14.387 Recitation 1
Expectations, Regressions, and Controls

Peter Hull

Fall 2014
Part 1: Expectations and their properties
One variable

Scalar random variable $x$:

Discrete $x$: $E[x] \equiv \sum_z zPr(x = z)$

Continuous $x$: $E[x] \equiv \int z f_x(z) \, dz$

Variance: $Var(x) \equiv E[(x - E[x])^2]$
Two variables

Scalar random variables $x$ and $y$:

Discrete $y$: \[ E[y|x] \equiv \sum z \Pr(y = z|x) \]

Continuous $y$: \[ E[y|x] \equiv \int z f_{y|x}(z) dz \]

Covariance: \[ \text{Cov}(x,y) \equiv E[(x - E[x])(y - E[y])] \]

Random or fixed?

- $x$ and $y$ are **uncorrelated** when $\text{Cov}(x,y) = 0$
- $y$ is **mean-independent** of $x$ when $E[y|x] = E[y]$

Which is stronger?
Two useful properties

- **Linearity**: for fixed $a, b, c,$ and $d$
  
  $$E[a + bx] = a + bE[x]$$
  
  $$\implies \text{Cov}(a + bx, c + dy) = bd\text{Cov}(x, y)$$

- **The Law of Iterated Expectations**:
  
  $$E[E[y|x]] = E[y]$$

  (Sloppy) proof of LIIE in continuous case:
  
  $$E[E[y|x]] \equiv \int \left( \int zf_y|x(z|w)dz \right) f_x(w)dw$$
  
  $$= \int z \int f_{x,y}(w,z)dwdz$$
  
  $$= \int zf_y(z)dz$$
  
  $$\equiv E[y]$$
Mean independence implies uncorrelatedness:

\[ E[(x - E[x])(y - E[y])] = E[E[(x - E[x])(y - E[y])|x]] \]
\[ = E[(x - E[x])(E[y|x] - E[y])] \]
\[ = E[(x - E[x]) \cdot 0] \]
\[ = 0 \]

Covariance with mean-zero r.v.s is the expectation of their product:

\[ E[(x - E[x])(y - E[y])] = E[xy - E[x]y - xE[y] + E[x]E[y]] \]
\[ = E[xy] - E[x]E[y] \]
\[ = E[xy], \text{ if either } E[x] = 0 \text{ or } E[y] = 0 \]
Part 2: Regressions, large and small
Bivariate regression

Scalar random variables $x_i$ and $y_i$:

$$(\alpha, \beta) = \arg\min_{a,b} E[(y_i - a - bx_i)^2]$$

FOC: $-2E[(y_i - \alpha - \beta x_i)] = 0$

$-2E[(y_i - \alpha - \beta x_i)x_i] = 0$

or

$$\alpha = E[y_i] - \beta E[x_i]$$

$$\beta E[x_i^2] = E[y_i x_i] - \alpha E[x_i]$$

Substituting:

$$\beta E[x_i^2] = E[y_i x_i] - E[y_i]E[x_i] + \beta E[x_i]^2$$

$$\beta = \frac{E[y_i x_i] - E[y_i]E[x_i]}{E[x_i^2] - E[x_i]^2} = \frac{\text{Cov}(y_i, x_i)}{\text{Var}(x_i)}$$
Multivariate regression

Scalar random variable $y_i$ and $k \times 1$ random vector $x_i$:

$$\beta = \arg \min_b E[(y_i - x_i'b)^2]$$

$\text{FOC}: -2E[x_i(y_i - x_i'\beta)] = 0$

(A useful matrix-’metrics resource: The Matrix Cookbook)

$$\beta = E[x_ix_i']^{-1}E[x_iy_i]$$

How do we reconcile this with the last slide? (Where did $\alpha$ go? What about $\text{Cov}()$ and $\text{Var}()$?)
Partialling out

