Assignment1

Philip & Philipp

23 Feb 2016

```
# Load packages and create BibTeX file for R-packages
PackagesUsed <- c("ggplot2", "repmis", "doBy")

setwd("~/Dropbox/4_Spring_2016/2_Collaborative Data Analysis/PandP_Ass1")

# Load PackagesUsed and create .bib BibTeX file
repmis::LoadandCite(PackagesUsed, file = "Ass1Packages.bib", install = FALSE)
```

```
## Loading required package: survival
```

Analysis of the dataset 'occupationalStatus'

The dataset consists of a contingency table between the occupational status measured on an 8-point scales for fathers and their sons.

```
dist_combined <- rbind(dist_fathers, dist_sons)
barplot(dist_combined,
        col=c("navyblue", "darkkhaki"),
        main="Distribution of occupational status among fathers and sons",
        xlab = "Occupational status categories",
        legend = c("Fathers", "Sons"),
        ylim = c(0, 1600),
        beside = TRUE
        )
```

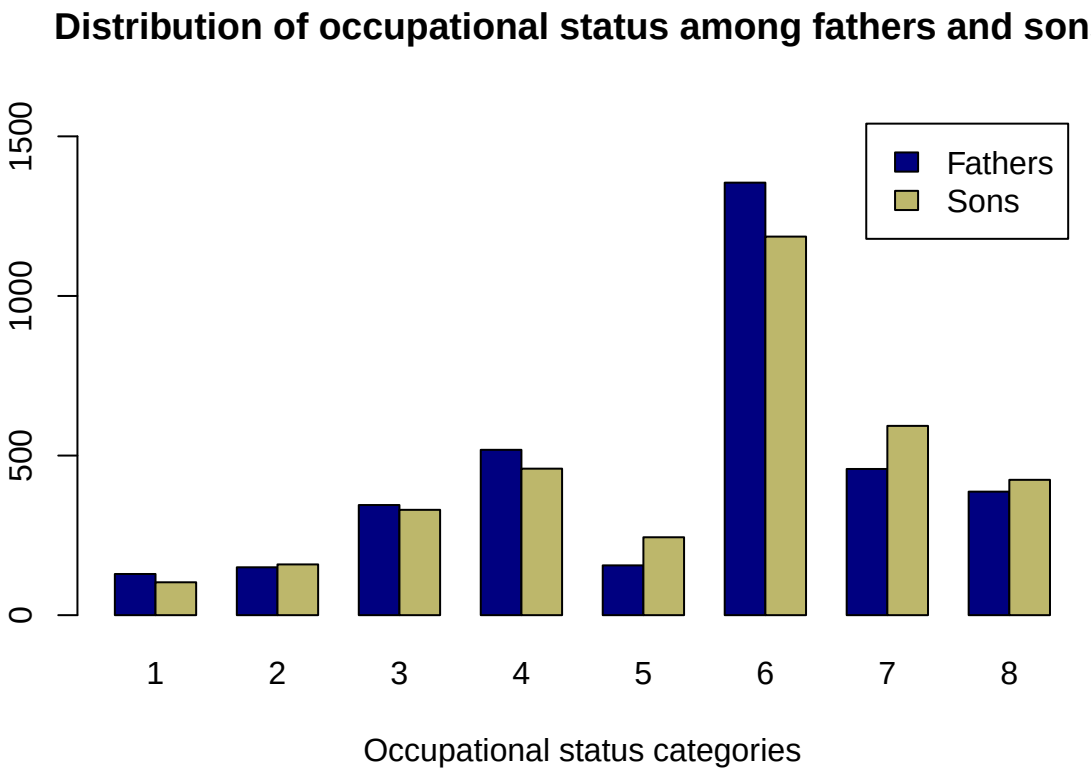


Figure 1 shows the distribution of occupational status for sons and fathers respectively. There does not appear to be any major generational shifts, though sons are slightly overrepresented in occupational status group 7 and 8.

```
knitr::kable(frequency_table, digits = 2)
```

