Part A: Regression and causality

A1: Key facts about regression. Statistical inference

Kirill Borusyak

ARE 213 Applied Econometrics

UC Berkeley, Fall 2025

Outline

1. Course intro
2. Some econometric vocabulary
3. What is regression and why do we use it?
4. Why linear regression?
5. Linear regression mechanics
6. Statistical inference

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 researchers in economics and related fields do rigorous empirical work
  - Requires understanding the underlying econometrics and how it links to economic theory and applied questions
  - Attention to intuition, not regularity conditions
  - But we will go much beyond a black-box toolkit, study key proofs
- Bonus: students interested in methodological work may get some inspiration: lots of open problems!

What is this course about (2)

- 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)

What is this course about (3)

- 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)

You should be able to

- Pick the empirical strategy that best suits your application
- Understand its assumptions/limitations
- Identify problems in the papers you read

Course outline (1)

A. Recap of regression and causality

- Key facts about regression; statistical inference; potential outcomes and RCTs

B. ~~Unconfoundedness (a.k.a. Selection on observables)~~

C. Instrumental variables (IVs)

- Linear IV; IV with treatment effect heterogeneity; Examiner designs (“judge IV”)

D. Panel data methods

- Linear panel data methods; Dynamic panels
- Diff-in-diffs and event studies; synthetic controls and factor models

Course outline (2)

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

F. Spillovers and equilibrium effects
   - Peer effects
   - Formula instruments and recentering; Shift-share IV designs

G. Miscellaneous topics: some of...
   - Uncovering causal mechanisms
   - Nonlinear methods: Poisson regression, multinomial choice, quantile regression
   - Machine learning and causality
   - Topics of your interest (email me in advance!)

The course in brief

Khoa Vu
@KhoaVuUmn

Readings

See link in syllabus for up-to-date readings for each topic. Useful textbooks:

MHE	Angrist and Pischke (2009) “Mostly Harmless Econometrics”
JW	Wooldridge (2002/2010) “Econometric Analysis of Cross Section and Panel Data”
IW	Imbens and Wooldridge (2009) “New developments in econometrics: Lecture notes”
CT	Cameron and Trivedi (2005) “Microeconometrics: Methods and Applications”
SW	Wager (2024) “Causal Inference: A Statistical Learning Approach”
PD	Ding (2024) “Linear Models and Extensions”
SC	Cunningham (2021) “Causal Inference: The Mixtape”

Check your background: True/False

Let X, Y be some random variables

1. $\mathrm{Cov}\left[Y, X\right] = \mathbb{E}\left[Y \cdot (X - \mathbb{E}\left[X\right])\right]$

2. If Y is continuous and X is discrete, $\mathbb{E}\left[Y \mid X\right]$ is a continuous random variable

3. If $X \in \{0, 1\}$ is binary, $\mathbb{E}\left[\mathbb{E}\left[Y \mid X = 1\right] \mid X = 0\right]$ is a well-defined constant

Outline

1. Course intro
2. Some econometric vocabulary
3. What is regression and why do we use it?
4. Why linear regression?
5. Linear regression mechanics
6. Statistical inference

Econometric vocabulary: Model

- An economic or statistical **model**:
  - Defines the relevant observed and unobserved variables
  - Imposes restrictions on their joint distribution (e.g., relationships between vars), sometimes via **parameters**
- Example 1: $Y = \beta_0 + \beta_1 X + \varepsilon$
  - Not useful without restrictions. Useful when adding $\mathbb{E}\left[\varepsilon \mid X\right] = 0$
- Example 2: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon, \qquad \mathbb{E}\left[\varepsilon \mid X\right] = 0$
- Example 3 (RCT w/ heterogeneous effects): $Y_i = \tau_i X_i + \varepsilon_i, \quad X_i \perp\!\!\!\perp (\tau_i, \varepsilon_i)$
- Example 4 (demand and supply):

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

Econometric vocabulary: Estimands and Identification

- **Estimand** is a non-stochastic population object within a model, property of the joint distribution of observed and unobserved variables and parameters
  - Example 1: in the model $Y = \beta_0 + \beta_1 X + \varepsilon$, besides individual parameters, we may be interested in $\beta_0 + 2\beta_1$: prediction of Y at $X = 2$
  - Example 2: in the heterogeneous effects model $Y_i = \tau_i X_i + \varepsilon_i, \quad X_i \perp\!\!\!\perp (\tau_i, \varepsilon_i)$, we may be interested in $\mathbb{E}[\tau_i]$
- A model parameter or estimand is (point-)**identified** if it is uniquely determined by the joint distribution of observed variables
  - (**Partially identified** if a strict subset of values is consistent with data)
  - Example 1 (cont.): β_1 is not identified without extra restrictions
    - ★ But under $\mathbb{E}[\varepsilon \mid X] = 0, \qquad \beta_1 = \mathrm{Cov}[Y, X] / \mathrm{Var}[X]$
  - Example 2 (cont.): $\mathbb{E}[\tau_i] = \mathrm{Cov}[Y, X] / \mathrm{Var}[X]$ *(prove this!)*

