Computer Lab 3

In this computer lab we will have to achieve the following tasks/learning outcomes:

- import data
- understand some important features of the dataset
- select a subset of data
- estimate and compare single and multiple regressions

Preparing your workfile

We add the basic libraries needed for this week's work:

```
library(tidyverse)    # for almost all data handling tasks
library(ggplot2)      # to produce nice graphics
library(stargazer)    # to produce nice results tables
library(haven)        # to import stata file
library(AER)          # access to HS robust standard errors
```

You should also save the separately supplied `stargazer_HC.r` file in your working directory. This will make it straightforward to estimate and compare regressions with robust standard errors. Once you have done that you should include the following line into your code which basically makes this function available to you.

```
source("stargazer_HC.r")  # includes the robust regression
```

Introduction

The data are an extract from the Understanding Society Survey (formerly the British Household Survey Panel).

Data Upload - and understanding data structure

Upload the data from `20222_USoc_extract.dta`. This is STATA datafile (extension `.dta`). There is a function which loads STATA file. It is called `read_dta` and is supplied by the `haven` package.

```
##  [1] "pidp"    "age"     "jbhrs"   "paygu"   "wave"    "cpi"     "year"
##  [8] "region"  "urate"   "male"    "race"    "educ"    "degree"  "mfsize9"
```

```
data_USoc <- read_XXXX(XXXX)
names(data_USoc)
```

Let us ensure that categorical variables are stored as `factor` variables as this will make working with them easiest.

```
data_USoc$region <- XXXX(data_USoc$region)
data_USoc$male <- XXXX(data_USoc$male)
data_USoc$degree <- XXXX(data_USoc$degree)
data_USoc$race <- XXXX(data_USoc$race)
```

Click on the little table symbol in your environment tab to see the actual data table.

The pay information (`paygu`) is provided as a measure of the (usual) gross pay per month. Let's calculate a typical hourly wage. In the variable `jbhrs` we find the typical number of weekly hours worked. To get an hourly pay we divide `paygu` by `4*jbhrs`. As the pay is from years 2009 to 2013 we shall also adjust for increasing price levels (as measured by `cpi`). To create a new variable we use the `mutate` function. We call the new variable `hrpay` and also calculate the natural log of this variable (`lnhrpay`).

```
data_USoc <- data_USoc XXXX
              XXXX(hrpay = paygu/(jbhrs*4)/(cpi/100)) %>%
              XXXX(lnhrpay = XXXX(hrpay))
```

As we wanted to save these additional variables we assign the result of the operation to `data_USoc` (which now has two additional variables, i.e. 16). You can check that your calcullations are correct if you obtain a mean value for `lnhrpay` of 2.28 and you have 74312 missing observations. Use `summary(data_USoc$lnhrpay)` to check for these stats.

First Analysis - Do Regions matter?

Have a look at the `region` variable. Establish what the different regions in the dataset are and how many observations we have in each region in each year. (Hint: we did something similar for Lecture 2):

You did it right if you find that for 2009 there were 1867 observations from the East Midlasnds region and for 2013 only 47 observations from Wales. In fact if look at the number of observations across the years you should realise that for the Year 2013 there are much fewer observations than for the other years. This could be an indication for some problem (or systematic selection) with the data from this year and hence we decide to remove all 2013 observations from the dataset.

We did achieve things like this in previous empirical work and you could look in previous files how we achieved this. There are of course different ways to do this and you could google for solutions ("R select observations", "R remove observations")

After doing this you should find that the `data_USoc` has 132,119 remaining observations.

Let's run a regression of `lnhrpay` as the dependent variable against `region` as an explanatory variable.

```
mod1 <- lm(XXXX~XXXX, data = XXXX)
stargazer_HC(mod1)
```

If you do it correctly you should obtain the following output:

```
##
## ===============================================================
##                                        Dependent variable:
##                                    ---------------------------
##                                               lnhrpay
## ---------------------------------------------------------------
## regionnorth west                              0.032**
##                                               (0.014)
##
## regionyorkshire and the humber                -0.012
##                                               (0.014)
##
## regioneast midlands                           -0.017
##                                               (0.014)
##
## regionwest midlands                            0.020
##                                               (0.014)
##
```

```
## regioneast of england                          0.108***
##                                                 (0.015)
##
## regionlondon                                    0.205***
##                                                 (0.014)
##
## regionsouth east                                0.167***
##                                                 (0.014)
##
## regionsouth west                                 0.036**
##                                                 (0.014)
##
## regionwales                                     -0.074***
##                                                 (0.017)
##
## regionscotland                                  0.062***
##                                                 (0.014)
##
## regionnorthern ireland                           0.010
##                                                 (0.016)
##
## Constant                                        2.215***
##                                                 (0.011)
##
## ------------------------------------------------------------------------
## Observations                                      58,399
## R2                                                 0.017
## Adjusted R2                                        0.017
## Residual Std. Error                          0.625 (df = 58387)
## F Statistic                               92.666*** (df = 11; 58387)
## ========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
##                                    Robust standard errors in parenthesis
```

What is the base region? The first level is the `north east`, (Check `levels(data_USoc$region)`) and that is the base reagion. For all other regions the above regression included a dummy variable. For instance, the variable called `regionwales` takes the value 1 if an observation is from Wales and 0 otherwise. In other words, R has taken your categorical `region` variable (with 12 regions) and turned it into 11 dummy variables and included these into the regression. All of this without you actually having to compute these variables.

