Econ 271/Mgtecon 604 - Section 1

Amar Venugopal

Stanford University

Winter 2026

Outline

1. General Info & Logistics

2. Statistical Background

3. Identification & Estimation

4. Properties of Estimators
   - Finite Sample
   - Large Sample

5. Practice Problems

General Info

- Email: amar.venugopal@stanford.edu
- Section: Fridays 1:30-3:20pm in 200-303
- Office Hours: Tuesdays 12pm-1pm, Wednesdays 3:30-4:30pm, in Econ 149
- Section Questions: Question Form
- Lecture/Section Feedback: Feedback Form
- Note: please use office hours to ask questions about the material, rather than over email.

For the entirety of the course, section materials are drawn heavily from those prepared by Lea Bottmer last year, as well as from some materials produced by Merrill Warnick in previous years.

Section Norms

In order to foster a healthy, inclusive classroom environment, we will all abide by the following **community norms**:

- One microphone: we will allow everyone to finish their thought without interruption
- Everyone got accepted to Stanford and is here for a reason. There is no need to prove anything to anyone.
- Asking questions is a great way to deepen your understanding and resolve any confusion you may have. It is not a signaling device.
- There is no such thing as a stupid or dumb question
- Questions asked during section will be answered by the TA

Homework

- First problem set due **Thursday Jan 15, 11:59pm**
- You may work in groups, but you must turn in your own submission and note who you collaborated with on any given problem
- For easy grading, please submit only typed or scanned/electronic handwritten submissions (no photos of paper please)
- For problem set-related questions, come to my office hours or reach out to me
  - I **will not** respond to problem set emails sent on the due date

Hints for Problem Set 1

- For problem 1, make sure you remember how to use matrices and matrix multiplications (e.g. matrix derivations if you use matrix notation for some of the subparts or generally how to take the inverse of a 2×2). In the last part, make sure you use the fact that x is binary.
- For problem 2, recall from the lecture to please not use Excel or Stata for this problem or even any packages in Matlab/R/etc. We want you to implement the OLS estimator **from scratch**!

1. General Info & Logistics

2. Statistical Background

3. Identification & Estimation

4. Properties of Estimators
   - Finite Sample
   - Large Sample

5. Practice Problems

Properties of Random Variables

Let X, Y, Z be random variables and M a (constant) matrix. We then have:

- **Linearity of Expectation:** $\mathbb{E}_{XY}[X+Y] = \mathbb{E}_X[X] + \mathbb{E}_Y[Y]$
- **Covariance:** $\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
  - $\text{Cov}(aX, Y) = a\,\text{Cov}(X,Y)$
  - $\text{Cov}(X+Z, Y) = \text{Cov}(X,Y) + \text{Cov}(Z,Y)$
- **Variance:** $\text{Var}(X) = \text{Cov}(X,X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
  - Given a matrix M, we have: $\text{Var}(MX) = M\,\text{Var}(X)\,M^T$
- **Law of Iterated Expectations (Tower Rule):** $\mathbb{E}_X[X] = \mathbb{E}_Y\left[\mathbb{E}_X[X|Y]\right]$
- **Law of Total Variance:** $\text{Var}(Y) = \text{Var}\left(\mathbb{E}[Y|X]\right) + \mathbb{E}\left[\text{Var}(Y|X)\right]$

Let X_n be a sequence of random variables, with corresponding CDFs F_{X_n}.

- **Almost-Sure Convergence:** $X_n \xrightarrow{a.s.} X$ (aka w.p. 1)

$$\forall \varepsilon > 0, \; P\left(\lim_{n\to\infty} |X_n - X| < \varepsilon\right) = 1$$

- **Convergence in Probability:** $X_n \xrightarrow{p} X$

$$\forall \varepsilon > 0, \; \lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$$

- **Convergence in Distribution:** $X_n \xrightarrow{d} F$

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

(at all $x \in \mathbb{R}$ where $F_X(x)$ continuous)

Lemma:
$(X_n \xrightarrow{d} a) \Rightarrow (X_n \xrightarrow{p} a)$

$\Rightarrow$ = pointwise convergence of CDFs

Key Theorems

- **Weak Law of Large Numbers (WLLN)**
  - Assumptions: $X_i \overset{iid}{\sim} F$, $E[X_i] = \mu$, $Var(X_i) = \sigma^2 < \infty$
  - Result: $\bar{X} = E_n[X_i] = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$

    (Strong LLN: $\bar{X} \xrightarrow{a.s.} \mu$, extra assumption: $E[X^4] < \infty$)
- **Continuous Mapping Theorem**
  - Assumptions: $X_n \xrightarrow{p} X$, $g(x)$ is continuous
  - Result: $g(X_n) \xrightarrow{p} g(X)$

- **Central Limit Theorem (CLT)**
  - Assumptions: $X_1, \dots, X_n$ iid, $E[X_i] = \mu$, $Var(X_i) = \sigma^2$, $\bar{X} = E_n[X_i]$
  - Result: $\sqrt{n}\left(\frac{\bar{X} - \mu}{\sigma}\right) \xrightarrow{d} N(0,1)$

↳ "asymptotic normality"

- **Slutsky's Theorem**
  - Assumptions: $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{p} a$
  - Result: $X_n + Y_n \xrightarrow{d} X + a$, $X_n Y_n \xrightarrow{d} Xa$

1. General Info & Logistics

2. Statistical Background

3. Identification & Estimation

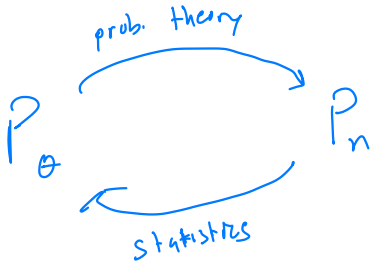
4. Properties of Estimators
   - Finite Sample
   - Large Sample

5. Practice Problems

Population vs. Sample Distributions

P_θ = population distribution
↳ population parameter of interest

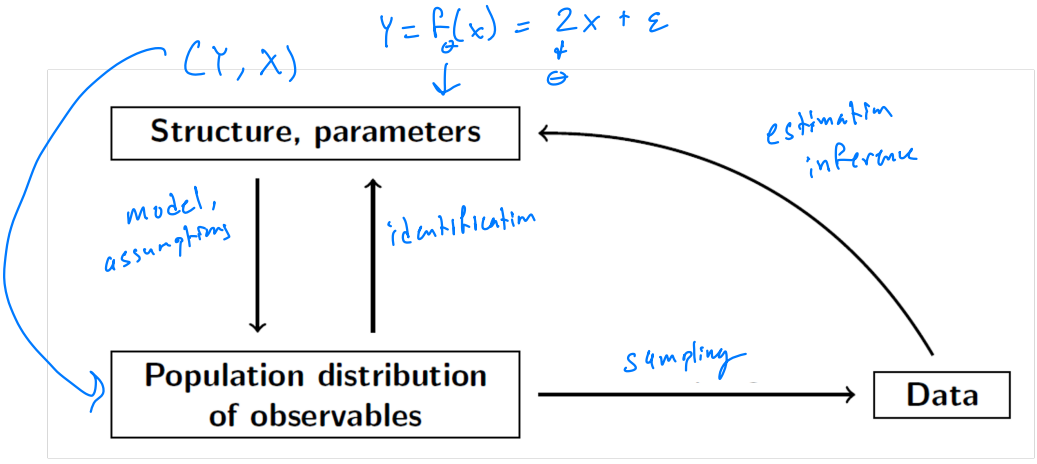
P_n = sampling distribution
↳ sample size



Definitions

- **Estimand:** Object of interest of estimation as a population quantity
  (function of P_θ)
- **Estimator:** Sample equivalent of the estimand
  (function of P_n)
- **Estimate:** Numeric value of estimator when we plug in data

Big Picture


$Y = f_\theta(x) = 2x + \varepsilon$
θ
(Y, X)
Structure, parameters
model, assumptions
identification
estimation inference
Population distribution of observables
sampling
Data

1. General Info & Logistics

2. Statistical Background

3. Identification & Estimation

4. Properties of Estimators
   - Finite Sample
   - Large Sample

5. Practice Problems

Running Example: Average Treatment Effect

- **Average Treatment Effect:**

$$\tau = \mathbb{E}\left[Y(1) - Y(0) \right]$$

- **Difference-in-Means Estimator:**

$$\hat{\tau} = \frac{1}{n_1} \sum_{d_i = 1} Y_i - \frac{1}{n_0} \sum_{d_i = 0} Y_i$$

1. General Info & Logistics
2. Statistical Background
3. Identification & Estimation
4. Properties of Estimators
   - Finite Sample
   - Large Sample
5. Practice Problems

Definition: $\hat{\mu}$ is unbiased for $\mu \iff \mathbb{E}[\hat{\mu}] = \mu$

Is $\hat{\tau}$ unbiased?

$$\mathbb{E}[\hat{\tau} \mid d_1, \dots, d_n] = \frac{1}{n_1} \sum_{d_i=1} \mathbb{E}[Y_i \mid d_1, \dots, d_n] - \frac{1}{n_0} \sum_{d_i=0} \mathbb{E}[Y_i \mid d_1, \dots, d_n]$$

(SUTVA)
$$= \frac{1}{n_1} \sum_{d_i=1} \mathbb{E}[Y_i \mid d_i] - \frac{1}{n_0} \sum_{d_i=0} \mathbb{E}[Y_i \mid d_i]$$

(defn. of P.O.)
$$= \frac{1}{n_1} \sum_{d_i=1} \mathbb{E}[Y_i(1)] - \frac{1}{n_0} \sum_{d_i=0} \mathbb{E}[Y_i(0)]$$

(identical dist. of P.O.)
$$= \left(\frac{1}{n_1} \cdot n_1\right) \mathbb{E}[Y(1)] - \left(\frac{1}{n_0} \cdot n_0\right) \mathbb{E}[Y(0)]$$

(linearity of $\mathbb{E}[\cdot]$)
$$= \mathbb{E}[Y(1) - Y(0)]$$
$$= \tau$$

$Var(\hat{\mu}) = E[(\hat{\mu} - E[\hat{\mu}])^2]$

Definition: $MSE(\hat{\mu}) = E_\theta[(\hat{\mu} - \mu)^2]$

Lemma: $MSE(\hat{\mu}) = Bias^2(\hat{\mu}) + Var(\hat{\mu})$

What is the MSE of $\hat{\tau}$?

Proof: $E[(\hat{\mu} - \mu)^2]$

$= E[((\hat{\mu} - E[\hat{\mu}]) - (\mu - E[\hat{\mu}]))^2]$

$= E[(\hat{\mu} - E[\hat{\mu}])^2] + E[(\mu - E[\hat{\mu}])^2] - 2E[(\hat{\mu} - E[\hat{\mu}])(\mu - E[\hat{\mu}])]$

$= Var(\hat{\mu})$

$= E[\mu^2 + E[\hat{\mu}]^2 - 2\mu E[\hat{\mu}]]$

$= \mu^2 + E[\hat{\mu}]^2 - 2\mu E[\hat{\mu}]$

$= (E[\hat{\mu}] - \mu)^2$

$= Bias^2(\hat{\mu})$

$= 0$

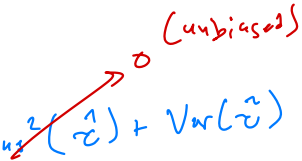
Tower rule

Mean Squared Error (MSE)

Definition: See above

What is the MSE of $\hat{\tau}$?