Econometric vocabulary: Estimators

- **Estimator** is a function of observed variables in the sample (⇒ a random variable)
  - OLS estimator: $\hat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} \equiv \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} X_i Y_i\right)$
- $\hat{\beta}$ corresponds to a linear **specification** $Y_i = \beta' X_i + \text{error}$
  - Just notational convention for `reg Y X`, not [necessarily] a model
- Assuming a random sample, as $N \to \infty$, most reasonable estimators converge to *some* population object
  - This is the **estimand corresponding to the estimator**
- E.g. OLS estimand: $\beta_{OLS} = \mathbb{E}\left[XX'\right]^{-1}\mathbb{E}\left[XY\right] \equiv \mathbb{E}\left[X_i X_i'\right]^{-1}\mathbb{E}\left[X_i Y_i\right]$
  - I.e., $\hat{\beta} \xrightarrow{p} \beta_{OLS}$ in a random sample under weak regularity conditions
  - This does <u>not</u> require imposing a model $Y = \beta_{OLS}' X + \varepsilon$ or assuming $\mathbb{E}\left[\varepsilon \mid X\right] = 0$
  - But OLS estimand can be interpreted under any given model...

Examples: OLS estimand in different models

1. Model $Y = \beta' X + \varepsilon$, $\mathbb{E}\left[\varepsilon \mid X\right] = 0 \quad \Longrightarrow \quad \beta_{OLS} = \beta$
2. Heterogeneous effects: Regressing Y_i on X_i and a constant yields $\beta_{OLS} = \mathbb{E}\left[\tau_i\right]$
3. Demand and supply: Regressing Q_i on P_i (and a constant) yields *(prove this!)*

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

We will analyze the estimands of simple estimators (e.g. OLS, IV) under different assumptions, usually in the potential outcomes model

- Search for other estimators when simple ones do not yield what we want

Outline

1. Course intro
2. Some econometric vocabulary
3. What is regression and why do we use it?
4. Why linear regression?
5. Linear regression mechanics
6. Statistical inference

Regression and its uses

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

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

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

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

- What is the best guess depends on the loss function
- 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 $g(X_i)$.
  - Proof by the law of iterated expectations (LIE)
  - $\mathbb{E}[Y_i - \mathbb{E}[Y_i \mid X_i]] = \mathbb{E}[\mathbb{E}[Y_i - \mathbb{E}[Y_i \mid X_i] \mid X_i]] = 0$
  - $\mathbb{E}[(Y_i - h(X_i))\, g(X_i)] = \mathbb{E}[\mathbb{E}[(Y_i - h(X_i))\, g(X_i) \mid X_i]] =$ $\mathbb{E}[\mathbb{E}[Y_i - h(X_i) \mid X_i] \cdot g(X_i)] = 0$

Proposition proof

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

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

- Proof:

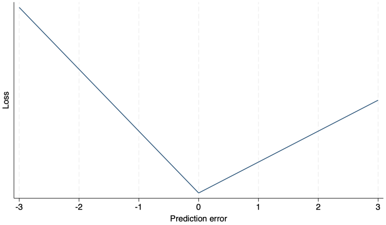
$$\begin{aligned}\mathbb{E}\left[(Y_i - g(X_i))^2\right] &= \mathbb{E}\left[\{(Y_i - h(X_i)) + (h(X_i) - g(X_i))\}^2\right] \\ &= \mathbb{E}\left[(Y_i - h(X_i))^2\right] + 2\mathbb{E}\left[(Y_i - h(X_i))\left(h(X_i) - g(X_i)\right)\right] + \mathbb{E}\left[(h(X_i) - g(X_i))^2\right] \\ &= \mathbb{E}\left[(Y_i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

Regression as optimal prediction: Exercise

- Show that conditional median is the best predictor with loss $|Y_i - g(X_i)|$, i.e.

$$\text{median}(Y_i \mid X_i) = \arg\min_{g(\cdot)} \mathbb{E}\left[|Y_i - g(X_i)|\right]$$

  - *Hint*: solve it first assuming X_i takes only one value
- What about the "check" loss function (slope $q \in (0,1)$ on the right, $q - 1$ on the left)?



