IQ_research

Taras Svystun, Dmytro Kalitin, Yaroslav Tsymbalista

12/24/2021

IQ prediction

Overview

In this mini-research project we decided to discover the dependencies between IQ and human factors (gender, country, education, urban status). Also we were looking for links between some random questions (I prefer to be barefoot or I had an imaginary friend as a child) and IQ.

Libraries

```
library(ggplot2)
library(dplyr)
library(ggpubr)
library(e1071)
library(fitdistrplus)
```

Reading the data

```
df <- read.csv(".\\cleaned.csv", sep = "\t")
df <- df[df$age < 94,]
df <- df[df$S27 > 0.5,]
head(df, 10)
```

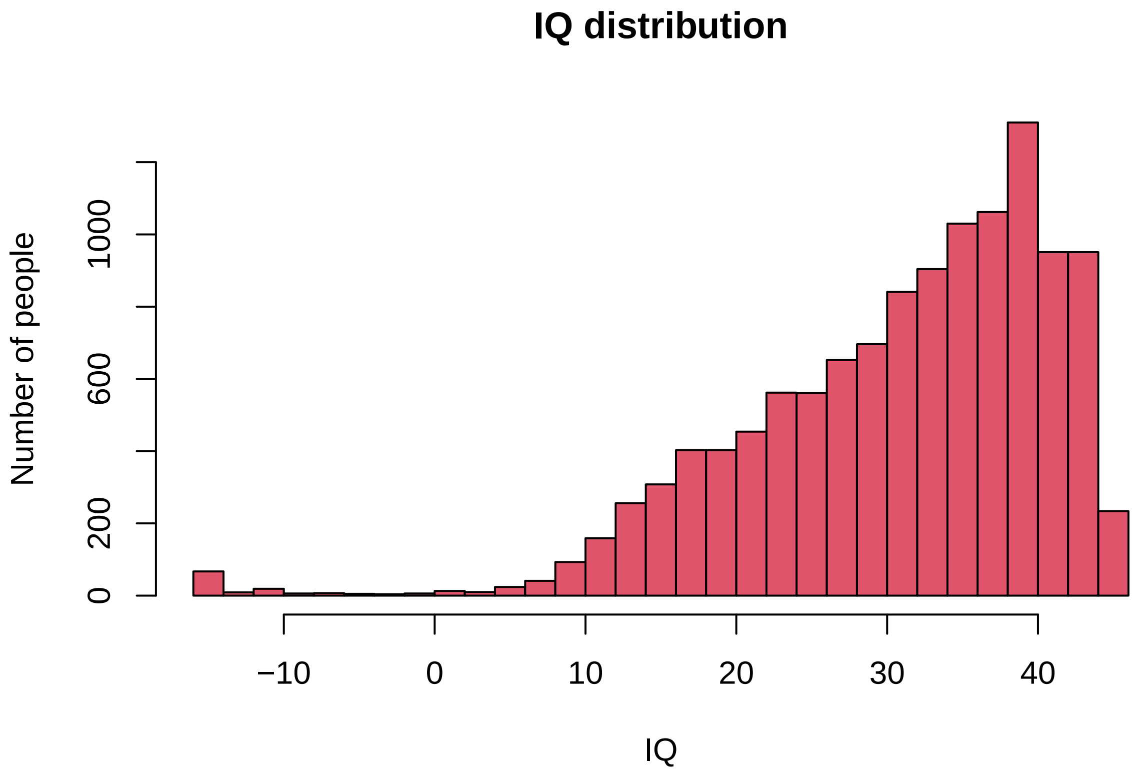
```
##    score_full country S2 S5 S26 S27 education age
## 1       25.25      US  2  4   2   5         3  54
## 2       20.80      US  4  4   3   1         2  16
## 3       32.55      US  5  1   4   5         4  32
## 4       16.65      IN  5  5   1   1         4  24
## 5       39.65      US  2  1   3   3         3  63
## 6       31.20      US  4  1   0   2         3  51
## 7       16.25      AU  4  1   4   4         2  40
## 8       19.45      CH  2  3   3   3         4  45
## 9       43.65      GB  5  1   3   1         3  71
## 10      32.25      US  4  4   4   2         1  14
```

```
IQ <- df$score_full
country <- df$country
education <- df$education
age <- df$age
s2 <- df$S2
s5 <- df$S5
s26 <- df$S26
s27 <- df$S27
```

What is s2, s5, s26, s27? That's random questions from the survey. S2 I avoid contacts with others. S5 I would be interested in getting my fortune told. S26 I always do the bare minimum I need to get by. S27 I like to play devils advocate. Respondents answered to them in range (1=Disagree, 5=Agree). Based on them we will show you unobvious dependency.