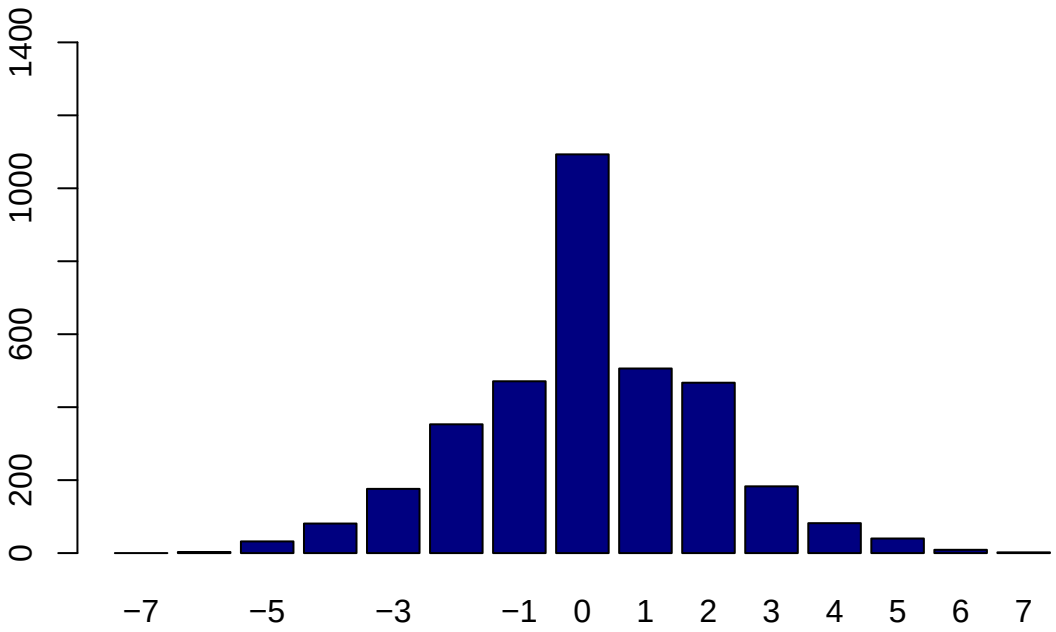
Category	Fathers	Sons
1	0.04	0.03
2	0.04	0.05
3	0.10	0.09
4	0.15	0.13
5	0.04	0.07
6	0.39	0.34
7	0.13	0.17
8	0.11	0.12

Table 1 is equivalent to figure 1 except that it shows the frequency distribution for each status category.

Generational mobility

Figure 1 and table 1 show the overall distribution of occupational status, but is not informative on the extent to which there are generational mobility. Figure 2 shows the distribution of generational mobility. -7 indicates observations where the father had social status 8 and the son 1. As such, a zero is when father and son had the same occupational status.

```
barplot(collapsed$Freq.sum,
        names = collapsed$difference,
        col = "Navyblue",
        ylim = c(0, 1400)
        )
```



Analysis of the dataset 'LifeCycleSavings'

Inspecting the data

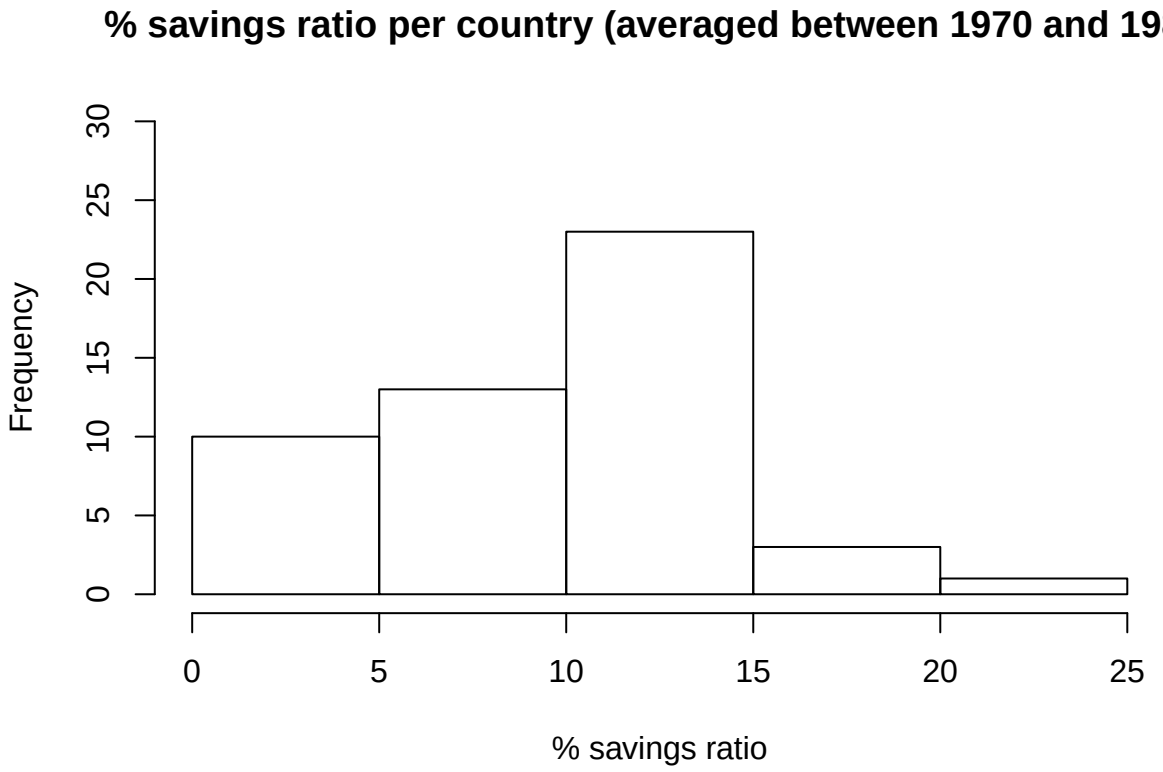
Table of Summary Statistics

```
knitr::kable(res_df, digits = 2)
```

