Lecture 4 - agenda

1. FW for the population regression (review)

2. FW for OLS

3. an example

4. Omitted variable formula

Frisch-Waugh theorem

Population regression: $y_i = x_i\beta^* + u_i$

Auxilliary regression of x_{ji} on all the other $x's$:

$$x_{ji} = x'_{(\sim j)i}\pi + \xi_i.$$

The j^{th} row of β^* is:

$$\beta_j^* = E[\xi_i^2]^{-1}E[\xi_i y_i]$$

β_j^* = coefficient from univariate regression of y_i on x_{ji}, after "partialling out" other $x's$.

How does the proof work? Use FOC for β^* and π!

$$\begin{aligned}
E[\xi_i y_i] &= E[\xi_i(\beta_1^* x_{1i} + \beta_2^* x_{2i} + ... + \beta_j^* x_{ji} + ... + \beta_K^* x_{Ki} + u_i)] \\
&= \beta_1^* E[\xi_i x_{1i}] + \beta_2^* E[\xi_i x_{2i}] + ... + \beta_j^* E[\xi_i x_{ji}] + ... + \beta_K^* E[\xi_i x_{Ki}] \\
&\quad + E[\xi_i u_i]
\end{aligned}$$

FOC for $\pi \Rightarrow E[x_{(\sim j)i}\xi_i] = 0 \Rightarrow E[\xi_i x_{ni}] = 0$ unless $n = j$

FOC for $\beta^* \Rightarrow E[x_i u_i] = 0 \Rightarrow E[\xi_i u_i] = E[(x_{ji} - x'_{(\sim j)i}\pi)u_i] = 0$

And: $E[\xi_i x_{ji}] = E[\xi_i(x'_{(\sim j)i}\pi + \xi_i)] = E[\xi_i^2]$ using the FOC for π (again). So

$$E[\xi_i y_i] = \beta_j^* E[\xi_i^2] \Rightarrow \beta_j^* = E[\xi_i^2]^{-1} E[\xi_i y_i]$$

Suppose we have a constant and one other x variable: $x_i' = (1, x_{2i})$:

$$y_i = \beta_1^* + \beta_2^* x_{2i} + u_i$$

In this case, auxilliary regression is

$$x_{i2} = 1 \bullet \pi + \xi_i$$

And we know $\pi = E[x_{i2}]$ (population regression = CEF). So in this case, $\xi_i = x_{i2} - E[x_{i2}]$. Therefore:

$$\begin{aligned} \beta_2^* &= E[(x_i - E[x_i])^2]^{-1} E[(x_i - E[x_i]) y_i] \\ &= Var[x_i]^{-1} Cov[x_i, y_i] \end{aligned}$$

Now let's move from the population regression to the OLS regression. Recall the OLS objective is

$$min_{\ \beta} \quad \sum_{i=1}^{N}(y_i - x_i'\beta)^2$$

The FOC (which defines $\widehat{\beta}$) is:

$$\begin{aligned}
\sum_{i=1}^{N} x_i(y_i - x_i'\widehat{\beta}) &= 0 \qquad \Rightarrow \frac{1}{N}\sum_{i=1}^{N} x_i(y_i - x_i'\widehat{\beta}) \\
\Rightarrow \frac{1}{N}\sum_{i=1}^{N} x_i y_i &= \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)\widehat{\beta} \\
\Rightarrow \widehat{\beta} &= \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)^{-1}\frac{1}{N}\sum_{i=1}^{N} x_i y_i
\end{aligned}$$

$$\widehat{\beta} = \left(\frac{1}{N} \sum_{i=1}^{N} x_i x_i' \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} x_i y_i$$

c/w population regression:

$$\beta^* = E[x_i x_i']^{-1} E[x_i y_i]$$

repl. $E[x_i x_i']$ with $S_{xx} = \frac{1}{N} \sum_{i=1}^{N} x_i x_i'$

repl. $E[x_i y_i]$ with $S_{xy} = \frac{1}{N} \sum_{i=1}^{N} x_i y_i$

The 3 properties of the population regression are also true of the OLS regression. For the pop. regression, these come from FOC: $E[x_i(y_i - x_i'\beta^*)] = 0$.