Quick data overview

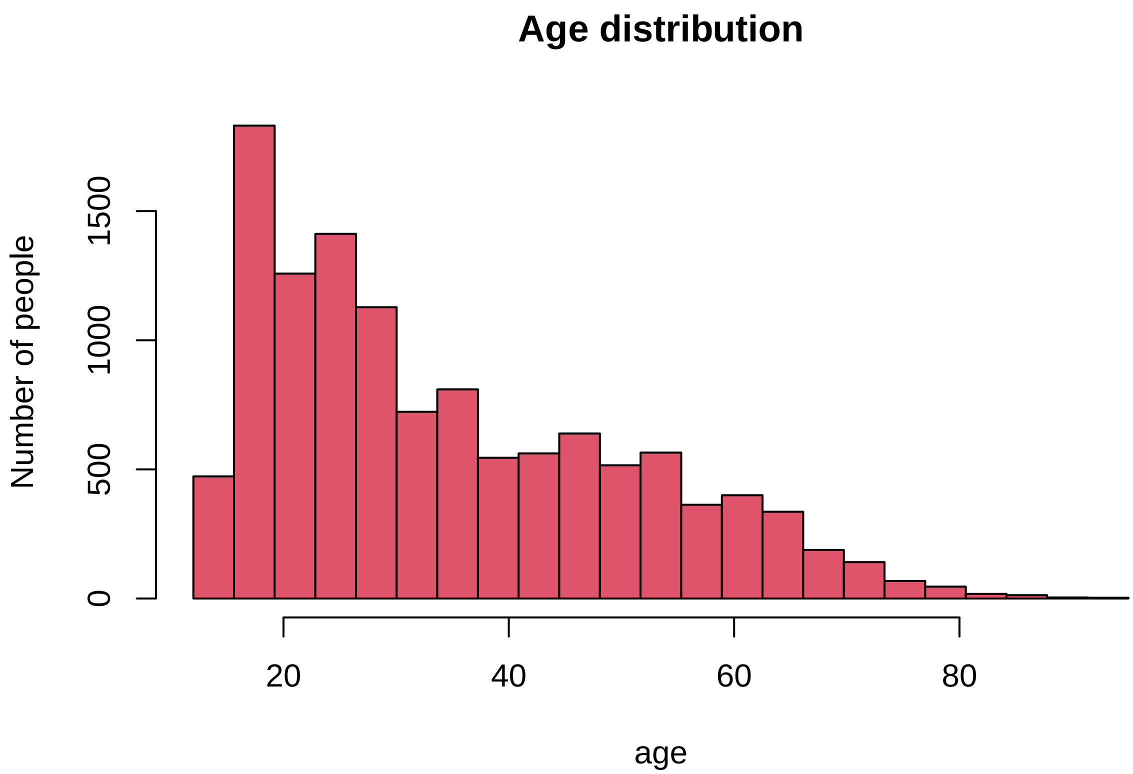
```
set.seed(2021)
hist(IQ, breaks = 24, main = "IQ distribution", col = "666", ylab = "Number of people")
```



```
summary(age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.00   21.00   30.00   34.52   46.00   93.00
```

```
hist(age, breaks = seq(12, 95, length = 24), main = "Age distribution", col = "666", ylab = "Number of
```



8 most popular countries in this data

```
tail(sort(table(country)), 8)
```

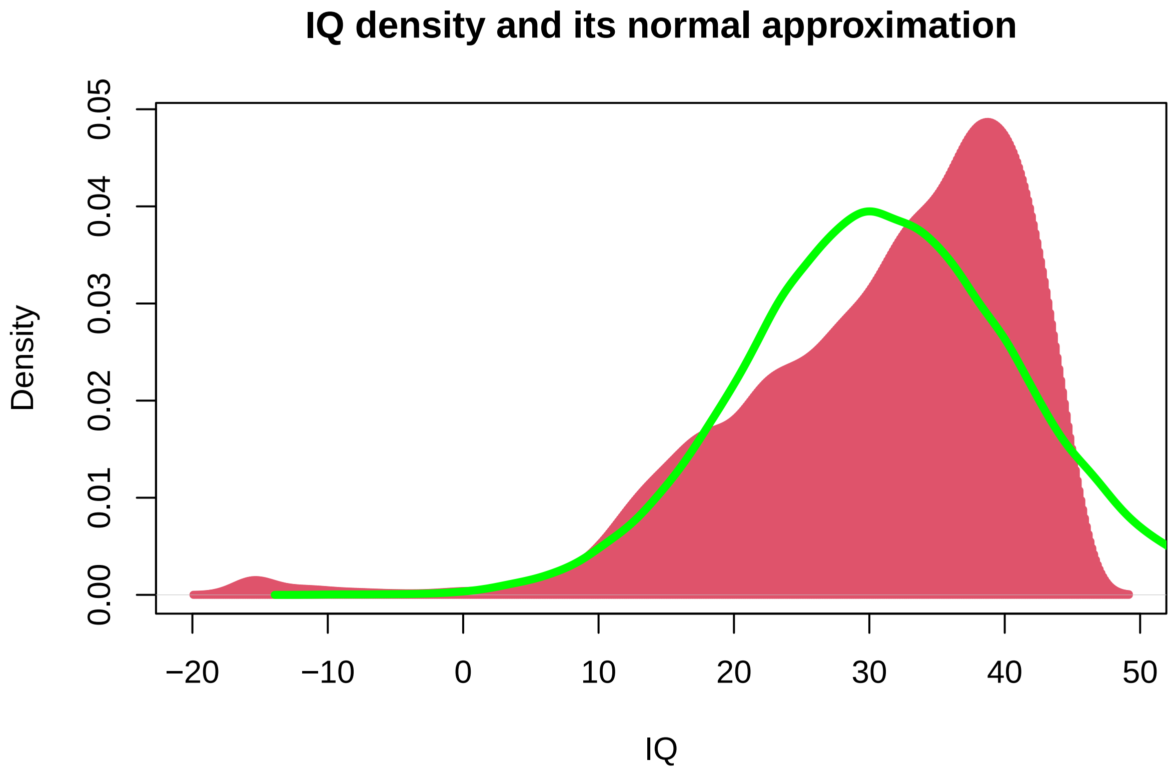
```
## country
##   NZ   DE   PH   IN   AU   CA   GB   US
##  124  184  202  232  534 1001 1186 6394
```