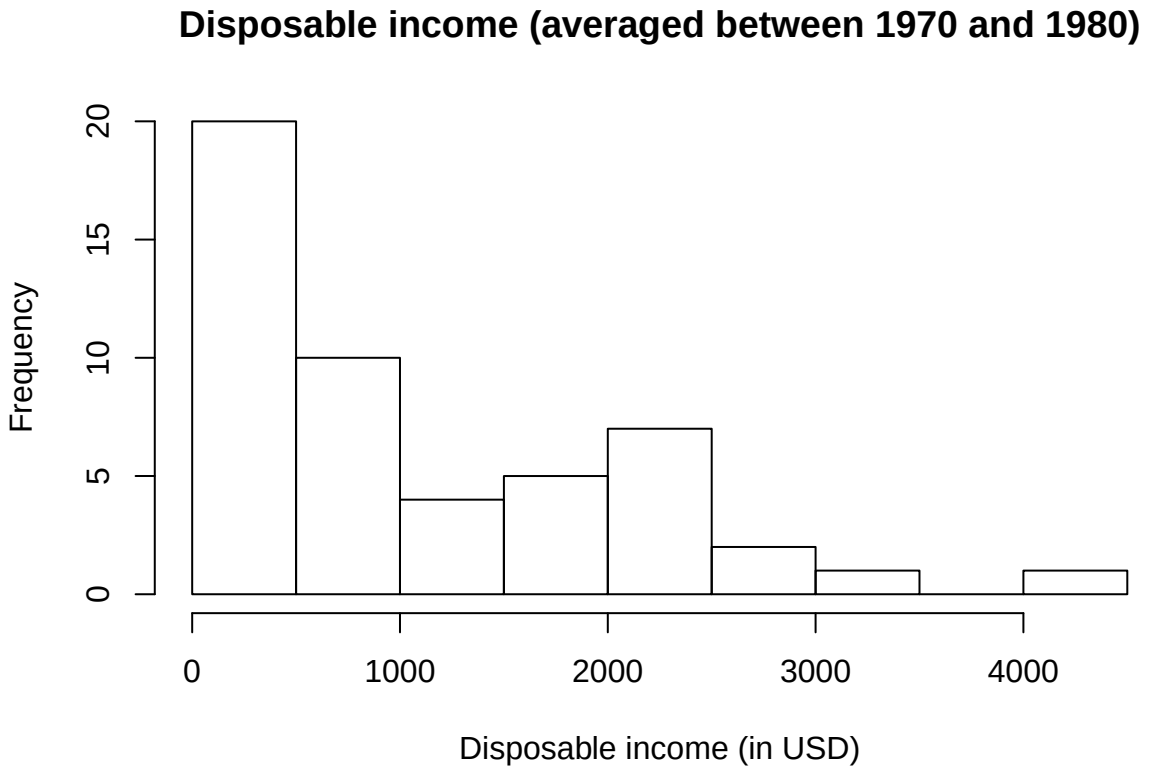
	sr	pop15	pop75	dpi	ddpi
mean	9.67	35.09	2.29	1106.76	3.76
sd	4.48	9.15	1.29	990.87	2.87
median	10.51	32.58	2.17	695.66	3.00
minimum	0.60	21.44	0.56	88.94	0.22
maximum	21.10	47.64	4.70	4001.89	16.71
s.size	50.00	50.00	50.00	50.00	50.00