Outline

1. Course intro
2. Some econometric vocabulary
3. What is regression and why do we use it?
4. Why linear regression?
5. Linear regression mechanics
6. Statistical inference

Definition

- In the population,

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

  - Interpretation: $X'\beta_{OLS}$ is the best linear predictor of Y with quadratic loss

- FOC: $\mathbb{E}\left[X(Y - X'\beta_{OLS})\right] = 0$

  - Interpretation: residuals are mean-zero (when X includes an intercept) and uncorrelated with X *(a property of OLS, not an assumption)*

  - Sample analog also holds

- Solution: $\beta_{OLS} = \mathbb{E}\left[XX'\right]^{-1} \mathbb{E}\left[XY\right]$

Five reasons for linear regression

How is linear regression (OLS) related to regression (CEF)? Why do we often prefer OLS?

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:

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

- 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 (in addition to an intercept), β_{OLS} is a convexly-weighted average of $\mathrm{d}\mathbb{E}[Y \mid X = x]/\mathrm{d}x$ (or its discrete analog)

What does #4 mean?

- Consider discrete X taking values $x_0 < \cdots < x_K$
- We have scalar (X, Y), with no assumptions imposed on them (no linear model!)
- By definition,
$$Y = h(X) + \varepsilon, \qquad \mathbb{E}\left[\varepsilon \mid X\right] = 0$$
- We want to show that population regression of Y on X yields
$$\beta_{OLS} = \sum_{k=1}^{K} \omega_k \frac{h(x_k) - h(x_{k-1})}{x_k - x_{k-1}}$$
for some $\omega_1, \ldots, \omega_K \geq 0$ that do not depend on $h(\cdot)$, with $\omega_1 + \cdots + \omega_K = 1$
- Gives interpretation of OLS as an average slope even when CEF is nonlinear

Proof of #4: Discrete X

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

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

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

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

- 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 scalar 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 $\mathrm{d}\mathbb{E}[Y \mid X = x] / \mathrm{d}x$
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 $X = (1, D)$,

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

Outline

1. Course intro
2. Some econometric vocabulary
3. What is regression and why do we use it?
4. Why linear regression?
5. Linear regression mechanics
6. Statistical inference

(Linear) regression mechanics: Key results

1. What happens under linear transformations of RHS variables
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. What happens under linear transformations of the LHS variable
5. Frisch-Waugh-Lovell (FWL) theorem
6. Omitted variable bias (OVB) formula

#1: Linear transformations of RHS variables

- Consider OLS estimation of two specifications:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \text{error};$$
$$Y = \gamma_1 X_1 + \gamma_2 (X_1 + X_2) + \text{error}.$$

Then $\gamma_1 + \gamma_2 = \beta_1$, $\gamma_2 = \beta_2$; fitted values/residuals are the same *(prove it!)*

  - *Application*: for binary X, it doesn't matter whether to run

$$Y = \beta_0 + \beta_1 X + \text{error} \qquad \text{vs.} \qquad Y = \gamma_0 (1 - X) + \gamma_1 X + \text{error}$$

#2: Regressing $Y = X_k$ on $X_1, \dots, X_K$

- Regressing $Y = X_k$ on $X_1, \dots, X_K$ produces coefficients $(0, \dots, 0, 1, 0, \dots 0)$
    - *Prove it using one of key OLS properties (min squared loss among linear predictors; residuals uncorrelated with X; β_{OLS} formula)*

#3: OLS is a linear estimator

- Given regressors $\boldsymbol{X}$, each $\hat{\beta}_k$ is linear in the outcomes, i.e.

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

  for some weights $\omega_{ki} \equiv \omega_{ki}(\boldsymbol{X})$ that do not depend on $\boldsymbol{Y}$ *(prove it!)*
- Moreover, weights ω_{ki} add up to zero (for $X_k \neq$ intercept)
  - I.e., adding 1 to the outcome doesn't change $\hat{\beta}_k$
  - $\hat{\beta}_k$ is a *contrast* between outcomes of different groups of observations
- Weights are orthogonal to non-X_k regressors ($\sum_i \omega_{ki} X_{\ell i} = 0$, $\ell \neq k$)
  - I.e., adding X_ℓ to the outcome doesn't change $\hat{\beta}_k$
- And $\sum_i \omega_{ki} X_{ki} = 1$: adding X_k to the outcome adds 1 to $\hat{\beta}_k$