By Lemma: $MSE(\hat{\tau}) = Bias^2(\hat{\tau}) + Var(\hat{\tau})$



Linearity

Definition: $\hat{\mu}$ linear $\iff \exists a_i : \hat{\mu} = \sum_{i=1}^{n} a_i X_i$

Is $\hat{\tau}$ linear? $\hat{\tau} = \frac{1}{n_1} \sum_{d_i=1} Y_i - \frac{1}{n_0} \sum_{d_i=0} Y_i$

$\rightarrow a_i = \begin{cases} \frac{1}{n_1} & \text{if } d_i = 1 \\ -\frac{1}{n_0} & \text{if } d_i = 0 \end{cases}$

1. General Info & Logistics
2. Statistical Background
3. Identification & Estimation
4. Properties of Estimators
   - Finite Sample
   - Large Sample
5. Practice Problems

Consistency

Definition: $\hat{\mu}$ consistent for $\mu \iff \hat{\mu} \xrightarrow{p} \mu$

Is $\hat{\tau}$ consistent? $\hat{\tau} = E_n[Y_i | d_i = 1] - E_n[Y_i | d_i = 0]$

(WLLN + Δ-inequality) $\xrightarrow{p}$ $E[Y|d=1] - E[Y_i|d=0]$

(linearity of $E[\cdot]$ + defn of P.O.) $= E[Y(1) - Y(0)]$

$= \tau$

Asymptotic Normality ($\sqrt{n}$ - consistent)

Definition: $\hat{\mu}$ asymptotically normal $\iff \sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2)$

Is $\hat{\tau}$ asymptotically normal?

$\sqrt{n}\left(\frac{1}{2}\hat{c} - \tau\right) = \sqrt{n}\left(\frac{1}{n_1}\right.$

1. General Info & Logistics
2. Statistical Background
3. Identification & Estimation
4. Properties of Estimators
   - Finite Sample
   - Large Sample
5. Practice Problems

Econ 271/Mgtecon 604 - Section 2

Amar Venugopal

Stanford University

Winter 2026

Outline

1. Announcements

2. Identification Revisited

3. Review: Matrices

4. OLS: Things are Nice

5. OLS: Things Fall Apart (?)

6. Practice Problems

Announcements

- Midterm exam: finalized for **Monday Feb 9** (during lecture)
- If you have OAE accommodations, please upload them to the form announced on canvas so that they can be taken into account for the exams
- Optional **Lecture** on **Friday Jan 30** (same room, time as regularly scheduled section)
  - Lecture content (econometrics with incentives) will not be on problem sets or exams
  - Makeup section earlier that week

Hints for Problem Set 2

- For problem 1, think about what possible values ϵ_n could take to obtain the full probability distribution for ϵ_i. Then use important properties of the expectation and variances of random variables (and their sums) for (b). To show that the variance is minimum for an estimator, there are two ways: (1) show it directly or (2) show or argue that it is the OLS estimator.
- For problem 2, think about our general idea for solving for estimands: 1) use the optimization problem directly and plug in your true model somewhere or 2) use what we know the estimator converges to and plug in the true model. You will need use the fact that $x \sim N(0, 1)$ somewhere.
- For problem 3, remember that there is a difference between deriving the asymptotic variance $\widehat{\Sigma}$ and the asymptotic variance of the estimator $\hat{\beta}$ (namely a factor of N).

1. Announcements

2. Identification Revisited

3. Review: Matrices

4. OLS: Things are Nice

5. OLS: Things Fall Apart (?)

6. Practice Problems

What is identified?

Ex. Data (Y_i, W_i)

$\theta = E[1\{Y(1) > Y(0)\}]$

θ is not identified

Even given joint dist'n of (Y, W), can't get joint of $(Y(1), Y(0))$

Note: $\tau = E[Y(1) - Y(0)]$

$\hat{\tau} \xrightarrow{p} E[Y|W=1] - E[Y|W=0]$

$= E[Y(1)]$ $\quad = E[Y(0)]$

→ requires randomization of W

Ex. $Y_i \mid X_i \sim N(\alpha + \beta X_i + \gamma(\frac{X_i}{2}), \sigma^2)$

→ can only ID $(\beta + \frac{\gamma}{2})$

1. Announcements

2. Identification Revisited

3. Review: Matrices

4. OLS: Things are Nice

5. OLS: Things Fall Apart (?)

6. Practice Problems

Matrix Notation

Data: $(y_i, \vec{x}_i)$ $i = 1, \dots, n$

$y_i \in \mathbb{R}$ $\vec{x}_i \in \mathbb{R}^k$ $x_i \in \mathbb{R}^{k \times 1}$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} — \ \vec{x}_1 \ — \\ \vdots \\ — \ \vec{x}_n \ — \end{pmatrix} \in \mathbb{R}^{n \times k}$$

$$y_i = x_i^T \beta + \varepsilon_i \longrightarrow y = X\beta + \varepsilon$$

Matrix Properties

Let **A** be a matrix.

- **A** symmetric $\iff A = A^T$
- **A** left invertible $\iff \exists A^{-1} : A^{-1}A = I$
- **A** right invertible $\iff \exists A^{-1} : AA^{-1} = I$
- **A** invertible $\iff \exists A^{-1} : A^{-1}A = AA^{-1} = I$ (A square)
- $\mathbf{A}^\dagger$ pseudoinverse of **A** $\iff$ ① $AA^+A = A$ ② $A^+AA^+ = A^+$ ③ $(AA^+)^T = AA^+$ ④ $(A^+A)^T = A^+A$
- **A** is positive semi-definite $\iff \forall x \geq 0, \; x^TAx \geq 0$

Properties of A^+:

① If columns of A linearly indep: $A^+ = (A^TA)^{-1}A^T$ $=$

② If rows linearly indep $A^+ = A^T(AA^T)^{-1}$ $=$

③ If A invertible, $A^+ = A^{-1}$

Matrix Rules

Let A, B, C be matrices.

- Addition is:
  - Commutative $A+B = B+A$
  - Associative $(A+B)+C = A+(B+C)$
- Multiplication is:
  - Distributive over addition $(A+B)C = AC+BC$
  - Associative $(AB)C = A(BC)$
- Transposition is:
  - Distributive over addition $(A+B)^T = A^T + B^T$
  - Distributive(ish) over multiplication $(AB)^T = B^T A^T$
- Inversion is:
  - Distributive(ish) over multiplication $(AB)^{-1} = B^{-1}A^{-1}$
  - Swappable with transposition $(A^T)^{-1} = (A^{-1})^T$

Matrix Derivatives

Let $\mathbf{y}, \mathbf{x}$ be (column) vectors and $\mathbf{A}$ be a matrix (and assume all the shapes work out).

- $\mathbf{y} = \mathbf{A}\mathbf{x} \implies \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = A$

$a = y^T x$

$\rightarrow \frac{\partial a}{\partial x} = y$

- $\alpha = \mathbf{y}^\top \mathbf{A}\mathbf{x} \implies \frac{\partial \alpha}{\partial \mathbf{x}} = y^T A$

- $\alpha = \mathbf{x}^\top \mathbf{A}\mathbf{x} \implies \frac{\partial \alpha}{\partial \mathbf{x}} = x^T (A + A^T)$

Setting

- **Data:** (y_i, x_i) for $i = 1, \dots, n$, $x_i \in \mathbb{R}^k$
- **Baseline model:** $y_i = x_i^T \beta + \varepsilon_i$, $\beta \in \mathbb{R}^k$
- **Modeling assumptions:** Correct specification: $\mathbb{E}[\varepsilon | x] = 0$ a.s.

  Homoskedasticity: $\mathrm{Var}(\varepsilon | x) = \sigma^2$ a.s.

Estimand: "Analytic" Solution

Note that we can write β in the population as:

$$\beta = \mathbb{E}^{-1}[xx']\mathbb{E}[xy]$$

$$= \mathbb{E}^{-1}[xx']\mathbb{E}[x(x'\beta + \varepsilon)]$$

$$= \mathbb{E}^{-1}[xx']\left(\mathbb{E}[xx']\beta + \mathbb{E}[x\varepsilon]\right)$$

$$= \beta$$

$$\mathbb{E}[x\varepsilon] = 0$$

$$= \mathbb{E}\left[\mathbb{E}(x\varepsilon|x)\right]$$

$$= \mathbb{E}\left[x\,\mathbb{E}[\varepsilon|x]\right]$$

$$\mathbb{E}[\varepsilon|x] = 0$$

Estimand: Optimization Solution

Recall that OLS solves the following optimization problem:

$$\beta = \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} \; E\left[(y - x^T b)^2 \right]$$

We can then write:

$$\text{FOC:} \quad \frac{\partial}{\partial b} = 2E\left[-x\left(y - x^T \hat{b}\right) \right] = 0$$

$$\rightarrow E[xy] = E\left[xx' \hat{b}\right]$$

$$\rightarrow \hat{b} = E[xx']^{-1} E[xy]$$

Estimator: Analogy Principle

Population: $\beta = E^{-1}[xx^T]\,E[xy]$

$\downarrow$

Sample: $\hat{\beta} = E_n^{-1}[xx^T]\,E[xy]$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right)$$

$$= \left(X^T X\right)^{-1} X^T y$$

Gauss-Markov Theorem

Under these assumptions: Model $y = X\beta + \varepsilon$, $E[\varepsilon|X] = 0$, $Var(\varepsilon|X) = \sigma^2 I_n$

$X^T X$ invertible

OLS is:

- **B** est (lowest variance) $Var(A) \geq Var(B) \leftrightarrow Var(A) - Var(B)$ PSD
- **L** inear
- **U** nbiased
- **E** stimator

Proof of Gauss-Markov

Let $\tilde{\beta} = Cy$, WLOG $C = (X^TX)^{-1}X^T + D$

Unbiased: $E[\tilde{\beta}|X] = E[Cy|X] = E[((X^TX)^{-1}X^T + D)(X\beta + \varepsilon)|X]$

$= ((X^TX)^{-1}X^T + D)X\beta + ((X^TX)^{-1}X^T + D)E[\varepsilon|X]$ ($E[\varepsilon|X] \to 0$)

$= (X^TX)^{-1}X^TX\beta + DX\beta = (I_k + DX)\beta \to DX = 0$

Variance: $Var(\tilde{\beta}|X) = Var(Cy|X) = C\,Var(y|X)\,C^T = \sigma^2 CC^T$

$= \sigma^2((X^TX)^{-1}X^T + D)(X(X^TX)^{-1} + D^T)$

$\vdots$

$= \sigma^2((X^TX)^{-1} + DD^T) = Var(\hat{\beta}|X) + \sigma^2 DD^T$ ($DD^T \geq 0$, PSD)

1. Announcements
2. Identification Revisited
3. Review: Matrices
4. OLS: Things are Nice
5. OLS: Things Fall Apart (?)
6. Practice Problems

Relaxing Assumptions

We made 3 key assumptions before:

1. $E[\varepsilon|x] = 0$ → correct specification
2. $Var(\varepsilon|x) = \sigma^2$ → homoskedasticity
3. $E[xx^T]$ invertible

2 of these can be relaxed to allow for:

1. Misspecification: $E[y|x] \neq x^T\beta$
2. Heteroskedasticity: $Var(\varepsilon|x) \neq \sigma^2$
   - ↳ new variance: Sandwich Formula: $H^{-1} J H^{-1} \rightarrow E^{-1}(xx^T)$
     - $J \rightarrow E[xx^T]$

Theorem: Best Linear Predictor

$$\beta = \underset{b}{\operatorname{argmin}}\, E\left[\left(E[y|x] - b^T x\right)^2\right] = E^{-1}[xx^T]\,E[xy]$$

What changed?