Looking for skewed data

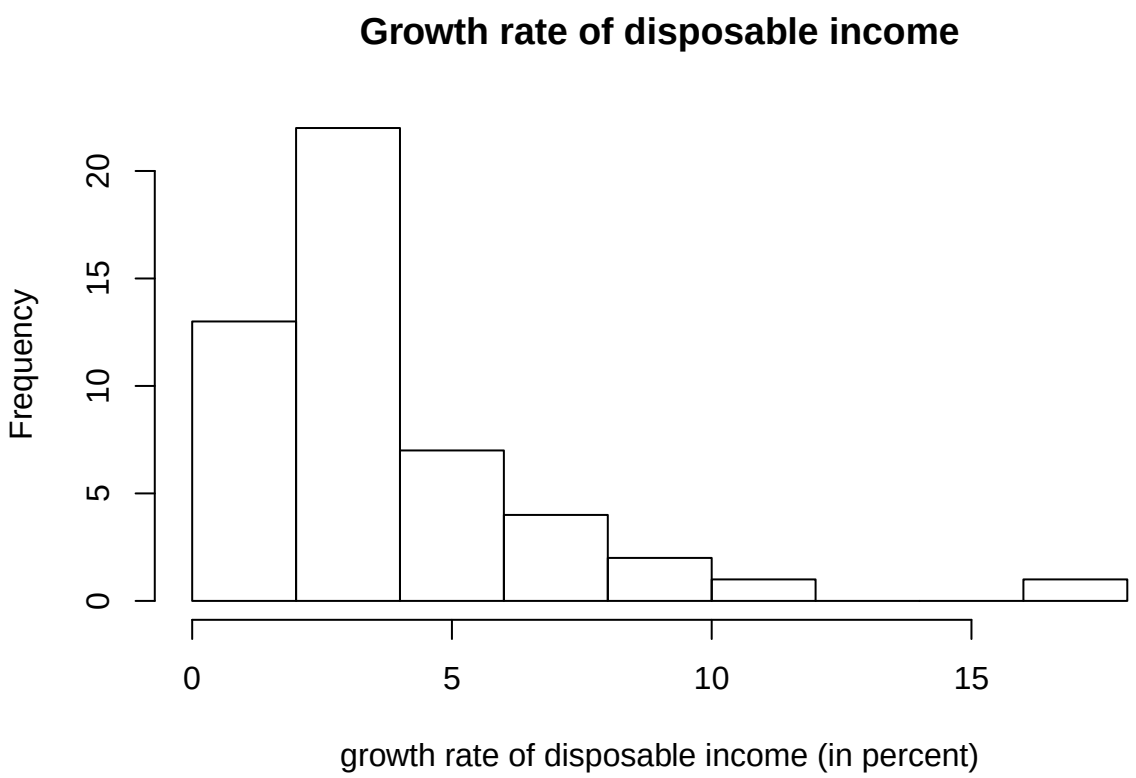
```
hist(LifeCycleSavings$sr,
     ylim = c(0,30),
     main = '% savings ratio per country (averaged between 1970 and 1980)',
     xlab = '% savings ratio',
     ylab = 'Frequency')
```




```
hist(LifeCycleSavings$dpi,
     ylim=c(0,20),
     main = 'Disposable income (averaged between 1970 and 1980)',
     xlab = 'Disposable income (in USD)')
```



```r
hist(LifeCycleSavings$ddpi, main = 'Growth rate of disposable income',
     xlab = 'growth rate of disposable income (in percent)')
```



The histrograms reveal that *savings ratio* is only slightly left-skewed, while *disposable income* and *growth of disposable incom* are strongly right-skewed. To correct for this, we will take the natural log of the latter two variables.

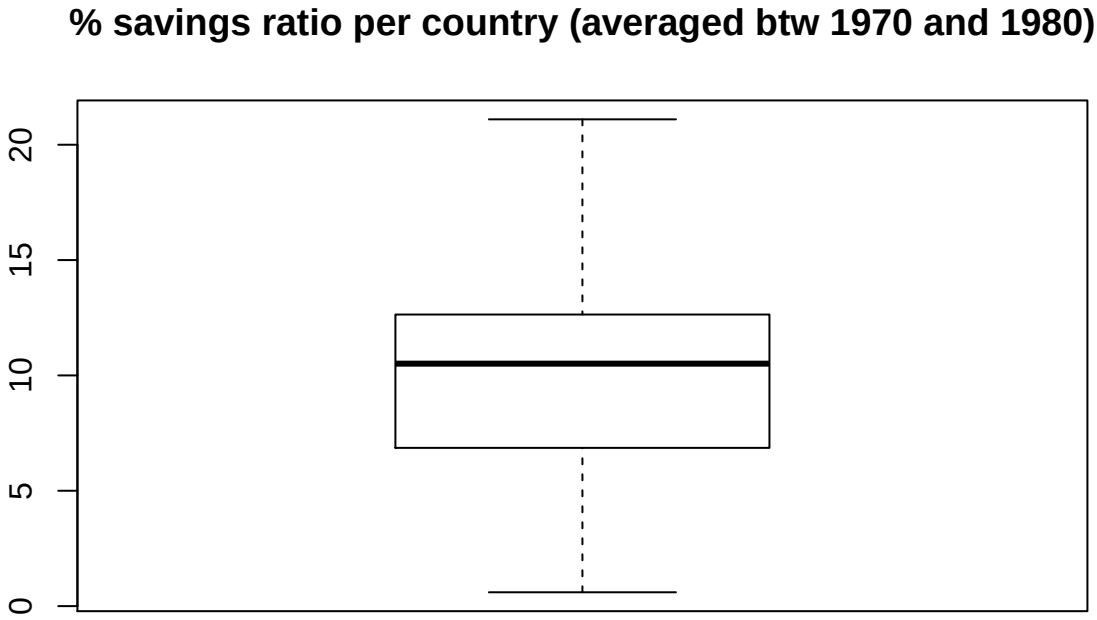
Additionally we create dummies which split the data set in above the median and below the median for the variables *percent of population under 15 (pop15)*, *precent of population over 75 (pop75)*, as well as for *disposable income* and *growth of disposable income*.

