Chapter 1: Regression Recap

Peter Hull

Applied Econometrics II
Brown University
Spring 2024

Outline

1. Estimators vs. Estimands vs. Parameters
2. Linear Regression and the CEF
3. Regression Anatomy and the OVB Formula
4. (Standard) Standard Errors

Regression Basics

The standard way regression/OLS is taught can be very confusing...

- Gauss-Markov? $E[\varepsilon \mid X] = 0$? Normal errors? Where's causality??!

OLS is an *estimator*: a simple algorithm applied to data, $(X_i, Y_i)_{i=1}^N$

- Specifically, $\hat{\beta} = \left(\sum_{i=1}^N X_i X_i'\right)^{-1} \sum_{i=1}^N X_i Y_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
- A.k.a. *reg y x1 x2 x3, r*; i.e. the thing you can actually "run"!
- Because data are random, $\hat{\beta}$ is random; it has a distribution

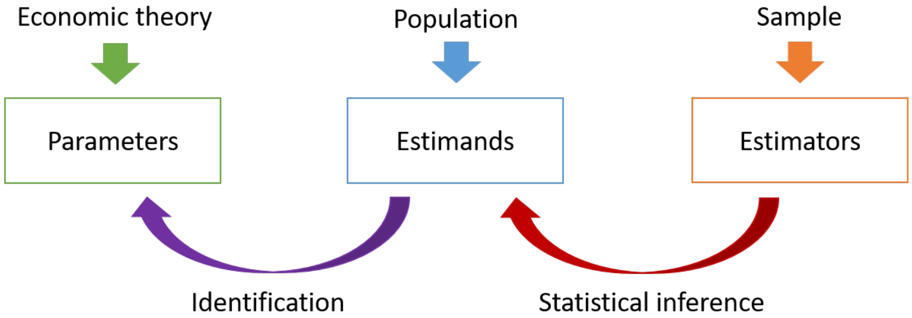
For big N, OLS is close to a (non-random) population regression *estimand*

- Specifically, $\hat{\beta} \xrightarrow{p} \beta = E\left[X_i X_i'\right]^{-1} E\left[X_i Y_i\right]$ under mild conditions
- We can make inferences about β from $\hat{\beta}$ using asymptotic statistics

But a separate question is whether β identifies a *parameter* of interest

- This is econometrics' real value-added vs. statistics
- To study identification, we need a *model* and *assumptions*...

Econometrics: The Big Picture



Make life easier by separating the *statistical task* (inferring estimands from data) from the *modeling task* (picking estimands that identify parameters)

Example: Estimating ATEs in an Experiment

Suppose we have an outcome Y_i and a binary treatment D_i

- Potential outcome model: $Y_i = Y_i(0)(1 - D_i) + Y_i(1)D_i$
- Assume as-good-as-random assignment: $D_i \perp (Y_i(0), Y_i(1))$

What does a regression of Y_i on D_i identify?

- "Saturated," so the slope coefficient is $E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$
- By the model, this is $E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]$
- By random assignment, this is $E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1) - Y_i(0)]$

OLS estimates $E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$ by the corresponding difference in *sample* means, $\frac{1}{N_1}\sum_{i:D_i=1} Y_i - \frac{1}{N_0}\sum_{i:D_i=0} Y_i$

- Under mild conditions (e.g. *iid* sampling) sample means plim to population means (LLN), with a known distribution (CLT)
- But this is *totally separate* from the model / assumptions

Some Things You Can Forget About (In This Class)

The Gauss-Markov Theorem

- We will rarely assume $E[\varepsilon_i \mid X_i] = 0$ or spherical standard errors
- We will care less about *efficiency* than about *robustness*

Homoskedastic SEs / testing for heteroskedasticity

- We will always just ", r" (at minimum)

Finite/small-sample inference tools (e.g. t-scores)

- We will usually assume we're close to "asymptopia"

Outline

1. Estimators vs. Estimands vs. Parameters✓
2. Linear Regression and the CEF
3. Regression Anatomy and the OVB Formula
4. (Standard) Standard Errors

Conditional Expectations and the Big LIE

The *conditional expectation function* (CEF) for a dependent variable Y_i given a $(K+1)\times 1$ vector X_i is written $E[Y_i \mid X_i = x]$ for $x \in Supp(X_i)$