Define: $\varepsilon = y - x^T\beta$. FOC gives: $E[x\varepsilon] = 0$

→ assumption by construction

↳ weaker than $E[\varepsilon|x] = 0$

Estimand Under Misspecification: Excluded Variable

What if there is some unobserved variable u that impacts y linearly?

True model: $y = \beta_0 + x\beta_1 + u\beta_2 + \varepsilon \qquad E[\varepsilon | x, u] = 0$

Misspecified reg.: $y = \tilde{\beta}_0 + x\tilde{\beta}_1 + \delta$

$$\tilde{\beta}_1 = \beta_1 + \beta_2 \frac{Cov(u,x)}{Var(x)}$$

$\rightarrow$ OVB unless $Cov(u,x) = 0$ or $\beta_2 = 0$

Frisch-Waugh-Lovell Theorem

Motivation: Multivariate regression, but only care about effect of 1 regressor
Want to control for others

Controlling: Let X_A is regressor of interest. X_B collects all controls
By controlling for X_B, account for variation in y explainable by X_B
and see how much of residual explainable by X_A.
May also be true $X_B \to X_A \to y$, so need to residualize X_A too.

What I want: $\hat{y} = \hat{\gamma}_1 X_B$, $\hat{X}_A = \hat{\gamma}_2 X_B \to (y - \hat{y}) = \beta(X_A - \hat{X}_A) + \varepsilon$
is param. of interest

If I run 1 big regression: $y = \tilde{\beta} X_A + \gamma_3 X_B + \delta$

By FWL: $\tilde{\beta} = \beta$

1. Announcements
2. Identification Revisited
3. Review: Matrices
4. OLS: Things are Nice
5. OLS: Things Fall Apart (?)
6. Practice Problems

Econ 271/Mgtecon 604 - Section 3

Amar Venugopal

Stanford University

Winter 2026

Outline

1. Announcements

2. Linear Regression: Cluster-Robust Inference

3. Intro to Maximum Likelihood Estimation (MLE)

4. Practice Problems

Announcements

- Optional **Lecture** on **Friday Jan 30** (same room, time as regularly scheduled section)
  - Lecture content (econometrics with incentives) will not be on problem sets or exams
- Makeup section on **Wednesday Jan 28, 4-6pm, GSB M104**
- Office hours next week: Tuesday 11:30am-1pm (Landau 149) (Wednesday OH canceled for section)
- No problem set this week!

1. Announcements

2. Linear Regression: Cluster-Robust Inference

3. Intro to Maximum Likelihood Estimation (MLE)

4. Practice Problems

Motivation: Suppose observe same individuals at multiple points in time

Individuals i, each observed T_i times

$(X_{i0}, X_{i1}, \dots, X_{iT_i})$ → correlated! no longer iid

$$\begin{pmatrix} X_{11}, & \dots, & X_{1T} \\ \vdots & & \vdots \\ X_{n1}, & \dots, & X_{nT} \end{pmatrix}$$

$n \times T$

→ effectively just have n "real" data points

the rest are correlated "versions"

Assumptions

- **Model:** $y_{it} = x_{it}'\beta + \varepsilon_{it}$

  $E[x_{it}\varepsilon_{it}] = 0$

  Assuming balanced panel, $T_i = T \quad \forall i$

- **Sample:** Sampling individuals, not time periods

  $(y_{01}, \ldots, y_{0T}, x_{01}, \ldots, x_{0T})$

  $(y_{11}, \ldots, y_{1T}, x_{11}, \ldots, x_{1T})$

  $\vdots$

  $(y_{n1}, \ldots, y_{nT}, x_{n1}, \ldots, x_{nT})$

  → sample

Estimation: Pooled OLS

Big Idea: ignore time component

$$\hat{\beta} = \left(\sum_{i=1}^{n} \sum_{t=1}^{T} X_{it} X_{it}' \right)^{-1} \left(\sum_{i=1}^{n} \sum_{t=1}^{T} X_{it} Y_{it} \right)$$

T_i in general

Assume iid across i, but not t

↓

sampling

Consistency

$$\hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} x_{it} x_{it}' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} x_{it} y_{it} \right)$$

$$= E_n^{-1}\left[\sum_{t=1}^{T} x_t x_t' \right] E_n\left[\sum_{t=1}^{T} x_t y_t \right]$$

$$\xrightarrow{p} E^{-1}\left[\sum_{t=1}^{T} x_t x_t' \right] \quad \text{by WLLN + CMT}$$

$$\xrightarrow{p} E\left[\sum_{t=1}^{T} x_t y_t \right] \quad \text{by WLLN}$$

$$\xrightarrow{p} E^{-1}\left[\sum_{t=1}^{T} x_t x_t' \right] E\left[\sum_{t=1}^{T} x_t y_t \right]$$

by CMT ($g(x,y) = xy$ is cts)

(WLLN: $E_n[x] \xrightarrow{p} E[x]$)

Asymptotic Normality

$$\sqrt{n}\left(\hat{\beta}-\beta\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} x_{it}x_{it}'\right)^{-1}\left(\sqrt{n}\cdot\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} x_{it}\underbrace{\left(y_{it}-x_{it}'\beta\right)}_{\varepsilon_{it}}\right)$$

$\xrightarrow{p} E^{-1}\left[\sum_{t=1}^{T} x_t x_t'\right]$ by WLLN, CMT

$\xrightarrow{d} N(0,\Sigma)$ by CLT

$\neq E\left[\sum x x'\varepsilon^2\right]$

Define $Q = E\left[\sum_{t=1}^{T} x_t x_t'\right]$, $\Sigma = E\left[\left(\sum_{t=1}^{T} x_t\varepsilon_t\right)\left(\sum_{t=1}^{T} x_t\varepsilon_t\right)'\right]$

By Slutsky: $\sqrt{n}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N\left(0, Q^{-1}\Sigma Q^{-1}\right)$

Did not assume ε_{it} uncorrelated across t

1. Announcements

2. Linear Regression: Cluster-Robust Inference

3. Intro to Maximum Likelihood Estimation (MLE)

4. Practice Problems

Big Idea (Normal Example)

- **Model:** $Z \sim N(\mu, \sigma^2)$
- **Sample:** $Z_1, \ldots, Z_n \overset{iid}{\sim} N(\mu, \sigma^2)$ — iid sampling; normal pdf
- **Likelihood:** $L(\mu, \sigma^2) = P(\text{data} \mid \text{model}) = \prod_{i=1}^{n} f_{(\mu,\sigma^2)}(z_i)$
- **Estimator:** $(\hat{\mu}, \hat{\sigma}^2) = \underset{\mu,\sigma^2}{\text{argmax}}\, L(\mu, \sigma^2)$

Generally write $\theta = (\mu, \sigma^2)$, $L(\theta)$

Conditional Log-Likelihood

Let $z = (y, x), y|x \sim f_\beta(y|x), x \sim g_\phi(x), \theta = (\beta, \phi)$

Definition of conditional probability: $f_\theta(y, x) = f_\beta(y|x)\, g_\phi(x)$

$$L(\theta) = \prod_{i=1}^{n} f_\theta(y_i, x_i) = \prod_{i=1}^{n} f_\beta(y_i|x_i)\, g_\phi(x_i)$$

Identification condition: $P_{\theta_0}\left(f_\theta(z) \neq f_{\theta_0}(z)\right) > 0 \quad \forall \theta \neq \theta_0$ $\quad \forall \theta \neq \theta_0$

Jensen's Inequality gives us Information Inequality: $E_{\theta_0}[\log f_{\theta_0}(z)] > E_{\theta_0}\left[\log f_\theta(z)\right]$

Take logs of everything:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f_\beta(y_i|x_i) + \log g_\phi(x_i)$$

$\log f_\beta(y_i|x_i)$: $\perp \phi$ $\quad$ $\log g_\phi(x_i)$: $\perp \beta$

If β, ϕ don't overlap, then $\hat{\beta} = \arg\max_\beta \sum_{i=1}^{n} \log f_\beta(y_i|x_i)$

An Algorithm for MLE

1. Transform population model into an equation for which you know the distribution if needed
2. Write down $L(\theta) = \prod_{i=1}^{n} f_\theta(y_i|x_i)$, where we know distribution of $y|x$, and θ is parameter of interest
3. Transform to log-likelihood, $\ell(\theta)$
4. Simplify objective function (ignore positive scalar multiples, extraneous terms, etc.)
5. Take FOC and solve for $\hat{\theta}$
6. Check second derivative to verify curvature

Estimation Example: OLS

Let $y = x'\beta + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2), \sigma^2$ known.

1. What is the likelihood?

$\varepsilon = y - x'\beta \rightarrow y|x \sim \mathcal{N}(x'\beta, \sigma^2) \rightarrow L(\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}}$

2. What is the log-likelihood?

$\ell(\beta) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}}\right)$

3. Simplify

$= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_i - x_i'\beta)^2}{2\sigma^2}$ ($\perp\beta$; $2\sigma^2 \rightarrow$ positive scalar)

$\rightarrow \hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 = \hat{\beta}^{OLS}$

OLS is MLE (in this setting)

Estimation Example: Logit

Let $y \in \{0, 1\}, P(y = 1|x) = (1 + \exp(-x^\top \beta))^{-1}$.

1. What is the likelihood?

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{1}{1+e^{-x_i'\beta}}\right)^{y_i} \left(1 - \frac{1}{1+e^{-x_i'\beta}}\right)^{1-y_i}$$

2. What is the log-likelihood?

$$\ell(\beta) = \sum_{i=1}^{n} \log\left[\left(\frac{1}{1+e^{-x_i'\beta}}\right)^{y_i}\left(1 - \frac{1}{1+e^{-x_i'\beta}}\right)^{1-y_i}\right]$$

3. Simplify

$$\ell(\beta) = \sum_{i=1}^{n} y_i \log\left(\frac{1}{1+e^{-x_i'\beta}}\right) + (1-y_i)\log\left(1 - \frac{1}{1+e^{-x_i'\beta}}\right)$$

$$= \sum_{i=1}^{n} -y_i \log\left(1+e^{-x_i'\beta}\right) + (1-y_i)\log\left(\frac{e^{-x_i'\beta}}{1+e^{-x_i'\beta}}\right)$$

$$\vdots$$

$$\rightarrow \hat{\beta}_{logit} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (1-y_i)(x_i'\beta) + \log\left(1+e^{-x_i'\beta}\right)$$

1. Announcements

2. Linear Regression: Cluster-Robust Inference

3. Intro to Maximum Likelihood Estimation (MLE)

4. Practice Problems

Econ 271/Mgtecon 604 - Section 4

Amar Venugopal

Stanford University

Winter 2026

Outline

1. Announcements
2. MLE: Theoretical Properties
3. Panel Data
   - Fixed Effects
   - Random Effects
   - Difference in Differences
4. Practice Problems

1. Announcements

2. MLE: Theoretical Properties

3. Panel Data
   - Fixed Effects
   - Random Effects
   - Difference in Differences

4. Practice Problems

Announcements

- Optional **Lecture** on **Friday Jan 30** (same room, time as regularly scheduled section)
- Next week, schedule goes back to normal
- **Midterm Monday Feb 9** - Section 5 will spend (at least) some time on review

