Assignment1

Philip & Philipp

23 Feb 2016

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

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

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

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.

Figure 1

```
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
        )
```

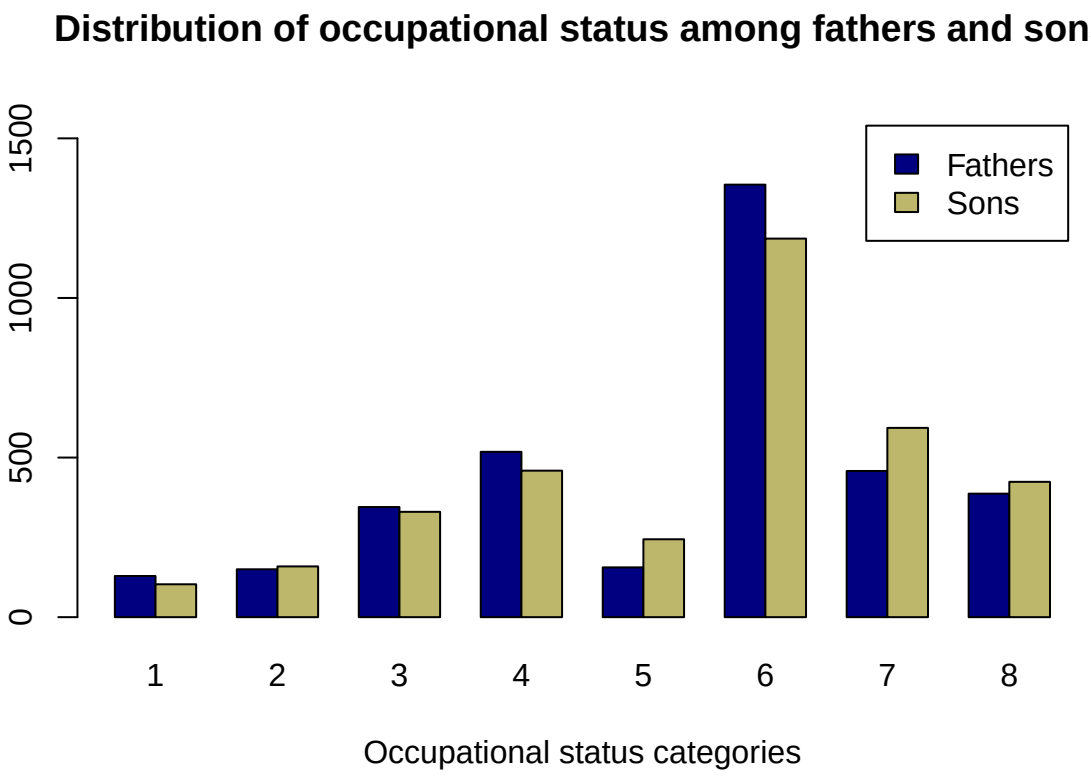


Figure 1 shows the distribution of occupational status for sons and fathers respectively. Overall there has been no major generational shift, but a slightly higher share of sons are in status group 7 and 8 relative to fathers.

