AS.180.633: Econometrics **Spring 2020**

Homework 3: Suggested Solutions

Instructor: Yingyao Hu *By: Tong Zhou*

3.22

You estimate a least-squares regression

$$y_i = \boldsymbol{x}_{1i}'\widetilde{\boldsymbol{\beta}}_1 + \widetilde{u}_i$$

and then regress the residuals on another set of regressors

$$\widetilde{u}_i = \boldsymbol{x}_{2i}'\widetilde{\boldsymbol{\beta}}_2 + \widetilde{e}_i.$$

Does this second regression give you the same estimated coefficients as from estimation of a least-squares regression on both set of regressors?

$$y_i = \boldsymbol{x}_{1i}'\widehat{\boldsymbol{\beta}}_1 + \boldsymbol{x}_{2i}'\widehat{\boldsymbol{\beta}}_2 + \widehat{e}_i$$

In other words, is it true that $\widetilde{\boldsymbol{\beta}}_2 = \widehat{\boldsymbol{\beta}}_2$? Explain your reasoning.

Proof. The residual from the regression of $\boldsymbol{y}$ on $\boldsymbol{X}_1$ is:

$$\widetilde{\boldsymbol{u}} = \boldsymbol{M}_1\boldsymbol{y},$$

where $\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{P}_1$ and $\boldsymbol{P}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$.
Then $\widetilde{\boldsymbol{\beta}}_2$ is obtained by the regression of $\widetilde{\boldsymbol{u}}$ on $\boldsymbol{X}_2$:

$$\widetilde{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\widetilde{\boldsymbol{u}} = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}. \tag{1}$$

By the *Frisch-Waugh-Lovell Theorem* (or *partialling-out operation*):

$$\widehat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}. \tag{2}$$

Thus, in general $\widetilde{\boldsymbol{\beta}}_2$ does not agree with $\widehat{\boldsymbol{\beta}}_2$.

There are two extreme conditions under which $\widetilde{\boldsymbol{\beta}}_2 = \widehat{\boldsymbol{\beta}}_2$:

- $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal, i.e. $\boldsymbol{X}_1'\boldsymbol{X}_2 = \boldsymbol{0}$. Under it, $\widetilde{\boldsymbol{\beta}}_2 = \widehat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y}$.
- $\boldsymbol{y}$ is in the column space of $\boldsymbol{X}_1$, i.e. $\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{0}$, under which $\widetilde{\boldsymbol{\beta}}_2 = \widehat{\boldsymbol{\beta}}_2 = \boldsymbol{0}$. □

3.26

Use the data set from Section 3.22.

- Estimate a log wage regression for the subsample of white male Hispanics. In addition to education, experience, and its square, include a set of binary variables for regions and marital status. For regions, create dummy variables for Northeast, South and West so that Midwest is the excluded group. For marital status, create variables for married, widowed or divorced, and separated, so that single (never married) is the excluded group.

- Repeat this estimation using a different econometric package. Compare your results. Do they agree?

We use *R* and *Stata* to do the regression respectively. Table 1 and Table 2 show that those estimates and standard errors are roughly equal. The only substantial blip is the significance level of the variable d_NE. In *Stata*, it is insignificantly different from 0, while in *R* it is significantly different from 0 at the level of 0.1. The reason is that its standard error is around the cut-off point between two significance levels. As a result, rounding errors lead to different inference about its parameter.

Note that some of you did not correctly transform the outcome variable "wage", which is defined to be ln(earnings/(hours × week)). See page 94 of Hanse's book.

Table 1: Regression Results Using Stata

	(1)
	wage
education	0.088^{***}
	(0.003)
exp_1	0.028^{***}
	(0.003)
exp_2	-0.036^{***}
	(0.005)
d_NE	0.062
	(0.038)
d_S	-0.068^{**}
	(0.031)
d_W	0.020
	(0.031)
d_married	0.178^{***}
	(0.024)
d_WD	0.086^{**}
	(0.042)
d_sep	0.017
	(0.058)
Constant	1.193^{***}
	(0.051)
Observations	4230

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2: Regression Results Using R

	Dependent variable:
	log_inc
edu	0.087***
	(0.003)
exp_1	0.028***
	(0.003)
exp_2	−0.036***
	(0.005)
d_NE	0.063*
	(0.038)
d_S	−0.066**
	(0.031)
d_W	0.018
	(0.030)
d_married	0.191***
	(0.022)
d_widow	0.091**
	(0.041)
d_sep	0.020
	(0.058)
Constant	1.208***
	(0.051)
Observations	4,230
Note:	*p<0.1; **p<0.05; ***p<0.01

R codes

```
# load data
data <- read.csv("cps09mar.csv",header = TRUE,sep=",")

# Transform raw variables: log(income), education, experience and
   sqaured experience
log_inc <- log(data$earnings/(data$hours*data$week))
edu <- data$education
exp_1 <- data$age - data$education - 6
exp_2 <- exp1^2/100

# dummies for regions
d_NE <- ifelse(data$region==1, 1, 0)
d_S <- ifelse(data$region==3, 1, 0)
d_W <- ifelse(data$region==4, 1, 0)

# dummies for marital status
d_married <- ifelse(data$marital==1 | data$marital==2, 1, 0)
d_widow <- ifelse(data$marital==4 | data$marital==5, 1, 0)
d_sep <- ifelse(data$marital==6, 1, 0)

# subsample of white male Hispanics
subsample <- data$race==1 & data$female==0 & data$hisp == 1

# Run regression
regression <- lm(log_inc ~ edu + exp_1 + exp_2 + d_NE+d_S+d_W+d_
   married+d_widow+d_sep,data,subsample)

# Report
stargazer(regression, title="Regression Results Using R", align=T)
```

Stata codes

```
** 3.26
gen wage = log(earnings/(hours*week))

gen exp_1 = age - education - 6

gen exp_2 = exp_1^2/100

gen d_NE = (region == 1)

gen d_S =  (region == 3)

gen d_W = (region == 4 )

gen d_married = (marital == 1 | marital == 2 | marital == 3 )

gen d_WD = ( marital == 4 | marital == 5 )

gen d_sep =  ( marital == 6 )

reg wage education exp_1 exp_2 d_NE d_S d_W d_married d_WD d_sep if
   race == 1 & female ==0 & hisp == 1

est store A

esttab A using example5.tex ,b(3) star(* 0.1 ** 0.05 *** 0.01) se(3)
    compress label ///
title(Regression Results Using Stata)
```