Analysis and Graphical Representation of Health and Labour Force Participation among the Elderly in Europe

Phi Nguyen
Julian Winkel
Claudia Guenther

November 27, 2017

Introduction

- relevance of exploring relationship between health and labour force participation due to changing demographic in Europe
- cite relevant study about ageing
- 1-2 sentences about relevant papers exploring relationship -> use paper from DIW
- share data set as rich data set for this purpose: 2 sentences about it
- introduction of journal article
- our approach: replicate results an enrich analysis
- especially: introduce graphical visualization tools for descriptive statistics -> ease interpretation of variables
- our aim: write code in a way that allows the user to work with easySHARE data set, even when working on different question

Theory and Desgin

- very short literature overview on relationship between health and labour force participation
-

Implementation

Empirical Study

Data cleaning and subgrouping

The easySHARE dataset released in spring 2017 is a panel dataset of 108 variables of more than 100.000 individuals covering data from six survey waves carried out between 2004 and 2005. As we are only concerned with a small subset of observations, an important task was to define appropriate functions for subsetting. In oder to make our subsetting process understandable to readers, we decided on using pipe-operator. This allows us to apply the filter and select option in a clever way, where we can select different criteria at once.

-> code sniplet here

In this example, with first filter for participation in wave 1 and the age group between 50 and 64 and the select the desired variables as described in Kalwij and Vermeulen (2005).

Although the overall response rate in the SHARE are comparably high, the data set still has numerous missing values. The reason for this is due to the fact that the study was carried out on a crossnational scale, with some national survey institutions deciding not to participate in all survey modules. This means that the majority of missing values are to be found within observations that have missing values for entire survey modules or waves. The reason for the missing values are documented well in the "Guide to easySHARE release 6.0.0" and specifically coded. For example, the numbers -13 and -14 refer to "not asked in this wave" and "not asked in this country". Since this coding scheme is not useful for the purpose of our analysis, we decided on recoding all of the missing values as "NA". To this end, we defined a function based on the missing codes provided by SHARE that finds the NAs in the data and declares them as such.

-> code snipet here

Since the study carried out by Kalwij and Vermeulen (2005) is based on the use of mostly binary data, we needed to construct numerous dummies based on the orginal data.

-> code snipet here

The resulting dataframe containes XXX observations of YYY variables.

- coded countries in more readable manner
- use packe ddd -> match offical ISO code with country name
- create country list with all countries in study (not Israel)
- defined dummies

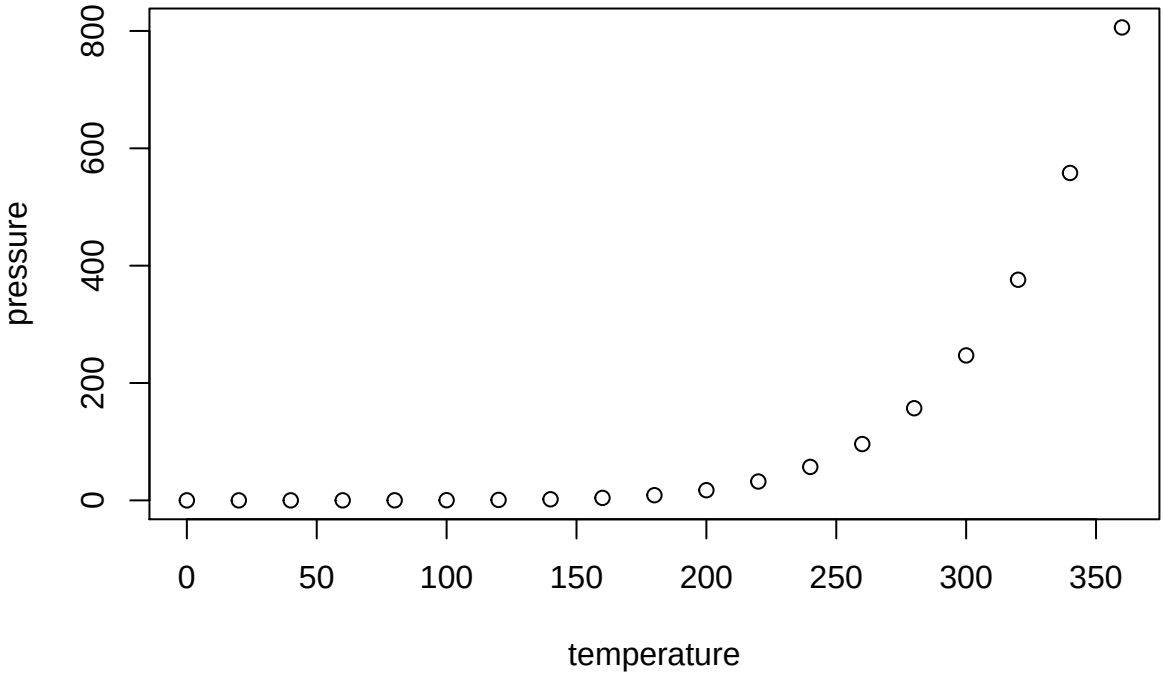
```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
```

```
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.