Part A: Regression and causality

A1: Key facts about regression

Kirill Borusyak

ARE 213 Applied Econometrics

UC Berkeley, Fall 2024

Acknowledgments

- These lecture slides draw on the materials by Michael Anderson, Peter Hull, Paul Goldsmith-Pinkham, and Michal Kolesar

- All errors are mine — please let me know if you spot them!

What is this course about (1)

- Goal: help you do rigorous empirical (micro)economic research

- Focus on causal inference / program evaluation / treatment effects

What is shared by [the causal] literature is [...] an explicit emphasis on credibly estimating causal effects, a recognition of the heterogeneity in these effects, clarity in the identifying assumptions, and a concern about endogeneity of choices and the role study design plays. (Imbens, 2010, "Better LATE Than Nothing")

What is this course about (2)

- Focus on most common research designs / identification strategies

  The econometrics literature has developed a small number of canonical settings where researchers view the specific causal models and associated statistical methods as well established and understood. [They are] referred to as identification strategies. [These] include unconfoundedness, IV, DiD, RDD, and synthetic control methods and are familiar to most empirical researchers in economics. The [associated] methods associated are commonly used in empirical work and are constantly being refined, and new identification strategies are occasionally added to the canon. Empirical strategies not currently in this canon, rightly or wrongly, are viewed with much more suspicion until they reach the critical momentum to be added. (Imbens, 2020)

- We will study target estimands, assumptions, tests, estimators, statistical inference
- Introduce multi-purpose econometric tools: e.g. randomization inference

Course outline (1)

A. Introduction: regression and causality (~4 lectures)

- Key facts about regression; potential outcomes and RCTs

B. Selection on observables (~4 lectures)

- Covariate adjustment via regression, via propensity scores, doubly-robust methods, double machine learning

C. Panel data methods (~7 lectures)

- Diff-in-diffs and event studies; synthetic controls and factor models

Course outline (2)

D. Instrumental variables (IVs) (~7 lectures)
   - Linear IV; IV with treatment effect heterogeneity
   - formula instruments, recentering, shift-share IV, spillovers
   - Examiner designs ("judge IVs")

E. Regression discontinuity (RD) designs (~3 lectures)
   - Sharp and fuzzy RD designs and various extensions

F. Miscellaneous topics (~3 lectures)
   - Nonlinear models: Poisson regression, quantile regression
   - Statistical inference: clustering, bootstrap
   - Topics of your interest (email me in advance!)

Course outline (3)

Currently not covered

- Descriptive statistics, data visualization
- Structural estimation
- Time series data
- Experimental design

Textbooks

MHE Angrist, Joshua and Jorn-Steffen Pischke (2009). Mostly Harmless Econometrics. Princeton University Press.

CT Cameron, A. Colin and Pravin Trivedi (2005). Microconometrics: Methods and Applications. Cambridge University Press.

IW Imbens, Guido and Jeffrey Wooldridge (2009). New developments in econometrics: Lecture notes. https://www.cemmap.ac.uk/resource/new-developments-in-econometrics/

JW Wooldridge, Jeffrey (2002). Econometric Analysis of Cross Section and Panel Data. MIT Press. *(Or second edition from 2010)*

Some econometric vocabulary

- OLS **estimator**: $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \equiv \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} X_i Y_i\right)$
  - Random variable, function of the observed sample
- OLS **estimand**: $\beta_{OLS} = \mathbb{E}[XX']^{-1}\mathbb{E}[XY] \equiv \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i]$ (assuming a random sample)
  - A non-stochastic population parameter
  - $\hat{\beta} \xrightarrow{p} \beta_{OLS}$ with a random sample under weak regularity conditions
  - This does not involve assuming a model, exogeneity conditions etc.
- $\hat{\beta}$ and β_{OLS} correspond to a linear **specification** $Y_i = \beta' X_i + \text{error}$
  - Just notational convention for `reg Y X`, not necessarily a model

Some econometric vocabulary (2)

- An economic or statistical **model** is needed to interpret β_{OLS} and other estimands
  - A model involves **parameters** (with economic meaning) and **assumptions** (restricting the DGP)
  - Assumptions hopefully make some parameters **identified**, i.e. possible to uniquely determine from everything the data contain — here, the distribution of (X, Y)

Some econometric vocabulary (3)