IQ distribution

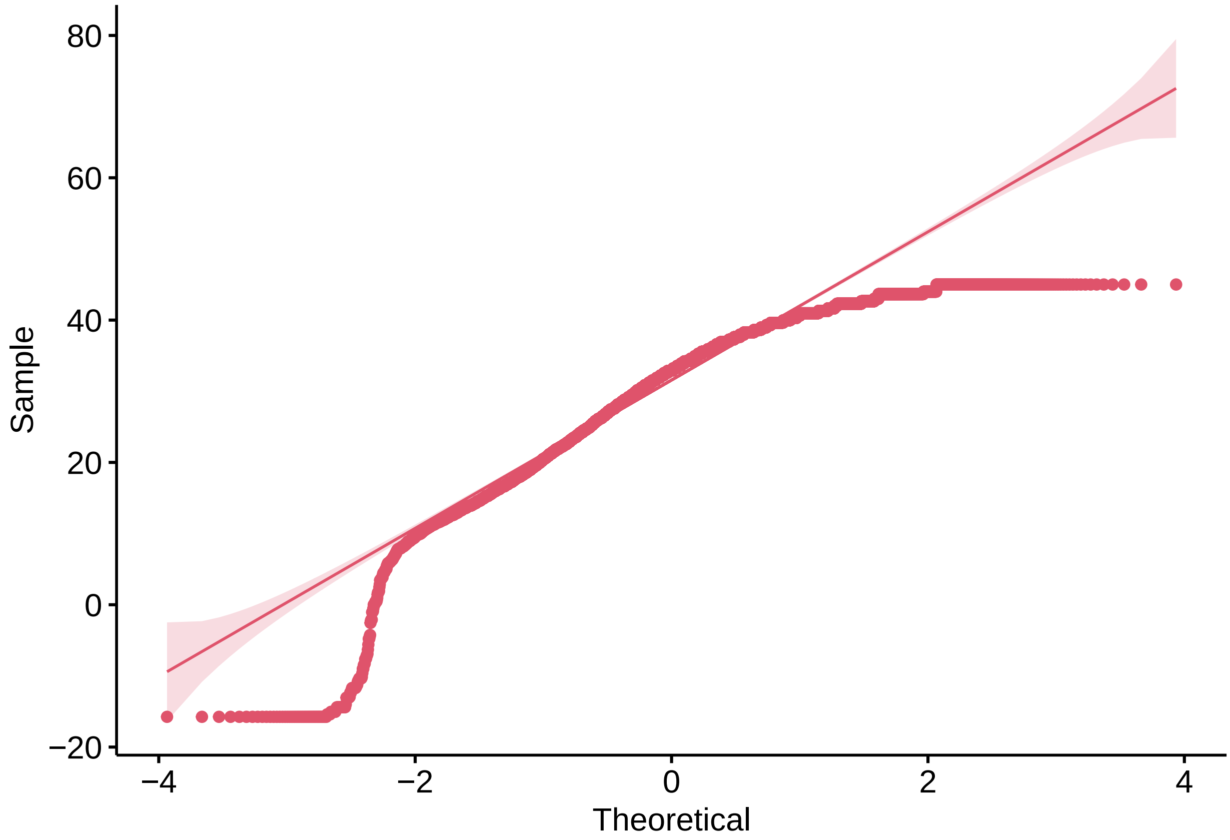
Now it's time to analyze the IQ distribution, especially whether it could be normalized.

```
set.seed(2021)
plot(density(IQ), main = "IQ density and its normal approximation",
     col="666", lwd=4, type="h", xlab="IQ")

normal <- rnorm(12000, mean = mean(IQ), sd=sd(IQ))
lines(density(normal),col="green", lwd=4)
```



```
ggqqplot(IQ, color = "666")
```



```
ks.test( IQ, normal, alternative = "t" )
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  IQ and normal
## D = 0.087307, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
shapiro.test(head(IQ, 5000))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  head(IQ, 5000)
## W = 0.94544, p-value < 2.2e-16
```

```
skewness(IQ)
```