1. Announcements

2. MLE: Theoretical Properties

3. Panel Data
   - Fixed Effects
   - Random Effects
   - Difference in Differences

4. Practice Problems

Extremum Estimators (M-Estimators)

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \; \hat{Q}_n(\theta)$$

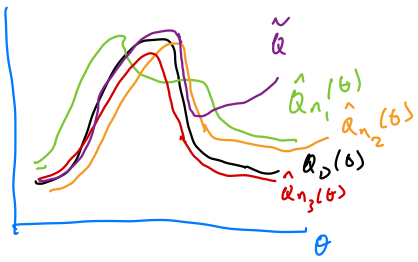
$$\hat{Q}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m(z_i; \theta)$$

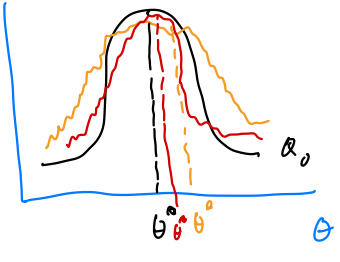
Ex. MLE:

$$m(z_i; \theta) = \log f_\theta(z_i)$$

Uniform Convergence in Probability: $\hat{Q}_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$ if:

$$\sup_{\theta \in \Theta} \left| \hat{Q}_n(\theta) - Q_0(\theta) \right| \xrightarrow{p} 0$$





Useful Lemma

Let $\hat{Q}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(z_i;\theta)$, $Q_0(\theta) = E\left[m(z;\theta)\right]$

Assume:

1. $\{z_1, \ldots, z_n\}$ iid
2. Θ compact
3. $m(z;\theta)$ is continuous at each $\theta \in \Theta$ w.p. 1
4. $\exists M(y)$ s.t. $|m(y;\theta)| \le M(y)$ $\forall \theta \in \Theta$ and $E[M(y)] < \infty$

"M-estimators"

Then: $Q_0(\theta)$ continuous and $\hat{Q}_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$

Theorem: Consistency

Assume there is a function $Q_0(\theta)$ and a vector $\theta_0 \in \Theta \cup \mathbb{R}^k$ such that:

1. $Q_0(\theta)$ is uniquely maximized at θ_0
2. Θ compact
3. $Q_0(\theta)$ continuous
4. $\hat{Q}_n(\theta)$ converge unif. in prob. to $Q_0(\theta)$

Then, $\hat{\theta} \xrightarrow{p} \theta_0$.

Asymptotic Linearity: $\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(z_i) + o_p(1)$

influence function; noise

avg. over small increments

$\hookrightarrow \mathbb{E}[\psi(z)] = 0$

$\mathbb{E}[\|\psi(z)\|^2] < \infty$

Thm, CLT gives:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$$

$$V = \mathrm{Var}(\psi(z)) = \mathbb{E}[\psi(z)\psi(z)']$$

Score Function: $s(z;\theta) = \frac{\partial \log f_\theta(z)}{\partial \theta}$

MLE FOC: $\frac{1}{n}\sum_{i=1}^{n} s(z_i;\hat{\theta}) = 0$

Taylor Expansion around θ_0: $s(z_i;\theta_0) + \frac{\partial}{\partial \theta'} s(z;\theta_0)'(\hat{\theta}-\theta_0) \approx 0$

$$\rightarrow \frac{1}{n}\sum s(z_i;\theta_0) + \frac{1}{n}\sum \frac{\partial}{\partial \theta'} s(z_i;\theta_0)(\hat{\theta}-\theta_0) \approx 0$$

$$H = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta} s(z_i;\theta)$$

$$\rightarrow \frac{1}{n}\sum_{i=1}^{n} s(z_i;\theta_0) + H(\hat{\theta}-\theta_0) \approx 0$$

$$\rightarrow \sqrt{n}(\hat{\theta}-\theta_0) \approx -H^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(z_i;\theta) \rightarrow \sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} N(0,V)$$

$$\psi(z_i) = -H^{-1} s(z_i;\theta)$$

$$V = H^{-1} J H^{-1}$$

$$J = E\left[s(z;\theta)s(z_i;\theta)'\right]$$

Kullback-Liebler Divergence:

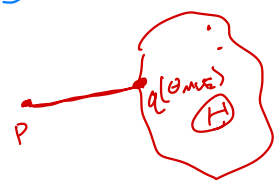
$$D_{KL}(p \,||\, q) = E_p\left[\log \frac{p(x)}{q(x)}\right]$$

p = true underlying distribution

q = our approx.

Maximizing Likelihood = Minimizing KL Divergence:

$$\begin{aligned}
\operatorname*{argmin}_\theta D_{KL}(p||q(\theta)) &= \operatorname*{argmin}_\theta E\left[\log \frac{p(x)}{q(x;\theta)}\right] \\
&= \operatorname*{argmin}_\theta E\left[\log p(x) - \log q(x;\theta)\right] \\
&= \operatorname*{argmin}_\theta -E\left[\log q(x;\theta)\right] \\
&= \operatorname*{argmax}_\theta E\left[\log q(x;\theta)\right] \\
&= \hat{\theta}_{MLE}
\end{aligned}$$



1. Announcements

2. MLE: Theoretical Properties

3. Panel Data
   - Fixed Effects
   - Random Effects
   - Difference in Differences

4. Practice Problems

1. Announcements

2. MLE: Theoretical Properties

3. Panel Data
   - Fixed Effects
   - Random Effects
   - Difference in Differences

4. Practice Problems

Modeling Assumptions: $y_{it} = x_{it}'\beta + \eta_i + \varepsilon_{it}$

is called an individual variable

Data iid across i, $E[\varepsilon_{it} | x_{i1}, \dots, x_{iT}, \eta_i] = 0$

May have $Cov(x_{it}, \eta_i) \neq 0$

Estimator: $(\hat{\beta}_{FE}, \hat{\eta}_1, \dots, \hat{\eta}_n) = \underset{\beta, \{\eta_i\}_{i=1}^n}{\arg\min} \sum_{i=1}^{n} \sum_{t=1}^{T} (y_{it} - x_{it}'\beta - \eta_i)^2$

FOC: $\frac{\partial}{\partial \beta}: \sum_i \sum_t -2 x_{it} (y_{it} - x_{it}'\hat{\beta} - \hat{\eta}_i) = 0$

$\frac{\partial}{\partial \eta_i}: \sum_t -2 (y_{it} - x_{it}'\hat{\beta} - \hat{\eta}_i) = 0$

$\to \hat{\eta}_i = \frac{1}{T} \sum_{t=1}^{T} y_{it} - x_{it}'\hat{\beta} = \bar{y}_i - \bar{x}_i'\hat{\beta}$

$\to \hat{\beta}_{FE} = \left(\sum_i \sum_t (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \left(\sum_i \sum_t (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) \right)$

Can also apply FVL, get same thing.

Asymptotic Properties

Small T, large n asymptotics

OLS machinery applies:

$\hat{\beta}_{FE} \xrightarrow{p} \beta_{FE}$

$$\sqrt{n}\left(\hat{\beta}_{FE} - \beta_{FE}\right) \xrightarrow{d} N\left(0, H^{-1} J H^{-1}\right)$$

$$H = E\left[\sum_t (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'\right]$$

$$J = E\left[\left(\sum_t (x_{it} - \bar{x}_i)(\varepsilon_{it} - \bar{\varepsilon}_i)\right)\left(\sum_t (x_{it} - \bar{x}_i)(\varepsilon_{it} - \bar{\varepsilon}_i)\right)'\right]$$

1. Announcements

2. MLE: Theoretical Properties

3. Panel Data
   - Fixed Effects
   - Random Effects
   - Difference in Differences

4. Practice Problems

Modeling Assumptions: $y_{it} = x_{it}'\beta + \eta_i + \varepsilon_t$

$E[\varepsilon_{it} | x_{i1}, \ldots, x_{iT}, \eta_i] = 0$ $\quad$ $E[\eta_i | x_{i1}, \ldots, x_{iT}] = \alpha$

diff. b/w i are random

Estimator: Treat η_i like error term, $u_{it} + \alpha = \varepsilon_{it} + \eta_i$

$$y_{it} = \alpha + x_{it}'\beta + u_{it}$$

$$E[u_{it} | x_{i1}, \ldots, x_{iT}] = 0$$

→ looks just like pooled regression!

inherits asymptotic properties

1. Announcements

2. MLE: Theoretical Properties

3. Panel Data
   - Fixed Effects
   - Random Effects
   - Difference in Differences

4. Practice Problems

Model: Two Way Fixed Effects

Suppose also want time FE

Consider causal setting, d_{it} = treatment indicator

TWFE: $y_{it} = \alpha_i + \beta_t + \tau d_{it} + \varepsilon_{it}$

---

2 units, 2 time periods, $d_{A2} = 1$

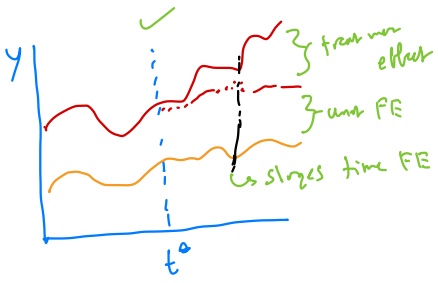
i \ t	1	2
A	$Y_A(0)$	$Y_A(1)$
B	$Y_B(0)$	$Y_B(0)$

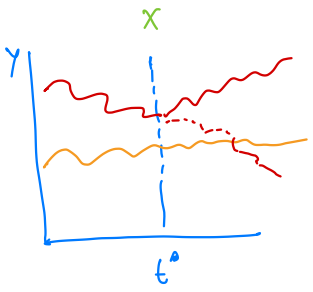
$$(Y_{A2} - Y_{A1}) - (Y_{B2} - Y_{B1}) = \tau^{DiD}$$

When does $\tau^{DiD} = \tau^{TWFE}$

Key Assumption: Parallel Trends

Big Idea: If treatment never happened, then the 2 groups' outcomes would have remained parallel





Testable? No

But, can get at suggestive evidence → pre-period

1. Announcements

2. MLE: Theoretical Properties

3. Panel Data
   - Fixed Effects
   - Random Effects
   - Difference in Differences

4. Practice Problems

Econ 271/Mgtecon 604 - Section 5

Amar Venugopal

Stanford University

Winter 2026

Outline

1. Announcements

2. DiD Extensions

3. Midterm Practice

Announcements

- **Midterm Monday Feb 9**
  - Usual lecture room/time
  - One-sided formula sheet permitted

1. Announcements

2. DiD Extensions

3. Midterm Practice

Big Idea: Different units start receiving treatment at different times

e_i = 1^{st} time period unit i is treated

Absorbing treatment

Ex. 2×2 **DiD with staggered adoption**

$$\hat{\tau} = \left(E_n[y \mid t=2, e_i=2] - E_n[y \mid t=1, e_i=2]\right) - \left(E_n[y \mid t=2, e_i=\infty] - E_n[y \mid t=1, e_i=\infty]\right)$$

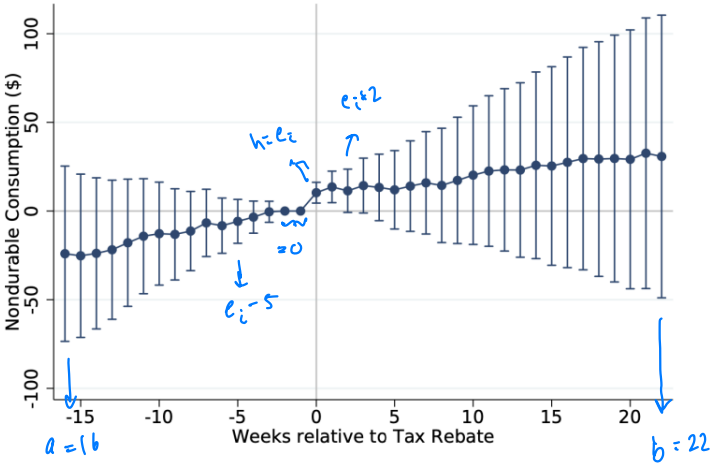
TWFE:

$$Y_{it} = \alpha_i + \beta_t + \sum_{\substack{h=-A \\ \neq -1}}^{b-1} \tau_h \mathbb{1}\{t = e_i + h\} + \tau_b \mathbb{1}\{t \geq e_i + b\} + \varepsilon_{it}$$

pre- / post-treatment effects per-period

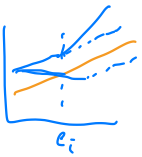
final period (lumping all future together)

Event Studies



What if we are misspecified?