- Example 1: demand and supply

$$Q_i = -\beta_d P_i + \varepsilon_d, \qquad Q_i = \beta_s P_i + \varepsilon_s, \qquad \mathrm{Cov}\left[\varepsilon_d, \varepsilon_s\right] = 0$$

  - Regressing Q_i on P_i and a constant yields *(prove this!)*

$$\beta_{OLS} = \frac{\mathrm{Var}\left[\varepsilon_s\right]}{\mathrm{Var}\left[\varepsilon_d\right] + \mathrm{Var}\left[\varepsilon_s\right]} \cdot (-\beta_d) + \frac{\mathrm{Var}\left[\varepsilon_d\right]}{\mathrm{Var}\left[\varepsilon_d\right] + \mathrm{Var}\left[\varepsilon_s\right]} \cdot \beta_s$$

- Example 2: heterogeneous effects

$$Y_i = \beta_i X_i + \varepsilon_i, \qquad X_i \perp\!\!\!\perp (\beta_i, \varepsilon_i)$$

  - Regressing Y_i on X_i and a constant yields *(prove this!)*

$$\beta_{OLS} = \mathbb{E}\left[\beta_i\right]$$

Outline

1. Course intro
2. What is regression and why do we use it?
3. Linear regression and its mechanics

Regression and its uses

Regression of Y on $X \equiv$ **conditional expectation function** (CEF):

$$h(\cdot)\colon x \mapsto h(x) \equiv \mathbb{E}\left[Y_i \mid X_i = x\right]$$

- Conditional expectation $\mathbb{E}\left[Y_i \mid X_i\right] = h(X_i)$ is a random variable because X_i is

Uses of regression:

- **Descriptive:** how Y on average covaries with X — *by definition*
- **Prediction:** if we know X_i, our best guess for Y_i is $h(X_i)$ — *prove next*
- **Causal inference:** what happens to Y_i if we manipulate X_i — *sometimes*

Regression as optimal prediction (1)

- What is the best guess is defined by a loss function
- Proposition: CEF is the best predictor with quadratic loss:

$$h(\cdot) = arg\min_{g(\cdot)} \mathbb{E}\left[\left(Y_i - g(X_i)\right)^2\right]$$

- Lemma: the **CEF residual** $Y_i - \mathbb{E}\left[Y_i \mid X_i\right]$ is mean-zero and uncorrelated with any $g(X_i)$.
  - Proof by the law of iterated expectations (LIE)
  - $\mathbb{E}\left[Y_i - \mathbb{E}\left[Y_i \mid X_i\right]\right] = \mathbb{E}\left[\mathbb{E}\left[Y_i - \mathbb{E}\left[Y_i \mid X_i\right] \mid X_i\right]\right] = 0$
  - $\mathbb{E}\left[\left(Y_i - h(X_i)\right) g(X_i)\right] = \mathbb{E}\left[\mathbb{E}\left[\left(Y_i - h(X_i)\right) g(X_i) \mid X_i\right]\right] =$ $\mathbb{E}\left[\mathbb{E}\left[Y_i - h(X_i) \mid X_i\right] \cdot g(X_i)\right] = 0$

Regression as optimal prediction (2)

- Proposition: CEF is the best predictor with quadratic loss:

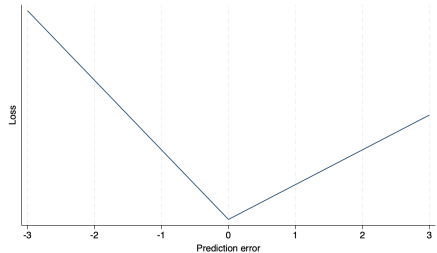
$$h(\cdot) = \arg\min_{g(\cdot)} \mathbb{E}\left[(Y_i - g(X_i))^2\right]$$

- Lemma: the CEF residual $Y_i - \mathbb{E}[Y_i \mid X_i]$ is mean-zero and uncorrelated with any $g(X_i)$.
- Proposition proof:

$$\begin{aligned}
\mathbb{E}\left[(Y_i - g(X_i))^2\right] &= \mathbb{E}\left[\{(Y_i - h(X_i)) + (h(X_i) - g(X_i))\}^2\right] \\
&= \mathbb{E}\left[(Y_i - h(X_i))^2\right] + 2\mathbb{E}\left[(Y_i - h(X_i))\,(h(X_i) - g(X_i))\right] + \mathbb{E}\left[(h(X_i) - g(X_i))^2\right] \\
&= \mathbb{E}\left[(Y_i - h(X_i))^2\right] + \mathbb{E}\left[(h(X_i) - g(X_i))^2\right] \geq \mathbb{E}\left[(Y_i - h(X_i))^2\right]
\end{aligned}$$

Regression as optimal prediction: Exercise

- What is the best predictor with loss $|Y_i - g(X_i)|$, i.e. $\arg\min_{g(.)} \mathbb{E}\left[|Y_i - g(X_i)|\right]$?
- Or with the "check" loss function (slope $q \in (0, 1)$ on the right, $q - 1$ on the left)?



