Part A: Regression and causality

A1: Key facts about regression

Kirill Borusyak

ARE 213 Applied Econometrics

UC Berkeley, Fall 2023

Acknowledgments

- These lecture slides draw on the notes by Michael Anderson and slides by Peter Hull and Paul Goldsmith-Pinkham

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

What is this course about?

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

- Formal language of causal inference
- 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

Informal pre-requisites

- You can derive the asymptotic variance of the OLS/IV estimator under heteroskedasticity

- You are familiar with GMM

Course outline (1)

A. Introduction: regression and causality

- Key facts about regression; potential outcomes and RCTs

B. Selection on observables

- Regression, propensity score, and doubly-robust methods; double machine learning

C. Panel data methods

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

Course outline (2)

D. Instrumental variables (IVs)

- Linear IV; IV with treatment effect heterogeneity; control function approaches
- Shift-share IV designs, formula instruments, recentering, spillovers
- Examiner designs ("judge IVs")

E. Regression discontinuity (RD) designs

- Sharp and fuzzy RD designs, RD extrapolation, spatial RD

F. Miscellaneous topics

- Nonlinear models: Poisson regression, quantile regression
- Statistical inference: clustering, bootstrap

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**: $\beta_{OLS} = \mathbb{E}\left[XX'\right]^{-1}\mathbb{E}\left[XY\right]$
  - With a random sample, can also write $\beta_{OLS} = \mathbb{E}\left[X_i X_i'\right]^{-1}\mathbb{E}\left[X_i Y_i\right]$
  - A non-stochastic population parameter
  - $\hat{\beta} \xrightarrow{p} \beta_{OLS}$ for a random sample under weak 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, not a model

Some econometric vocabulary (2)

- An economic or statistical **model** is needed to interpret $\beta_{OLS} = \mathbb{E}\left[X_i X_i'\right]^{-1} \mathbb{E}\left[X_i Y_i\right]$
  - 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)
- Examples:
  1. $Y_i = \beta X_i + \varepsilon_i$, $\mathbb{E}\left[\varepsilon_i \mid X_i\right] = 0 \implies \beta_{OLS} = \beta$
  2. $Y_i = \beta X_i + \varepsilon_i$, $\mathbb{E}\left[\varepsilon_i \mid X_i\right] \neq 0 \implies \beta_{OLS} = \beta + \mathbb{E}\left[X_i X_i'\right]^{-1} \mathbb{E}\left[X_i \varepsilon_i\right]$
  3. $Y_i = X_i Y_i(1) + (1 - X_i) Y_i(0)$, $(Y_i(0), Y_i(1)) \perp X_i \implies$ $\beta_{OLS} = ATE \equiv \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$

Outline

Regression and its uses

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

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

- 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:

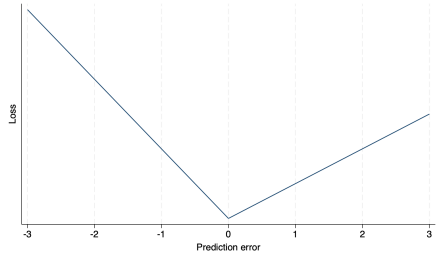
$$h(\cdot) = \arg\min_{g(\cdot)} \mathbb{E}\left[(Y_i - g(X_i))^2\right]$$

- 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

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 (2)

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 (3)

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 $\partial \mathbb{E}[Y \mid X = x] / \partial x$ (or its discrete analog)

Proof of #4: Discrete X (with values $x_0, \ldots, x_K$)

- 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,

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

- 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:

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

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

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

  - *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 (4)

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. Residuals are mean-zero (when an intercept is included) and uncorrelated with regressors: $\mathbb{E}\left[X(Y - \beta'_{OLS}X)\right] = 0$
   - This is in the population; 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, $\frac{1}{N}\sum_i \hat{\beta}' X_i = \frac{1}{N}\sum_i Y_i$
2. Frisch-Waugh-Lovell (FWL) theorem
3. Omitted variable bias (OVB) formula
4. Asymptotic distribution and robust standard errors for OLS estimator