Relied heavily on 2 assumptions:



1. Parallel trends
2. Homogeneous TE $\rightarrow$ relax

seems plausible this would not hold under staggered adoption...

Effect Heterogeneity - Simple Example

Specification: $y_i = d_i \tau + x_i' \beta + \varepsilon_i$

Estimator:

$$\hat{\tau} = \frac{\sum_i (d_i - x_i'\hat{\gamma}) y_i}{\sum_i (d_i - x_i'\hat{\gamma})^2}$$

$\hat{\gamma}$ = coeff. of reg. $d_i \sim x_i$

$\sum_i (d_i - x_i'\hat{\gamma}) y_i$: $\mathrm{cov}(y - \hat{y}, d - \hat{d})$

$\sum_i (d_i - x_i'\hat{\gamma})^2$: $\mathrm{Var}(d - \hat{d})$

Alternative Model: $y_i = d_i \tilde{\tau}_i + x_i'\tilde{\beta} + \tilde{\varepsilon}_i$

Estimand (Under Alternative):

$$E\left[\hat{\tau} \,\middle|\, \begin{matrix} d_1, \dots, d_n \\ x_1, \dots, x_n \\ \tilde{\tau}_1, \dots, \tilde{\tau}_n \end{matrix}\right] = E\left[\frac{\sum_i (d_i - x_i'\hat{\gamma})(d_i \tilde{\tau}_i + x_i'\tilde{\beta}_1 \tilde{\varepsilon}_i)}{\sum_i (d_i - x_i'\hat{\gamma})^2}\right] = \frac{\sum (d_i - x_i'\hat{\gamma}) d_i \tilde{\tau}_i}{\sum (d_i - x_i'\hat{\gamma}) d_i}$$

w_i

$$= \frac{\sum w_i \tilde{\tau}_i}{\sum w_i}$$

Effect Heterogeneity - TWFE

Binary treatment

Specification: $Y_{it} = \alpha_i + \beta_t + d_{it}\,\tau + \varepsilon_{it}$

Estimator: $\hat{\tau}$ = very complicated

Alternative Model: $Y_{it} = \alpha_i + \beta_t + d_{it}\,\tau_{it} + \varepsilon_{it}$

Estimand (Under Alternative): $E\left[\hat{\tau} \mid (d_{it}, \tau_{it})_{it}\right] = \sum_{d_{it}=1} \tau_{it}\, v_{it}$

(Borusyak, Jaravel, Spiess 2024)

$\sum_{d_{it}=1} v_{it} = 1$, but can be negative

→ spawned massive literature on "fixes"

1. Announcements

2. DiD Extensions

3. Midterm Practice

Econ 271/Mgtecon 604 - Section 6

Amar Venugopal

Stanford University

Winter 2026

Outline

1. Announcements

2. Synthetic Control

3. Machine Learning Intro

4. Practice Problems

Announcements

- No class Monday Feb 16 (Presidents' Day)

1. Announcements

2. Synthetic Control

3. Machine Learning Intro

4. Practice Problems

What kind of panel must we have?

Proper panel

$$W = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{matrix} \} N \\ \\ \\ \\ \underbrace{\qquad\qquad}_{T} \end{matrix}$$

Estimand:

ATT

$$\tau^{ATT} = \frac{\sum_{i,t} W_{it}\left(Y_{it}(1) - Y_{it}(0)\right)}{\sum_{i,t} W_{it}} = Y_{NT}(1) - Y_{NT}(0) = \tau^{N}$$

$Y_{NT}(1)$: observed

$Y_{NT}(0)$: need to estimate

Vertical vs. Horizontal Regression

Goal: Impute / estimate $Y_{NT}(0)$

$$Y = \begin{pmatrix} Y_{11} & Y_{12} & \cdots & Y_{1T} \\ \vdots & & & \\ Y_{N1} & Y_{N2} & \cdots & Y_{NT} \end{pmatrix}$$

horizontal unit of obs.

vertical unit of obs.

Vertical Regression: $T-1$ obs, $N-1$ regressors

eg. SC

Main idea: similar units behave similarly

Horizontal Regression: $N-1$ obs., $T-1$ regressors

eg. uncomfortableness

Main idea: history explains all

$Y(1)$

$\hat{Y}(0)$

$$Y_{Nt} = \sum_{j \neq N} w_j Y_{jt}$$

w_j ?

Factor Model: Assume movement in outcome summarized by shared latent factors

$$Y_{it}(0) = \mu_i^T \lambda_t + \varepsilon_{it}$$

μ_i: loadings, λ_t: factors

With covariates:

$$Y_{it}(0) = \delta_t + \theta_t z_i + \mu_i^T \lambda_t + \varepsilon_{it}$$

δ_t: common factor constant load; $\theta_t z_i$: obs cov unaffected by treatment

Synthetic Control - Estimation (No Covariates)

Ideal: Match on factor loadings

$$\sum_{j=1}^{N-1} \hat{w}_j \mu_j = \mu_N \longrightarrow \mu_i \text{ unobserved}$$

In Practice: Match on pre-treatment outcomes

$$\hat{Y}_{NT}(0) = \sum_{j=1}^{N-1} \hat{w}_j Y_{jT}$$

$$\hat{w} = \underset{w \in \Delta^{N-1}}{\operatorname{argmin}} \sum_{t=1}^{T-1} \left(Y_{Nt} - \sum_{j=1}^{N-1} w_j Y_{jt} \right)^2$$

Challenge: With 1 treated unit, no averaging → no CLT

Open question

Placebo analyses generally used, but still have a choice to make:

leave out true treated unit

1. **Placebo on Units:** Iteratively pretend each unit is treated, estimate SC for treated period

$$\hat{V} = \frac{1}{N-1} \sum_{j \neq 1}^{N-1} \left(\hat{Y}_{j,T} - Y_{j,T} \right)^2$$

except for true treated period

2. **Placebo on Time Periods:** Iteratively pretend each period is the treatment period, estimate SC for true treated unit

$$\hat{V} = \frac{1}{T - T_0 - 1} \sum_{t=T_0+1}^{T-1} \left(\hat{Y}_{1,t} - Y_{1,t} \right)^2$$

1. Announcements

2. Synthetic Control

3. Machine Learning Intro

4. Practice Problems

Training Data: $(Y_1, X_1), \dots, (Y_n, X_n)$ (assume iid)

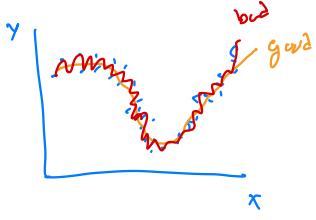
Prediction: Suppose true $f: X \to Y$

Goal: $\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ \ell(f(x), y)$ → loss function

e.g. MSE (regression)

cross-entropy (classification)

(Main?) Concern: Overfitting



Given new (y', x') want $\hat{f}(x') \approx y'$

If I do "too well" in-sample, I may not generalize

Big Idea: Prevent models from getting too complex by penalizing complexity

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \, \text{Penalty}(\beta)$$

↳ penalty "strength"

Ridge (L_2): $\text{Penalty}(\beta) = \|\beta\|_2^2 = \sum_{j=1}^{k} \beta_j^2$

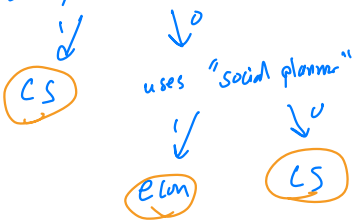
Shrink towards 0, but never all the way. Distributive

Lasso (L_1): $\text{Penalty}(\beta) = \|\beta\|_1 = \sum_{j=1}^{k} |\beta_j|$

Shrink to 0. Selective → yields sparse solutions

Big Idea: Come up with "split" rules to fit the data

e.g. economist or CS?

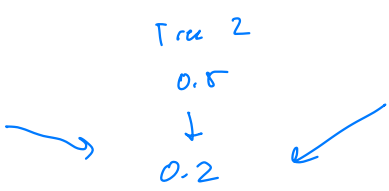


Choose splits to minimize entropy / maximize info. gain at each step

How to regularize? Limit depth, min # of units in node before split, ...

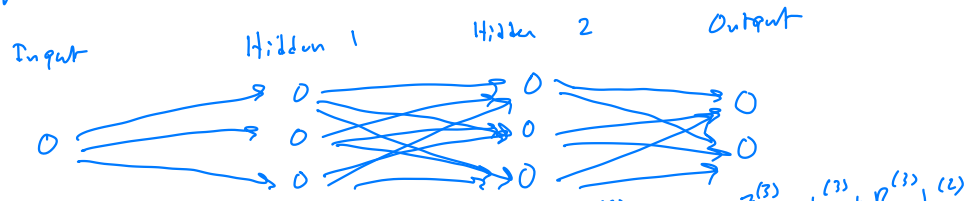
Big Idea: Instead of believing just 1 tree, "grow" a whole "forest" and have them "vote"

- each gets random subset of features, observations
- aggregate via "bagging"



How to regularize? Change # trees, # features given to each, ...

Big Idea: Mimic "brain" structure via lots of linear (with nonlinear components) allowing for very flexible functional forms



$\sigma^{(\ell)} \in \{ \text{ReLU}, \text{sigmoid}, \tanh, \dots \}$

How to regularize? # layers, # neurons/layer, penalty terms, dropout

1. Announcements

2. Synthetic Control

3. Machine Learning Intro

4. Practice Problems

Econ 271/Mgtecon 604 - Section 7

Amar Venugopal

Stanford University

Winter 2026

Outline

1. Announcements

2. Unconfoundedness, Propensity Scores, & IPW

3. Model Tuning

4. Multimodal Data

5. Practice Problems

Announcements

- Problem Set 4 due Thursday Feb 26
- Jann's last lecture Monday Feb 23
- Luigi starts Wednesday Feb 25 (time series!)

1. Announcements

2. Unconfoundedness, Propensity Scores, & IPW

3. Model Tuning

4. Multimodal Data

5. Practice Problems

How can we identify causal effects?

- **Experiments:** Randomization of treatment

- **Observational Data:** Treatment no longer random

Goal: "as-if" randomization

unconfoundedness + overlap

Unconfoundedness Assumption

Also known as "selection on observables" or "exogeneity".