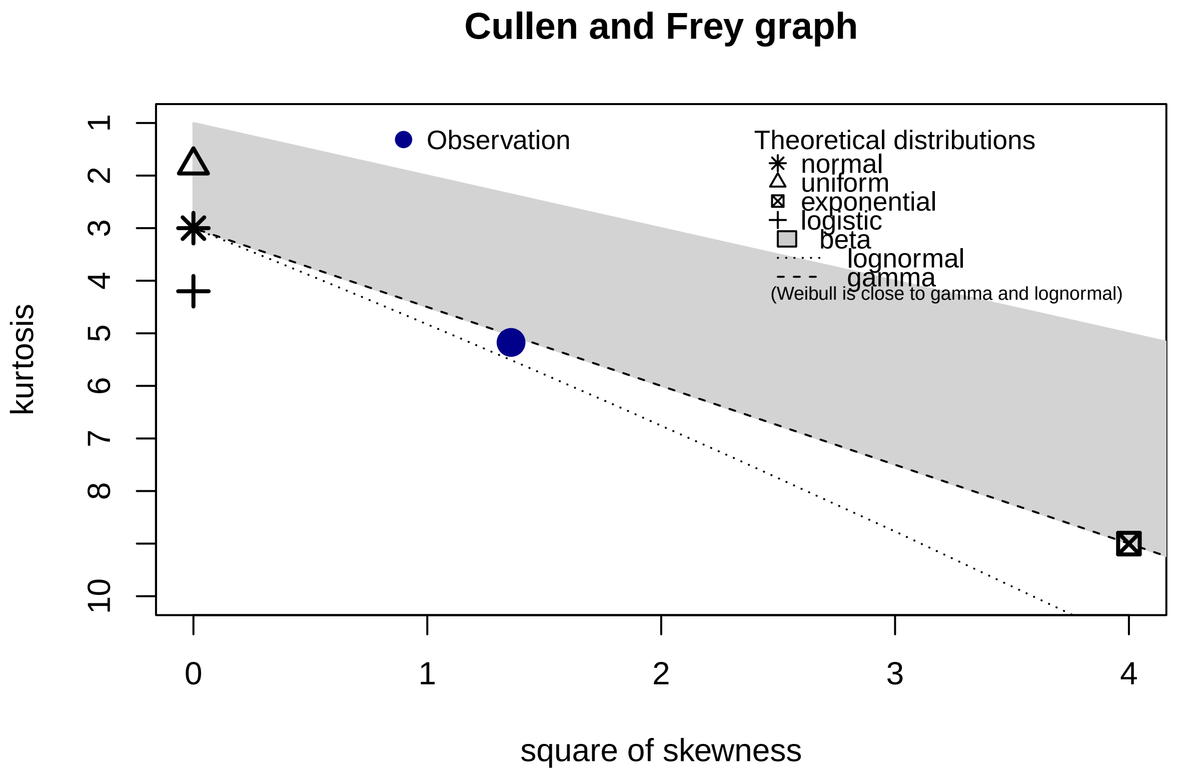
```
## [1] -1.165102
```

As we can see from the graphs, IQ fluctuates from the normal below -10 and after 47. It could be seen in 2 different ways. First one is the ecdf plot. We see that IQ distribution has heavy left tail, which means negative skewness. We proved that by calculating skew. So it drops quickly to the right hand side. It also has some values below -10, which is not observed for normal distribution. Second one is by using qq plot, which shows the correlation between a given sample and the normal distribution. Two tests approve our upper words. In both cases (Shapiro test and Kolmogorov-Smirnov test) the p_value is almost 0, so we cannot consider IQ as normally distributed data.

Skewness-kurtosis graph for the choice of distributions

The function descdist() provides a skewness-kurtosis graph to help to choose the best candidate(s) to fit a given dataset. If we want to use it for discrete distributions we may use argument discrete=TRUE

```
descdist(IQ)
```