- The function is non-random (i.e. a series of fixed numbers). We sometimes write $E[Y_i \mid X_i]$ as the CEF evaluated at the random X_i
- Since X_i is random, $E[Y_i \mid X_i]$ is also random

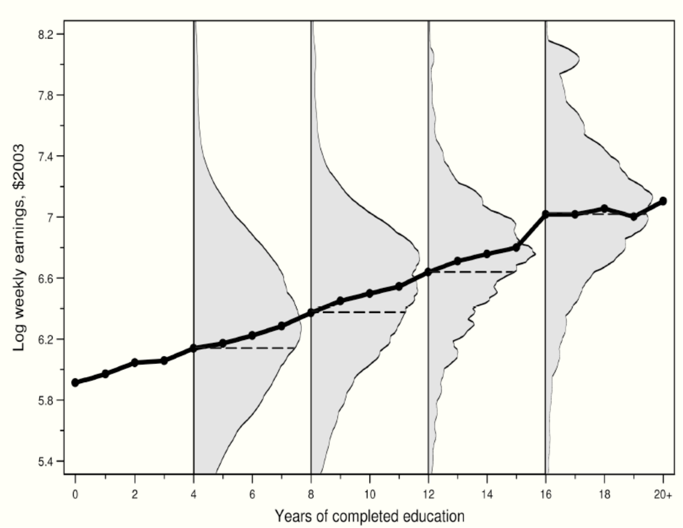
The *law of iterated expectations* (LIE) is a very useful fact about CEFs:

$$E[Y_i] = E[E[Y_i \mid X_i]]$$

Note that this only makes sense if we understand $E[Y_i \mid X_i]$ as random!

- Also useful to recall conditioning on X_i makes any function $h(X_i)$ "fixed", i.e. $E[h(X_i)Y_i] = E[E[h(X_i)Y_i \mid X_i]] = E[h(X_i)E[Y_i \mid X_i]]$

A Famous CEF for Labor Economists



Notes: distribution and CEF of average log weekly wages given schooling for white men aged 40-49 from the 1980 IPUMS 5% sample

LIEing Practice

Prop: The CEF residual $R_i = Y_i - E[Y_i \mid X_i]$ is uncorrelated with any $h(X_i)$

Proof: Note

$$E[R_i \mid X_i] = E[Y_i - E[Y_i \mid X_i] \mid X_i] = E[Y_i \mid X_i] - E[Y_i \mid X_i] = 0$$

Thus, by the LIE, $E[R_i] = E[E[R_i \mid X_i]] = 0$. Moreover,

$$E[R_i h(X_i)] = E[E[R_i h(X_i) \mid X_i]] = E[E[R_i \mid X_i] h(X_i)] = 0.$$

Thus $Cov(R_i, h(X_i)) = E[R_i h(X_i)] - E[R_i]E[h(X_i)] = 0 - 0 \cdot E[h(X_i)] = 0$

In particular, this shows the CEF residual is *orthogonal* to X_i: $E[R_i X_i] = 0$

Regression as CEF Approximation

Suppose we want to learn the CEF $E[Y_i \mid X_i = x]$. This is straightforward when X_i takes on few values (i.e. we can use a small # of sample means)

- But this quickly becomes hard as X_i gets continuous/high-dimensional

Regression can be understood as an attractive way of approximating CEFs

- Consider: MSE-minimizing linear approximation to $\mu(X_i) = E[Y_i \mid X_i]$

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

- It turns out this β coincides with the regression *least squares problem*:

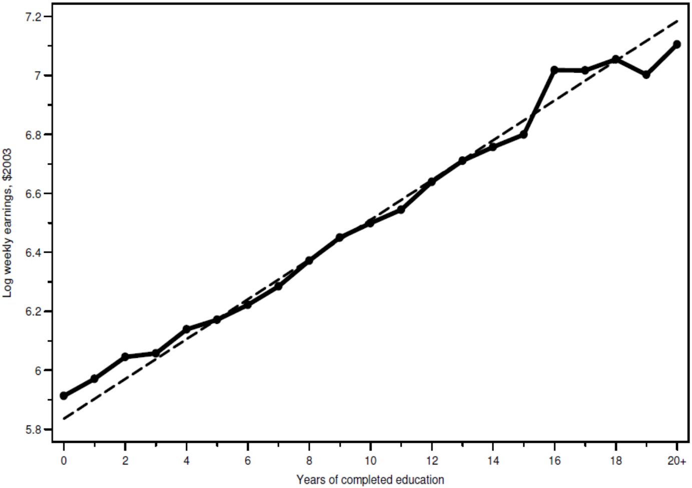
$$\beta = \arg\min_b E[(Y_i - X_i' b)^2]$$

