Labor Force Survey 2017

Bieske, Kibsgaard, Laouiti, Tan, Wenyu, Zhang

26/09/2017

1. Introduction

```
library(haven)
lfsp <- read_dta("lfsp_jm17_eul.dta")
```

2. Exploratory Data Analysis

2.1 Clean dataset

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
selected <- lfsp %>% select(MARSTA,RELIG11,SEX,AGE,INDE07M,
                            BANDG,ETHGBEUL,HIQUL15D,HIGHO)

#AGE

#SEX
selected <- selected[!selected$RELIG11==-8,]

# MARSTA
selected <- selected[!selected$MARSTA==-9,]

#RELIG11
selected <- selected[!selected$RELIG11==-8,]

#INDE07M
selected <- selected[!selected$INDE07M==-8,]

#ETHGBEUL
selected <- selected[!selected$ETHGBEUL==-8 & !selected$ETHGBEUL==-9,]

#HIQUL15D
selected <- selected[!selected$HIQUL15D==-8 & !selected$HIQUL15D==-9 & !selected$HIQUL15D==7,]
```

```
#HIGHO
selected <- selected[!selected$HIGHO==-8,]

#BANDG
selected <- selected[!selected$BANDG==1 & !selected$BANDG==2 & !selected$BANDG==3 & !selected$BANDG==-8
```

2.2 Label categorical variables

```
#AGE

#SEX
selected$SEX <- factor( selected$SEX,
                        levels=c(1,2),
                        labels = c('Male','Female'))

# MARSTA
selected$MARSTA <- ifelse(!selected$MARSTA==1 &
                          !selected$MARSTA==2, 3 , selected$MARSTA)
selected$MARSTA <- factor( selected$MARSTA,
                           levels=c(1,2,3),
                           labels = c('Single','Married','Other'))

#RELIG11
selected$RELIG11 <- ifelse(selected$RELIG11 == 1, 1 ,2)
selected$RELIG11 <- factor(selected$RELIG11,
                           levels=c(1,2),
                           labels = c('No','Yes'))

#INDEO7M
selected$INDEO7M <- ifelse( selected$INDEO7M == 1
                           |selected$INDEO7M == 2
                           |selected$INDEO7M == 3
                           |selected$INDEO7M == 4
                           |selected$INDEO7M == 6
                           |selected$INDEO7M == 9, 4, selected$INDEO7M)
selected$INDEO7M <- ifelse( selected$INDEO7M == 5, 1, selected$INDEO7M)
selected$INDEO7M <- ifelse( selected$INDEO7M == 7, 2, selected$INDEO7M)
selected$INDEO7M <- ifelse( selected$INDEO7M == 8, 3, selected$INDEO7M)

selected$INDEO7M <- factor( selected$INDEO7M,
                            levels = c(1,2,3,4),
                            labels = c('Distribution, hotels and restaurants',
                                       'Banking and Finance',
                                       'Public admin, education, health',
                                       'Other'))

#ETHGBEUL
selected$ETHGBEUL <- ifelse(  selected$ETHGBEUL== 1
                             |selected$ETHGBEUL== 2
                             |selected$ETHGBEUL== 3, 1, selected$ETHGBEUL)
selected$ETHGBEUL <- ifelse( selected$ETHGBEUL == 5
                             |selected$ETHGBEUL== 6
```

```
                              |selected$ETHGBEUL== 7
                              |selected$ETHGBEUL== 8
                              |selected$ETHGBEUL== 9,  2, selected$ETHGBEUL)
selected$ETHGBEUL <- ifelse( selected$ETHGBEUL == 10, 3, selected$ETHGBEUL)
selected$ETHGBEUL <- ifelse( selected$ETHGBEUL == 4
                              |selected$ETHGBEUL== 11, 4, selected$ETHGBEUL)
selected$ETHGBEUL <-factor( selected$ETHGBEUL,
                            levels = c(1,2,3,4),
                            c('White','Asian','Black','Other'))

#HIQUL15D
selected$HIQUL15D <- ifelse( selected$HIQUL15D == 1
                            &selected$HIQUL15D == 2, 1, selected$HIQUL15D )
selected$HIQUL15D <- ifelse( selected$HIQUL15D == 3, 2 ,selected$HIQUL15D)
selected$HIQUL15D <- ifelse( selected$HIQUL15D == 4, 3, selected$HIQUL15D)
selected$HIQUL15D <- ifelse( selected$HIQUL15D == 5, 4, selected$HIQUL15D)
selected$HIQUL15D <- ifelse( selected$HIQUL15D == 6, 5, selected$HIQUL15D)

selected$HIQUL15D <- factor( selected$HIQUL15D,
                             levels = c(1,2,3,4,5),
                             labels = c('University',
                                        'College',
                                        'Secondary School',
                                        'Other',
                                        'Non'))

#HIGHO
selected$HIGHO <- ifelse( selected$HIGHO  == 1
                           |selected$HIGHO == 2
                           |selected$HIGHO == 3, 1, selected$HIGHO)
selected$HIGHO <- ifelse( selected$HIGHO  == 4
                           |selected$HIGHO == 5
                           |selected$HIGHO ==-9, 2, selected$HIGHO)
selected$HIGHO <- factor(selected$HIGHO,
                         levels=c(1,2),
                         c('Yes','Other'))


