Introduction to Regression Analysis 1

Preparing your workfile

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

```r
library(tidyverse)    # for almost all data handling tasks
library(readxl)       # to import Excel data
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
library(estimatr)     # use robust se
source("stargazer_HC.r")
```

Introduction

The data are a classic dataset used in econometrics.

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.

```r
mroz <- read_dta("mroz.dta")
mroz <- as.data.frame(mroz)    # ensure data frame structure
names(mroz)
```

```
##  [1] "inlf"     "hours"    "kidslt6"  "kidsge6"  "age"      "educ"
##  [7] "wage"     "repwage"  "hushrs"   "husage"   "huseduc"  "huswage"
## [13] "faminc"   "mtr"      "motheduc" "fatheduc" "unem"     "city"
## [19] "exper"    "nwifeinc" "lwage"    "expersq"
```

The variables have short descriptions: 1. inlf =1 if in labor force, 1975
2. hours hours worked, 1975
3. kidslt6 # kids < 6 years
4. kidsge6 # kids 6-18
5. age woman's age in yrs
6. educ years of schooling
7. wage estimated wage from earns., hours
8. repwage reported wage at interview in 1976
9. hushrs hours worked by husband, 1975
10. husage husband's age
11. huseduc husband's years of schooling
12. huswage husband's hourly wage, 1975
13. faminc family income, 1975
14. mtr fed. marginal tax rate facing woman
15. motheduc mother's years of schooling
16. fatheduc father's years of schooling
17. unem unem. rate in county of resid.

18. city =1 if live in SMSA
19. exper actual labor mkt exper
20. nwifeinc (faminc - wage*hours)/1000
21. lwage log(wage)
22. expersq exper^2

A standard regression

Let's start by running a standard regression of log wages (`lwage`) as dependent variable and a respondents education (`educ`) as the explanatory variable.

But before we do this we shall ensure that we remove those observations from the dataset for which we do not have a measure of wage (or log(wage)).

```
mroz <- mroz %>% filter(!is.na(lwage))
```

```
ols <- lm(lwage~educ,data = mroz)
stargazer_HC(ols)
```

```
##
## =========================================================
##                               Dependent variable:
##                      -------------------------------------
##                                     lwage
## ---------------------------------------------------------
## educ                              0.109***
##                                    (0.014)
##
## Constant                           -0.185
##                                    (0.185)
##
## ---------------------------------------------------------
## Observations                         428
## R2                                  0.118
## Adjusted R2                         0.116
## Residual Std. Error           0.680 (df = 426)
## F Statistic                56.929*** (df = 1; 426)
## =========================================================
## Note:                         *p<0.1; **p<0.05; ***p<0.01
##                      Robust standard errors in parenthesis
```

The IV estimator

Let's consider a respondent's father's education as an instrument for education. We therefore run a first stage regression:

```
iv_s1 <- lm(educ~fatheduc, data = mroz)
stargazer_HC(iv_s1)
```

```
##
## =========================================================
##                               Dependent variable:
##                      -------------------------------------
##                                      educ
## ---------------------------------------------------------
## fatheduc                          0.269***
##                                    (0.029)
##
## Constant                          10.237***
##                                    (0.276)
##
## ---------------------------------------------------------
## Observations                         428
## R2                                  0.173
## Adjusted R2                         0.171
## Residual Std. Error           2.081 (df = 426)
## F Statistic                88.841*** (df = 1; 426)
## =========================================================
## Note:                         *p<0.1; **p<0.05; ***p<0.01
##                     Robust standard errors in parenthesis
```

What we learn from this is that the (`fatheduc`) is indeed related to the `educ` variable. Hence we feel justified in using this in our IV regression. But do remember that you will have to make an argument why `fatheduc` is a valid instrument, we cannot formally show that it is unrelated to the error term.

```r
iv <- ivreg(lwage~educ|fatheduc,data=mroz)
stargazer_HC(iv)
```

```
##
## =========================================================
##                               Dependent variable:
##                     -------------------------------------
##                                     lwage
## ---------------------------------------------------------
## educ                               0.059*
##                                    (0.035)
##
## Constant                            0.441
##                                    (0.446)
##
## ---------------------------------------------------------
## Observations                         428
## R2                                  0.093
## Adjusted R2                         0.091
## Residual Std. Error           0.689 (df = 426)
## =========================================================
## Note:                         *p<0.1; **p<0.05; ***p<0.01
##                     Robust standard errors in parenthesis
```

We can show all three estimates in the same table (omitting the F statistic as this would make the table very wide).

```r
stargazer_HC(ols,iv,iv_s1, omit.stat = "f")
```

```
##
## =======================================================================
##                                          Dependent variable:
##                                ----------------------------------------
##                                          lwage                educ
##                                     OLS      instrumental     OLS
##                                                variable
##                                     (1)          (2)          (3)
## ----------------------------------------------------------------------
## educ                             0.109***       0.059
##                                   (0.013)      (0.037)
##
## fatheduc                                                    0.269***
##                                                              (0.029)
##
## Constant                          -0.185        0.441      10.237***
##                                   (0.171)      (0.465)      (0.272)
##
## ----------------------------------------------------------------------
## Observations                        428          428          428
## R2                                 0.118        0.093        0.173
## Adjusted R2                        0.116        0.091        0.171
## Residual Std. Error (df = 426)     0.680        0.689        2.081
## ======================================================================
## Note:                                      *p<0.1; **p<0.05; ***p<0.01
##                                  Robust standard errors in parenthesis
```

