:5,1]/summ$coefficients[2:5,2]
p_value[c(1:4)] <- 2*(1- pt(test[c(1:4)], n-p-1))

names(p_value) <- p_names
print(p_value)
```

```
##       Health          GDP      Freedom    Sociality
## 9.652171e-04 2.927459e-03 2.054785e-08 5.414976e-04
```

Basically every beta has an extremely low p-value, below the 5% threshold: this means that there's very strong evidence against the null hypothesis of the associated beta being zero, therefore we should accept the predictors in the model.

Testing multiple regressors [point 12]

First, let's compute the two models: **null** is the empty model, without predictors but the intercept; **nested** is the model without Health and GDP as predictors. The reason of this choice is simple: Score, Freedom and Sociality are averages of answers provided by users and therefore subjective. One could assume that Health and GDP have no influence: why should I evaluate my own happiness considering the number of years people in my country live on average or the medium wealth? They are numbers somewhat unrelated to the single individual.

```r
null <- lm(Score ~ 1, data=happy, subset=(cook<0.03))
nested <- lm(Score ~ Freedom + Sociality, data=happy, subset=(cook<0.03))
```

So, here's the test:

Null test:

H0 : β_i=0 for every i!=0

H1 : β_i!=0 for every i!=0

Nested model:

H0 : β_i=0 for i=1 (Health), i=2 (GDP)

H1 : β_i!=0 for i=1 (Health), i=2 (GDP)

```r
anova(nested, ols)
```

```
## Analysis of Variance Table
##
## Model 1: Score ~ Freedom + Sociality
## Model 2: Score ~ Health + GDP + Freedom + Sociality
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
```

```
## 1    144 5449.3
## 2    142 3840.1  2    1609.2 29.753 1.614e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(null, ols)
```

```
## Analysis of Variance Table
##
## Model 1: Score ~ 1
## Model 2: Score ~ Health + GDP + Freedom + Sociality
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    146 17117.9
## 2    142  3840.1  4     13278 122.75 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In both cases the p-values are extremely low, therefore there's strong evidence against the null hypothesis: the group composed by the 4 predictors and the one composed by Health and GDP cannot be discarded. The model we used so far should be kept.

Prediction [point 13]

To make a prediction, I need a new data first: a Country called Utopia. Life expectancy is around 100 years, the logarithm of the GDP is 12 (so it's wealthy), 100% of people is free to make life choices and 100% of people has a least a loved one or a friend to rely on.

```
x.new <- data.frame(Health = 90,GDP = 12,Freedom = 100, Sociality = 100)

predicted_score = predict.lm(ols, x.new, interval="confidence")
print(predicted_score)
```

```
##        fit      lwr      upr
## 1 81.89738 78.23575 85.55902
```

The predicted Happiness Score is 81.89%. The variability of this number is obviously 5.2%, which is the Residual Standard Error we computed before. Therefore the 95% Confidence Interval is [78.23 , 85.55]