Table 1

```
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 based on the same data as figure 1, but rather than showing the absolute number of fathers and sons in each status category, it shows the frequency distribution.

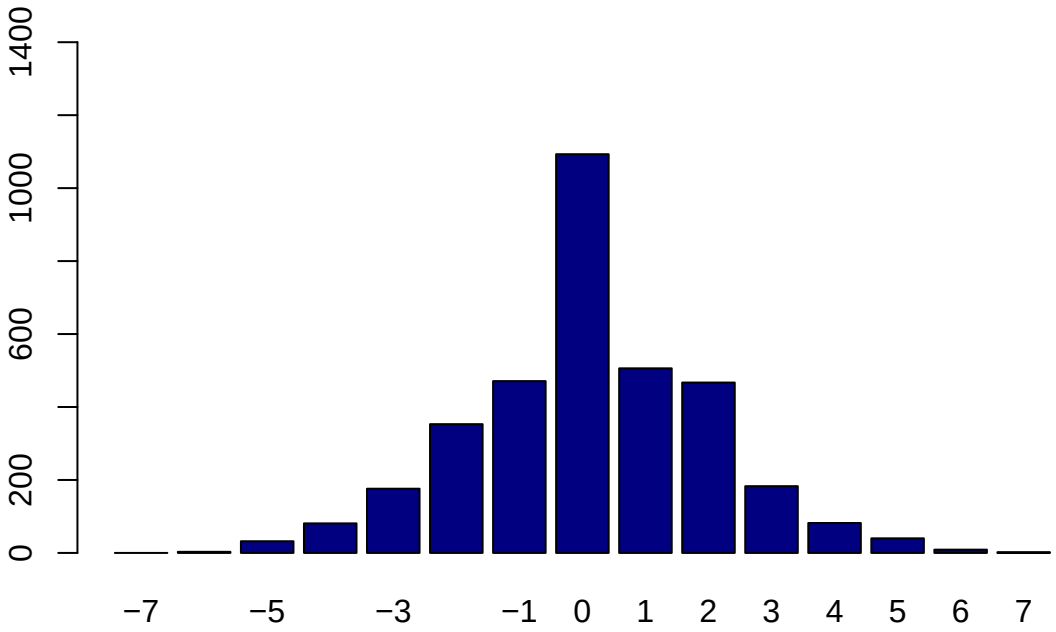
Generational mobility

Figure 1 and table 1 show the overall distribution of occupational status, but is not informative on the extent of generational mobility. To get an indication of generational mobility, we have calculated the difference in occupational status for each father-son pair. Thus, a '-7' indicates observations where the father had social

status 8 and the son 1, and '0' is when father and son had the same occupational status. The number of observations with each difference is reported in figure 2.

Figure 2

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



Note, the density is by construction higher around '0', as only a subset of observation pairs can experience the highest and lowest scores. For instance, if a father has occupational status 7, the father-son pair can only have a difference score between [-6;1]. Nonetheless, the graph indicates that occupational status differs in a substantial share of the sample.

Analysis of the dataset 'LifeCycleSavings'

The dataset 'LifeCycleSavings' consists of a data frame with 5 variables observed for 50 countries. The variables measure the savings rate (sr), real per-capita disposable income (dpi), the growth rate of dpi (ddpi) as well as the percentage of the population over 75 (pop75) and under 15 (pop15). All variables are numeric and averaged over the decade 1960 to 1970.

Inspecting the data

Table 2: Summary Statistics (Life Cycle Savings)

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

	mean	sd	median	minimum	maximum	s.size
sr	9.67	4.48	10.51	0.60	21.10	50

	mean	sd	median	minimum	maximum	s.size
pop15	35.09	9.15	32.58	21.44	47.64	50
pop75	2.29	1.29	2.17	0.56	4.70	50
dpi	1106.76	990.87	695.66	88.94	4001.89	50
ddpi	3.76	2.87	3.00	0.22	16.71	50

Table 2 shows that the variables dpi (disposable income) and ddpi (growth of disposable income) are skewed. Consequently, we take a closer look at the distributions.

Figure 3: Histogram of disposable income