For the OLS regession, these come from FOC:

$\sum_{i=1}^{N} x_i(y_i - x_i'\widehat{\beta}) = 0$

a. if x_i contains a constant, then $\bar{y} = \bar{x}'\widehat{\beta}$: the regression model "fits the mean of y"

b. if x_i contains a dummy variable for membership in group g then $\bar{y}_g = \bar{x}_g'\widehat{\beta}$: the regression model "fits the mean of y for subgroup g"

c. Frisch-Waugh (FW) for OLS: The j^{th} row of $\widehat{\beta}$ is:

$$\widehat{\beta}_j = [\frac{1}{N}\sum_{i=1}^{N}\widehat{\xi}_i^2]^{-1}[\frac{1}{N}\sum_{i=1}^{N}\widehat{\xi}_i y_i]$$

where $\widehat{\xi}_i$ is the *estimated residual* from an OLS regression of x_{ji} on all the other $x's$:

$$x_{ji} = x'_{(\sim j)i}\widehat{\pi} + \widehat{\xi}_i.$$

How are we going to prove FW for OLS?

(i) define $\widehat{u}_i = y_i - x_i'\widehat{\beta}$. We know $\frac{1}{N}\sum_{i=1}^{N} x_i\widehat{u}_i = 0$

(ii) define $\widehat{\xi}_i = x_{ji} - x'_{(\sim j)i}\widehat{\pi}$. We know $\frac{1}{N}\sum_{i=1}^{N} x_{(\sim j)i}\widehat{\xi}_i = 0$

(iii) write: $y_i = \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + ... + \widehat{\beta}_j x_{ji} + ... + \widehat{\beta}_K x_{Ki} + \widehat{u}_i$

and form

$$\frac{1}{N}\sum_{i=1}^{N}\widehat{\xi}_i y_i = \frac{1}{N}\sum_{i=1}^{N}\widehat{\xi}_i(\widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + ... + \widehat{\beta}_j x_{ji} + ... + \widehat{\beta}_K x_{Ki} + \widehat{u}_i)$$

What terms are equal to 0 from the 2 FOC?

An example: calculating the relationship between gender and self-reported health, controlling for wage

Sample: people age 30-60 in March 2014 Current Pop Survey

Questions: age, education, race, gender, earnings last year, hours worked last year, self-reported health (1-5 scale)

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
age	41754	44.23804	8.63941	1847115	30.00000	60.00000
educ	41754	13.99313	2.91045	584269	0	20.00000
twage	41754	25.93441	19.65316	1082865	6.00000	200.00000
logwage	41754	3.05289	0.61457	127470	1.79176	5.29832
healthr	41754	3.90564	0.92125	163076	1.00000	5.00000
female	41754	0.47653	0.49945	19897	0	1.00000

Pearson Correlation Coefficients, N = 41754

	age	educ	twage	logwage	healthr	female
age	1.00000	-0.02521	0.07538	0.07627	-0.13714	0.00752
educ	-0.02521	1.00000	0.39451	0.45478	0.19646	0.07658
twage	0.07538	0.39451	1.00000	0.89902	0.13500	-0.14397
logwage	0.07627	0.45478	0.89902	1.00000	0.16880	-0.16728
healthr	-0.13714	0.19646	0.13500	0.16880	1.00000	-0.02178
female	0.00752	0.07658	-0.14397	-0.16728	-0.02178	1.00000

Self reported Health and Gender
Age 30-60 and worked last year, March 2014)

	males	females	all
Health = 1 (poor)	0.9	1.1	1.0
Health = 2 (fair)	4.8	5.6	5.2
Health = 3 (good)	26.0	26.3	26.2
Health = 4 (very good)	37.6	37.7	37.7
Health = 5 (excellent)	30.7	29.3	30.0