Partialling out: Frisch-Waugh-Lovell 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: Write $Y = \beta'_{OLS} X + \varepsilon$ and plug this in noting that $\tilde{X}_k$ is uncorrelated with ε, with other regressors (and mean-zero), and with $Y - \tilde{Y}$

Implication (weight representation of the OLS estimator):

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

ω_{ki} are mean-zero, orthogonal to non-X_k regressors, and $\sum_i \omega_{ki} X_{ki} = \sum_i \omega_{ki} \tilde{X}_{ki} = 1$

- Intuition: regressing X_k on X yields $\beta_k = 1$; regression other X_ℓ on X yields $\beta_k = 0$

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:** $\delta_1 = \frac{\text{Cov}[X_1, Y]}{\text{Var}[X_1]} = \frac{\text{Cov}[X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon]}{\text{Var}[X_1]} = \beta_1 + \beta_2 \frac{\text{Cov}[X_1, X_2]}{\text{Var}[X_1]}$.
- Generalizes to multiple omitted variables (with OVB $= \beta_2' \rho$) and additional controls X_3 (with the auxiliary regression controlling for them, too)
- When included X_1 is uncorrelated with omitted X_2, OVB $= 0$

Asymptotic distribution of the OLS estimator

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

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

Outline

Causality vs. prediction

- Economists obsess with causality but sometimes prediction is the relevant goal
- The right choice depends on the ultimate goal: decision making
- Two scenarios (see Kleinberg et al., 2015):

1. The action $D \in \{0, 1\}$ affects the outcome Y, and the payoff (i.e., utility) π depends on Y
   - E.g. $D =$ rain dance in a drought, $Y =$ it rains

$$\pi(D) = aY(D) - bD \quad \Longrightarrow \quad \mathbb{E}\left[\pi(1) - \pi(0)\right] = a\mathbb{E}\left[Y(1) - Y(0)\right] - b$$

   - Optimal decision: $D = \mathbf{1}\left[\mathbb{E}\left[Y(1) - Y(0)\right] \geq b/a\right]$
   - This is a causal problem
   - Better knowledge of heterogeneous causal effects $\mathbb{E}\left[Y(1) - Y(0) \mid X\right]$ based on observed covariates X yields better decisions

Causality vs. prediction (2)

2. Y is unaffected by D but the marginal payoff of actions, $\partial\pi/\partial D$, depends on Y
   - E.g. $D =$ take an umbrella, $Y =$ it rains

$$\pi(D) = aY \cdot D - bD \quad \Longrightarrow \quad \mathbb{E}\left[\pi(1) - \pi(0)\right] = a\mathbb{E}\left[Y\right] - b$$

   - Optimal decision: $D = \mathbf{1}\left[\mathbb{E}\left[Y\right] \geq b/a\right]$
   - This is a prediction problem. Better prediction $\mathbb{E}\left[Y \mid X\right]$ yields better decisions

- *Note:* This scenario can also be recast as an unusual causal problem:
  - D affects $\tilde{Y}(D) =$ you get wet $= Y \cdot (1 - D)$
  - But we know potential outcome $\tilde{Y}(1) = 0$
  - And we have data on $\tilde{Y}(0) = Y$ to make a *prediction* of $\tilde{Y}(1) - \tilde{Y}(0)$

Policy-relevant prediction problems: Examples

1. Eliminating futile hip and knee replacement surgeries
   - Surgery has costs: monetary + painful recovery
   - Benefits depend on life expectancy
   - Kleinberg et al. (2015) show 10% (1%) of patients have *predictable* probability of dying within a year of 24% (44%) for reasons unrelated to this surgery
2. Improving admissions by predicting college success
   - Geiser and Santelices (2007) show that high-school GPA is a better predictor of performance at UC colleges than SAT
   - If UC had to reduce admissions, rejecting applicants with marginal GPAs would result in losing fewer good students than rejecting marginal SAT applicants
3. See Kleinberg et al. "Human Decisions and Machine Predictions" (2018) for a more subtle example on bail decisions by judges