Lecture 3

Today's agenda

1. recap ideas from Lecture 2:

$E[y_i|x_i]$, population regression $x_i'\beta^*$, and OLS

2. an example where $E[y_i|x_i]{\neq}x_i'\beta^*$

3. properties of the population regression (F-W)

4. parallel properties of OLS

Recap

- vector notation: $y_i = x_i'\beta + u_i \qquad x_i' = (1, x_{1i}, x_{2i} ... x_{Ji})$

- $CEF \equiv E[y_i|x_i]$, a (possibly messy) function of x

- forecast error $\epsilon_i = y_i - E[y_i|x_i]$

$$E[\epsilon_i|x_i] = 0 \text{ and } E[\epsilon_i h(x_i)] = 0 \text{for } \textit{any } h(x_i)$$

Aside: what if $x_i = 1$ (only constant)$\Rightarrow E[y_i|1] = E[y_i]$

- showed CEF minimizes $E[(y_i - m(x_i))^2]$ among all possible $m(.)$ functions

Next: the *population regression function (PRF)*

- for a particular set of $x's$, a *regression function* is just a linear combination $x_i'\beta$

- PRF: $\beta^* = argmin_\beta \, E[(y_i - x_i'\beta)^2]$

- FOC: $E[x_{ji}(y_i - x_i'\beta^*)] = 0$, one row for each covariate $j = 1...J$.

- re-write as $E[x_{ji}\, x_i'\beta^*] = E[x_{ji}\, y_i]$

- e.g., 3-covariate case:

$$\begin{array}{lcl} E[x_{1i}x_{1i}\beta_1^* + x_{1i}x_{2i}\beta_2^* + x_{1i}x_{3i}\beta_3^*] & & E[x_{1i}y_i] \\ E[x_{2i}x_{1i}\beta_1^* + x_{2i}x_{2i}\beta_2^* + x_{2i}x_{3i}\beta_3^*] & = & E[x_{2i}y_i] \\ E[x_{3i}x_{1i}\beta_1^* + x_{3i}x_{2i}\beta_2^* + x_{3i}x_{3i}\beta_3^*] & & E[x_{2i}y_i] \end{array}$$

$$\begin{array}{rcl} E[x_{1i}x_{1i}\beta_1^* + x_{1i}x_{2i}\beta_2^* + x_{1i}x_{3i}\beta_3^*] & & E[x_{1i}y_i] \\ E[x_{2i}x_{1i}\beta_1^* + x_{2i}x_{2i}\beta_2^* + x_{2i}x_{3i}\beta_3^*] & = & E[x_{2i}y_i] \\ E[x_{3i}x_{1i}\beta_1^* + x_{3i}x_{2i}\beta_2^* + x_{3i}x_{3i}\beta_3^*] & & E[x_{2i}y_i] \end{array}$$

Now use matrix notation:

$$E\begin{pmatrix} x_{1i}x_{1i} & x_{1i}x_{2i} & x_{1i}x_{3i} \\ x_{2i}x_{1i} & x_{2i}x_{2i} & x_{2i}x_{3i} \\ x_{3i}x_{1i} & x_{3i}x_{1i} & x_{3i}x_{3i} \end{pmatrix}\begin{pmatrix} \beta_1^* \\ \beta_2^* \\ \beta_3^* \end{pmatrix} = E\begin{pmatrix} x_{1i}y_i \\ x_{2i}y_i \\ x_{3i}y_i \end{pmatrix}$$

Or

$$\begin{array}{rcl} E[x_i x_i']\beta^* & = & E[x_i y_i] \\ \Rightarrow \beta^* & = & E[x_i x_i']^{-1}E[x_i y_i] \end{array}$$

Good properties of PRF

1. if $E[y_i|x_i] = x_i'\beta^e$ then $\beta^* = \beta^e$ (i.e., PRF = CEF)

2. $x_i'\beta^*$ is the *best linear approximation* to $E[y_i|x_i]$

How did we prove that? Define $\epsilon_i \equiv y_i - E[y_i|x_i]$.