Analyzing the Data

Data Distribution

First, we take a closer look at the dependent variable *savings rate*:

```r
boxplot(combined_df$sr, main = '% savings ratio per country (averaged btw 1970 and 1980)')
```



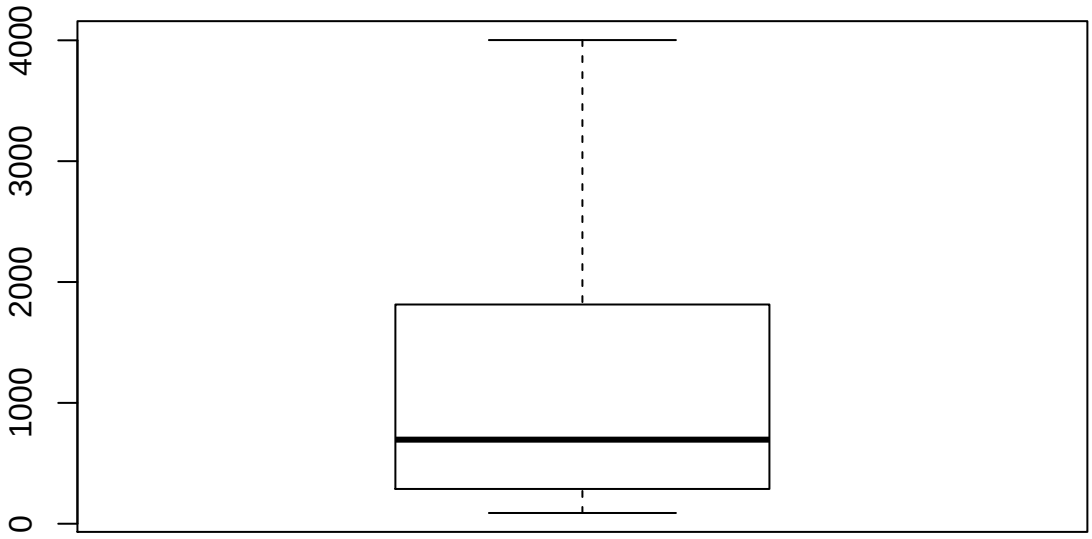

```
summary(combined_df$sr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   6.970  10.510   9.671  12.620  21.100
```

The boxplot shows that 75 percent of the country observations have an average savings rate between 6.97 and 12.62 percent. The average savings rate is 9.671 percent (mean).The distribution is slightly left-skewed (higher median than mean).

Second, we inspect the main explanatory variable *disposable income*:

```
boxplot(combined_df$dpi, main = 'Distribution of disposable income (in USD)')
```



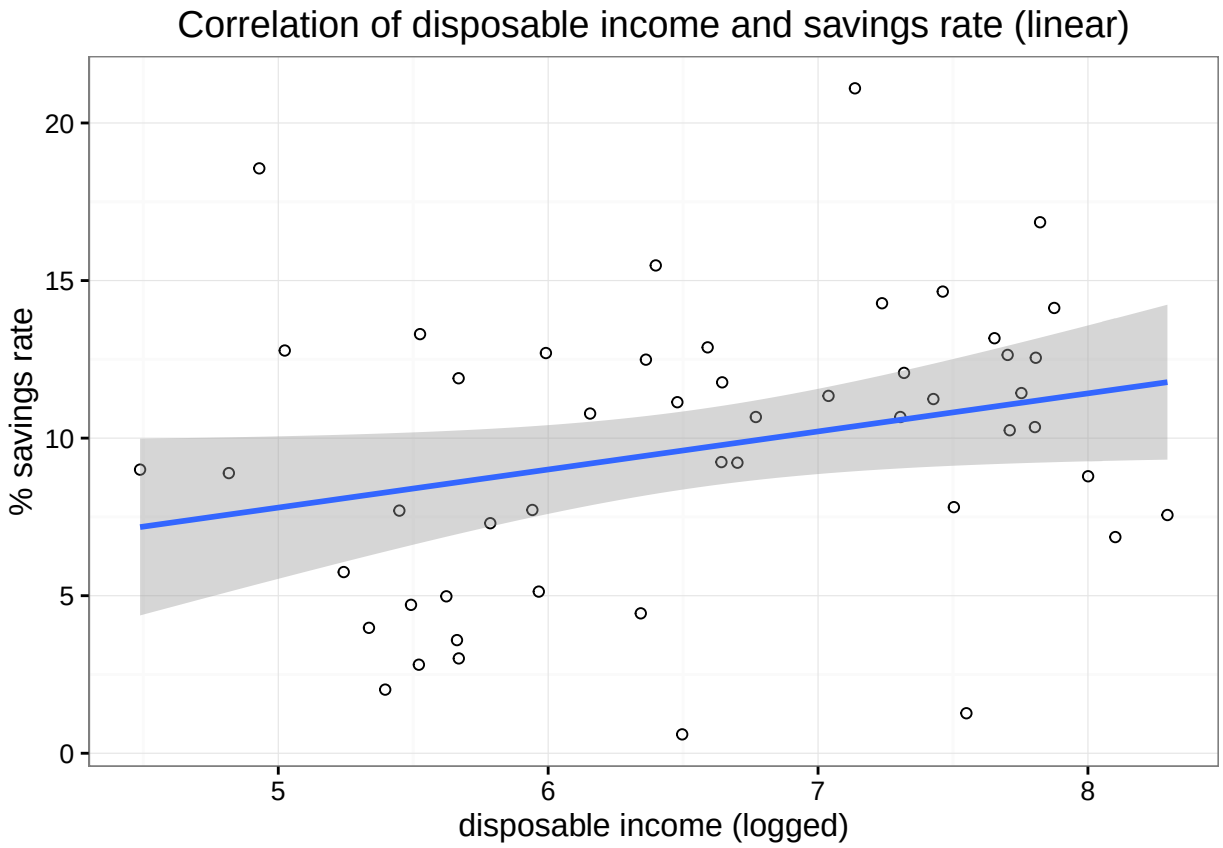
```r
summary(combined_df$dpi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   88.94  288.20  695.70 1107.00 1796.00 4002.00
```