#3: Proof

- $\hat{\beta} = \Omega \boldsymbol{Y}$ for an $K \times N$ matrix $\Omega = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$
  - Thus $\hat{\beta}_k = \Omega_k \boldsymbol{Y} \equiv \sum_i \omega_{ki} Y_i$ for Ω_k being kth row of Ω
- The weights don't depend on $\boldsymbol{Y}$ so let's consider $Y_i \equiv 1$. Regression yields $\hat{\beta}_k = 0$ for $X_k \neq$ intercept
  - But $0 = \hat{\beta}_k = \sum_i \omega_{ki} Y_i = \sum_i \omega_{ki}$
- Now let's consider $Y_i \equiv X_{i\ell}$ for $\ell \neq k$. Again, we get $\hat{\beta}_k = 0$
  - So $0 = \sum_i \omega_{ki} Y_i = \sum_i \omega_{ki} X_{\ell i}$: weights for X_k are orthogonal to other regressions
- Finally consider $Y_i = X_{ik}$, which implies $\hat{\beta}_k = 1$
  - So $1 = \sum_i \omega_{ki} Y_i = \sum_i \omega_{ki} X_{ki}$

A note on weights

- We now have two weight representations of OLS:
  - Average $d\mathbb{E}\left[Y \mid X = x\right] / dx$ with non-negative weights adding up to 1
  - $\mathbb{E}\left[\omega(X) \cdot Y\right]$ for $\mathbb{E}\left[\omega(X)\right] = 0$ (some positive and some negative weights)
- Both of these reflect that OLS makes outcome contrasts and aggregates them
- Later we will have more weighting results linking OLS to averages of causal effects

#4: Linear transformations of the LHS variable

- 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 *(prove it!)*

- This yields a popular decomposition of regression coefficients

    - E.g. regress total deaths on PM2.5 pollution, then decompose into different causes (cf. Deryugina et al. (2019, Table A2))

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

Theorem: Let $\tilde{X}_k$ be the residual from regressing X_k on all other regressors X_{-k}; and $\tilde{Y}$ be the residual from regressing Y on X_{-k}. Then k'th element of β_{OLS} can be obtained in two ways:

$$\beta_k = \frac{\mathrm{Cov}\left[\tilde{X}_k, Y\right]}{\mathrm{Var}\left[\tilde{X}_k\right]} = \frac{\mathrm{Cov}\left[\tilde{X}_k, \tilde{Y}\right]}{\mathrm{Var}\left[\tilde{X}_k\right]}$$

Importance: characterizes the estimand of each coef in multiple regression without matrix algebra

Proof idea: Define $\varepsilon = Y - \beta'_{OLS}X$. Plug in $Y = \beta'_{OLS}X + \varepsilon$ to $\mathrm{Cov}\left[\tilde{X}_k, Y\right]$ or $\mathrm{Cov}\left[\tilde{X}_k, \tilde{Y}\right]$; verify you get $\beta_k \mathrm{Var}\left[\tilde{X}_k\right]$

#5: Proof

- We have by definition: $Y = \beta'_{OLS} X + \varepsilon$ with $\mathbb{E}\left[X\varepsilon\right] = 0$
  - And $X_k = \gamma' X_{-k} + \tilde{X}_k$ with $\mathbb{E}\left[X_{-k}\tilde{X}_k\right] = 0$
  - And $Y = \delta' X_{-k} + \tilde{Y}$ with $\mathbb{E}\left[X_{-k}\tilde{Y}\right] = 0$
- Then

$$\begin{aligned}
\mathrm{Cov}\left[\tilde{X}_k, Y\right] &= \mathrm{Cov}\left[\tilde{X}_k, \beta'_{-k}X_{-k} + \beta_k X_k + \varepsilon\right] \\
&= \mathrm{Cov}\left[\tilde{X}_k, \beta'_{-k}X_{-k}\right] + \beta_k \mathrm{Cov}\left[\tilde{X}_k, \gamma' X_{-k} + \tilde{X}_k\right] + \mathrm{Cov}\left[X_k - \gamma' X_{-k}, \varepsilon\right] \\
&= 0 + \beta_k \mathrm{Var}\left[\tilde{X}_k\right] + 0
\end{aligned}$$

- Moreover,

$$\mathrm{Cov}\left[\tilde{X}_k, \tilde{Y}\right] = \mathrm{Cov}\left[\tilde{X}_k, Y - \delta' X_{-k}\right] = \mathrm{Cov}\left[\tilde{X}_k, Y\right]$$

Implications of FWL

1. If $\mathrm{Cov}\left[X_1, X_2\right] = 0$ (e.g. X_1 = randomized treatment, X_2 = predetermined covariate), the "long" and "short" specifications have the same estimand $\beta_1 = \delta_1$:

$$Y = \beta_0 + \beta_1' X_1 + \beta_2' X_2 + \text{error}; \qquad\qquad Y = \delta_0 + \delta_1' X_1 + \text{error}$$

2. Explicit characterization of the weights ω_{ki} in any regression:

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

- *Exercise:* which observations are weighted positively by a regression on scalar X_i and an intercept?