$$\begin{aligned} y_i - x_i\beta &= y_i - E[y_i|x_i] + E[y_i|x_i] - x_i\beta \\ &= \epsilon_i + E[y_i|x_i] - x_i\beta \\ \Rightarrow E[(y_i - x_i\beta)^2] &= E[\epsilon_i^2] + E[(E[y_i|x_i] - x_i\beta)^2] \end{aligned}$$

min LHS $\Leftrightarrow$ min RHS $\Leftrightarrow$min $E[(E[y_i|x_i] - x_i\beta)^2]$

A direct approach:

$$E[(y_i - x_i\beta)^2] = E[\epsilon_i^2] + E[(E[y_i|x_i] - x_i\beta)^2]$$

To minimize RHS w.r.t. β we have FOC:

$$\begin{aligned} -2E[x_i(E[y_i|x_i] - x_i'\beta)] &= 0 \\ \Rightarrow E[x_i x_i'\beta] &= E[x_i E[y_i|x_i]] \\ &= E[E[x_i y_i|x_i] \\ &= E[x_i y_i] \end{aligned}$$

which is the FOC for the PRF!

Special case: groups $0, 1, 2$; indicators D_{1i}, D_{2i}; $x_i' = (1, D_{1i}, D_{2i})$;

$E[y_i | i \in group\ g] = \mu_g$. Then:

$$E[y_i|x_i] = \mu_0 + D_{1i}(\mu_1 - \mu_0) + D_{2i}(\mu_2 - \mu_0)$$

which means $E[y_i|x_i]$ is linear in x_i. Thus $x_i\beta^* = E[y_i|x_i]$, so we know:

$$\beta^* = \begin{pmatrix} \mu_0 \\ \mu_1 - \mu_0 \\ \mu_2 - \mu_0 \end{pmatrix}$$

If $E[y_i|x_i]$ is not truly linear, we end up with an approximation error at each x_i.

Write:

$$\begin{aligned} y_i &= E[y_i|x_i] + \epsilon_i \\ E[y_i|x_i] &= x_i'\beta^* + v_i \\ \Rightarrow y_i &= x_i'\beta^* + v_i + \epsilon_i \\ &= x_i'\beta^* + u_i \end{aligned}$$

So the error in the population regression is $v_i + \epsilon_i$, where v_i is function of x_i. Note that $E[x_i v_i] = 0$ but $E[v_i|x_i] \neq 0$ (in general).