	Dependent variable:			
			Female	
	Health	Health	(Aux. Regr.)	Health
Constant	3.9248 (0.0062)	3.1220 (0.0239)	0.8916 (0.0122)	3.9056 (0.0045)
Female	-0.0402 (0.0091)	0.0123 (0.0090)	--	--
log(wage)	--	0.2547 (0.0073)	-0.1359 (0.0039)	--
resid-Female	--	--	--	0.0123 (0.0090)

Omitted Variables Formula

FW: if you add a regressor, the coefficient is "as if" you added only the part of that regressor that is *unexplained by the other regressors:* $x_{ji} - x'_{(\sim j)i}\pi$ (or $\widehat{\pi}$)

What about the opposite direction? What happens if you forget a regressor?

$$y_i = \beta_1^* x_{1i} + \beta_2^* x_{2i} + \beta_3^* x_{ji} + u_i$$

Suppose we don't include x_{3i}?

a. Population version

Aux. model for the *omitted variable*: $x_{3i} = \pi_1 x_{1i} + \pi_2 x_{2i} + \xi_i$

Then:

$$
\begin{aligned}
y_i &= \beta_1^* x_{1i} + \beta_2^* x_{2i} + \beta_3^* (\pi_1 x_{1i} + \pi_2 x_{2i} + \xi_i) + u_i \\
&= (\beta_1^* + \beta_3^* \pi_1) x_{1i} + (\beta_2^* + \beta_3^* \pi_2) x_{2i} + \beta_3^* \xi_i + u_i \\
&= \beta_1^0 x_{1i} + \beta_2^0 x_{2i} + \eta_i
\end{aligned}
$$

Notice that $E[(x_{1i}, x_{2i})' \eta_i] = E[(x_{1i}, x_{2i})' (\beta_3^* \xi_i + u_i)] = (0, 0)'$.

So (β_1^0, β_2^0) satisfy FOC for population regression of y_i on (x_{1i}, x_{2i})

Conclusion: If

$$y_i = \beta_1^* x_{1i} + \beta_2^* x_{2i} + \beta_3^* x_{ji} + u_i$$

and we don't include x_{3i}, the coefficient on x_{2i} is:

$$\beta_2^0 = \beta_2^* + \beta_3^* \pi_2$$

where π_2 is the coefficient on x_{2i} from the regression of the omitted variable on the remaining $x's$:

$$x_{ji} = \pi_1 x_{1i} + \pi_2 x_{2i} + \xi_i$$

Intuition: you forgot x_{3i} so the house elf is doing the best he can to predict y given what he has to work with. The best he can do is use the other $x's$ to predict x_{3i}.

b. OLS (sample) version

Aux. model for the omitted variable: $x_{3i} = \hat{\pi}_1 x_{1i} + \hat{\pi}_2 x_{2i} + \hat{\xi}_i$

Then:

$$\begin{aligned} y_i &= \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3(\hat{\pi}_1 x_{1i} + \hat{\pi}_2 x_{2i} + \hat{\xi}_i) + \hat{u}_i \\ &= (\hat{\beta}_1 + \hat{\beta}_3 \hat{\pi}_1) x_{1i} + (\hat{\beta}_2 + \hat{\beta}_3 \hat{\pi}_2) x_{2i} + \hat{\beta}_3 \hat{\xi}_i + \hat{u}_i \\ &= \hat{\beta}_1^0 x_{1i} + \hat{\beta}_2^0 x_{2i} + \hat{\eta}_i \end{aligned}$$

Notice $\frac{1}{N}\sum_{i=1}^{N}(x_{1i}, x_{2i})'\hat{\eta}_i = \frac{1}{N}\sum_{i=1}^{N}(x_{1i}, x_{2i})'(\hat{\beta}_3\hat{\xi}_i + \hat{u}_i) = (0, 0)'$.