#6: Omitted variable "bias"

OVB formula is a mechanical relationship between β_1 from a "long" specification and δ_1 from a short specification:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \text{error}, \qquad\qquad Y = \delta_0 + \delta_1 X_1 + \text{error}$$

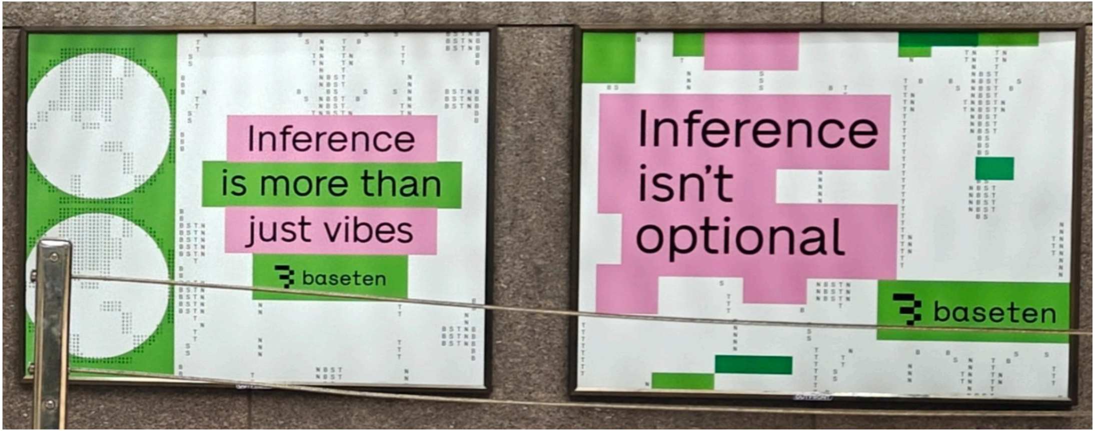
Claim: $\delta_1 = \beta_1 + \beta_2 \rho$, where $\rho = \text{Cov}\left[X_1, X_2\right] / \text{Var}\left[X_1\right]$ is the auxiliary regression slope of X_2 ("omitted") on X_1 ("included")

- **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 $= \beta_2' \rho$)
    - Gelbach (2016) uses it to measure contributions of each extra group of variables to $\hat{\beta}_1 - \hat{\delta}_1$, i.e. how much the coef changes from including them
- Applies with extra controls X_3 included in long, short, and auxiliary regression

Outline

1. Course intro
2. Some econometric vocabulary
3. What is regression and why do we use it?
4. Why linear regression?
5. Linear regression mechanics
6. Statistical inference

As seen on the BART...



(From Karthik Tadepalli's Twitter)

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

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

Heteroskedasticity-robust standard errors

- We estimate V by its sample analog ("sandwich formula") :

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

  where $\frac{N}{N - \dim(X)}$ = degree-of-freedom correction for small-sample downward bias

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

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

- Never use homoskedastic standard errors in practice!
- Note: this only works when $\dim(X) \ll N$
  - We'll discuss this issue for panel regressions with fixed effects

FWL for standard errors

- **Theorem:** robust standard errors for a subset of coefficients $\hat{\beta}_1$ from specification

$$Y = \beta_1' X_1 + \beta_2' X_2 + \text{error}$$

are numerically the same as from the FWL version (residualized on X_2):

$$\tilde{Y} = \beta_1' \tilde{X}_1 + \text{error}.$$

  - Note: LHS has to be residualized, too
  - Same holds for homoskedastic standard errors

Beyond random samples

We assumed $(X_i, \varepsilon_i, Y_i)$ are independent and identically distributed *(iid)* across $i = 1, \ldots, N$ — not always the case

- E.g. you sampled villages, then surveyed all adults in selected villages ("clustered sampling") $\implies$ ε_i likely correlated within village

- Or you sampled people but randomly assigned treatment by state ("clustered assignment") $\implies$ X_i is perfectly correlated within state

- Or dataset consists of all counties in a country
  $\implies$ ε_i likely correlated in nearby counties

Statistical inference beyond random samples

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

- LLN $\frac{1}{N}\sum_i X_i X_i' \xrightarrow{p} \mathbb{E}\left[\frac{1}{N}\sum_i X_i X_i'\right]$ applies much beyond random samples

