R Notebook

Sanne Meijering
Hanne Oberman
Gerbrich Ferdinands

Initialization

```
# Load packages
require(survey)
```

```
## Loading required package: survey
```

```
## Warning: package 'survey' was built under R version 3.4.4
```

```
## Loading required package: grid
```

```
## Loading required package: Matrix
```

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

```
##
## Attaching package: 'survey'
```

```
## The following object is masked from 'package:graphics':
##
##     dotchart
```

```
require(sampling)
```

```
## Loading required package: sampling
```

```
## Warning: package 'sampling' was built under R version 3.4.4
```

```
##
## Attaching package: 'sampling'
```

```
## The following objects are masked from 'package:survival':
##
##     cluster, strata
```

```
# Load society dataset
society <- readRDS("Understanding Society innovation pnel wave A.RDS")
society$a_dvage <- as.numeric(society$a_dvage)
```

Sampling design

Question 2

Investigation of design weights

```
# Calculate variance of design weight
Var <- var(society$a_psnenip_xd)
# Combine variance with other descriptive statistics
Descr <- cbind(t(summary(society$a_psnenip_xd)), Var)
# Print with up to 2 decimals
round(Descr, 2)
```

```
##      Min. 1st Qu. Median Mean 3rd Qu. Max.  Var
## [1,]    1       1      1 1.01       1    4 0.02
```

Question 3

```r
# First, make the dataset smaller by removing unnecessary columns
society_3 <- society[,-c(5:7,9:56,60:75,77:89)]

# Investigate levels of government office region variable
levels(society_3$a_gor_dv)
```

```
##  [1] "missing"                  "north east"
##  [3] "north west"               "yorkshire and the humber"
##  [5] "east midlands"            "west midlands"
##  [7] "east of england"          "london"
##  [9] "south east"               "south west"
## [11] "wales"                    "scotland"
## [13] "northern ireland"
```

```r
nrow(society_3[society_3$a_gor_dv == "missing",])
```

```
## [1] 0
```

```r
nrow(society_3[society_3$a_gor_dv == "northern ireland",])
```

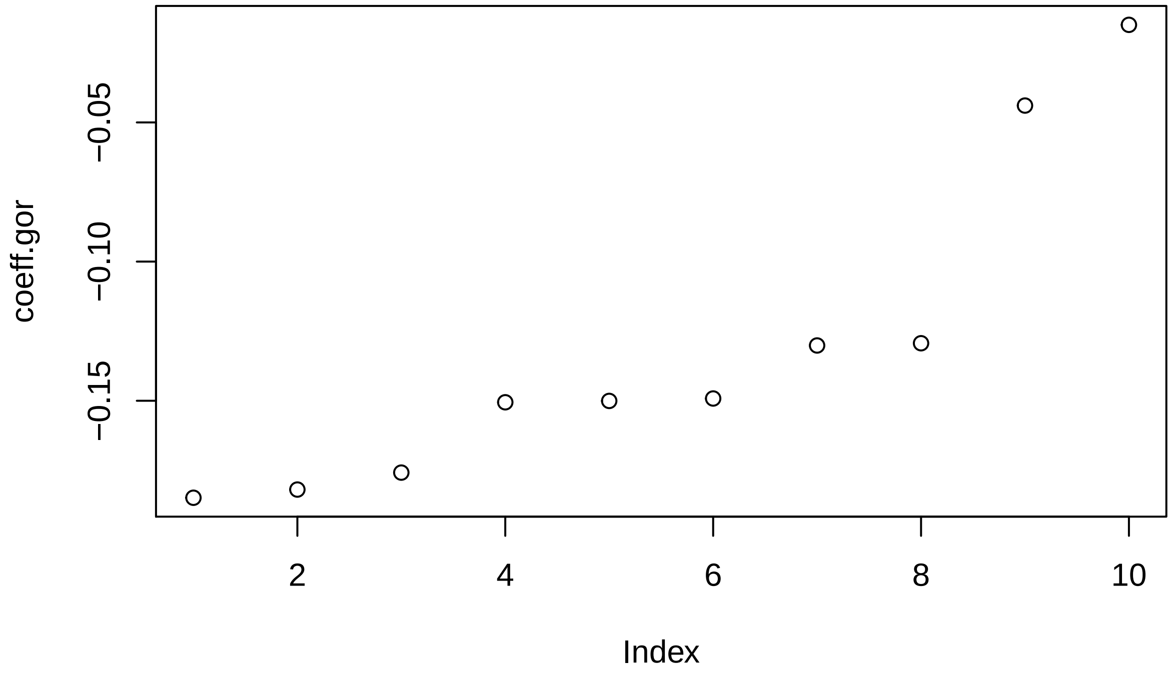
```
## [1] 0
```

```r
# None of the value are missing or northern ireland, so those can be ignored

# Create dummy variables for government office regions
society_3$NE <- society_3$a_gor_dv == "north east"
society_3$NW <- society_3$a_gor_dv == "north west"
society_3$Y <- society_3$a_gor_dv == "yorkshire and the humber"
society_3$EM <- society_3$a_gor_dv == "east midlands"
society_3$WM <- society_3$a_gor_dv == "west midlands"
society_3$EOE <- society_3$a_gor_dv == "east of england"
society_3$L <- society_3$a_gor_dv == "london"
society_3$SE <- society_3$a_gor_dv == "south east"
society_3$SW <- society_3$a_gor_dv == "south west"
society_3$W <- society_3$a_gor_dv == "wales"
society_3$S <- society_3$a_gor_dv == "scotland"

# Run linear regression (Scotland is the baseline for government office region)
coeff <- with(society_3, lm(a_psnenip_xw ~ a_psnenip_xd + a_sex + a_dvage + NE + NW + Y + EM + WM + EOE

# Get coefficients
coeff <- coeff$coefficients
coeff.gor <- sort(coeff[5:15])
plot(coeff.gor)
```