Example: data on education and earnings from Am. Community Survey

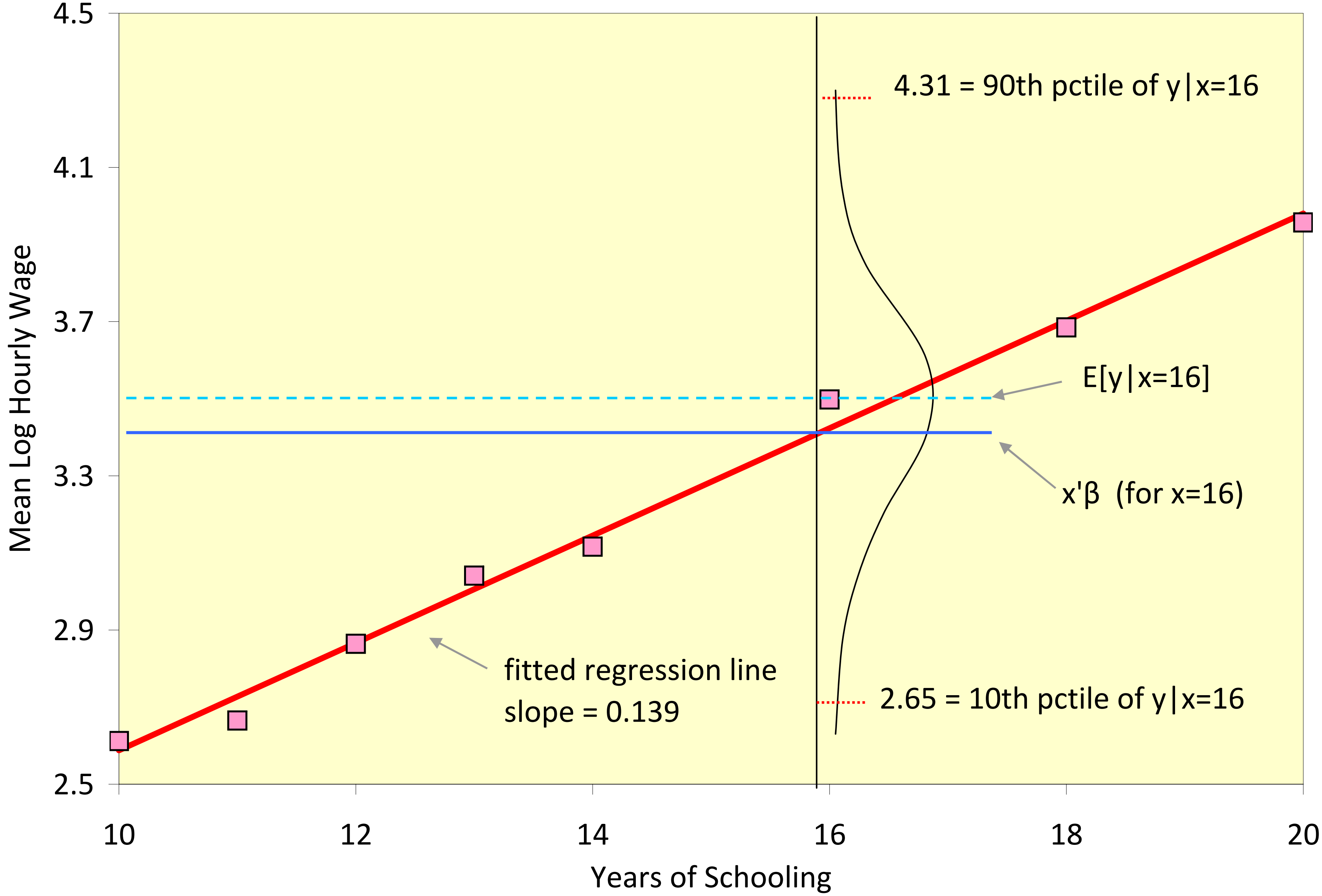
-1.5 million obs per year; we use 2011/2012 to get "big sample"

- native born men; 25-30 years since they should have finished school

  i.e., age-education-8 $\in [25, 30]$

- worked >20 weeks last year; 20+ hours week, earned $> \$2800$

- pretend we can ignore estimation errors



4.5
4.1
3.7
3.3
2.9
2.5
Mean Log Hourly Wage
4.31 = 90th pctile of y|x=16
E[y|x=16]
x'β (for x=16)
fitted regression line
slope = 0.139
2.65 = 10th pctile of y|x=16
10
12
14
16
18
20
Years of Schooling

Implications of $E[x_i u_i] = 0$ (defining property of PRF)

a) If x_i includes a constant, then $E[u_i] = 0$.

why? $E[1 \cdot u_i] = E[u_i] = 0$.

A nice implication is that when x includes a constant:

$$\begin{aligned} E[y_i] &= E[x_i'\beta^* + u_i] \\ &= E[x_i'\beta^*] + E[u_i] \\ &= E[x_i'\beta^*]. \end{aligned}$$

The mean of the PRF is the mean of y! This is why regressions should (almost) always include a constant.

b) If x_i includes a dummy for membership in subgroup g then $E[u_i|i \in g] = 0$.

why? Let $D_i = 1$ if $i \in g$, and 0 otherwise; let $\mu_g = E[y_i|i \in g]$.

Now:

$$\begin{aligned} E[u_i D_i] &= E[u_i D_i | D_i = 1] \times P(D_i = 1) + E[u_i D_i | D_i = 0] \times P(D_i = 0) \\ &= E[u_i | D_i = 1] \times P(D_i = 1) \end{aligned}$$

If the dummy D_i is included in x_i, we know $E[u_i D_i] = 0$, so

$$E[u_i | D_i = 1] = E[u_i | i \in g] = 0.$$

This is useful because it means:

$$E[u_i | i \in g] = E[y_i - x_i' \beta^* | i \in g] = \mu_g - E[x_i | i \in g]' \beta^* = 0$$

So the population regression *fits the mean* of group g exactly.

So to recap:

- if x_i includes a dummy for membership in subgroup g then $E[x_i|i \in g]'\beta^* = E[y_i|i \in g]$.