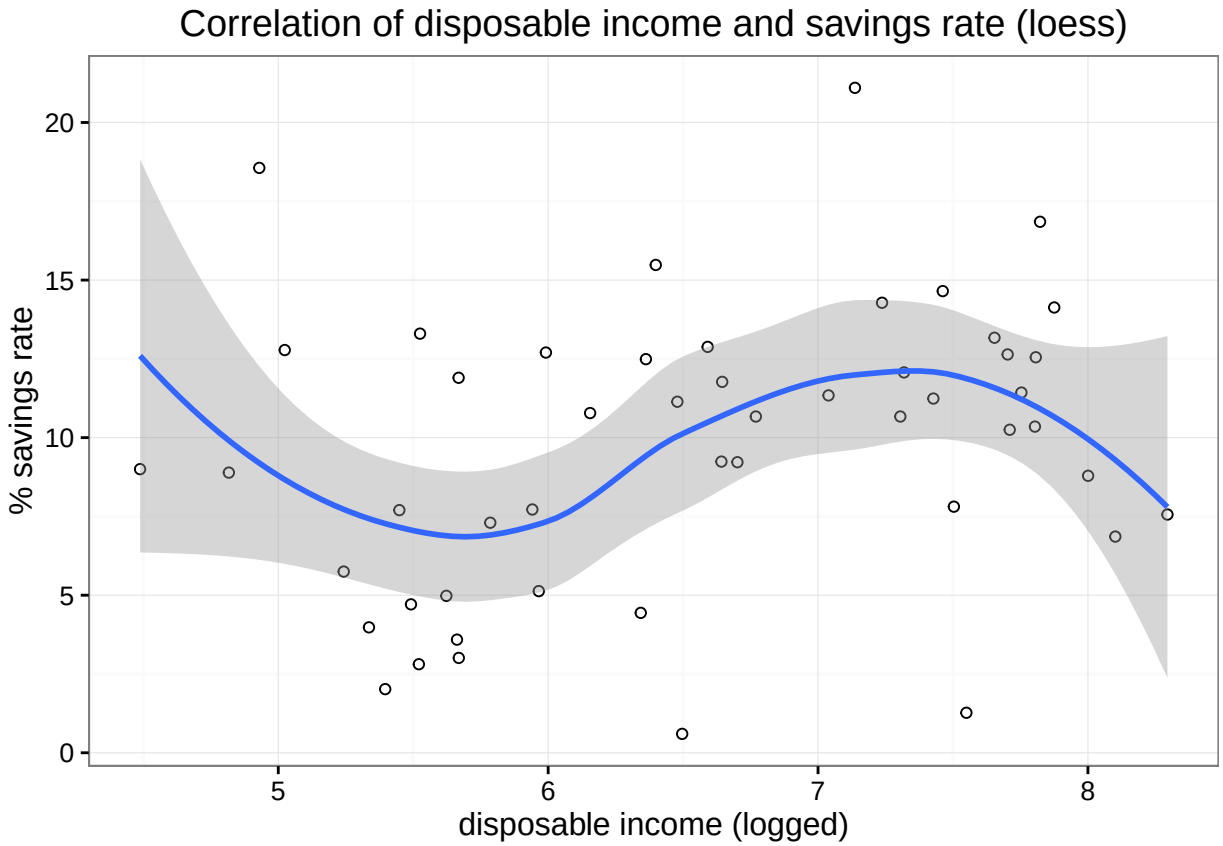
The boxplot shows that 75 percent of the country observations have an average disposable income between 288,2 and 1796 USD a month. The average disposable income is 1106.7584 USD. The distribution is strongly right-skewed (higher mean than median). Therefore, we will use the log-transformation.

Are savings rates and disposable income correlated?

```r
# Plot out the variables
ggplot2::ggplot(combined_df, aes(ln.dpi, sr))+
  geom_point(shape=1) + geom_smooth(method=lm) + theme_bw() +
  labs(title = "Correlation of disposable income and savings rate (linear)",
       x = "disposable income (logged)", y = "% savings rate")
```




```r
# Using loess instead of linear regression line
ggplot2::ggplot(combined_df, aes(ln.dpi, sr)) +
  geom_point(shape=1) + geom_smooth() + theme_bw() +
  labs(title = "Correlation of disposable income and savings rate (loess)",
       x = "disposable income (logged)", y = "% savings rate")
```