Definition: $W_i \perp (Y_i(0), Y_i(1)) \mid X_i$

Observables may impact treatment assignment

But, conditional on them, "as-if" random

Propensity Scores

- **Motivation:** $\forall m$ continuous X's, $P(\exists i,j : X_i = X_j) = 0$
- **Propensity Score:** $e(x) = P(W_i = 1 \mid X_i = x)$
  → "sufficient statistic" for selection on observables
- **(Alternative) Unconfoundedness:** $W_i \perp (Y_i(0), Y_i(1)) \mid e_i$

Assessing Unconfoundedness

While unconfoundedness is not *testable*, its plausibility can be *assessed*.

2 assessment strategies:

1. Estimate effect on pseudo outcome
2. Estimate effect on pseudo treatment

Conditional Average Treatment Effect (CATE): $\tau(X) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$

Goal for ATE ID: $\tau = \mathbb{E}[\tau(X_i)]$

Let's see: $\mathbb{E}[\tau(X_i)] = \mathbb{E}\left[\mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]\right]$ (defn of $\tau(x)$)

$= \mathbb{E}\left[\mathbb{E}[Y_i(1)|X_i = x] - \mathbb{E}[Y_i(0)|X_i = x]\right]$ (linearity of $\mathbb{E}[\cdot]$)

$= \mathbb{E}\left[\mathbb{E}[Y_i(1)|W_i = 1, X_i = x] - \mathbb{E}[Y_i(0)|W_i = 0, X_i = x]\right]$ (unconfoundedness)

$= \mathbb{E}\left[\mathbb{E}[Y_i|W_i = 1, X_i = x] - \mathbb{E}[Y_i|W_i = 0, X_i = x]\right]$ (defn of PO)

$\stackrel{?}{=} \tau$

may not be separately estimable → need something more

Overlap Assumption

Definition: $P(W_i = 1 \mid X_i) = e_i \in (0,1)$

For any X_i, we can find both treated and control units

This assumption is *testable*. Some strategies:

1. Normalized differences
2. Plotting 
3. Log odds ratio

$$\log\left(\frac{e(x)}{1-e(x)}\right)$$

What if we have only limited overlap?

Challenge: Cannot get good estimates of separate conditional outcomes

→ need a "better" sample

2 strategies:

1. Matching
2. Subsampling

Identification!

$$
\begin{aligned}
\mathbb{E}[\tau(X_i)] &= \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]] \quad \text{by definition} \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|X_i = x] - \mathbb{E}[Y_i(0)|X_i = x]] \quad \text{by linearity} \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|W_i = 1, X_i = x] - \mathbb{E}[Y_i(0)|W_i = 0, X_i = x]] \quad \text{by unconfoundedness} \\
&= \mathbb{E}[\mathbb{E}[Y_i|W_i = 1, X_i = x] - \mathbb{E}[Y_i|W_i = 0, X_i = x]] \quad \text{by definition} \\
&= \tau \quad \text{by overlap}
\end{aligned}
$$

Inverse Propensity Weighting (IPW)

Starting from our standard difference-in-means estimator:

$$\hat{\tau} = \mathbb{E}_n[Y_i | W_i = 1] - \mathbb{E}_n[Y_i | W_i = 0]$$

$$= \frac{1}{n_1} \sum_{W_i=1} W_i Y_i - \frac{1}{n_0} \sum_{W_i=0} (1-W_i) Y_i$$

$$= \frac{\sum_i W_i Y_i}{\sum_i W_i} - \frac{\sum_i (1-W_i) Y_i}{\sum_i 1-W_i}$$

$$= \frac{\mathbb{E}_n[W_i Y_i]}{\mathbb{E}_n[W_i]} - \frac{\mathbb{E}_n[(1-W_i) Y_i]}{\mathbb{E}_n[1-W_i]}$$

$\hat{p} \leftarrow \mathbb{E}_n[W_i]$

$$= \mathbb{E}_n\left[\frac{W_i Y_i}{\hat{p}} - \frac{(1-W_i) Y_i}{1-\hat{p}} \right]$$

→ each group weighted by inverse probability

→ IPW estimator (non-random assignment):

$$\hat{\tau}_{IPW} = \mathbb{E}_n\left[\frac{W_i Y_i}{\hat{e}(x_i)} - \frac{(1-W_i) Y_i}{(1-\hat{e}(x_i)} \right]$$

↘ ↙

probability of treatment depends on covariates

1. Announcements

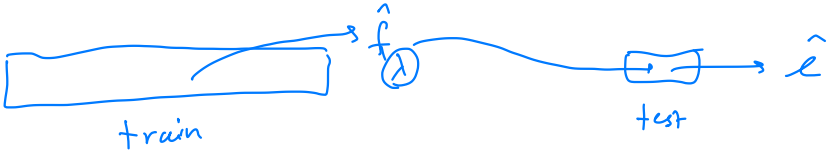
2. Unconfoundedness, Propensity Scores, & IPW

3. Model Tuning

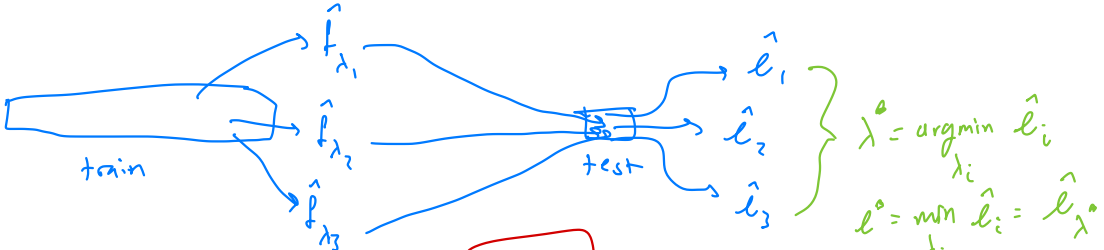
4. Multimodal Data

5. Practice Problems

How do we choose a value for our (regularization) parameters? λ



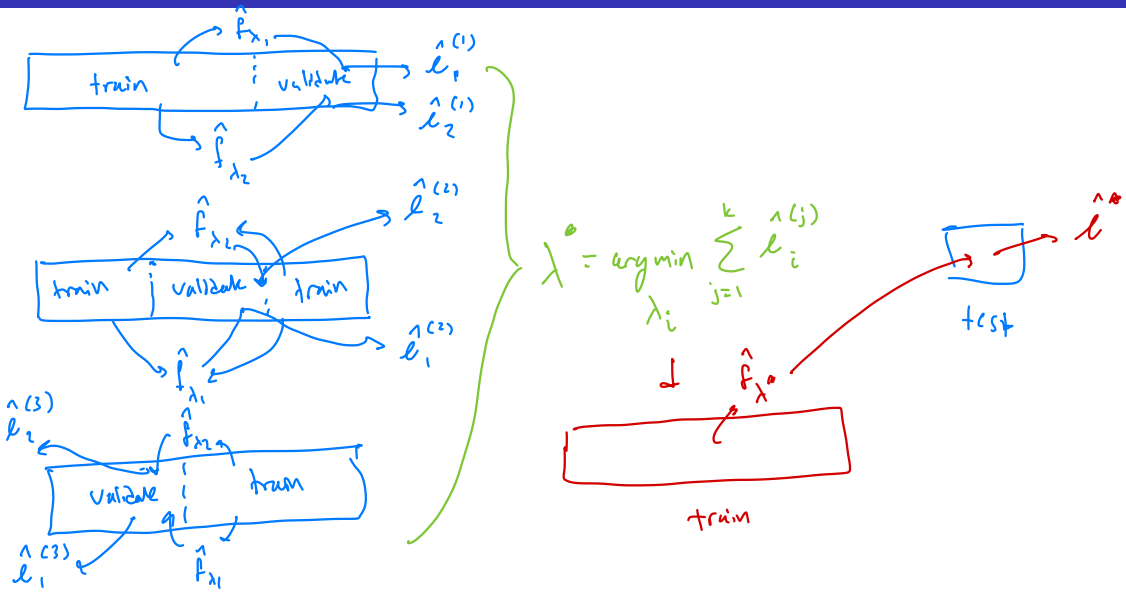
One strategy:



$$\lambda^* = \underset{\lambda_i}{\operatorname{argmin}}\ \hat{\ell}_i$$

$$\ell^* = \min_{\lambda_i} \hat{\ell}_i = \hat{\ell}_{\lambda^*}$$

Cross-Validation



1. Announcements
2. Unconfoundedness, Propensity Scores, & IPW
3. Model Tuning
4. Multimodal Data
5. Practice Problems

Image Data: Computer Vision

Big Idea: Model images as collections of pixels, each of which has an RGB value

→ learn to predict pixel values from neighboring pixels

Convolutional Neural Networks (CNNs):

Use "sliding window" (assign neurons to specific regions of space in image)

→ takes into account "grid topology" of image

→ allows for dense representation ("embedding") of image based learned inter-pixel relationships

Text Data: NLP

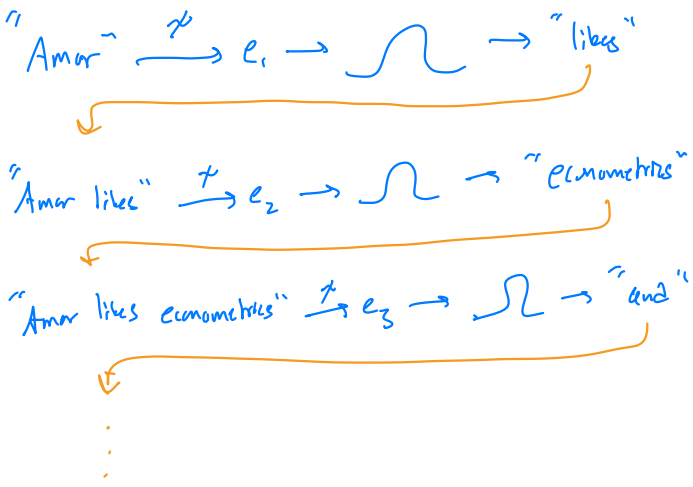
Big Idea: Model text as sequence of words
→ learn to predict next word from previous word(s)

Progression of methods:

1. Bag of words – word count matrix
2. Topic model (LDA, ...) – probabilistic "clustering" in text space
3. Embeddings (word2vec) – dense vector for predicting words
4. Transformers – fancy architecture that captures complex inter-word relationships (attention)
5. LLMs – transformer-based models that generate text (iteratively)

$\mathcal{F}$: text → embedding

Iterative Sampling:



At each step, model's embedding is "sufficient statistic" for next token/word.

But, generating that embedding requires knowledge of all prior tokens/words.