#BANDG
value_vec <- c('1.1','1.10','1.11','1.12','1.13','1.14','1.15','1.16','1.17',
               '1.18','1.19','1.2','1.20','1.21','1.22','1.23','1.24','1.25',
               '1.26','1.27','1.28','1.29','1.3','1.30','1.31','1.32','1.33',
               '1.34','1.4','1.5','1.6','1.7','1.8','1.9',
               '2.1','2.10','2.11','2.12','2.13','2.14','2.15','2.16','2.17',
               '2.18','2.19','2.2','2.20','2.21','2.22','2.23','2.24','2.25',
               '2.26','2.27','2.28','2.29','2.3','2.30','2.31','2.32','2.33',
               '2.34','2.4','2.5','2.6','2.7','2.8','2.9',
               "3.10","3.11","3.12","3.13","3.14","3.15","3.16","3.17","3.18",
               "3.19","3.20","3.21","3.22","3.23","3.24","3.25","3.26","3.27",
               "3.28","3.29","3.30","3.31","3.32","3.3","3.4","3.5","3.6",
               "3.7","3.8","3.9","3.2","3.34")
# Annually
a1  <- 0
```

```r
a2  <- seq(4249.5,4749.5,500)
a3  <- seq(5499.5,12499.5,1000)
a4  <- 249.5
a5  <- seq(13499.5,19499.5,1000)
a6  <-seq(21499.5,27499.5,3000)
a7  <- 749.5
a8  <- seq(30499.5,39499.5,3000)
a9  <- 41000
a10 <- seq(1249.5,3749.5,500)
annual <- c(a1,a2,a3,a4,a5,a6,a7,a8,a9,a10)

# Monthly
m1  <- 0
m2  <- seq(424.5, 674.5, 50)
m3  <- seq(749.5, 1049.5, 100)
m4  <- 24.5
m5  <- seq(1149.5, 1949.5, 100)
m6  <- 2099.5
m7  <- 74.5
m8  <- 2349.5
m9  <- seq(2749.5, 3749.5, 500)
m10 <- 4000
m11 <- seq(124.5, 374.5, 50)
month_values <- c(m1, m2, m3, m4, m5, m6, m7, m8, m9, m10, m11)
months_annual <- month_values*12

# Weekly
w1 <- seq(84.5,104.5,10)
w2 <- 117
w3 <- seq(137,487,25)
w4 <- seq(524.5,674.5,50)
# for 3.3 to 3.9
w5 <- seq(14.5,74.5,10)
# for 3.2
w6 <- 5
# for 3.34 weekly 750 or more
w7 <- 750
week_values <- c(w1,w2,w3,w4,w5,w6,w7)
week_annual <- week_values * 52

# Combine annually, monthly and weekly
salay_all <-c(annual,months_annual,week_annual)
salary_dict <- data.frame(key=value_vec,value=salay_all)
salary_dict$key <- as.character(salary_dict$key)

# Assign mean salary to selected$salary according to salary_dict
for (i in 1:nrow(selected)){
  selected$Salary[i] <- salary_dict$value[salary_dict$key==selected$BANDG[i]]
}
```

```
## Warning: Unknown or uninitialised column: 'Salary'.
```

3. Linear Regression

```
selected.lm <- lm(Salary ~ SEX,data=selected)
summary(selected.lm)
```

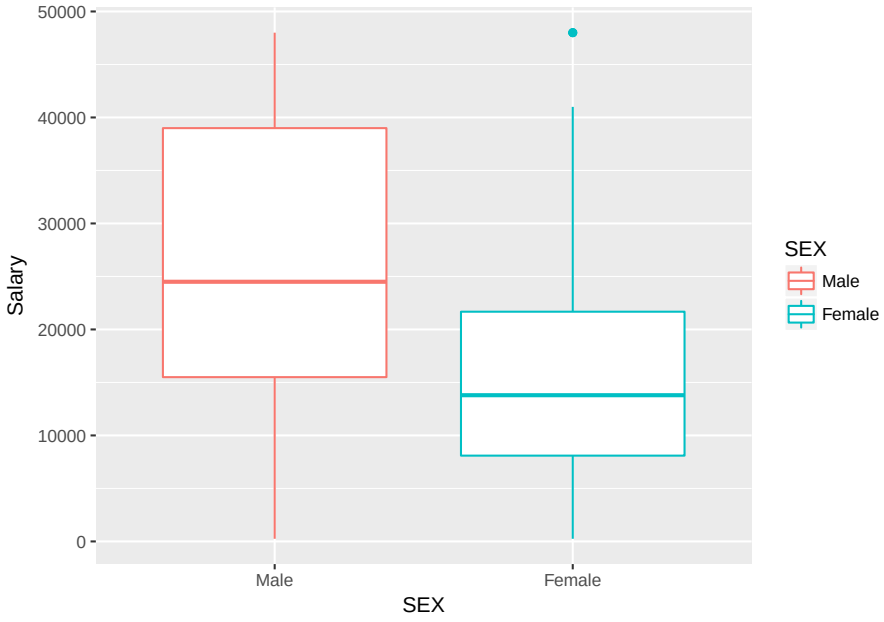
```
##
## Call:
## lm(formula = Salary ~ SEX, data = selected)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24992  -9017  -1242   8264  31764
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25241.1      713.9   35.35   <2e-16 ***
## SEXFemale    -9005.4      990.7   -9.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11820 on 568 degrees of freedom
## Multiple R-squared:  0.127,  Adjusted R-squared:  0.1255
## F-statistic: 82.62 on 1 and 568 DF,  p-value: < 2.2e-16
```

4. Analysis and Results

```
# Visualisation
library(ggplot2)
ggplot(selected, aes(x = AGE, y = Salary, col = SEX)) + geom_point() +
    labs(title = "Salary distribution over age") + theme(plot.title = element_text(hjust = 0.5))
```

```
## Don't know how to automatically pick scale for object of type labelled. Defaulting to continuous.
```

```r
ggplot(selected, aes(x = SEX, y = Salary, col = SEX)) + geom_boxplot()
```



5. Conclusion

6. References