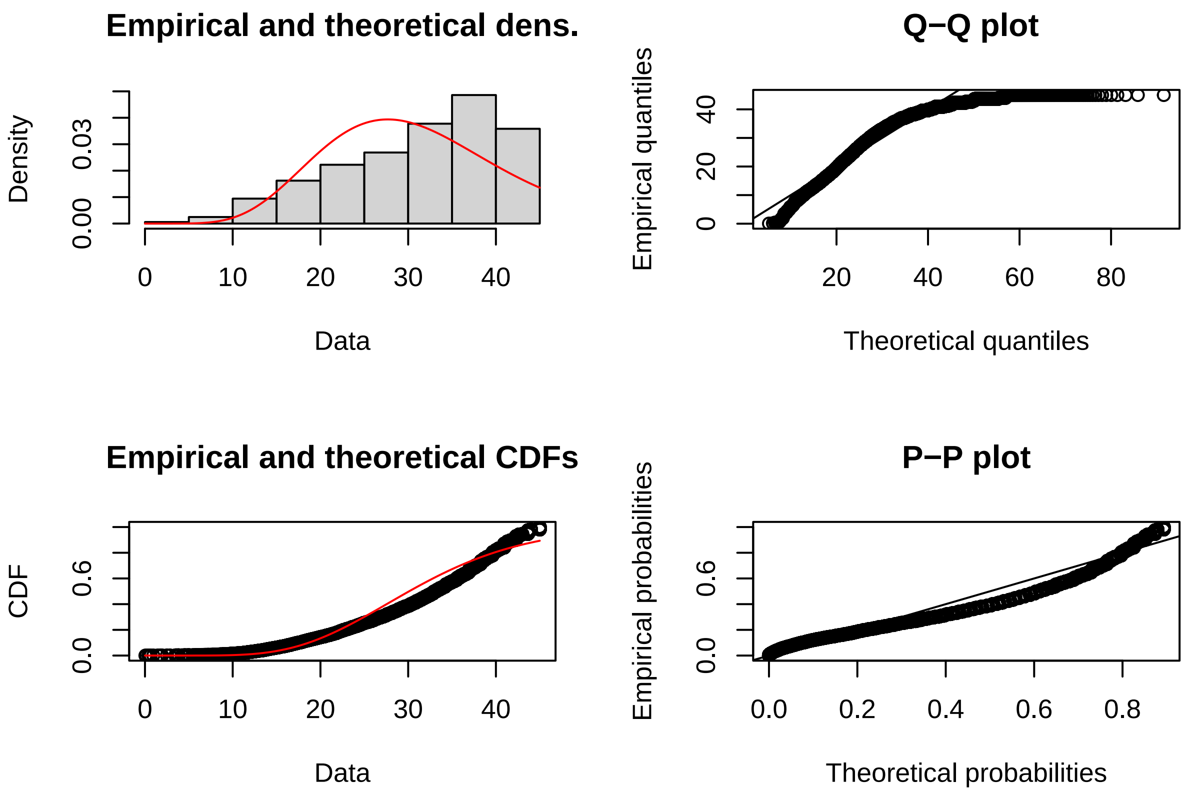
```
## summary statistics
## ------
## min:  -15.75   max:  45
## median:  32.9
## mean:  30.8637
## estimated sd:  10.12304
## estimated skewness:  -1.165392
## estimated kurtosis:  5.173669
```

We see that our IQ is close to Gamma distribution. Let's check this.

```
IQ.positive <- IQ[IQ > 0]
fit.gamma <- fitdist(IQ.positive, distr="gamma", method="mle")
fit.gamma
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters:
##        estimate  Std. Error
## shape 8.6283251 0.109660147
## rate  0.2755992 0.003606489
```

```
plot(fit.gamma)
```



As our testing shows us, MLE method works the best and IQ distribution is pretty close to Gamma distribution

Finding the correlations

In order to search for any dependency we used the following methods: Pearson, Kendall, Spearman.

Pearson correlation test

```
s2.cor1 <- cor.test(IQ, s2, method = "pearson")
s5.cor1 <- cor.test(IQ, s5, method = "pearson")
s26.cor1 <- cor.test(IQ, s26, method = "pearson")
s27.cor1 <- cor.test(IQ, s27, method = "pearson")
age.cor1 <- cor.test(IQ, age, method = "pearson")
education.cor1 <- cor.test(IQ, education, method = "pearson")
s2.cor1
```

```
##
##  Pearson’s product-moment correlation
##
## data:  IQ and s2
## t = -2.0575, df = 12040, p-value = 0.03966
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03659701 -0.00088749
## sample estimates:
##         cor
## -0.01874823
```

```
s5.cor1
```

```
##
##  Pearson’s product-moment correlation
##
## data:  IQ and s5
## t = -19.805, df = 12040, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1948678 -0.1602724
## sample estimates:
##       cor
## -0.177625
```

```
s26.cor1
```

```
##
##  Pearson’s product-moment correlation
##
## data:  IQ and s26
## t = -19.64, df = 12040, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1934444 -0.1588310
## sample estimates:
##        cor
## -0.1761922
```

```
s27.cor1
```

```
##
##  Pearson’s product-moment correlation
```

```
##
## data:  IQ and s27
## t = 24.307, df = 12040, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1991891 0.2332407
## sample estimates:
##       cor
## 0.2162806
```

```
age.cor1
```

```
##
##  Pearson's product-moment correlation
##
## data:  IQ and age
## t = 48.333, df = 12040, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3880441 0.4179629
## sample estimates:
##       cor
## 0.4031112
```

```
education.cor1
```

```
##
##  Pearson's product-moment correlation
##
## data:  IQ and education
## t = 36.024, df = 12040, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2957084 0.3279559
## sample estimates:
##      cor
## 0.311922
```

As we can see, predicted question2 "I avoid contacts with others" has no correlation with IQ. But also some unexpected results were obtained. Question5 "I would be interested in getting my fortune told." Question26 "I always do the bare minimum I need to get by." Question27 "I like to play devils advocate." These 3 questions correlate with IQ. S27 has possitive cor => the more I go against society - the higher my IQ is. S5 and S26 have negative cor => the more I buy - the higher my IQ is. And the more I am interested in my fortune - the higher my IQ is. And 2 obvious correlations: age and education. Both have possitive cor => the higher education or age - the higher IQ.

