Section: Module 4

OLS

Kentaro Nakamura

GOV 2003

February 20th, 2026

Today's agenda

- OLS
  - QR Decomposition / OLS Computation (PS1)
- Hypothesis testing
  - Basic Principle
  - Derivation of finite-sample test (T-test / F-test)
  - Asymptotic inference

- Due to time constraints, I put some materials I cannot cover as Appendix
  - OLS as Best Linear Approximation
  - OLS estimates without ith observation
  - FWL theorem and within transformation for fixed effect regression

Orthonormal Matrix

Definition (Orthonormal Matrix)

An $m \times m$ matrix $\boldsymbol{P} = (\boldsymbol{p}_1, \cdots, \boldsymbol{p}_m) = (\boldsymbol{q}_1, \cdots, \boldsymbol{q}_m)^\top$ is called an **orthonormal matrix** if

$$\boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}_m$$

- As is clear from the definition, $\boldsymbol{P}^{-1} = \boldsymbol{P}^\top$

Example

$$\boldsymbol{Q} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

is orthogonal matrix, satisfying $\boldsymbol{Q}^\top \boldsymbol{Q} = I$.

Instability of Matrix Inversion

- While we derived $\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$, this is NOT how `lm` function computes the regression.
  - This is because the computation of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ is numerically unsatble when columns of $\boldsymbol{X}$ are numerically colinear
  - To understand this, we need to understand **QR decomposition**

Theorem (QR Decomposition)

Let $\boldsymbol{X}$ be an $n \times p$ matrix with full column rank ($p \leq n$). Then there exist

- *an $n \times p$ matrix $\boldsymbol{Q}$ with orthonormal columns ($\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}_p$),*
- *a $p \times p$ upper-triangular matrix $\boldsymbol{R}$ with positive diagonal entries,*

such that

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}.$$

- NOTE: $\boldsymbol{Q} \in \mathbb{R}^{n \times p}$ is not a square matrix (unless $n = p$), and this satisfies $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}$ but in general $\boldsymbol{Q}\boldsymbol{Q}^\top \neq \boldsymbol{I}$

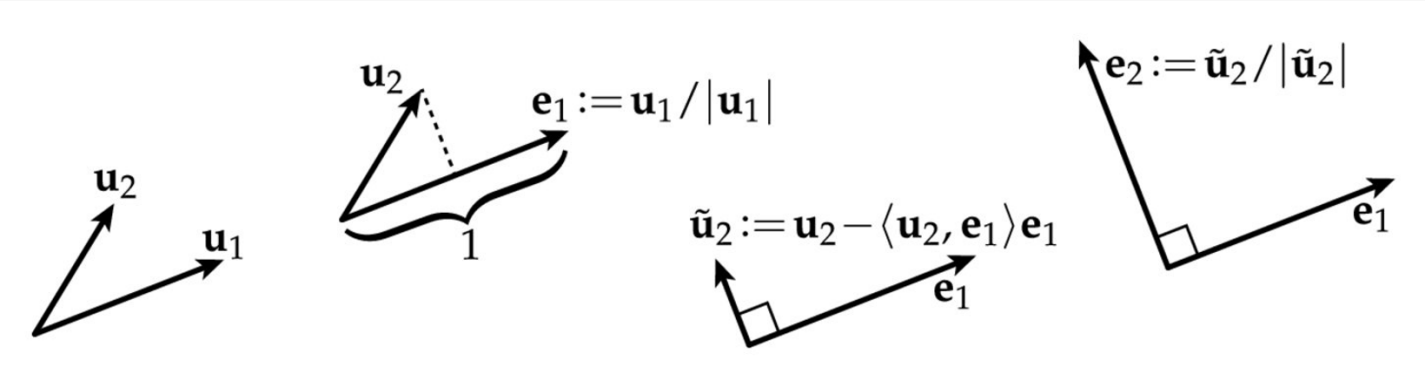
Gram–Schmidt Orthogonalization

Theorem (Gram–Schmidt)

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p]$ *with linearly independent columns. Define recursively:*

$$\boldsymbol{v}_1 = \boldsymbol{x}_1, \quad \boldsymbol{v}_k = \boldsymbol{x}_k - \sum_{j=1}^{k-1} \frac{\boldsymbol{v}_j^\top \boldsymbol{x}_k}{\boldsymbol{v}_j^\top \boldsymbol{v}_j} \boldsymbol{v}_j, \quad \boldsymbol{q}_k = \frac{\boldsymbol{v}_k}{\|\boldsymbol{v}_k\|}.$$

Then $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_p\}$ *is an orthonormal basis of* $S(\boldsymbol{X})$.