- *Hint*: solve it first assuming X_i takes only one value
- *Note*: this exercise is linked to quantile regression

Outline

1. Course intro
2. What is regression and why do we use it?
3. Linear regression and its mechanics

Five reasons for linear regression

What does CEF have to do with least squares estimand $\beta_{OLS} = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$? And why do we use it instead of $\mathbb{E}[Y \mid X]$?

1. Curse of dimensionality: $\mathbb{E}[Y \mid X]$ is hard to estimate when X is high-dimensional [but machine learning methods make it easier]
2. OLS and CEF solve similar problems: $X'\beta_{OLS}$ is the best *linear predictor* of Y, i.e.

$$\beta_{OLS} = \arg\min_b \mathbb{E}\left[(Y - X'b)^2\right]$$

3. OLS is also the best *linear approximation* to the CEF:

$$\beta_{OLS} = \arg\min_b \mathbb{E}\left[(\mathbb{E}[Y \mid X] - X'b)^2\right]$$

Five reasons for linear regression (cont.)

1. Curse of dimensionality: $\mathbb{E}[Y \mid X]$ is hard to estimate when X is high-dimensional
2. OLS and CEF solve similar problems: $X'\beta_{OLS}$ is the best linear predictor of Y, i.e.

$$\beta_{OLS} = \arg\min_b \mathbb{E}\left[(Y - X'b)^2\right]$$

3. OLS is also the best *linear approximation* to the CEF:

$$\beta_{OLS} = \arg\min_b \mathbb{E}\left[(\mathbb{E}[Y \mid X] - X'b)^2\right]$$

- Proof by FOC: $\mathbb{E}[X(\mathbb{E}[Y \mid X] - X'b)] = 0 \implies$
  $b = \mathbb{E}[XX']^{-1}\mathbb{E}[X\mathbb{E}[Y \mid X]] = \mathbb{E}[XX']^{-1}\mathbb{E}[XY] = \beta_{OLS}$

Five reasons for linear regression (cont.)

1. Curse of dimensionality: $\mathbb{E}[Y \mid X]$ is hard to estimate when X is high-dimensional
2. OLS and CEF solve similar problems: $X'\beta_{OLS}$ is the best *linear predictor* of Y
3. OLS is also the best *linear approximation* to the CEF
4. With scalar X, β_{OLS} is a convexly-weighted average of $\mathrm{d}\mathbb{E}[Y \mid X = x] / \mathrm{d}x$ (or its discrete analog)

Proof of #4: Discrete X (with values $x_0 < \cdots < x_K$)

- Rewrite $\mathbb{E}[Y \mid X = x] \equiv h(x) = h(x_0) + \sum_{k=1}^{K} (h(x_k) - h(x_{k-1}))\, \mathbf{1}[x \geq x_k]$
- Thus $\operatorname{Cov}[Y, X] = \operatorname{Cov}[\mathbb{E}[Y \mid X], X] = \sum_{k=1}^{K} (h(x_k) - h(x_{k-1})) \operatorname{Cov}[\mathbf{1}[X \geq x_k], X]$ and

$$\beta_{OLS} = \frac{\operatorname{Cov}[Y, X]}{\operatorname{Var}[X]} = \sum_{k=1}^{K} \omega_k \frac{h(x_k) - h(x_{k-1})}{x_k - x_{k-1}}, \quad \omega_k = \frac{(x_k - x_{k-1}) \operatorname{Cov}[\mathbf{1}[X \geq x_k], X]}{\operatorname{Var}[X]}$$

- Here $\omega_k \geq 0$ because $\mathbf{1}[X \geq x_k]$ is monotone. Specifically *(prove it!)*:

$$\operatorname{Cov}[\mathbf{1}[X \geq x_k], X] = (\mathbb{E}[X \mid X \geq x_k] - \mathbb{E}[X \mid X < x_k])\, P(X \geq x_k)\, P(X < x_k)$$