Kendall rank correlation test

```
s27.cor2 <- cor.test(IQ, s27, method = "kendall")
age.cor2 <- cor.test(IQ, age, method = "kendall")
education.cor2 <- cor.test(IQ, education, method = "kendall")
s27.cor2
```

```
##
##  Kendall's rank correlation tau
##
## data:  IQ and s27
## z = 22.046, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.1482618
```

```
age.cor2
```

```
##
##  Kendall's rank correlation tau
##
## data:  IQ and age
## z = 54.34, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.3350525
```

```
education.cor2
```

```
##
##  Kendall's rank correlation tau
##
## data:  IQ and education
## z = 39.887, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.2756653
```

As we can see, similar results to Pearson tests.

Spearman rank correlation coefficient

```
s27.cor2 <- cor.test(IQ, s27, method = "spearman")
age.cor2 <- cor.test(IQ, age, method = "spearman")
education.cor2 <- cor.test(IQ, education, method = "spearman")
s27.cor2
```

```
##
##  Spearman's rank correlation rho
##
## data:  IQ and s27
## S = 2.3275e+11, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2002774
```

```
age.cor2
```

```
##
##  Spearman's rank correlation rho
##
## data:  IQ and age
## S = 1.5028e+11, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.4836376
```

```
education.cor2
```

```
##
##  Spearman's rank correlation rho
##
## data:  IQ and education
## S = 1.8745e+11, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3559014
```