- Why? Note by adding and subtracting $\mu(X_i)$

$$\begin{aligned} E[(Y_i - X_i' b)^2] &= E[(\mu(X_i) - X_i' b + R_i)^2] \\ &= E[(\mu(X_i) - X_i' b)^2] + 2\underbrace{E[R_i(\mu(X_i) - X_i' b)]}_{=0} + E[R_i^2] \end{aligned}$$

So min'ing $E[(Y_i - X_i' b)^2]$ is the same as min'ing $E[(\mu(X_i) - X_i' b)^2]$

Regression Linearly Approximates the True CEF



Notes: CEF and linear regression of average log weekly wages given schooling for white men aged 40-49 from the 1980 IPUMS 5% sample

Solving the Least Squares Problem

The population regression of Y_i on X_i is $\beta = \arg\min_b E[(Y_i - X_i'b)^2]$

- Using the first-order condition, $E[X_i(Y_i - X_i'\beta)] = 0$
- Solving out, $\beta = E[X_i X_i']^{-1} E[X_i Y_i]$
- OLS estimates this by sample analogue: $\hat{\beta} = (\frac{1}{N}\sum_i X_i X_i')^{-1} \frac{1}{N}\sum_i X_i Y_i$

By construction, the regression residual $\varepsilon_i = Y_i - X_i'\beta$ is orthogonal to X_i: $E[X_i \varepsilon_i] = 0$, and $Cov(X_i, \varepsilon_i) = 0$ when X_i includes a constant

- Note ε_i has no life of its own: it owes its meaning and existence to β
- The analogous result holds for OLS: $\frac{1}{N}\sum_i X_i \hat{\varepsilon}_i = 0$ for $\varepsilon_i = Y_i - X_i'\hat{\beta}$

Sometimes we care about getting the best linear predictor of Y_i, but more often we like regression b/c it gives the best linear approx to the CEF

- When the CEF is truly linear, $E[Y_i \mid X_i] = X_i'\beta$, regression gives it!

The CEF is All You Need

The LIE tells us that if X_i only varies at some group "level," we can estimate β with a weighted regression at that level:

$$\beta = E[X_i X_i']^{-1} E[X_i Y_i] = E[X_i X_i']^{-1} E[X_i E[Y_i \mid X_i]]$$

This holds true for the OLS estimator as well

The "grouped-data" version of regression is useful when working on a project that precludes analysis of the "microdata"

- It can also help speed up OLS when the microdata is large

Grouped-Data Regression

A - Individual-level data

```
. regress earnings school, robust

      Source |       SS       df       MS          Number of obs =  409435
-------------+------------------------------       F(  1,409433) =49118.25
       Model | 22631.4793      1  22631.4793       Prob > F      =  0.0000
    Residual |  188648.31 409433  .460755019       R-squared     =  0.1071
-------------+------------------------------       Adj R-squared =  0.1071
       Total | 211279.789 409434   .51602893       Root MSE      =  .67879
-------------+------------------------------------------------------------
             |              Robust                     Old Fashioned
    earnings |      Coef.  Std. Err.       t             Std. Err.      t
-------------+------------------------------------------------------------
      school |   .0674387  .0003447   195.63           .0003043    221.63
      const. |   5.835761  .0045507  1282.39           .0040043   1457.38
--------------------------------------------------------------------------
```

B - Means by years of schooling

```
. regress average_earnings school [aweight=count], robust
(sum of wgt is   4.0944e+05)

      Source |       SS       df       MS          Number of obs =      21
-------------+------------------------------       F(  1,    19) =  540.31
       Model |  1.16077332     1  1.16077332       Prob > F      =  0.0000
    Residual |  .040818796    19  .002148358       R-squared     =  0.9660
-------------+------------------------------       Adj R-squared =  0.9642
       Total |  1.20159212    20  .060079606       Root MSE      =  .04635
-------------+------------------------------------------------------------
     average |              Robust                     Old Fashioned
   _earnings |      Coef.  Std. Err.       t             Std. Err.      t
-------------+------------------------------------------------------------
      school |   .0674387  .0040352    16.71          .0029013     23.24
      const. |   5.835761  .0399452   146.09          .0381792    152.85
--------------------------------------------------------------------------
```