- But CLT for "sandwich meat" relies on $\mathrm{Cov}\left[X_i\varepsilon_i, X_j\varepsilon_j\right] = 0$ for $i \neq j$:

$$\mathrm{Var}\left[\frac{1}{\sqrt{N}}\sum_i X_i\varepsilon_i\right] = \frac{1}{N}\sum_i \mathrm{Var}\left[X_i\varepsilon_i\right] + \frac{1}{N}\sum_{i\neq j}\mathrm{Cov}\left[X_i\varepsilon_i, X_j\varepsilon_j\right] = \frac{1}{N}\sum_i \mathrm{Var}\left[X_i\varepsilon_i\right]$$

  - Let's consider three relaxations: clustering, multi-way clustering, and spatial clustering

Cluster-robust inference

- Suppose units belong to clusters $g(i) \in \{1, \ldots, G\}$, and we believe $\text{Cov}\left[X_i\varepsilon_i, X_j\varepsilon_j\right] = 0$ if $g(i) \neq g(j)$
  - While arbitrary correlations are allowed within clusters
- Then we can use cluster-robust (“clustered”) inference (cf. MacKinnon et al. (2023))
  - Two equivalent derivations...

Derivation #1

Rewrite the estimator as

$$\hat{\beta} - \beta_{OLS} = \left(\sum_i X_i X_i'\right)^{-1} \left(\sum_i X_i \varepsilon_i\right) = \left(\sum_{g=1}^{G} \left(\sum_{i \in g} X_i X_i'\right)\right)^{-1} \left(\sum_{g=1}^{G} \left(\sum_{i \in g} X_i \varepsilon_i\right)\right)$$

Denoting $R_g = \sum_{i \in g} X_i \varepsilon_i$, apply CLT at the cluster level:

$$\frac{1}{\sqrt{G}} \sum_{g=1}^{G} R_g \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, V\right), \qquad V \equiv \mathrm{Var}\left[R_g\right]$$

Plug in sample analogs, provided G is large (not enough to have large N!):

$$\hat{V} = \left(\frac{1}{G} \sum_i X_i X_i'\right)^{-1} \cdot \left(\frac{1}{G} \sum_g \left(\sum_{i \in g} X_i \hat{\varepsilon}_i\right) \left(\sum_{i \in g} X_i \hat{\varepsilon}_i\right)'\right) \cdot \left(\frac{1}{G} \sum_i X_i X_i'\right)^{-1}$$

and set $SE(\hat{\beta}_k) = \sqrt{\hat{V}_{kk}/G}$

Derivation #2

- Sandwich meat satisfies

$$\mathrm{Var}\left[\sum_i X_i \varepsilon_i\right] = \sum_{i,j} \mathrm{Cov}\left[X_i\varepsilon_i, X_j\varepsilon_j\right] = \sum_{i,j:\ \ g(i)=g(j)} \mathrm{Cov}\left[X_i\varepsilon_i, X_j\varepsilon_j\right]$$

which can be estimated by

$$\sum_{i,j:\ \ g(i)=g(j)} X_i\hat{\varepsilon}_i X_j'\hat{\varepsilon}_j = \sum_g \left(\sum_{i\in g} X_i\hat{\varepsilon}_i\right)\left(\sum_{j\in g} X_j\hat{\varepsilon}_j\right)'$$

More on clustering

- DoF correction is more of an art. Stata uses $\frac{G}{G-1} \cdot \frac{N-1}{N-\dim(X)}$
- Having many clusters is important. In the extreme of $G = 1$,

$$\sum_{i,j} X_i \hat{\varepsilon}_i X_j' \hat{\varepsilon}_j = \left(\sum_i X_i \hat{\varepsilon}_i \right) \left(\sum_j X_j \hat{\varepsilon}_j \right)' = 0$$

  (and so $SE = 0$) which obviously does not approximate $\mathrm{Var}\left[\sum_i X_i \varepsilon_i\right]$ well

  - And $\frac{1}{\sqrt{N}} \sum_i X_i \varepsilon_i$ may not even converge if all observations are strongly correlated
- With few clusters, see guide by MacKinnon et al. (2023); with few treated clusters, see Alvarez et al. (2025)

At what level to cluster?

With individual-level data, when should you cluster by village? state? not at all?