Clearly the estimates for the `educ` variable are substantially different when comparing `ols` and `iv`. We really only want to revert to the `iv` model if there is evidence that the `educ` variable is indeed endogenous. The standard test applied in thsi context is the Wu-Hausmann test of endogeneity (H0: `educ` is exogenous). The easiest way to obtain this is to call `summary(iv, , diagnostics = TRUE)` where `iv` is the name we have given our IV regresison output:

```
summary(iv, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = lwage ~ educ | fatheduc, data = mroz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0870 -0.3393  0.0525  0.4042  2.0677
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.44110    0.44610   0.989   0.3233
## educ         0.05917    0.03514   1.684   0.0929 .
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   1 426     88.84  <2e-16 ***
## Wu-Hausman         1 425      2.47   0.117
## Sargan             0  NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6894 on 426 degrees of freedom
## Multiple R-Squared: 0.09344, Adjusted R-squared: 0.09131
## Wald test: 2.835 on 1 and 426 DF,  p-value: 0.09294
```

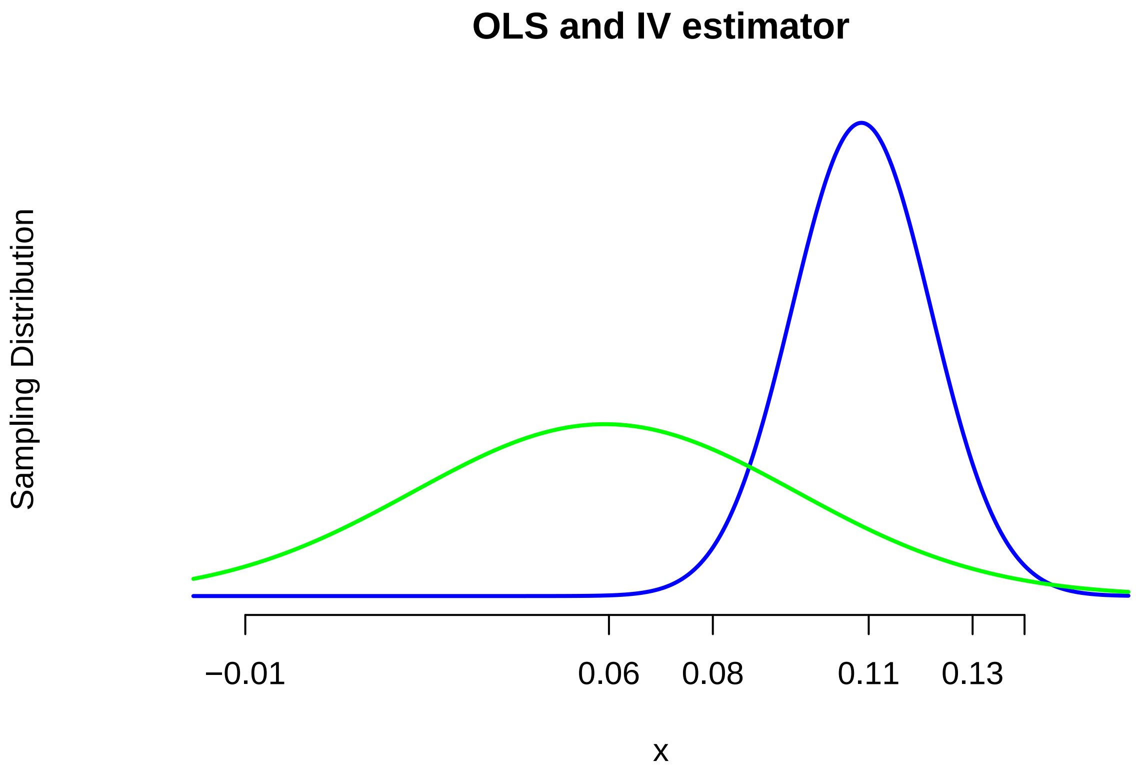
Note that from here you can read that the p-value for the Wu-Hausmann test is 0.117. So, for instance, at a

5% significance level we would not reject the null hypothesis that `educ` is actually exogenous.

Implications of the different estimators

Recall that the estimated coefficients are merely one draw from an underlying random distribution. The sampling distributions (i.e. our sample estimates of these unknown distributions) are shown in the following graph. The distributions for both are normal distributions where the mean is equal to the respective sample estimate and the sd, is taken from the regression outputs.

```r
#pdf("Lecture8plot_R.pdf",width = 5.5, height = 4) # uncomment to save as pdf
x <- seq(-0.02, 0.16, length=1000)
y_ols <- dnorm(x, mean=0.1086, sd=0.0134)
y_iv <- dnorm(x, mean = 0.0592, sd = 0.0369)
plot(x, y_ols, type="l", col="blue", lwd=2, axes = FALSE,
     ylab = "Sampling Distribution", main = "OLS and IV estimator")
lines(x,y_iv,col="green", lwd = 2)
axis(side = 1, at = c(-0.01,0.06,0.08,0.11,0.13,0.14))
```



```r
#dev.off() # uncomment to save as pdf
```