Scalar, mean-zero random variables $y_i$, $x_{1i}$, and $x_{2i}$:

$$ (\beta, \gamma) = \arg\min_{b,c} E[(y_i - bx_{1i} - cx_{2i})^2] $$

FOC $\gamma$: $-2E[x_{2i}(y_i - bx_{1i} - \gamma x_{2i})] = 0$

IFT: $\gamma(b) = \frac{E[x_{2i}(y_i - bx_{1i})]}{E[x_{2i}^2]}$

Plug $\gamma(b)$ back in (sometimes called “concentrating out” $\gamma$):

$$ \beta = \arg\min_{b} E \left[ \left( y_i - bx_{1i} - \frac{E[x_{2i}(y_i - bx_{1i})]}{E[x_{2i}^2]} x_{2i} \right)^2 \right] $$

$$ = \arg\min_{b} E \left[ \left( \left( y_i - \frac{E[x_{2i}y_i]}{E[x_{2i}^2]} x_{2i} \right) - b \left( x_{1i} - \frac{E[x_{2i}x_{1i}]}{E[x_{2i}^2]} x_{2i} \right) \right)^2 \right] $$

A bivariate regression! But of what on what?
Partialling out (cont.)

- Special case of the Frisch-Waugh (sometimes -Lovell) theorem: If \( x_i = [x'_{1i}, x'_{2i}] \)', \( \tilde{x}_{1i} \) is the residual (vector) from regressing (each component of) \( x_{1i} \) on \( x_{2i} \), and \( \tilde{y}_i \) is the residual from regressing \( y_i \) on \( x_{2i} \), then all three are equivalent:
  1. The component \( \beta_1 \) of \( \beta = [\beta_1', \beta_2'] \)' from regressing \( y_i \) on \( x_i \)
  2. \( \tilde{\beta}_1 \) from regressing \( y_i \) on \( \tilde{x}_i \)
  3. \( \tilde{\beta}_1 \) from regressing \( \tilde{y}_i \) on \( \tilde{x}_i \)

- Partialling out \( x_{2i} \) from \( y_i \) is unnecessary! Why? Back to our example:

\[
y_i = \beta x_{1i} + \gamma x_{2i} + e_i \\
\tilde{y}_i = \beta \tilde{x}_{1i} + \tilde{e}_i \\
y_i = \beta \tilde{x}_{1i} + \tilde{e}_i + y_i - \tilde{y}_i \\
y_i = \beta \tilde{x}_{1i} + \left( \tilde{e}_i + \frac{E[x_{2i}y_i]}{E[x^2_{2i}]} x_{2i} \right)
\]

Why must the last line be a regression (and not just an equation)?
From population to sample

- Regression is a **feature of data**: just like expectation, correlation, etc.
- It’s a function of population second moments: so easy to estimate!

\[ \hat{\beta} = E_n[x_i x_i']^{-1} E_n[x_i y_i] \]

- A more matrix-y way to write \( \hat{\beta} \):

\[ E_n[x_i x_i']^{-1} E_n[x_i y_i] = \left( \frac{1}{n} \sum_i x_i x_i' \right)^{-1} \left( \frac{1}{n} \sum_i x_i y_i \right) \]
\[ = (X'X)^{-1} X' Y \]

where

\[ X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \]
Regression subtlety

- $\beta$ is a **feature of data**. We know what it is and we know that it (probably) exists, given any $y_i$ and $x_i$.
- We also how to estimate it; we know that (probably) $\hat{\beta} \xrightarrow{P} \beta$ (why?)
  (where “probably” $\equiv$ “given some innocuous technical conditions”)

- ...ok...but then.... what’s all the fuss about?
- Some common examples of fuss: “endogeneity,” “simultaneity,”
  “omitted variable bias,” “selection bias,” “measurement error,” “division bias,” etc. etc. etc.

The fuss.
Part 3: Controls: good and bad
You can’t always get what you want

- *reg y x* is **always** going to give you a $\hat{\beta}$ estimating the $\beta$ satisfying $E[x_i(y_i - x_i'\beta)] = 0$

- But what if this isn’t what you want? (When might you want it?)