- And $\sum_{k=1}^{K} \omega_k = 1$ because $X = x_0 + \sum_{k=1}^{K} (x_k - x_{k-1})\, \mathbf{1}[X \geq x_k]$

Proof of #4: Continuous X

- Similarly for continuous X:

$$\beta_{OLS} = \int_{-\infty}^{\infty} \omega(x) h'(x) dx, \qquad \omega(x) = \frac{\mathrm{Cov}\left[\mathbf{1}\left[X \geq x\right], X\right]}{\mathrm{Var}\left[X\right]}$$

  with $\omega(x) \geq 0$ and $\int_{-\infty}^{\infty} \omega(x) = 1$

- Exercise: if X is Gaussian, $\beta_{OLS} = \mathbb{E}\left[h'(X)\right]$ *(prove it!)*

  - *Hint*: use $\mathbb{E}\left[Z \mid Z \geq a\right] = \frac{\varphi(a)}{1-\Phi(a)}$ for $Z \sim \mathcal{N}(0,1)$

Five reasons for linear regression (cont.)

1. Curse of dimensionality: $\mathbb{E}[Y \mid X]$ is hard to estimate when X is high-dimensional
2. OLS and CEF solve similar problems: $X'\beta_{OLS}$ is the best linear predictor of Y
3. OLS is also the best linear approximation to the CEF
4. With scalar X, β_{OLS} is a convexly-weighted average of $\partial \mathbb{E}[Y \mid X = x] / \partial x$
5. If $\mathbb{E}[Y \mid X]$ happens to be linear, $\mathbb{E}[Y \mid X] = X'\beta_{OLS}$
   - Linearity is guaranteed when (X, Y) are jointly normally distributed
   - or when X is "saturated": dummies for all values of a discrete variable. E.g. for binary D and $X = (1, D)$,

$$\mathbb{E}[Y \mid X] = \mathbb{E}[Y \mid D] = \underbrace{\mathbb{E}[Y \mid D = 0]}_{\text{intercept}} \cdot 1 + \underbrace{(\mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0])}_{\text{slope}} \cdot D$$

(Linear) regression mechanics: Key results

1. When an intercept is included, residuals are mean-zero and uncorrelated with regressors
2. Regressing $Y = X_k$ on $X_1, \ldots, X_K$ produces coefficients $(0, \ldots, 0, 1, 0, \ldots 0)$
3. $\hat{\beta}$ is a linear estimator
4. Frisch-Waugh-Lovell (FWL) theorem
5. Omitted variable bias (OVB) formula
6. Asymptotic distribution and robust standard errors for OLS estimator

Linear regression results (cont.)

- When an intercept is included, population residuals are mean-zero and uncorrelated with regressors: $\mathbb{E}\left[X\left(Y - \beta'_{OLS}X\right)\right] = 0$
  - A simple result, not an assumption *(prove it!)*
  - The sample analog also holds: $\frac{1}{N}\sum_i X_i \left(Y_i - \hat{\beta}' X_i\right) = 0$
  - Since residuals are mean-zero, average fitted value equals average outcome, $\frac{1}{N}\sum_i \hat{\beta}' X_i = \frac{1}{N}\sum_i Y_i$
- Regressing $Y = X_k$ on $X_1, \ldots, X_K$ produces coefficients $(0, \ldots, 0, 1, 0, \ldots 0)$
  - *Prove it!*

OLS is a linear estimator

- Given the regressors $\boldsymbol{X}$, each $\hat{\beta}_k$ is linear in the outcomes, i.e. $\exists \{\omega_{ki}\}_{i=1}^{N}$ such that

$$\hat{\beta}_k = \sum_i \omega_{ki} Y_i$$

  for some weights $\omega_{ki} \equiv \omega_{ki}(\boldsymbol{X})$ *(prove it!)*

  - Weights ω_{ki} are mean-zero (for $X_k \neq$ intercept), orthogonal to non-X_k regressors, and $\sum_i \omega_{ki} X_{ki} = 1$ *(prove it!)*

- **Implication:** Regression coefficients can be decomposed

  - If $Y_i = Y_{1i} + \cdots + Y_{Pi}$, regressing each Y_{pi} on X_i and adding up the coefficient estimates is numerically the same as regressing Y_i on X_i