How would you interpret the estimated parameter value for `regionwales`? The coefficient value is -0.0743675 and indicates that on average hourly pay is 7.4 percent lower than those in the North East.

Which region is the highest paying region?

Now estimate two more models. One in which the only explanatory variables is `educ` which measures the completed years of formal education.

```
mod2 <- lm(XXXX~XXXX, data = XXXX)
stargazer_HC(XXXX)
```

You got it right if you get a slope coefficient of 0.0935909. The result is ever so slightly different to that in the lecture as, here, we removed the 2013 observations.

Then also estimate a model which contains both, the `educ` and the `region` variables. Then display all three models in one table.

```r
mod3 <- lm(XXXX~XXXX+XXXX, data = XXXX)
stargazer_HC(mod2,mod1,mod3)
```

You should be able to replicate the following table (in R you should be able to see the complete 3rd column):

```
##
## ===========================================================================================================
##                                                               Dependent variable:
##                                    ------------------------------------------------------------------------
##                                                                     lnhrpay
##                                                (1)                           (2)                          (3
## -----------------------------------------------------------------------------------------------------------
## educ                                         0.094***                                                   0.09
##                                              (0.001)                                                    (0.0
##
## regionnorth west                                                          0.032**                        0.0
##                                                                           (0.014)                       (0.0
##
## regionyorkshire and the humber                                             -0.012                       -0.0
##                                                                           (0.014)                       (0.0
##
## regioneast midlands                                                        -0.017                       -0.0
##                                                                           (0.014)                       (0.0
##
## regionwest midlands                                                         0.020                        0.0
##                                                                           (0.014)                       (0.0
##
## regioneast of england                                                     0.108***                      0.087
##                                                                           (0.015)                       (0.0
##
## regionlondon                                                              0.205***                      0.089
##                                                                           (0.014)                       (0.0
##
## regionsouth east                                                          0.167***                      0.120
##                                                                           (0.014)                       (0.0
##
## regionsouth west                                                          0.036**                        0.0
##                                                                           (0.014)                       (0.0
##
## regionwales                                                              -0.074***                      -0.0
##                                                                           (0.017)                       (0.0
##
## regionscotland                                                            0.062***                      0.047
##                                                                           (0.014)                       (0.0
##
## regionnorthern ireland                                                      0.010                       -0.0
##                                                                           (0.016)                       (0.0
##
## Constant                                     1.032***                     2.215***                      1.025
##                                              (0.014)                      (0.011)                       (0.0
##
## -----------------------------------------------------------------------------------------------------------
## Observations                                  58,381                       58,399                        58,3
## R2                                            0.128                        0.017                         0.1
```

```
## Adjusted R2                              0.128                  0.017                  0.1
## Residual Std. Error                0.589 (df = 58379)      0.625 (df = 58387)      0.586 (df
## F Statistic                   8,600.210*** (df = 1; 58379) 92.666*** (df = 11; 58387) 762.131*** (df
## ===========================================================================================
## Note:                                                                          *p<0.1; **p<0
##                                                                       Robust standard errors i
```

You can not see the entire table as the columns are very wide. The reason being the reported F-statistic at the bottom of each column. You can omit that statistic using the `omit.stat` option.

```
stargazer_HC(mod2,mod1,mod3, omit.stat="f")
```

```
##
## ========================================================================================
##                                                   Dependent variable:
##                                    -----------------------------------------------------
##                                                         lnhrpay
##                                          (1)                (2)                (3)
## ----------------------------------------------------------------------------------------
## educ                                   0.094***                              0.091***
##                                        (0.001)                               (0.001)
##
## regionnorth west                                          0.032**             0.016
##                                                           (0.014)            (0.013)
##
## regionyorkshire and the humber                            -0.012             -0.021
##                                                           (0.014)            (0.013)
##
## regioneast midlands                                       -0.017             -0.011
##                                                           (0.014)            (0.013)
##
## regionwest midlands                                        0.020              0.016
##                                                           (0.014)            (0.013)
##
## regioneast of england                                     0.108***           0.087***
##                                                           (0.015)            (0.014)
##
## regionlondon                                              0.205***           0.089***
##                                                           (0.014)            (0.013)
##
## regionsouth east                                          0.167***           0.126***
##                                                           (0.014)            (0.013)
##
## regionsouth west                                          0.036**             0.016
##                                                           (0.014)            (0.013)
##
## regionwales                                              -0.074***          -0.069***
##                                                           (0.017)            (0.016)
##
## regionscotland                                            0.062***           0.047***
##                                                           (0.014)            (0.013)
##
## regionnorthern ireland                                     0.010             -0.001
##                                                           (0.016)            (0.015)
##
```

```
## Constant                            1.032***          2.215***          1.025***
##                                     (0.014)           (0.011)           (0.017)
##
## ------------------------------------------------------------------------------------
## Observations                         58,381            58,399            58,381
## R2                                   0.128             0.017             0.135
## Adjusted R2                          0.128             0.017             0.135
## Residual Std. Error           0.589 (df = 58379) 0.625 (df = 58387) 0.586 (df = 58368)
## ====================================================================================
## Note:                                                    *p<0.1; **p<0.05; ***p<0.01
##                                                 Robust standard errors in parenthesis
```