- General principle: if you believe $\mathbb{E}\left[X_i \varepsilon_i \varepsilon_j X_j'\right] = 0$ for $g(i) \neq g(j)$, it's enough to cluster by g
- If treatment assignment is independent across g, enough to cluster by g
  - If $\mathbb{E}\left[\tilde{X}_i \tilde{X}_j \mid \boldsymbol{\varepsilon}\right] = 0$ for $g(i) \neq g(j)$ where $\tilde{X}$ are residuals on the intercept and (at most a small number of) included covariates,

    $$\mathbb{E}\left[\tilde{X}_i \varepsilon_i \varepsilon_j \tilde{X}_j'\right] = \mathbb{E}\left[\mathbb{E}\left[\tilde{X}_i \tilde{X}_j' \mid \varepsilon_i, \varepsilon_j\right]\right] = 0, \qquad \text{for } g(i) \neq g(j)$$

    even if ε have correlations across clusters
- If $\mathbb{E}\left[\varepsilon_i \varepsilon_j \mid \boldsymbol{X}\right] = 0$ for $g(i) \neq g(j)$, enough to cluster by g even if X_i is correlated across clusters

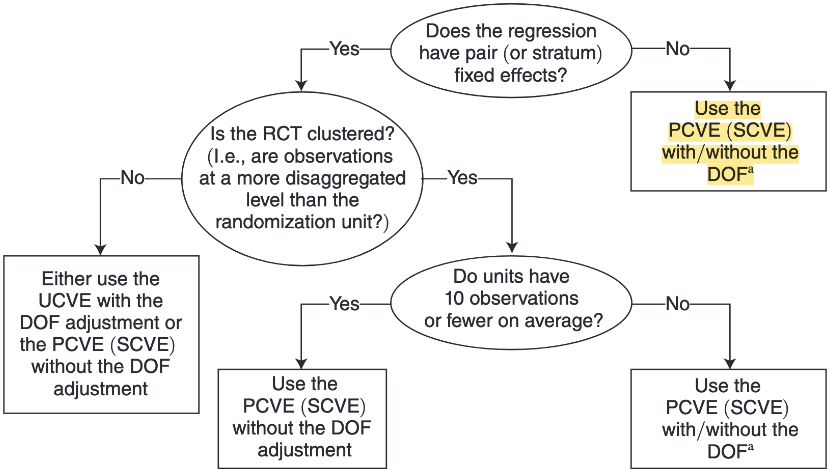
At what level to cluster? (2)

- If you are not sure, picking larger (“more conservative”) clusters relaxes the clustering assumption
  - Note: overclustering does not necessarily increase SE
  - E.g. if clustering by county is correct, clustering by state will give very similar SE: both approximate the true variance
- But avoid too large/few clusters which makes SE downward biased

Exercises

1. In a sample of twin siblings, you do a pair-randomization experiment, treating exactly one of each twins
   - Do you have to cluster SE in an individual-level regression of the outcome on the treatment?
   - Is it worth clustering by city in which they live if unobserved city factors can affect all residents?
2. To trace the demand curve for strawberries, you randomize prices across stores into 50 distinct values: $1.10, 1.20, ..., 6.00 per pound
   - $Y_i =$ quantity sold in a week, $X_i =$ 50 price dummies
   - Should you cluster SE by the price level?

Regressions with pair fixed effects: It's complicated



Does the regression have pair (or stratum) fixed effects?

Yes

No

Use the PCVE (SCVE) with/without the DOF[a]

Is the RCT clustered? (I.e., are observations at a more disaggregated level than the randomization unit?)

No

Yes

Either use the UCVE with the DOF adjustment or the PCVE (SCVE) without the DOF adjustment

Do units have 10 observations or fewer on average?

Yes

No

Use the PCVE (SCVE) without the DOF adjustment

Use the PCVE (SCVE) with/without the DOF[a]

(De Chaisemartin and Ramirez-Cuellar, 2024, Fig. 1B)

PCVE = Randomization pair-clustered; UCVE = Randomization unit-clustered

Multi-way clustering

- Assume each unit belongs to group $g(i) \in \{1, \dots, G\}$ and also to non-nested group $h(i) \in \{1, \dots, H\}$
  - E.g. workers i belong to state $g(i)$ and industry $h(i)$
  - Or bilateral trade flows: $i = (g, h)$ corresponds to exporter $g(i)$ and importer $h(i)$
- Two-way (a.k.a. double) clustering assumption *(cf. Cameron et al. (2011))*:

$$\mathbb{E}\left[X_i \varepsilon_i \varepsilon_j X_j'\right] = 0 \quad \text{unless } g(i) = g(j) \textbf{ or } h(i) = h(j), \text{ or both}$$

  - Allows correlation of $X_i \varepsilon_i$ between unit pairs that share at least one cluster