Partialling out: Frisch-Waugh-Lovell (FWL) theorem

Theorem: The k'th element of β_{OLS} can be obtained as $\beta_k = \frac{\mathrm{Cov}\left[\tilde{X}_k, Y\right]}{\mathrm{Var}\left[\tilde{X}_k\right]}$ or $\beta_k = \frac{\mathrm{Cov}\left[\tilde{X}_k, \tilde{Y}\right]}{\mathrm{Var}\left[\tilde{X}_k\right]}$ where $\tilde{X}_k$ is the residual from regressing X_k on all other regressors (and same for $\tilde{Y}$)

Proof:

- Define $\varepsilon = Y - \beta'_{OLS}X$. Plug in $Y = \beta'_{OLS}X + \varepsilon$ to $\frac{\mathrm{Cov}\left[\tilde{X}_k, Y\right]}{\mathrm{Var}\left[\tilde{X}_k\right]}$
- Note that $\tilde{X}_k$ is uncorrelated with ε; with other regressors; and with $Y - \tilde{Y}$

Implication: Explicit characterization of the weights ω_{ki}:

$$\hat{\beta}_k = \frac{\sum_i \tilde{X}_{ki} Y_i}{\sum_i \tilde{X}_{ki}^2} = \sum_i \omega_{ki} Y_i \qquad \text{for } \omega_{ki} = \frac{\tilde{X}_{ki}}{\sum_{j=1}^{N} \tilde{X}_{kj}^2}.$$

Omitted variable "bias"

OVB formula is a mechanical relationship between β_{OLS} from a "long" specification

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

and δ_{OLS} from a "short" specification

$$Y = \delta_0 + \delta_1 X_1 + \text{error}$$

Claim: $\delta_1 = \beta_1 + \beta_2 \rho$, where $\rho = \text{Cov}\left[X_1, X_2\right] / \text{Var}\left[X_1\right]$ is the regression slope of X_2 ("omitted") on X_1 ("included")

- **Proof:** $\delta_1 = \frac{\text{Cov}[X_1, Y]}{\text{Var}[X_1]} = \frac{\text{Cov}[X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon]}{\text{Var}[X_1]} = \beta_1 + \beta_2 \frac{\text{Cov}[X_1, X_2]}{\text{Var}[X_1]}$.
- When included X_1 is uncorrelated with omitted X_2, OVB $= 0$
- Generalizes to multiple omitted variables (with OVB $= \beta_2' \rho$)
- Applies with extra controls X_3 included in long, short, and auxiliary regression

Asymptotic distribution of the OLS estimator

$$\hat{\beta} = \left(\frac{1}{N}\sum_i X_i X_i'\right)^{-1}\left(\frac{1}{N}\sum_i X_i Y_i\right) = \beta_{OLS} + \left(\frac{1}{N}\sum_i X_i X_i'\right)^{-1}\left(\frac{1}{N}\sum_i X_i \varepsilon_i\right)$$

where by definition $\varepsilon = Y - \beta_{OLS}' X$. Thus,

$$\sqrt{N}\left(\hat{\beta} - \beta_{OLS}\right) = \left(\frac{1}{N}\sum_i X_i X_i'\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_i X_i \varepsilon_i\right)$$

- By LLN, $\frac{1}{N}\sum_i X_i X_i' \xrightarrow{p} \mathbb{E}[XX']$ (assumed non-singular)
- In a random sample, by CLT (using $\mathbb{E}[X\varepsilon] = 0$), $\frac{1}{\sqrt{N}}\sum_i X_i \varepsilon_i \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathrm{Var}[X\varepsilon])$
- By the continuous mapping theorem,

$$\sqrt{N}\left(\hat{\beta} - \beta_{OLS}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V), \qquad V = \mathbb{E}[XX']^{-1}\,\mathrm{Var}[X\varepsilon]\,\mathbb{E}[XX']^{-1}$$

Robust standard errors

- We estimate V by its sample analog ("sandwich formula"), up to a degree-of-freedom correction:

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

- Heteroskedasticity-robust (Eicker-Huber-White) standard error is

$$SE\left(\hat{\beta}_k\right) = \sqrt{V_{kk}/N}$$

- Never use homoskedastic standard errors!
- For later: standard errors outside iid samples, e.g. clustered SE in panels