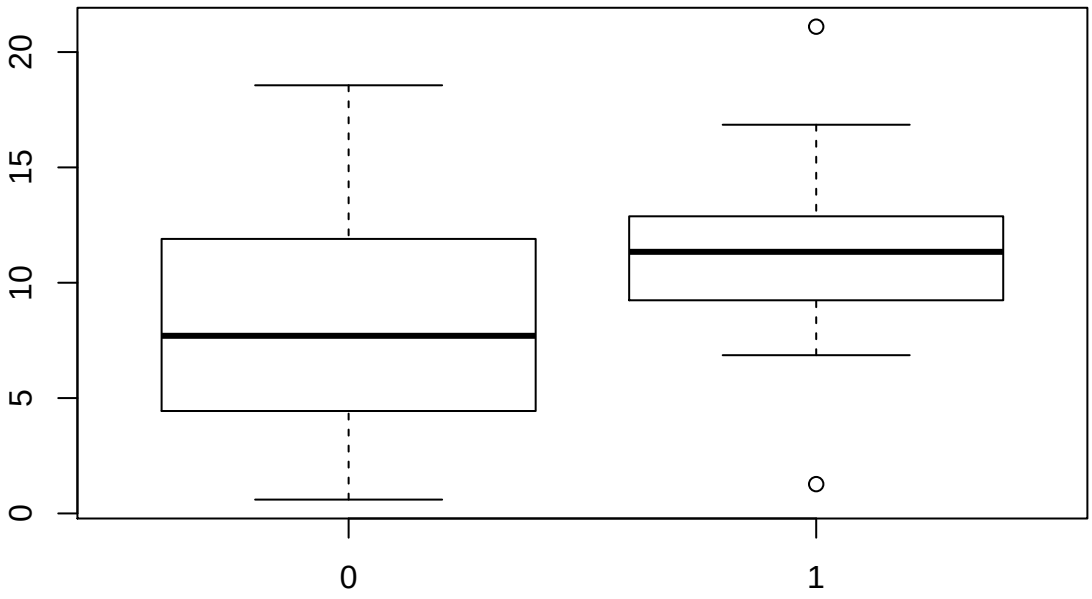

```
# Testing for correlation
cor.test(combined_df$ln.dpi, combined_df$sr)
```

```
##
##  Pearson's product-moment correlation
##
## data:  combined_df$ln.dpi and combined_df$sr
## t = 1.9922, df = 48, p-value = 0.05206
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.002166184  0.515075795
## sample estimates:
##       cor
## 0.2763482
```

The analysis shows that disposable income is postively associated with savings rates. The correlation coefficient of 0.28 is however not statistically significant at the 95% level. The second plot suggests a sinus curve relationship.

In order to understand the correlation better we assess the distribution with a splitted sample into richer and poorer countries (poorer = 0, richer = 1).

```
boxplot(combined_df$sr~combined_df$dum.rich)
```



The boxplot shows that richer countries have a much more compressed distribution that does not extend to very low savings rates. On average richer countries have a higher savings rate.

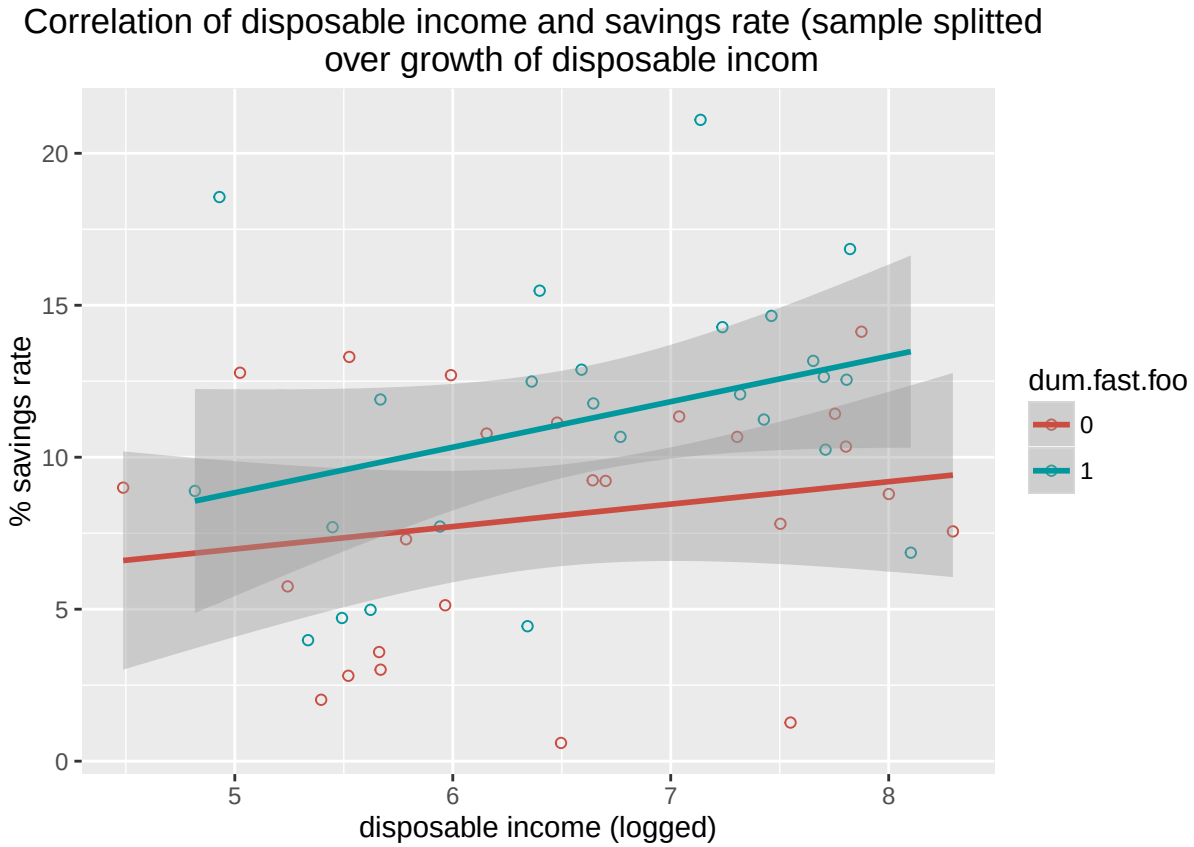
Inspecting the influence of control variables