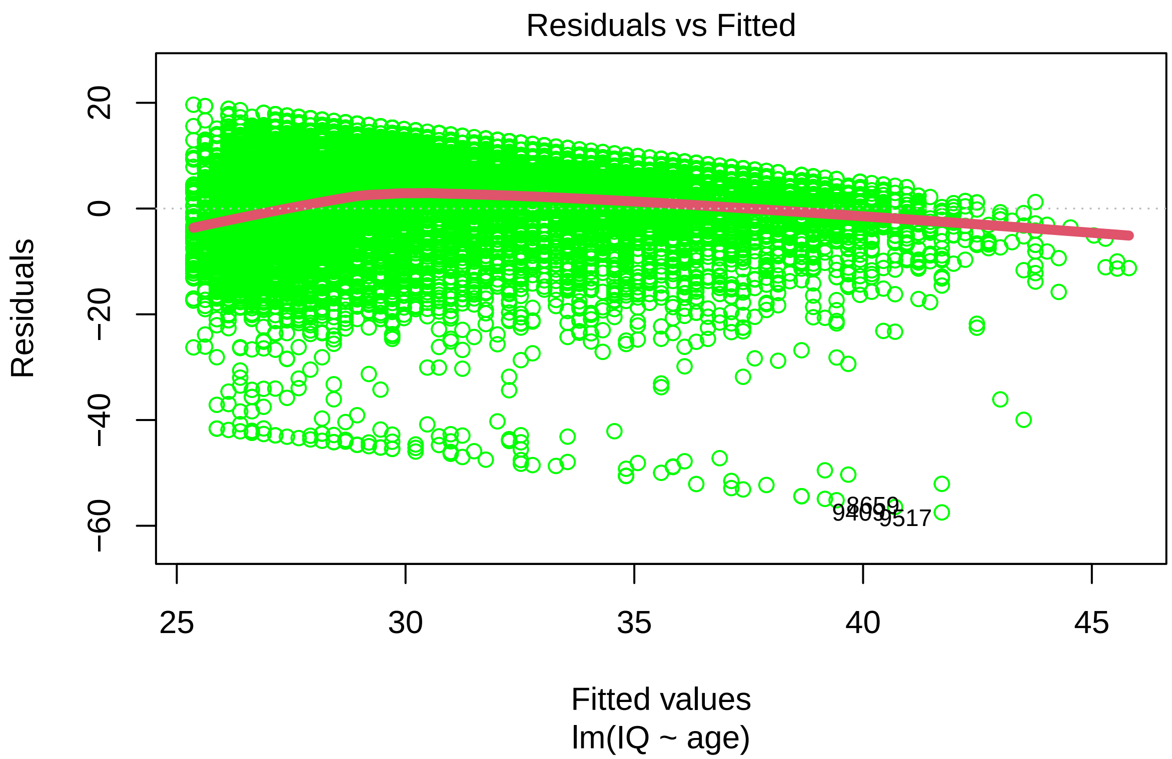
Regression analysis

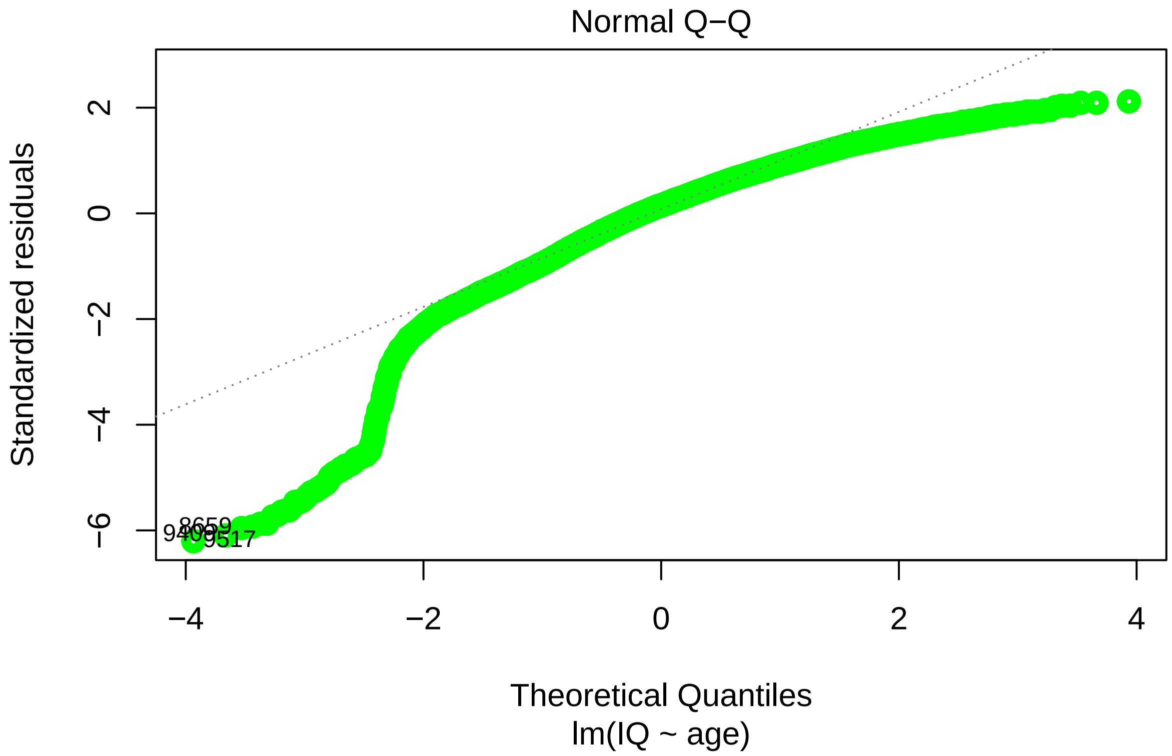
Now it's time to build some regressions. At the beginning we'll try to build regression model on devil's advocate question. One essential notion is how to measure the goodness of our model. Answer: by residuals or better to say determination coefficient. One measure very used to test how good is your model is the coefficient of determination or R^2. This measure is defined by the proportion of the total variability explained by the regression model. $Rˆ2 = $Explained Variation of the model / Total variation of the model.

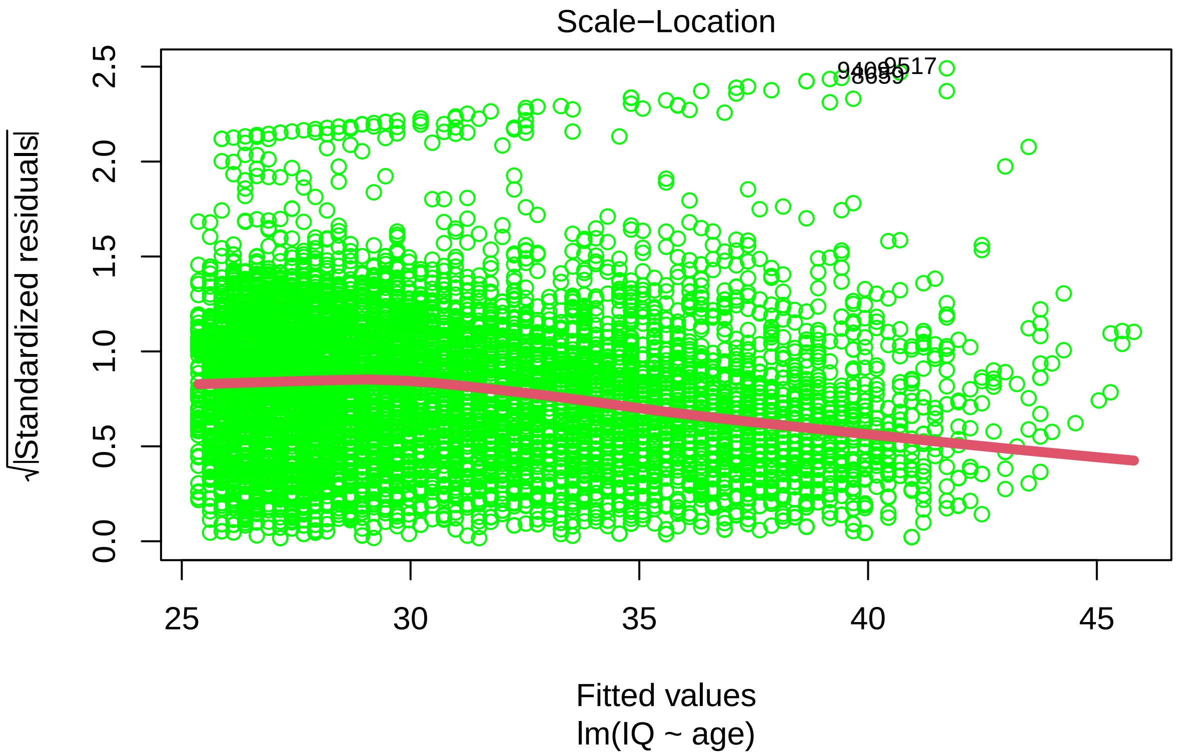
```
age.linear <- lm(IQ~age)
summary(age.linear)
```