Population Estimates

Question 4

```
# Investigate the a_employ variable.
levels(society$a_employ[society$a_dvage > 15 & society$a_dvage < 64])
```

```
## [1] "missing"           "inapplicable"      "proxy respondent"
## [4] "refuse"            "don't know"        "yes"
## [7] "no"
```

```
# The variable a_employ has seven levels.

summary(society$a_dvage[society$a_employ=="yes"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.00   38.00   48.00   47.55   57.00  102.00
```

```
summary(society$a_dvage[society$a_employ=="no"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.00   45.00   69.00   62.38   79.00  102.00
```

```
# Yes and no contain people of over 21 years of age.

nrow(society[society$a_employ=="missing",])
```

```
## [1] 0
```

```r
nrow(society[society$a_employ=="proxy respondent",])
```

```
## [1] 0
```

```r
# Missing and proxy respondent do not appear in the data

summary(society$a_dvage[society$a_employ=="inapplicable"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   10.00   14.00   13.79   18.00   21.00
```

```r
# Inapplicable seems to contain all children and youths of 21 years and younger
# It cannot be assumed that none of them is employed. It can however be assumed that only a
# small part of them is employed, as children under 15 cannot be employed legally
# and most would still be going to a school or university.

nrow(society[society$a_employ=="refuse",])
```

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

```r
nrow(society[society$a_employ=="don\'t know",])
```

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

```r
# Both refuse and don't know contain one row. These can be treated as missing data.

# Thus, our goal is to compare the proportion of employed people (yes) of working age against
# the number of unemployed people (no, inapplicable), excluding the missing data (refuse, don't know)

with(society, nrow(society[a_employ=="yes" & a_dvage >64,]))
```

```
## [1] 160
```

```r
# There are people older than 64 still working, we should exclude those.
with(society, nrow(society[a_employ=="yes" & a_dvage <15,]))
```

```
## [1] 0
```

```r
# No one younger than 15 years is reported to be working, which is to be expected as it was not
# a question asked to people under 21 years of age.

# Since we wish to know the proportion of employed people of working age, we need 2 groups, one with em
society$employ_dv <- as.numeric(0)
society$employ_dv[society$a_employ=='yes' & society$a_dvage <= 65] <- 1

# Create design
# Don't remove the missing values yet, as the weights are calculated including missing values
Design <- svydesign(ids=~a_hidp, strata=~a_strata, data=society, weights=~a_psnenip_xw)
# Make a subset of non-missing values
Nonmiss <- with(Design, subset(Design, a_employ!="refuse" & a_employ!="don\'t know"))
svymean(~employ_dv, Nonmiss)
```

```
##              mean     SE
## employ_dv 0.43342 0.0097
```

```r
confint(svymean(~employ_dv, Nonmiss))
```

```
##               2.5 %    97.5 %
## employ_dv 0.4144994 0.4523372
```

```
# 43,3% of the population is employed, with a 95% confidence interval of 41.4%-45.2%
```