In a final step we make a graphical inspection of the following control variables: 1. Growth of disposable income 2. Share of old population (above 75)

For the inspection we split the sample again with dummy variables in above and below median groups.

1. Growth of disposable income (slow growth = 0, fast growth = 1)

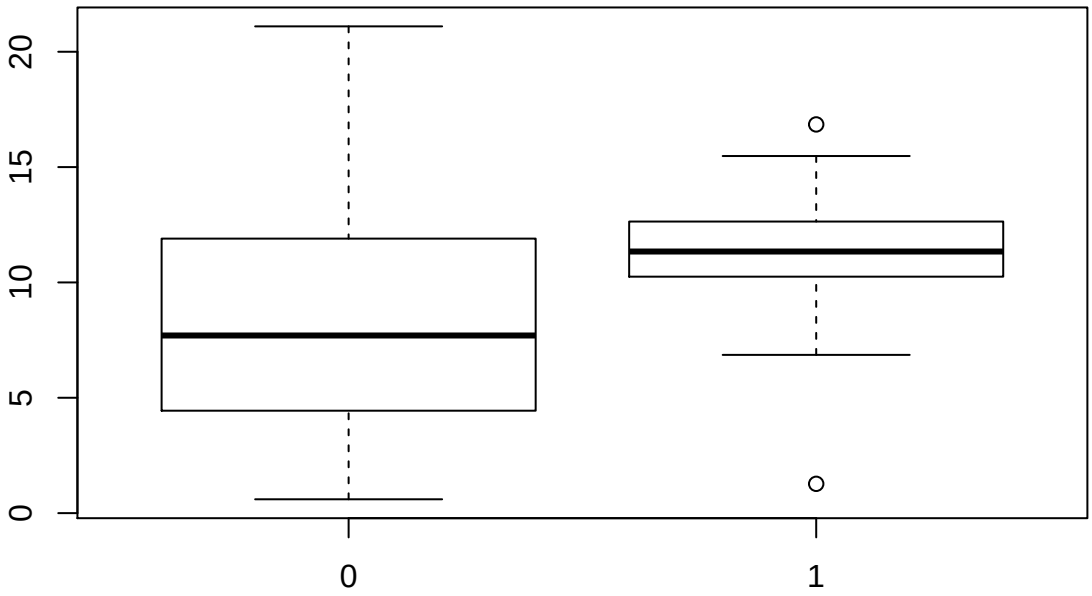
```
ggplot2::ggplot(combined_df, aes(ln.dpi, sr, color=dum.fast.foo)) +
  geom_point(shape=1) + scale_colour_hue(l=50) +
  geom_smooth(method=lm) +
  labs(title = "Correlation of disposable income and savings rate (sample splitted
       over growth of disposable incom",
       x = "disposable income (logged)", y = "% savings rate")
```



Countries with higher growth rates seem to have a slightly stronger correlation between savings rate and disposable income.

2. Share of older population (lower percentage = 0, higher percentage = 1)

```
boxplot(combined_df$sr~combined_df$dum.old)
```



Countries with a larger share of an older population seem to have a more compressed distribution of savings rates. On average savings rates are higher. This resembles the sample split for rich and poor countries.

Packages used (R Core Team 2015), (Højsgaard and Halekoh 2015), (Wickham and Chang 2015) and (Gandrud 2016).

References

Gandrud, Christopher. 2016. *Repmis: Miscellaneous Tools for Reproducible Research*. https://CRAN.R-project.org/package=repmis.

Højsgaard, Søren, and Ulrich Halekoh. 2015. *DoBy: Groupwise Statistics, LSmeans, Linear Contrasts, Utilities*. https://CRAN.R-project.org/package=doBy.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, and Winston Chang. 2015. *Ggplot2: An Implementation of the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.