```
hist(LifeCycleSavings$dpi,
     ylim=c(0,20),
     xlim=c(0,5000),
     col = "navyblue",
     border = "white",
     main = 'Disposable income (averaged between 1970 and 1980)',
     xlab = 'Disposable income (in USD)',
     ylab = 'Frequency')
```

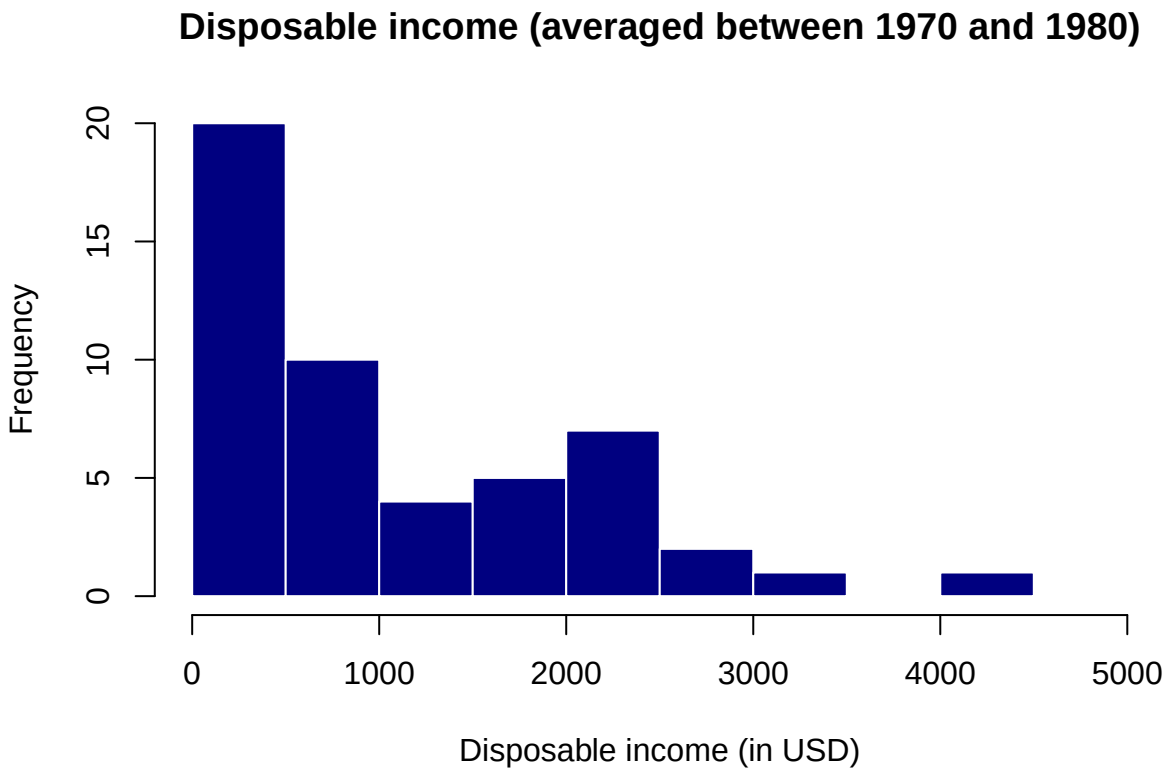


Figure 3 shows *disposable income* is strongly right-skewed. This also applies for *growth of disposable income*. To correct for the right-skewed data, we take the natural log of the two variables.

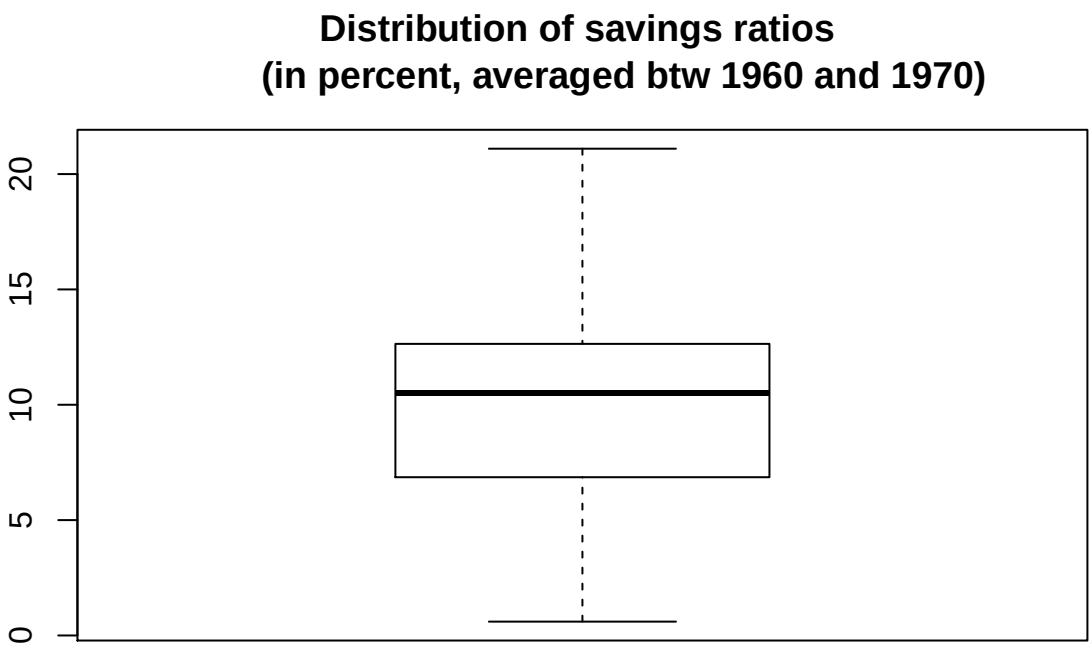
Additionally we create dummies which split the data set in above the median and below the median for the variables *precent of population over 75 (pop75)*, as well as for *disposable income* and *growth of disposable income*.

Analyzing the Data

First, we take a closer look at the dependent variable *savings rate*. The boxplot in figure 4 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).

Figure 4