1. Announcements
2. Unconfoundedness, Propensity Scores, & IPW
3. Model Tuning
4. Multimodal Data
5. Practice Problems

Econ 271/Mgtecon 604 - Section 8

Amar Venugopal

Stanford University

Winter 2026

Outline

1. Announcements

2. Double Machine Learning

3. CATE Estimation

4. Introduction to Linear Stationary Processes

5. Practice Problems

Announcements

- Final Exam: March 18, 3:30-6:30pm
- Last problem set: due March 13, 11:59pm
- Final review session: March 13, 1:30-3:20pm (usual section time/place)

1. Announcements

2. Double Machine Learning

3. CATE Estimation

4. Introduction to Linear Stationary Processes

5. Practice Problems

$$\hat{\tau}^{AIPW} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)) + W_i\frac{Y_i - \hat{\mu}_1(X_i)}{\hat{e}(X_i)} - (1 - W_i)\frac{Y_i - \hat{\mu}_0(X_i)}{1 - \hat{e}(X_i)}$$

Double Robustness: $\hat{\tau}^{AIPW} \xrightarrow{p} \tau$ if:

(1) $\hat{e}$ correctly specified (and estimated consistently)

or

(2) $\hat{\mu}_1, \hat{\mu}_0$ correctly specified (and estimated consistently)

Influence Function/AIPW - Correctly Specified Outcomes

$$\hat{\tau}^{AIPW} = \left[\frac{1}{N}\sum_{i=1}^{N}(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))\right] + \left[\frac{1}{N}\sum_{i=1}^{N} W_i \frac{Y_i - \hat{\mu}_1(X_i)}{\hat{e}(X_i)} - (1 - W_i)\frac{Y_i - \hat{\mu}_0(X_i)}{1 - \hat{e}(X_i)}\right]$$

$\to 0$ $\to 0$ $\to$ something $\to$ something

$\to E_n[\hat{\mu}_1(x) - \hat{\mu}_0(x)]$

$\to \mu_1(x)$ $\to \mu_0(x)$

$\to E[\mu_1(x) - \mu_0(x)]$

$= E[E[Y(1) - Y(0) \mid X]]$

$= \tau(x)$

$= \tau$

$\to$ 0. limit

$\to 0$

Influence Function/AIPW - Correctly Specified Propensity

$$\hat{\tau}^{AIPW} = \left[\frac{1}{N}\sum_{i=1}^{N} W_i \frac{Y_i}{\hat{e}(X_i)} - (1-W_i)\frac{Y_i}{1-\hat{e}(X_i)}\right] + \left[\frac{1}{N}\sum_{i=1}^{N}(\hat{e}(X_i) - W_i)\left(\frac{\hat{\mu}_1(X_i)}{\hat{e}(X_i)} + \frac{\hat{\mu}_0(X_i)}{1-\hat{e}(X_i)}\right)\right]$$

$\hat{e}$ pw

$\xrightarrow{p} \tau$

$E[\cdot] = E[Y_i(W_i=1)] - E[Y_i(W_i=0)]$

$E[e(X_i)] = N_1$

$E[\cdot] = 0$

$\xrightarrow{?} 0$

Cross-Fitting/Sample Splitting

Idea: Deriving formal properties easier if we estimate nuisances on separate sample from x (create independence)

Procedure:

1. Split sample into K folds, $J_i = \{1, .., K\}$
2. Estimate functions $\hat{\mu}_w^j(x)$, $\hat{e}^j(x)$ on sample with $J_i \neq j$
3. Estimate $\hat{\tau}^j(\hat{\mu}_w^j, \hat{e}^j)$ (AIPW)
4. Average over all $\hat{\tau}^j$ to get an efficient estimator

1. Announcements

2. Double Machine Learning

3. CATE Estimation

4. Introduction to Linear Stationary Processes

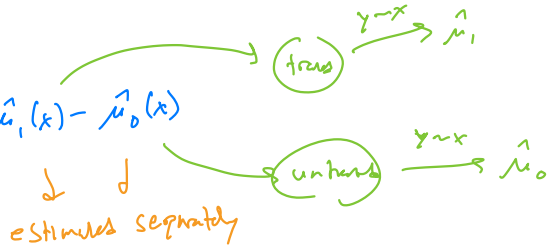
5. Practice Problems

CATEs

Big Idea: Maybe observables play some role in TE

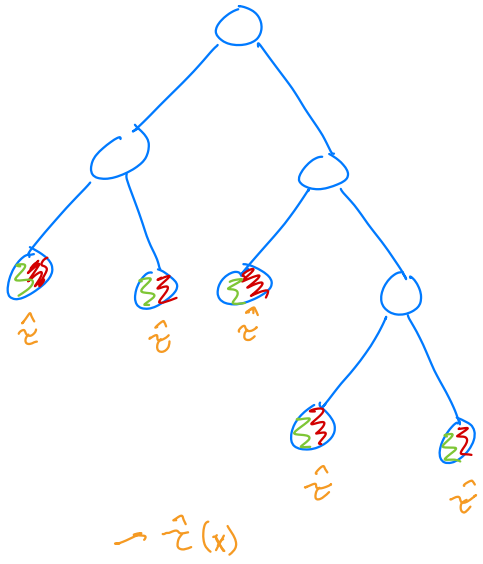
$$\tau(x) = E[Y(1) - Y(0) \mid X]$$

Idea: $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$



Problem: may not be comparing "alike" individuals

Causal Trees/Forests



1. Announcements

2. Double Machine Learning

3. CATE Estimation

4. Introduction to Linear Stationary Processes

5. Practice Problems

- **Stochastic Process:** sequence of RVs ordered by discrete time index

$$\{Y_t\} = [Y_0, Y_1, \cdots, Y_T] \rightarrow \text{1 observation}$$

- **Expectation of t-th observation:**

$$E[Y_t] = \mu_t = \int_{-\infty}^{\infty} y_t f_{Y_t}(y_t)\, dy_t$$

- **Variance of t-th observation:**

$$E[(Y_t - \mu_t)^2] = \gamma_{0,t} = \int_{-\infty}^{\infty} (y_t - \mu_t)^2 f_{Y_t}(y_t)\, dy_t$$

- **Autocovariance (of order j):**

$$\gamma_j = E[(Y_t - \mu_t)(Y_{t-j} - \mu_{t-j})] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (y_t - \mu_t)(y_{t-j} - \mu_{t-j})\, f(y_t, y_{t-1}, \ldots, y_{t-j}) \times dy_t\, dy_{t-1} \cdots dy_{t-j}$$

Stationarity

A time series Y_t is **covariance-stationary** (or **weakly stationary**) if the following hold:

1. $\forall t,\ \mu_t = \mu$
2. $\forall t,\ \gamma_{j,t} = \gamma_j$
3. $\gamma_0 < \infty$

Why is this useful?

Reduces # moments have to estimate

Can now estimate mean from 1 observation

Stationarity: Examples

Are each of the following covariance-stationary or not?

No.

Mean changes over time

(levels)

Yes

Means roughly constant

(changes)

Ergodicity

- A covariance-stationary process is **ergodic for the mean** if:

$$\operatorname*{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} y_t = \mu$$

Sufficient condition for ergodicity:

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty$$

- A covariance-stationary process is **ergodic for second moments** if:

$$\operatorname*{plim}_{T\to\infty} \frac{1}{T-j}\sum_{t=j+1}^{T} (y_t - \bar{y}(T))(y_{t-j} - \bar{y}(T)) = \gamma_j$$

Ergodicity over time

Why is this useful? Allows us to swap ensemble averages for time averages.

Need info. to accumulate over time.

Ergodicity: Examples

Are each of the following ergodic or not?

- Long-term inflation: inflation rates in an economy where the central ~~rate~~ Bank has a specific target in mind

Yes! Mean-reversion means things don't "blow up"

- Multiplicative gambling: individuals play a game where each round they win 50% of their wealth if the coin is heads and lose 40% of their wealth if the coin is tails

$w > 1$

No!

(1) $\mathbb{E}[\text{payoff}] = \frac{1}{2} \cdot 1.5 + \frac{1}{2} \cdot 0.6 = 1.05$

(2) $\text{Payoff}(T) = \prod_{t=1}^{T} r_t \, v(t) \to \bar{r}_T \cdot v \to \bar{r}_T \to \left(1.5^{1/2} \cdot 0.6^{1/2}\right) \approx 0.95$

White Noise

A basic building block of time series, white noise has the following properties:

- $E[\varepsilon_t] = 0$
- $E[\varepsilon_t^2] = Var(\varepsilon_t) = \sigma^2$
- $E[\varepsilon_t \varepsilon_s] = 0 \quad \forall t \neq s$

Moving Average (MA) Processes: MA(1)

Definition: $Y_t = \mu + \varepsilon_t + \theta \varepsilon_{t-1}, \quad \varepsilon_t \sim WN(0, \sigma^2)$

Mean: $E[Y_t] = \mu$

Variance: $E[(Y_t - \mu)^2] = (1+\theta^2) E[\varepsilon_t^2] = (1+\theta^2)\sigma^2$

***j*th Autocovariance:** $\gamma_j = Cov(Y_t, Y_{t-j}) = \theta\sigma^2 \cdot 1\{j \in \{-1, 1\}\}$

Lag Operator: L

Definition: Given Y_t, $LY_t = Y_{t-1}$

For higher orders: $L^j Y_t = Y_{t-j}$

Polynomials: $\psi(L) = \psi_0 + \psi_1 L + \psi_2 L^2 + \dots$

Rewrite MA(1): $Y_t = (1 + \theta L)\varepsilon_t + \mu$

$= \theta(L)\varepsilon_t + \mu$

Moving Average (MA) Processes: MA(q)

Definition: $Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} = \mu + \theta(L)\varepsilon_t$

Mean: μ

Variance: $(1 + \theta_1^2 + \cdots + \theta_q^2)\sigma^2 = \gamma_0$

***j*th Autocovariance:**

$$\gamma_j = \begin{cases} 0 & \text{if } j > q \\ (\theta_j + \theta_{j+1}\theta_1 + \cdots + \theta_q \theta_{q-j})\sigma^2 & \text{if } j = 1, 2, \ldots, q \end{cases}$$

Why is this useful?

A necessary assumption for covariance-stationarity:

$$\sum_{j=1}^{q} \theta_j^2 < \infty$$

Autoregressive (AR) Processes: AR(1)

Definition: $Y_t = c + \phi Y_{t-1} + \varepsilon_t$

2 strategies for finding MA(∞) representation:

1. Recursive substitution
2. Inverse of polynomial

Strategy #2: Inverse of Polynomial

Define $\Theta_q(L) = 1 + \theta_1 L + \dots + \theta_q L^q$

Inverse: $\psi(L) = \Theta_q^{-1}(L) \rightarrow \psi(L)\Theta_q(L) = 1$

Note: $Y_t = c + \phi L Y_t + \varepsilon_t \rightarrow (1 - \phi L) Y_t = c + \varepsilon_t$

Need: $(1 + \psi_1 L + \psi_2 L^2 + \dots)(1 - \phi L) = 1$

$\rightarrow 1 + (\psi_1 - \phi)L + (\psi_2 - \psi_1 \phi) L^2 + (\psi_3 - \psi_2 \phi) L^3 + \dots = 1$

Yields system of equations:

$\psi_1 - \phi = 0$

$\psi_2 - \psi_1 \phi = 0$

$\psi_3 - \psi_2 \phi = 0$

$\Big\} \; \psi_j = \phi^j$

$\rightarrow Y_t = \frac{c}{1-\phi} + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$

