

# In-Class Lab 14

*ECON 4223 (Prof. Tyler Ransom, U of Oklahoma)*

*October 23, 2018*

The purpose of this in-class lab is to use R to practice with two-stage least squares (2SLS) estimation. The lab should be completed in your group. To get credit, upload your .R script to the appropriate place on Canvas.

## For starters

Open up a new R script (named ICL14\_XYZ.R, where XYZ are your initials) and add the usual “preamble” to the top:

```
# Add names of group members HERE
library(tidyverse)
library(wooldridge)
library(broom)
library(AER)
library(magrittr)
library(stargazer)
```

## Load the data

We’re going to use data on working women.

```
df <- as_tibble(mroz)
```

## Summary statistics

Like last time, let’s use `stargazer` to get a quick view of our data:

```
df %>% as.data.frame %>% stargazer(type="text")
```

1. Is it a problem that `wage` and `lwage` have 428 observations, but all of the other variables have 753 observations?

## Drop missing wages

Using the `filter()` and `is.na()` functions, drop the observations with missing wages. (I suppress the code, since you should know how to do this.)

## The model

We want to estimate the return to education for women who are working, using mother’s and father’s education as instruments:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$$

where *wage* is the hourly rate of pay, *educ* is years of education, and *exper* is labor market experience (in years).

## First stage regression

Let's estimate the first stage regression, which is a regression of the endogenous variable (*educ*) on the instrument(s) (*motheduc* and *fatheduc*) and the exogenous explanatory variables (*exper* and *exper*<sup>2</sup>).<sup>1</sup>

Run this regression (again, I suppress the code). Call the estimation object `est.stage1`.

2. Double check that *motheduc* and *fatheduc* are jointly significant with an F-stat larger than 10:

```
linearHypothesis(est.stage1, c("motheduc", "fatheduc"))
```

## Second stage regression

In the second stage, we estimate the log wage equation above, but this time we include  $\widehat{educ}$  on the right hand side instead of *educ*, where  $\widehat{educ}$  are the fitted values from the first stage.

In R, we can easily access the fitted values by typing `fitted(est.stage1)`.

Let's estimate the second stage regression:

```
est.stage2 <- lm(log(wage) ~ fitted(est.stage1) + exper + I(exper^2), data=df)
```

## Both stages at once

The standard errors from the above second stage regression will be incorrect.<sup>2</sup> Instead, we should estimate these at the same time. We can do this with the `ivreg()` function, just like in the previous lab.

```
est.2sls <- ivreg(log(wage) ~ educ + exper + I(exper^2) | motheduc + fatheduc + exper + I(exper^2), data=df)
```

3. Estimate the OLS model (where *educ* is not instrumented). Then compare the output for all three models (OLS, 2SLS “by hand”, 2SLS “automatic”).

```
stargazer(est.ols, est.stage2, est.2sls, type="text")
```

4. Comment on the IV estimates. Do they make sense, relative to what we think would bias the returns to education? Is the exogeneity condition on *motheduc* and *fatheduc* plausible?

---

<sup>1</sup>Note that you can easily include the quadratic in experience as `I(exper^2)` without having to create this variable in a `mutate()` statement.

<sup>2</sup>The reason is that error term in the second stage regression includes the residuals from the first stage, but the standard errors fail to take this into account.