```
boxplot(combined_df$sr, main = 'Distribution of savings ratios
        (in percent, averaged btw 1960 and 1970)')
```



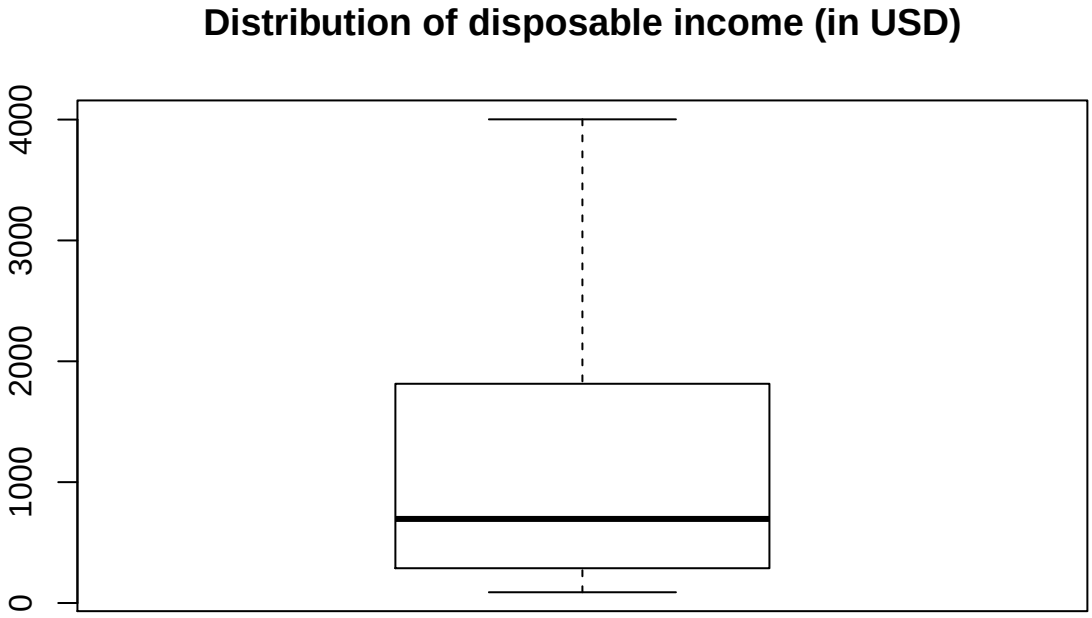