Multi-way clustering (2)

- Variance estimator: for $G, H \to \infty$,

$$\widehat{Var}(\hat{\beta}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\,\Omega\,(\boldsymbol{X}'\boldsymbol{X})^{-1}, \quad \Omega = \sum_{i,j=1}^{N} X_i \hat{\varepsilon}_i \hat{\varepsilon}_j X_j' \cdot \mathbf{1}\left[g(i) = g(j) \text{ or } h(i) = h(j)\right]$$

  with $\mathbb{E}\left[X_i \varepsilon_i \varepsilon_j X_j'\right] = 0$ plugged in for pairs that do not share any cluster

- Simple implementation:

$$\Omega = \Omega_g + \Omega_h - \Omega_{gh}$$

  where Ω_g is g-clustered, Ω_h is h-clustered, and Ω_{gh} is clustered by (g, h) pair

Multi-way clustering (3)

- *Warning:* do not confuse it with one-way clustering by $(g(i), h(i))$ pair:

$$\Omega = \sum_{i,j=1}^{N} X_i \hat{\varepsilon}_i \hat{\varepsilon}_j X_j' \cdot \mathbf{1}\left[g(i) = g(j) \textbf{ and } h(i) = h(j)\right]$$

- E.g. two-way clustering by state and by industry $\neq$ clustering by state-industry
  - The former is more conservative than clustering by state; the latter is less conservative
  - Avoid ambiguous "I cluster by state and industry"
- *Exercise:* how would you cluster when units belong to counties, which belong to states, and you think $X_i \varepsilon_i$ are correlated within counties and within states?

Spatially-clustered standard errors

- Suppose units are located in space (e.g., villages)
  - And we believe nearby units may have correlated $X_i\varepsilon_i$
- Clustering by geographic areas (e.g., states) would ignore correlations between places across area borders
- Spatially-clustered standard errors (Conley, 1999) use

$$\widehat{Var}\left(\sum_i X_i\varepsilon_i\right) = \sum_{i,j:\ d(i,j)<d_{max}} \kappa\left(\frac{d(i,j)}{d_{max}}\right) \cdot X_i\hat{\varepsilon}_i\hat{\varepsilon}_j X_j'$$

  - $d(i,j)$ is geographic distance
  - d_{max} is the distance cutoff such that $\mathrm{Cov}\left[X_i\varepsilon_i, X_j\varepsilon_j\right] = 0$ if $d(i,j) > d_{max}$
  - $\kappa(\cdot)$ is a kernel function: e.g. uniform kernel $\kappa(x) = \mathbf{1}\left[|x| \leq 1\right]$
    - Or Bartlett kernel: $\kappa(x) = \max\left\{1 - |x|, 0\right\}$

Spatially-clustered standard errors: Notes

- The structure of spatial correlation for units closer than d_{max} is not restricted
  - E.g. doesn't have to be constant even if uniform kernel is used
- d_{max} should be small enough, such that $d_{ij} > d_{max}$ for most i, j pairs
  - Similar to having many clusters with cluster-robust SE
  - If spatial correlation is too wide-ranging, see Conley and Kelly (2025) for other inference methods
- The issue solved here is spatial correlation that biases SE
  - Not spatial spillovers (i.e., nearby X_j affects Y_i) that bias estimates

References I

Alvarez, L., B. Ferman, and K. Wüthrich (2025): "Inference with few treated units," *arXiv preprint*.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011): "Robust Inference With Multiway Clustering," *Journal of Business & Economic Statistics*, 29, 238–249.

Conley, T. G. (1999): "GMM estimation with cross sectional dependence," *Journal of Econometrics*, 92, 1–45.

Conley, T. G. and M. Kelly (2025): "The Standard Errors of Persistence," *Journal of International Economics*, 153, 104027.

De Chaisemartin, C. and J. Ramirez-Cuellar (2024): "At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?" *American Economic Journal: Applied Economics*, 16, 193–212.

Deryugina, T., G. Heutel, N. H. Miller, D. Molitor, and J. Reif (2019): "The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction," *American Economic Review*, 109, 4178–4219.

References II

Gelbach, J. B. (2016): "When Do Covariates Matter? And Which Ones, and How Much?" *Journal of Labor Economics*, 34, 509–543.

Imbens, G. W. (2010): "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, 48, 399–423.

——— (2020): "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics," *Journal of Economic Literature*, 58, 1129–1179.

MacKinnon, J. G., M. O. Nielsen, and M. D. Webb (2023): "Cluster-robust inference: A guide to empirical practice," *Journal of Econometrics*, 232, 272–299.