Companion Form: Overview

$$\xi_t = \begin{pmatrix} y_t \\ \vdots \\ y_{t-p+1} \end{pmatrix} \qquad F = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & & 0 & 0 \\ 0 & 1 & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & & 1 & 0 \end{pmatrix} \qquad \eta_t = \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\rightarrow \xi_t = F\xi_{t-1} + \eta_t \quad \rightarrow \text{our AR}(p) \text{ in top left corner}$$

$$\rightarrow \xi_t = \sum_{j=0}^{T} F^j \eta_{t-j} + F^{T+1}\xi_{t-T-1}$$

$\hookrightarrow$ gets ugly

Companion Form: Eigenvalue Decomposition

$$F = T \Lambda T^{-1} \rightarrow F^j = T \Lambda^j T^{-1}$$

As long as $\max_i |\lambda_i| < 1$, stationary

1. Announcements

2. Double Machine Learning

3. CATE Estimation

4. Introduction to Linear Stationary Processes

5. Practice Problems

Econ 271/Mgtecon 604 - Section 9

Amar Venugopal

Stanford University

Winter 2026

Outline

1. Announcements

2. Forecasting & Mean Dynamics

3. Wold Representation Theorem

4. VAR and DSGE

5. State Space Representation & Kalman Filter

6. Practice Problems

Announcements

- Final Exam: March 18, 3:30-6:30pm
- Last problem set: due March 13, 11:59pm
- First half review session: March 11, 4-5:30pm, GSB BC104

1. Announcements

2. Forecasting & Mean Dynamics

3. Wold Representation Theorem

4. VAR and DSGE

5. State Space Representation & Kalman Filter

6. Practice Problems

Conditional Mean Dynamics

Goal: Given unemployment so far, what will unemployment be tomorrow?

$$E[Y_{t+1} \mid Y_t, Y_{t-1}, \dots]$$

(Linear) Forecasting:

$$X_t = [Y_t, Y_{t-1}, \dots] \to Y^t$$

Forecast: $g(X_t) = \alpha' X_t$ → linear restriction

Implied loss: $E[(Y_{t+1} - g(X_t))^2]$

→ BLP $\alpha = E^{-1}[X_t' X_t] E[X_t' Y_{t+1}]$

Linear Projection: MA(∞)

$Y_t = \mu + \psi(L)\varepsilon_t$

Goal: Since all other models have MA(∞) representation, solving this solves many

$$X_t = \begin{pmatrix} \varepsilon_t \\ \varepsilon_{t-1} \\ \vdots \end{pmatrix}$$

$$E^{-1}\left[X_t' X_t\right] = \left(\sigma^2 I\right)^{-1}$$

$$E\left[X_t' Y_{t+s}\right] = E\left[X_t'\left(\mu + \varepsilon_{t+s} + \psi_1 \varepsilon_{t+s-1} + \cdots\right)\right] = \begin{pmatrix} \psi_s \sigma^2 \\ \psi_{s+1}\sigma^2 \\ \vdots \end{pmatrix}$$

$$P(Y_{t+s} \mid X_t) = \mu + X_t'\alpha = X_t'\left(\sigma^2 I\right)^{-1}\begin{pmatrix} \psi_s \sigma^2 \\ \psi_{s+1}\sigma^2 \\ \vdots \end{pmatrix} = \mu + \psi_s \varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \cdots$$

$E\left[\varepsilon_j \cdot \left(\mu + \varepsilon_{t+s} + \psi_1 \varepsilon_{t+s-1} + \cdots\right)\right]$

Linear Projection: MA(∞)

Goal:

$$X_t = \begin{pmatrix} \varepsilon_t \\ \varepsilon_{t-1} \\ \vdots \end{pmatrix}$$

$$Y_{t+s} = \mu + \varepsilon_{t+s} + \psi_1 \varepsilon_{t+s-1} + \cdots$$

$$X_t' Y_{t+s} = \left(\varepsilon_t \left(\mu + \sum_{j=0}^{\infty} \psi_j \varepsilon \right) \right)$$

Forecasting: MA(∞)

Convert $P(Y_{t+s} \mid \varepsilon_t, \varepsilon_{t-1}, \ldots)$ to $P(Y_{t+s} \mid Y_t, Y_{t-1}, \ldots)$

Forecast error: $Y_{t+s} - P(Y_{t+s} \mid X_t) = \varepsilon_{t+s} + \psi_1 \varepsilon_{t+s-1} + \ldots + \psi_{s-1} \varepsilon_{t+1}$

Define: $\frac{\psi(L)}{L^s} = L^{-s} + \psi_1 L^{-(s-1)} + \ldots$

$$\left(\frac{\psi(L)}{L^s}\right)\varepsilon_t = L^{-s}\varepsilon_t + \psi_1 L^{-(s-1)}\varepsilon_t + \ldots$$
$$= \varepsilon_{t+s} + \psi_1 \varepsilon_{t+s-1} + \ldots$$

$$P(Y_{t+s} \mid X_t) = \mu + \left[\frac{\psi(L)}{L^s}\right]_+ \varepsilon_t$$

(+) → sets negative powers of L to 0

From polynomial inversion:

$\eta(L)(Y_t - \mu) = \varepsilon_t$ (where $\eta(L) = \psi^{-1}(L)$) $\rightarrow P(Y_{t+s} \mid Y_t, \ldots) = \mu + \left[\frac{\psi(L)}{L}\right]_+ \eta(L)(Y_t - \mu)$

↳ linear projection for MA(∞)

1. Announcements

2. Forecasting & Mean Dynamics

3. Wold Representation Theorem

4. VAR and DSGE

5. State Space Representation & Kalman Filter

6. Practice Problems

Wold Representation Theorem

Theorem: Let $\{Y_t\}$ be a covariance-stationary process with $\mathbb{E}[Y_t] = 0$ and $\gamma_j = \mathbb{E}[Y_t Y_{t-j}]$. Then: $Y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + \eta_t$

Where:

- $\varepsilon_t \sim$ white noise
- $\psi_0 = 1$, $\sum_{j=0}^{\infty} \psi_j^2 < \infty$
- $\forall\, t, s \quad \mathbb{E}[\varepsilon_t \eta_s] = 0$
- η_{t+s} perfectly predictable given Y_t

Wold Decomposition

A consequence of this is that we can decompose any covariance-stationary process into 2 components:

$$Y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + \eta_t$$

1. $\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$: stochastic component, impossible to predict perfectly
2. η_t : deterministic, perfectly predictable from history (seasonality)

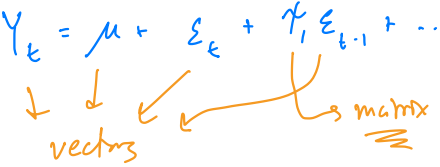
→ holds for <u>any</u> covariance-stationary process (not just linear ones)

Properties of Vector Processes

- **Vector Process:** $\{Y_t\} = [Y_{1,t}, Y_{2,t}, \cdots]$
- **Mean:** $\mu_t = E[Y_t]$ → independent of t if covariance-stationary
- **Variance:** $\Gamma_{0,t} = E[(Y_t - \mu_t)(Y_t - \mu_t)']$
- j^{th} **Order Autocovariance:** $\Gamma_{j,t} = E[(Y_t - \mu_t)(Y_{t-j} - \mu_{t-j})']$

Vector MA(∞)

Definition: $y_t = \mu + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \dots$



No restrictions ψ_t → can have nonzero coefficients on innovations from other sub-processes

When is it covariance stationary?

$$\forall i,j \quad , \quad \sum_{s=0}^{\infty} \left| \psi_s^{(i,j)} \right| < \infty$$

Vector AR(p)

Definition: $Y_t = c + \Phi_1 Y_{t-1} + \dots + \Phi_p Y_{t-p} + \varepsilon_t$

$\Phi_1, \dots, \Phi_p$: matrices

No restrictions on Φ_j

When is it covariance stationary?

$$F = \begin{pmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_{p-1} & \Phi_p \\ I_n & 0 & \cdots & 0 & 0 \\ 0 & I_n & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & I_n & 0 \end{pmatrix}$$

F: $(np) \times (np)$

→ all eigenvalues must lie in unit circle

Impulse Response Function

Goal: See how other variables respond to an unexpected shock in one variable.

Consider MA(∞) representation: $Y_t = \mu + \Psi(L)\,\varepsilon_t$

Can interpret $\Psi_s = \dfrac{\partial Y_{t+s}}{\partial \varepsilon_t^j}$ → how change in innovation to j at t impacts the whole process s periods later

Can we use this to compare across outcome variables?

$\varepsilon_t \sim N(0, \Omega)$

No! This holds all other innovations fixed

But $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ may be correlated

Orthogonalizing IRFs

If Ω diagonal, then ε's orthogonal

Cholesky Factorization: $\Omega = PP'$, P lower triangular

↳ implies hierarchy in innovations ε_t (causal flow)

→ order matters

Define: $u_t = P^{-1}\varepsilon_t$

$E[u_t u_t'] = P^{-1}PP'(P')^{-1} = I$ → orthogonalized

Can now write $MA(\infty)$ as: $y_t = \mu + Pu_t + \Psi_1 Pu_{t-1} + \dots$

→ here, IRFs orthogonalized

Does this really help with interpretation?

Not really ... u_t have no economic meaning (unless Cholesky ordering is true)

Big Idea: Many macro models can be represented as VAR using laws of motion for state variables as function of structural parameters

$\exists$ mapping : structural model → VAR (restricted)

Justifies 2 main uses of VAR in applied macro:

1. Reference models
2. Structural VARs

Structural VARs

Goal: Approximate "structure" by imposing restrictions on covariance matrix of VAR innovations

Identification challenge:

1. Announcements

2. Forecasting & Mean Dynamics

3. Wold Representation Theorem

4. VAR and DSGE

5. State Space Representation & Kalman Filter

6. Practice Problems

Gaussian State Space Models

Let $\mathbf{z}_t$ be the state vector and $\mathbf{y}_t$ be the vector of observations.

Gaussian state space model has 2 parts:

- **Law of motion for observables:**

$$y_t = \mu + A z_t + H \eta_t \quad , \quad \eta_t \sim N(0, I)$$

- **Law of motion for state:**

$$z_t = \Phi z_{t-1} + R \varepsilon_t \quad , \quad \varepsilon_t \sim N(0, \Sigma)$$

↳ Markov

Example: MA(1)

$z_t = (\epsilon_t, \epsilon_{t-1})$

$$y_t = \mu + \epsilon_t + \theta\epsilon_{t-1} \quad \epsilon_t \sim \mathcal{WN}(0, \sigma^2)$$

$y_t = \underline{\mu} + \underline{A} z_t + \underline{H}\eta_t$

$\mu = \mu$

$\Phi = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$

$A = (1 \quad \theta)$

$R = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$z_t = \underline{\Phi} z_{t-1} + \underline{R}\underline{\epsilon}_t$

(ϵ)

$\begin{pmatrix} \epsilon_t \\ \epsilon_{t-1} \end{pmatrix} = \Phi \begin{pmatrix} \epsilon_{t-1} \\ \epsilon_{t-2} \end{pmatrix} + R\epsilon_t$

$H = 0$

$\Sigma = \sigma^2$

$= \begin{pmatrix} 0 \\ \epsilon_{t-1} \end{pmatrix} + R\epsilon_t$

Big Idea: Extract state from observed time series via Bayesian updating

Consider the following simple example:

$$y_t = z_t + \sigma_y \eta_t, \quad \eta_t \sim N(0,1)$$

$$z_t = \phi z_{t-1} + \sigma_\varepsilon \varepsilon_t, \quad \varepsilon_t \sim N(0,1)$$

Belief over state: $p(z_1) \sim N(\hat{z}_{1|0}, P_{1|0})$

There are 3 steps to produce an update: Given state $\hat{z}_{t|t-1}, P_{t|t-1}$

1. Predict y_t ← law of motion for observable
2. Update beliefs $\hat{z}_{t|t}, P_{t|t}$
3. Update beliefs about $\hat{z}_{t+1|t}, P_{t+1|t}$ ← law of motion for state