```
summary(combined_df$sr)
```

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

Second, we inspect the main explanatory variable *disposable income*. The boxplot in figure 5 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). W will again use the log-transformation.

Figure 5

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



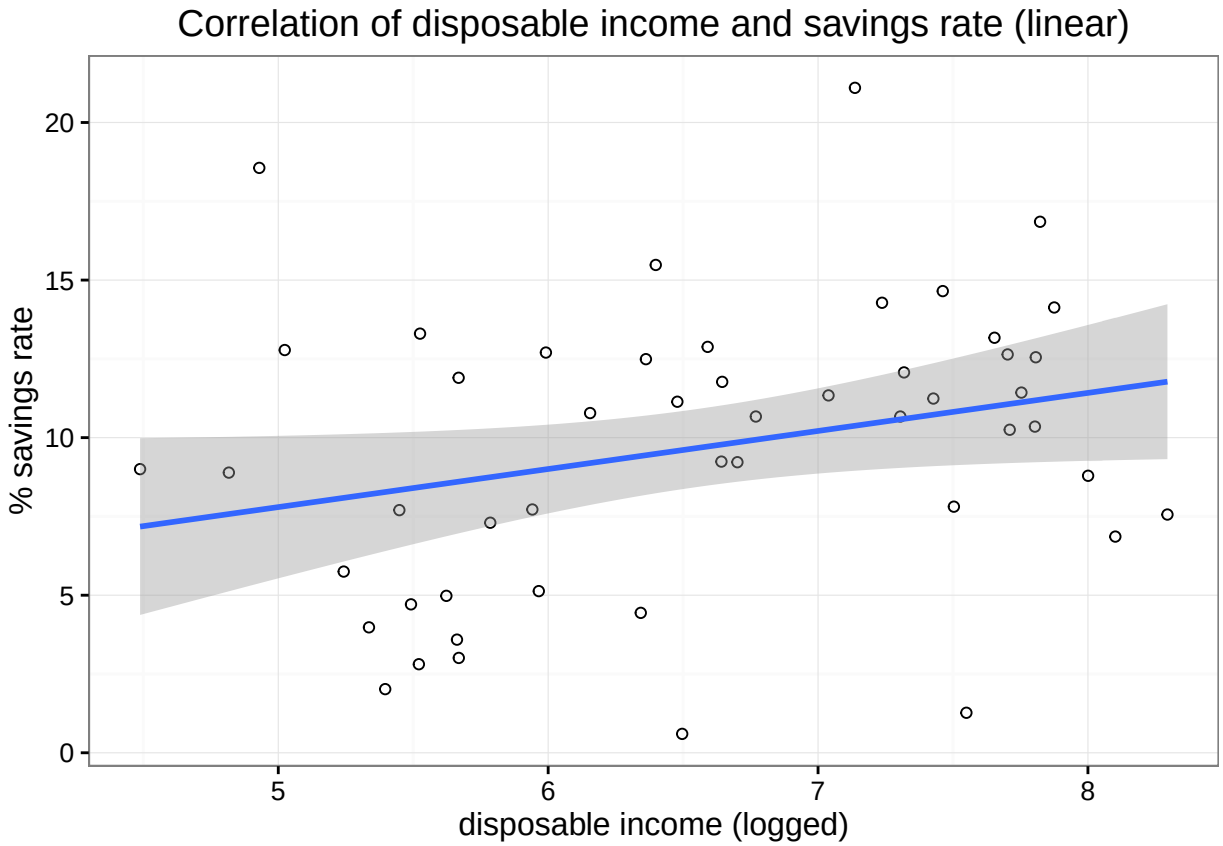

```
summary(combined_df$dpi)
```

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

In figure 6 we take a further look at the correlation between *savings rates* and *disposable income*. The two variables are positively correlated with a correlation coefficient of 0.28, which is borderline insignificant at the 95% level (p-value of 0.052).

Figure 6