Figure 3.1.3: Micro-data and grouped-data estimates of returns to schooling. Source: 1980 Census - IPUMS, 5 percent sample. Sample is limited to white men, age 40-49. Derived from Stata regression output. Old-fashioned standard errors are the default reported. Robust standard errors are heteroscedasticity-consistent. Panel A uses individual-level data. Panel B uses earnings averaged by years of schooling.

"Saturated" Regressions

Regression is sure to coincide with the CEF in saturated specifications, where all values of X_i are "dummied out"

- This follows because $E[Y_i \mid X_i]$ is always linear in dummies for X_i

E.g. the binary regression we saw before, where $X_i = [1, D_i]'$, has

$$E[Y_i \mid X_i] = \underbrace{E[Y_i \mid D_i = 0]}_{\text{constant}} \cdot 1 + \underbrace{(E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0])}_{\text{slope}} \cdot D_i$$

- We'll see more examples of this (e.g. canonical DiD) soon

Outline

1. Estimators vs. Estimands vs. Parameters✓
2. Linear Regression and the CEF✓
3. Regression Anatomy and the OVB Formula
4. (Standard) Standard Errors

Regression Anatomy

When $X_i = [1, X_{1i}]'$, the two elements of $E[X_i X_i']^{-1} E[X_i Y_i]$ are:

- Slope coefficient: $\beta_1 = \frac{Cov(X_{1i}, Y_i)}{Var(X_{1i})}$, intercept: $\beta_0 = E[Y_i] - \beta_1 E[X_{1i}]$

The Frisch-Waugh-Lovell Theorem tells us that more generally the k-th non-constant slope coefficient is $\beta_k = \frac{Cov(\tilde{X}_{ki}, Y_i)}{Var(\tilde{X}_{ki})}$, where $\tilde{X}_{ki}$ is the residual from regressing X_{ki} on all other elements of X_i

- Equivalently, $\beta_k = \frac{Cov(\tilde{X}_{ki}, \tilde{Y}_i)}{Var(\tilde{X}_{ki})}$ where $\tilde{Y}_i$ are analogous residuals

To prove, substitute the "long regression" $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + X_{Ki} + \varepsilon_i$ into $\beta_k = \frac{Cov(\tilde{X}_{ki}, Y_i)}{Var(\tilde{X}_{ki})}$ and use the facts that:

- $Cov(\tilde{X}_{ki}, X_{ji}) = 0$ for $j \neq k$ (why?) & $Cov(\tilde{X}_{ki}, X_{ki}) = Var(\tilde{X}_{ki})$ (why?)

The OVB Formula

The omitted variables bias (OVB) formula describes the relationship between regression coefficients in specifications with different controls

- It is a *mechanical* result about the regression estimand (w/analogous results for the OLS estimator); not about any model motivating it

The simplest / canonical version considers a "long regression" of Y_i (e.g. wages) on S_i (schooling) and A_i (ability):

$$Y_i = \alpha + \rho S_i + \gamma A_i + \varepsilon_i$$

Ability is hard to measure. What if we omit it? The "short regression" is:

$$\frac{Cov(S_i, Y_i)}{Var(S_i)} = \frac{Cov(S_i, \alpha + \rho S_i + \gamma A_i + \varepsilon_i)}{Var(S_i)} = \rho + \gamma\delta$$

where $\delta = \frac{Cov(S_i, A_i)}{Var(S_i)}$ comes from regressing A_i on S_i

- MHE: *"Short equals long plus the effect of omitted times the regression of omitted on included"* (catchy, right?)

The OVB Formula (Cont.)