- Ex: suppose we want $\beta$ from $y_i = \alpha + \beta x_i + \gamma a_i + \varepsilon_i$, where we know $E[\varepsilon_i|x_i, a_i] = 0$
  - We *reg y x* (maybe throw on a $, r$).
  - What do we get? What does $\hat{\beta}$ plim to? Could it be $\beta$?

- Obvious solution: just control for $a_i$. But why stop there?
Bad controls

- **Goal:** add *right* controls so that the regression $\beta$ you get is the $\beta$ you want (i.e. approximates the CEF you want)
- **Ex:** We randomly assign schooling $s_i \in \{0, 1\}$. Want the *causal* effect of schooling on income $y_i$ (a *causal* CEF)
  - Also measure race $b_i \in \{0, 1\}$ and post-schooling occupation $x_i \in \{0, 1\}$.
  - What regression should we run?
- **Natural choice:** $\beta$ satisfying $E[s_i(y_i - \alpha - \beta s_i)] = 0$
  - Another choice: $\beta$ satisfying $E[s_i(y_i - \alpha - \beta s_i - \gamma b_i)] = 0$. Better?
  - How about $\beta$ satisfying $E[s_i(y_i - \alpha - \beta s_i - \delta x_i)] = 0$?
Controlling composition

- **Potential outcomes**: \( \{y_{0i}, y_i\} \). Observe \( y_i = y_{0i} + (y_i - y_{0i})s_i \)
- **Bivariate regression**:

\[
E[y_i|s_i = 1] - E[y_i|s_i = 0] \\
= E[y_{0i} + (y_i - y_{0i})s_i|s_i = 1] - E[y_{0i} + (y_i - y_{0i})s_i|s_i = 0] \\
= E[y_{0i} + (y_i - y_{0i})|s_i = 1] - E[y_{0i}|s_i = 0] \\
= E[y_{1i} - y_{0i}|s_i = 1] + (E[y_{0i}|s_i = 1] - E[y_{0i}|s_i = 0]) \\
= \underbrace{E[y_{1i} - y_{0i}]}_{\text{why?}}
\]

Average treatment effect

- Recover the CEF, and the CEF is *causal*. 
Potential occupations: \{x_{0i}, x_i\}. Observe \(x_i = x_{0i} + (x_i - x_{0i})s_i\).

Suppose three types \(T_i\):

1. **Always-zeros** (\(T_i = AZ\)): \(x_{0i} = 0\), \(x_{1i} = 0\)
2. **Always-ones** (\(T_i = AO\)): \(x_{0i} = 1\), \(x_{1i} = 1\)
3. **Switchers** (\(T_i = SW\)): \(x_{0i} = 0\), \(x_{1i} = 1\)

\(\beta\) satisfying \(E[s_i(y_i - \alpha - \beta s_i - \delta x_i)] = 0\) will be a weighted average of

1. \(\beta_0\) satisfying \(E[s_i(y_i - \alpha_0 - \beta_0 s_i)|x_i = 0] = 0\)
2. \(\beta_1\) satisfying \(E[s_i(y_i - \alpha_1 - \beta_1 s_i)|x_i = 1] = 0\)

Why? Think fixed-effects, or work through Frisch-Waugh algebra.
Part 3: Controls: good and bad

Controlling composition (cont.)

- $\beta$ (similar for $\beta_0$):

\[ E[y_i|s_i = 1, x_i = 1] - E[y_i|s_i = 0, x_i = 1] \]
\[ = E[y_{0i} + (y_i - y_{0i})|s_i = 1, x_i = 1] - E[y_{0i}|T_i = AO] \]
\[ = E[y_i - y_{0i}|T_i = AO \lor (T_i = SW \land s_i = 1)] \]

Weighted avg. of type-specific treatment effects

+ $E[y_{0i}|T_i = AO \lor (T_i = SW \land s_i = 1)] - E[y_{0i}|T_i = AO]$

Bias (no causal interpretation)

- Recover the CEF (why?), but it’s not a CEF we want (not causal)
- When would this CEF be causal?