Computer Lab 3

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 graphiscs
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)
data_USoc <- as.data.frame(data_USoc)    # ensure data frame structure
names(data_USoc)
```

Let us ensure that categorical variables are stored as `factor` variables. It is easiest to work with these in R.

```
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. As workers work for dy we shall also adjust for increasing price levels (as measured`mutate` function. We call this 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
              XXXX(lnhrpay = XXXX(hrpay))
```

As we wanted to save these additional variables we assign the result of the operation to `data_USoc`.

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 sometrhing similar for Lecture 2.

```
## # A tibble: 12 x 6
## # Groups:   region [12]
##    region                   `2009` `2010` `2011` `2012` `2013`
##    <fct>                     <int>  <int>  <int>  <int>  <int>
##  1 north east                 1011   1764   1741    805     47
##  2 north west                 2628   4762   4406   2186    113
##  3 yorkshire and the humber   2009   3636   3585   1807    102
##  4 east midlands              1867   3345   3288   1679     78
##  5 west midlands              2168   3936   3678   1866     99
##  6 east of england            2121   3976   3754   1872    137
##  7 london                     3538   6793   6244   3169    250
##  8 south east                 3021   5521   5269   2518    132
##  9 south west                 1978   3469   3388   1633     86
## 10 wales                      1166   2211   2187   1065     47
## 11 scotland                   1827   3158   2908   1366     62
## 12 northern ireland           2029   1924   1765     82     NA
```

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`.

```
##
## ===================================================================
##                                         Dependent variable:
##                                  -------------------------------------
##                                               lnhrpay
## -------------------------------------------------------------------
## regionnorth west                              0.032**
##                                               (0.015)
##
## regionyorkshire and the humber                -0.012
##                                               (0.016)
##
## regioneast midlands                           -0.017
##                                               (0.016)
##
## regionwest midlands                            0.020
##                                               (0.016)
##
## regioneast of england                         0.108***
##                                               (0.016)
##
```

```
## regionlondon                                 0.205***
##                                               (0.015)
##
## regionsouth east                             0.167***
##                                               (0.015)
##
## regionsouth west                              0.036**
##                                               (0.016)
##
## regionwales                                  -0.074***
##                                               (0.018)
##
## regionscotland                               0.062***
##                                               (0.016)
##
## regionnorthern ireland                         0.010
##                                               (0.018)
##
## Constant                                     2.215***
##                                               (0.013)
##
## --------------------------------------------------------------------
## 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
```

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

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.

How would you interpret the estimated parameter value for `regionwales`? The coefficient value is -0.0743675 and indicates that on average hourly pay is 7.5 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.

```
##
## =========================================================
##                              Dependent variable:
##                     -------------------------------------
##                                    lnhrpay
## ---------------------------------------------------------
## educ                              0.094***
##                                    (0.001)
##
## Constant                          1.032***
##                                    (0.014)
```

```
##
## ---------------------------------------------------------
## Observations                       58,381
## R2                                  0.128
## Adjusted R2                         0.128
## Residual Std. Error          0.589 (df = 58379)
## F Statistic             8,600.210*** (df = 1; 58379)
## =========================================================
## Note:                         *p<0.1; **p<0.05; ***p<0.01
##                       Robust standard errors in parenthesis
```

```
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.

```
##
## ====================================================================================================
##                                                                 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***                   -0.00
##                                                                      (0.017)                     (0.0
##
## regionscotland                                                       0.062***                   0.04
##                                                                      (0.014)                     (0.
##
## regionnorthern ireland                                                0.010                      -0.
##                                                                      (0.016)                     (0.
##
## Constant                                 1.032***                    2.215***                   1.02
##                                          (0.014)                     (0.011)                     (0.
##
## ----------------------------------------------------------------------------------------------------
## Observations                              58,381                      58,399                     58,3
## R2                                        0.128                       0.017                       0.
## Adjusted R2                               0.128                       0.017                       0.
## 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*** (d
## ====================================================================================================
## Note:                                                                                   *p<0.1; **p<0
##                                                                              Robust standard errors
```

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