The simple formula generalizes to multiple omitted variables

- If $Y_i = \alpha + \rho S_i + A_i'\gamma + \varepsilon_i$ then $\frac{Cov(S_i, Y_i)}{Var(S_i)} = \rho + \gamma'\delta$ where δ contains coefficients from regressing each element of A_i on S_i

It also generalizes to specifications with included controls:

- If $Y_i = \alpha + \rho S_i + \phi X_i + \gamma A_i + \varepsilon_i$ then $\frac{Cov(\tilde{S}_i, Y_i)}{Var(\tilde{S}_i)} = \rho + \gamma\tilde{\delta}$ where $\tilde{\delta} = \frac{Cov(\tilde{S}_i, A_i)}{Var(\tilde{S}_i)}$ comes from regressing A_i on S_i controlling for X_i

An important consequence of the OVB formula is "short" equals "long" when "included" and "omitted" are uncorrelated

- E.g. adding pre-randomization controls to our RCT regression of Y_i on D_i won't change the estimand

Illustrating the OVB Formula

Estimates of the returns to education for men in the NLSY

Controls:	(1) None	(2) Age Dummies	(3) Col. (2) and Additional Controls*	(4) Col. (3) and AFQT Score	(5) Col. (4), with Occupation Dummies
	.132 (.007)	.131 (.007)	.114 (.007)	.087 (.009)	.066 (.010)

Notes: Data are from the National Longitudinal Survey of Youth (1979 cohort, 2002 survey). The table reports the coefficient on years of schooling in a regression of log wages on years of schooling and the indicated controls. Standard errors are shown in parentheses. The sample is restricted to men and weighted by NLSY sampling weights. The sample size is 2,434.

*Additional controls are mother's and father's years of schooling, and dummy variables for race and census region.

Outline

1. Estimators vs. Estimands vs. Parameters✓
2. Linear Regression and the CEF✓
3. Regression Anatomy and the OVB Formula✓
4. (Standard) Standard Errors

Asymptotic Behavior of OLS

In matrix form, the OLS estimator is $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Suppose *iid* data

Substitute in the population regression $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$, rearrange terms, and multiply both sides by $\sqrt{N}$ to get:

$$\sqrt{N}(\hat{\beta} - \beta) = (\mathbf{X}'\mathbf{X}/N)^{-1}\sqrt{N}(\mathbf{X}'\varepsilon/N)$$

- $(\mathbf{X}'\mathbf{X}/N)$ is a matrix of sample means, $\frac{1}{N}\sum_i X_{ki}X_{ji}$.
  The LLN tells us $(\mathbf{X}'\mathbf{X}/N) \xrightarrow{p} E[X_i X_i']$ as $N \to \infty$
- $(\mathbf{X}'\boldsymbol{\varepsilon}/N)$ is a vector of sample means, $\frac{1}{N}\sum_i X_{ki}\varepsilon_i$, where $E[X_i\varepsilon_i] = 0$.
  The CLT tells us $\sqrt{N}(\mathbf{X}'\varepsilon/N) \xrightarrow{d} \mathrm{N}(0, Var(X_i\varepsilon_i))$ as $N \to \infty$

Putting these two pieces together with the continuous mapping theorem,

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} \mathrm{N}(0, V)$$

where $V = E[X_i X_i']^{-1} Var(X_i\varepsilon_i) E[X_i X_i']^{-1}$. I.e., $\hat{\beta} \approx \mathrm{N}(\beta, V/N)$

Robust Standard Errors

We estimate V by sample analogue ("sandwich formula"):

$$\hat{V} = \left(\frac{1}{N} \sum_i X_i X_i' \right)^{-1} \frac{1}{N} \sum_i X_i X_i' \hat{\varepsilon}_i^2 \left(\frac{1}{N} \sum_i X_i X_i' \right)^{-1}$$

- The "robust" standard error for $\hat{\beta} k$ is then $\hat{SE} = \sqrt{\hat{V}_{kk}/N}$

Under homoskedasticity, $E[\varepsilon_i^2 \mid X_i] = \sigma^2$, this formula simplifies to $\hat{V} = \left(\frac{1}{N} \sum_i X_i X_i' \right)^{-1} \frac{1}{N} \sum_i \hat{\varepsilon}_i^2$.... useful for intuition-building, but not useful

- Why would we assume $E[\varepsilon_i^2 \mid X_i] = \sigma^2$, especially when we're just approximating the CEF?
- Always ", r" in Stata (or whatever you guys do in R) at minimum
- We'll talk about clustering and "non-standard" SEs later in the course