```
# 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")
```



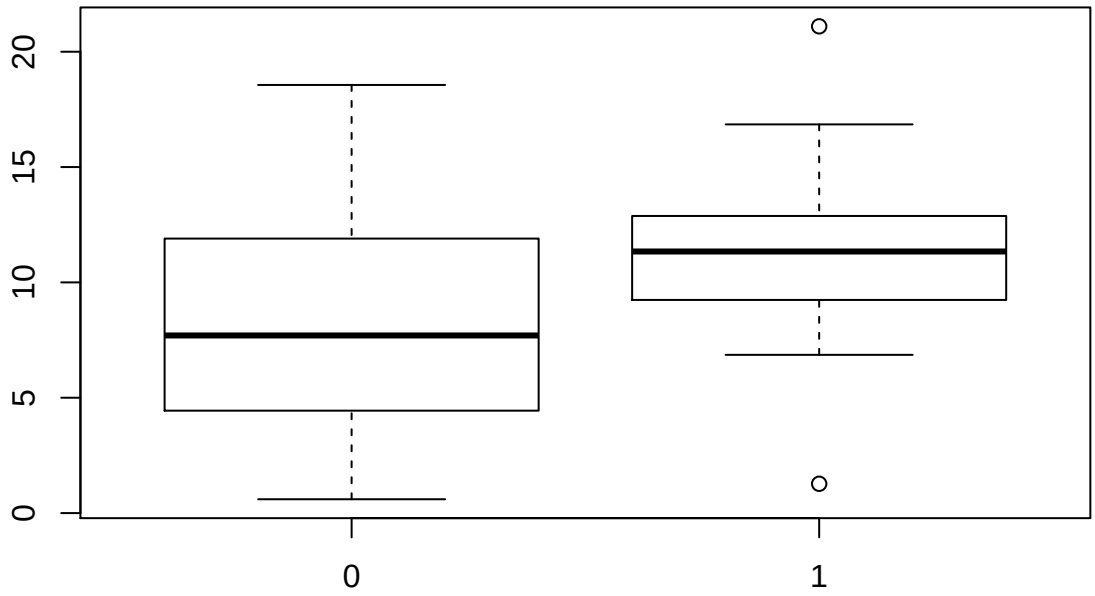

```
# Testing for correlation
cor.test(combined_df$ln.dpi, combined_df$sr)
```

```
##
##  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
```

In order to understand the correlation better we assess the distribution of savings rates between richer and poorer countries (poorer = 0, richer = 1). Figure 7 shows that richer countries have a much more compressed distribution that does not extend to very low savings rates. Further, richer countries have, on average, higher savings rates.

Figure 7

```
boxplot(combined_df$sr~combined_df$dum.rich, main =
        'Distribution of savings rates for relatively poor and relatively rich countries')
```



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 again split the sample with dummy variables in above and below median groups. Figure 8 shows that countries with higher growth rates seem to have a slightly stronger correlation between savings rate and disposable income. However, the difference is not statisitiaclly significant.

Figure 8: 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")
```