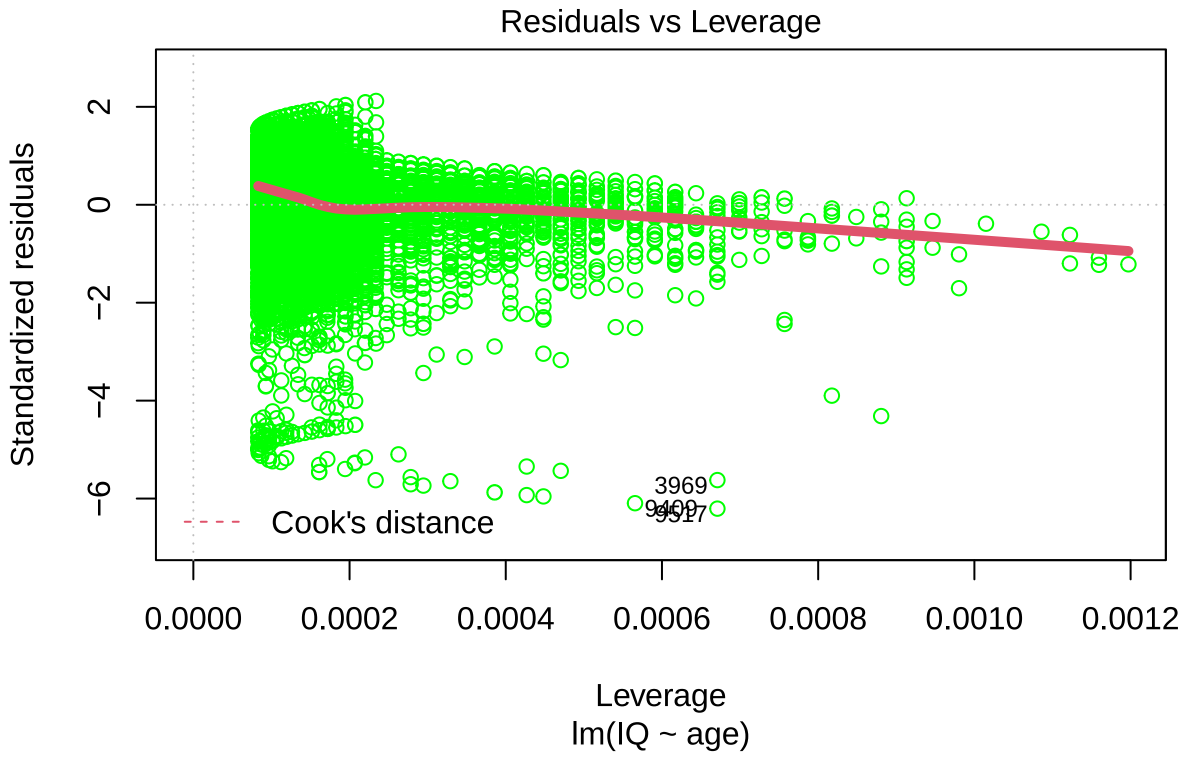
```
##
## Call:
## lm(formula = IQ ~ age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.471  -5.059   1.397   6.470  19.637
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.040747   0.201122  109.59   <2e-16 ***
## age          0.255589   0.005288   48.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.264 on 12040 degrees of freedom
## Multiple R-squared:  0.1625, Adjusted R-squared:  0.1624
## F-statistic:  2336 on 1 and 12040 DF,  p-value: < 2.2e-16
```

```
plot(age.linear, col=c("green"), lwd = 5)
```




Normal Q−Q
Standardized residuals
2
0
−2
−4
−6
8659
9409
9517
−4
−2
0
2
4
Theoretical Quantiles
lm(IQ ~ age)


Scale−Location
9409
8659
9517
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
25
30
35
40
45
Fitted values
lm(IQ ~ age)


Residuals vs Leverage
Standardized residuals
2
0
−2
−4
−6
Cook's distance
3969
9409
9517
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012
Leverage
lm(IQ ~ age)