- if you have lots of data, you probably want to include a dummy for each separate subgroup in the data (or something "close" to that)

Next up: "Frisch-Waugh"

This shows how we can think of multivariate regression as a univariate regression after we "regress out" the other X's

c) the "Frisch-Waugh" theorem

The j^{th} row of β^* is:

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

where ξ_i is the residual from a population regression of x_{ji} on all the other $x's$:

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

Note that $E[\xi_i^2]^{-1} E[\xi_i y_i]$ is the formula for the population regression of y_i on ξ_i : So FW says that you can think of β_j^* as the coefficient from a univariate regression of y_i on x_{ji}, after "partialling out" all the other $x's$.

Proof: $x_i' = (x_{1i}, x_{2i} ... x_{ji} ... x_{Ki})$ has K elements.

Let $x_{(\sim j)i}$ be x_i *after removing row* j.

Now write the "auxilliary" regression of x_{ji} on $x_{(\sim j)i}$:

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

As usual, the FOC for π require $E[x_{(\sim j)i}\xi_i] = 0$.

Finally, since $y_i = x_i'\beta^* + u_i$ we can write:

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

Now notice that from the FOC for π, $E[\xi_i x_{mi}] = 0$ unless $m = j$.

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

So $E[\xi_i x_{mi}] = 0$ unless $m = j$

Also: $E[\xi_i u_i] = E[(x_{ji} - x'_{(\sim j)i}\pi)u_i] = 0$ because u_i is orthogonal to all the $x's$. So the *only nonzero term* on the r.h.s. is $\beta_j^* E[\xi_i x_{ji}] \Rightarrow$

$$E[\xi_i y_i] = \beta_j^* E[\xi_i x_{ji}]$$

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

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

One extremely useful version of FW: Suppose we have a constant and one other x variable: $x_i' = (1, x_{2i})$. Consider the population regression:

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

Then

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

Why? From FW, we can get β_2^* from a '2 step' approach: first regress x_{2i} on the other regressor (i.e., a constant), then regress y_i on the residual from the first regression. But what is the auxilliary regression of x_{i2} on a constant? This is:

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

And $\pi = E[x_{i2}]$ is the solution. So in this case, $\xi_i = x_{i2} - E[x_{i2}]$.

In fact, there is a slightly more general version of FW. Suppose we are interested in a subset of regressors, e.g., (x_{1i}, x_{2i}). Then the coefficients (β_1^*, β_2^*) can be expressed as the outcome of a two-step process: first consider the population regression of (x_{1i}, x_{2i}) on all the other regressors, then consider the population regression of y_i on the pair of residuals.

A version of this result: suppose that $x_i' = (1, x_{2i}, x_{3i}, ...x_{Ki})$. Then we can get the coefficients on the non-constant regressors by considering the population regression of y on the set of variables $(x_{2i} - E[x_{2i}], x_{3i} - E[x_{3i}]...)$. But this is just:

$$\begin{pmatrix} \beta_2^* \\ \beta_3^* \\ ... \\ \beta_K^* \end{pmatrix} = Var[x_{2i}, x_{3i}, ...x_{Ki}]^{-1} Cov[(x_{2i}, x_{3i}, ...x_{Ki})', y_i]$$

People often express the pop. regression in terms of variances and covariances, but this is a little sloppy unless y_i and *all the elements* of x_i have mean 0. In that case, you can write:

$$y_i = x_i'\beta^* + u_i$$

$$\beta^* \;=\; Var[x_i]^{-1}Cov[x_i, y_i]$$

which is certainly very nice looking!

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

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

The FOC 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:

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

So we are "matching moments":

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

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

Computer programs compute S_{xx}, S_{xy} and invert S_{xx} very efficiently

The 3 properties of the (infeasible) 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): The j^{th} row of $\widehat{\beta}$ is:

$$\widehat{\beta}_j = E[\widehat{\xi}_i^2]^{-1} E[\widehat{\xi}_i y_i]$$

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) OLS: get $\widehat{\beta}$, 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) OLS for auxilliary model: $\widehat{\xi}_i = x_{ji} - x'_{(\sim j)i}\widehat{\pi}$. We know $\frac{1}{N}\sum_{i=1}^{N} x_{(\sim j)i}\widehat{\xi}_i = 0$

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

Now 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?