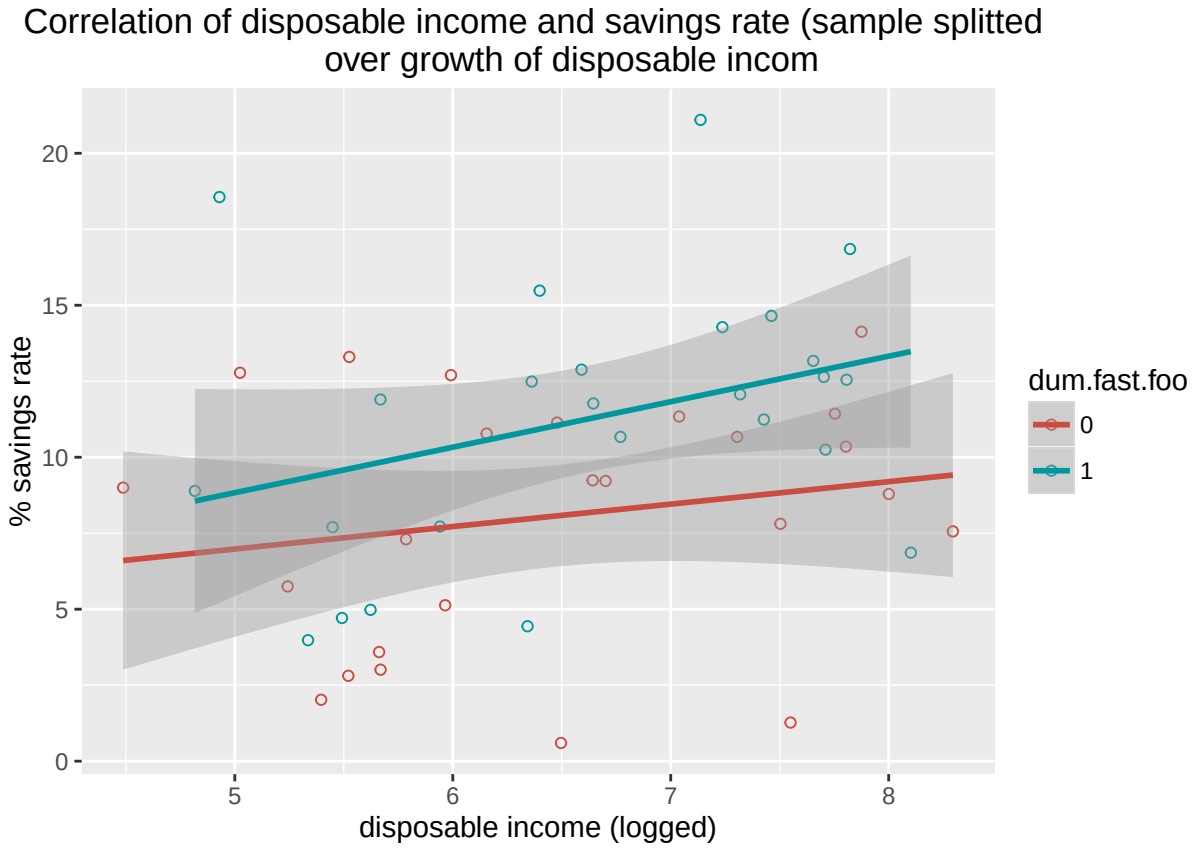
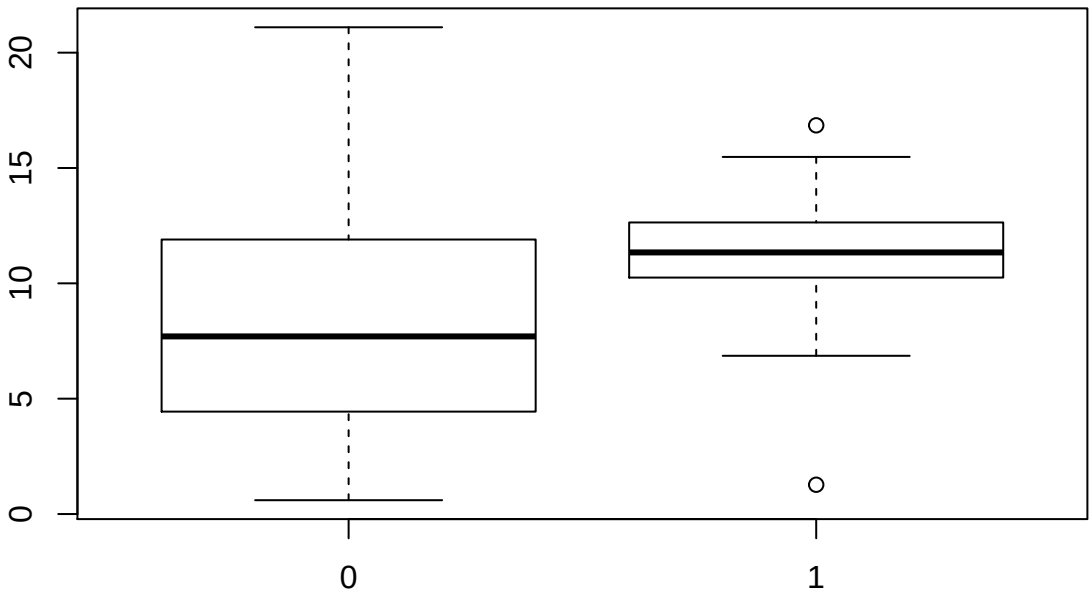


Figure 9 shows that countries with a larger share of an older population have a more compressed distribution of savings rates, and on average savings rates are higher. This resembles the sample split for rich and poor countries.

Figure 9: Share of older population (lower percentage = 0, higher percentage = 1)

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



Software and packages used for the analysis

The analysis is done in R (R Core Team 2015) with the use of the following packages: "doBy" (Højsgaard and Halekoh 2015), "ggplot2" (Wickham and Chang 2015), "repmis" (Gandrud 2016) and "dplyr" (Wickham and Francois 2015).

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.

Wickham, Hadley, and Romain Francois. 2015. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.