So $(\hat{\beta}_1^0, \hat{\beta}_2^0)$ satisfy FOC for OLS regression of y_i on (x_{1i}, x_{2i})

Summary of OLS version: OLS if you used x_{3i}:

$$y_i = \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \widehat{\beta}_3 x_{3i} + \widehat{u}_i$$

If we don't include x_{3i}, the OLS coefficient on x_{2i} is:

$$\widehat{\beta}_2^0 = \widehat{\beta}_2 + \widehat{\beta}_3 \widehat{\pi}_2$$

where $\widehat{\pi}_2$ is the coefficient on x_{2i} from the OLS regression of the omitted variable on the remaining $x's$:

$$x_{ji} = \widehat{\pi}_1 x_{1i} + \widehat{\pi}_2 x_{2i} + \widehat{\xi}_i$$

Same intuition as for population version.

An example: Suppose we are interested in comparing wages of immigrants and natives. We will use a sample of women age 35-44 who were surveyed in the March 2012 CPS about earnings last year. We consider two models:

$$log(wage) = \beta_1 + \beta_2 Immigrant + \beta_3 Education \tag{1}$$

and a simpler model:

$$log(wage) = \beta_1^0 + \beta_2^0 Immigrant \tag{2}$$

Using the OVF, we can relate β_2^0 to β_2, β_3, and the coefficient from an auxilliary regression:

$$Education = \pi_1 + \pi_2 Immigrant$$

We know that

$$\beta_2^0 = \beta_2 + \beta_3 \pi_2$$

This holds both for the population regression and for the OLS estimates. What is π_2? In this regression we are including a dummy for immigrant status. So in the population version

$$\pi_2 = E[education|immigrant] - E[education|native]$$

and in the sample version π_2 will equal the difference in mean education between immigants and natives. This will be a pretty big negative number! And since β_3 is a number like 0.11 we can conclude that if you "leave out" education, you will tend to find that immigrants earn less.

Table 1: Relationships Between Wages, Education and Immigrant Status
for Working Women Age 35-44 in March 2012 Current Population Survey

	Log Wage (1)	Log Wage (2)	Immigrant Status (3)	Years of Education (4)	Log Wage (5)
Immigrant Status	-0.1800	--	--	-1.4920	-0.0101
	(0.0165)			(0.0674)	(0.0129)
Years of Education	--	0.1141	-0.0297	--	0.1139
		(0.0021)	(0.0013)		(0.0021)
R-squared of model	0.0111	0.2239	0.0442	0.0442	0.2239

Notes: Each column reports separate regression of dependent variable in column heading on regressors shown in rows, plus a constant (coefficient not reported). Sample is females age 35-44 in March 2012 CPS who reported earnings for the last year. "Wage" refers to average hourly earnings last year, n=10,601. Means and standard deviations of dependent variables are: for log wage 2.8527 and 0.6677; for immigrant status 0.1872 and 0.3901; for education 14.1724 and 2.7676.

NOTE: $-0.0101 + 0.1139 \times (-1.4920) = -0.1800$

$0.1139 - 0.0101 \times (-0.0297) = 0.1141$

The real importance of the OVF is that we can often think about how the omission of a variable affects the estimated coefficient of variables we include, even if we can't estimate the auxilliary regression. Classic example: suppose the true model is:

$$log(wage) = \beta_1 + \beta_2 Education + \beta_3 Ability$$

But we don't know "ability", and estimate the simpler model:

$$log(wage) = \beta_1^0 + \beta_2^0 Education$$

We know that

$$\beta_2^0 = \beta_2 + \beta_3 \pi_2$$

where in this case, π_2 is the coefficient from a regression of ability on education. Many people (especially those with high education) think that $\beta_3 > 0$ and $\pi_2 > 0$, which leads them to believe that a model that does not control for ability "overstates" the effect of education.