Introduction to Regression Analysis - Multivariate Regression

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

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, which are saved in a STATA datafile (extension `.dta`). There is a function which loads STATA file. It is called `read_dta` and is supplied by the `haven` package.

```
data_USoc <- read_dta("20222_USoc_extract.dta")
data_USoc <- as.data.frame(data_USoc)    # ensure data frame structure
names(data_USoc)
```

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

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

```
data_USoc$region <- as_factor(data_USoc$region)
data_USoc$male <- as_factor(data_USoc$male)
data_USoc$degree <- as_factor(data_USoc$degree)
data_USoc$race <- as_factor(data_USoc$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 %>%
              mutate(hrpay = paygu/(jbhrs*4)/(cpi/100)) %>%
              mutate(lnhrpay = log(hrpay))
```

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

Let's run a simple regression of `lnhrpay` on `educ` which is a variable which counts the years of formal education.

```
mod1 <- lm(lnhrpay~educ,data = data_USoc)
stargazer_HC(mod1)
```

```
##
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                       lnhrpay
## ---------------------------------------------------------------
## educ                                  0.094***
##                                        (0.001)
##
## Constant                              1.033***
##                                        (0.014)
##
## ---------------------------------------------------------------
## Observations                           58,942
## R2                                      0.128
## Adjusted R2                             0.128
## Residual Std. Error               0.589 (df = 58940)
## F Statistic                  8,651.315*** (df = 1; 58940)
## ===============================================================
## Note:                               *p<0.1; **p<0.05; ***p<0.01
##                         Robust standard errors in parenthesis
```

Before we continue we create a squared age variable

```
data_USoc <- data_USoc %>% mutate(agesq = age*age/100)
```

Now we allow for age as an additional explanatory variable. But we will allow for the effect of `age` to be nonlinear, in particular quadratic.

```
mod2 <- lm(lnhrpay~educ+age+agesq,data = data_USoc)
stargazer_HC(mod1,mod2)
```

```
##
## ===========================================================================
##                                        Dependent variable:
##                     ------------------------------------------------------
##                                             lnhrpay
##                                (1)                          (2)
## ---------------------------------------------------------------------------
## educ                         0.094***                     0.092***
##                              (0.001)                      (0.001)
##
## age                                                       0.072***
##                                                           (0.001)
##
## agesq                                                    -0.075***
##                                                           (0.001)
##
## Constant                     1.033***                    -0.495***
##                              (0.013)                      (0.025)
##
```

```
## ------------------------------------------------------------------------
## Observations                    58,942                        58,942
## R2                              0.128                         0.215
## Adjusted R2                     0.128                         0.215
## Residual Std. Error      0.589 (df = 58940)           0.559 (df = 58938)
## F Statistic         8,651.315*** (df = 1; 58940) 5,395.272*** (df = 3; 58938)
## ========================================================================
## Note:                                           *p<0.1; **p<0.05; ***p<0.01
##                                      Robust standard errors in parenthesis
```