QR Decomposition and Gram Schmidt

- Start with the columns of $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p]$.
- **Gram–Schmidt step:** for each $k = 1, \ldots, p$,

$$\boldsymbol{v}_k = \boldsymbol{x}_k - \sum_{j=1}^{k-1} \underbrace{(\boldsymbol{q}_j^\top \boldsymbol{x}_k)}_{\text{scalar}} \boldsymbol{q}_j, \qquad \boldsymbol{q}_k = \frac{\boldsymbol{v}_k}{\|\boldsymbol{v}_k\|}.$$

- Rearrange the equation for $\boldsymbol{x}_k$:

$$\boldsymbol{x}_k = (\boldsymbol{q}_1^\top \boldsymbol{x}_k)\boldsymbol{q}_1 + (\boldsymbol{q}_2^\top \boldsymbol{x}_k)\boldsymbol{q}_2 + \cdots + (\boldsymbol{q}_k^\top \boldsymbol{x}_k)\boldsymbol{q}_k.$$

- Notice that each column satisfies $\boldsymbol{x}_k = \sum_{j=1}^{k} r_{jk}\boldsymbol{q}_j$, where $r_{jk} := \boldsymbol{q}_j^\top \boldsymbol{x}_k$ for $j \le k$ and $r_{jk} := 0$ for $j > k$.

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{q}_1 & \cdots & \boldsymbol{q}_p \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ 0 & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{pp} \end{bmatrix} = \boldsymbol{QR}.$$

Gram–Schmidt Orthogonalization (Example 1/3)

- Let

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = [\boldsymbol{x}_1, \boldsymbol{x}_2], \quad \boldsymbol{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{x}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}.$$

- **Step 1**:

$$\boldsymbol{v}_1 = \boldsymbol{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad \|\boldsymbol{v}_1\| = \sqrt{1^2 + 0^2 + 0^2} = 1$$

and thus

$$\boldsymbol{q}_1 = \frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

Gram–Schmidt Orthogonalization (Example 2/3)

- **Step 2**: compute the projection of $\boldsymbol{x}_2$ onto $\boldsymbol{q}_1$. The projection is

$$\operatorname{proj}_{\boldsymbol{q}_1}(\boldsymbol{x}_2) = (\boldsymbol{q}_1^\top \boldsymbol{x}_2)\boldsymbol{q}_1 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

- Subtract it:

$$\boldsymbol{v}_2 = \boldsymbol{x}_2 - \operatorname{proj}_{\boldsymbol{q}_1}(\boldsymbol{x}_2) = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

- Normalize:

$$\|\boldsymbol{v}_2\| = \sqrt{0^2 + 1^2 + 0^2} = 1, \qquad \boldsymbol{q}_2 = \frac{\boldsymbol{v}_2}{\|\boldsymbol{v}_2\|} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

Gram–Schmidt Orthogonalization (Example 3/3)

- Thus,

$$\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad \boldsymbol{Q}^\top \boldsymbol{Q} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \boldsymbol{I}_2.$$

- Compute $\boldsymbol{R} = \boldsymbol{Q}^\top \boldsymbol{X}$. Since

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

- Verify:

$$\boldsymbol{QR} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = \boldsymbol{X}.$$

OLS via QR (Why `lm()` does this)

- Recall the least squares problem:

$$\hat{\beta} = \arg\min_{\beta} \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2.$$

- Suppose $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}$ and $\boldsymbol{R}$ is upper triangular.

$$\begin{aligned}
\|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 &= \|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{R}\beta\|_2^2 \\
&= \|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{Y} + \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{R}\beta\|_2^2 \\
&= \|(\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^\top)\boldsymbol{Y} + \boldsymbol{Q}(\boldsymbol{Q}^\top \boldsymbol{Y} - \boldsymbol{R}\beta)\|_2^2 \\
&= \|(\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^\top)\boldsymbol{Y}\|_2^2 + \|\boldsymbol{Q}(\boldsymbol{Q}^\top \boldsymbol{Y} - \boldsymbol{R}\beta)\|_2^2 \\
&= \|(\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^\top)\boldsymbol{Y}\|_2^2 + \|\boldsymbol{Q}^\top \boldsymbol{Y} - \boldsymbol{R}\beta\|_2^2
\end{aligned}$$

  where the second equality is by adding and subtracting the same quantity, the fourth equality is because of orthogonality (i.e., $(\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^\top)\boldsymbol{Q} = 0$), and the fifth equality is because $\boldsymbol{Q}$ is orthonormal (i.e., $\boldsymbol{Q}^\top \boldsymbol{Q} = I$).

- The first term does not involve β, and since $\boldsymbol{R}$ is invertible (as $\boldsymbol{X}$ is full rank), the second component is minimized when $\boldsymbol{Q}^\top \boldsymbol{Y} = \boldsymbol{R}\beta$.

OLS via QR

- Suppose

$$\boldsymbol{R} = \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad \boldsymbol{z} = \boldsymbol{Q}^\top \boldsymbol{Y} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

- Then $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{z}$ means:

$$\begin{cases} r_{11}\beta_1 + r_{12}\beta_2 = z_1 \\ \qquad\qquad r_{22}\beta_2 = z_2 \end{cases}$$

- Start from the bottom equation:

$$r_{22}\beta_2 = z_2 \quad \Rightarrow \quad \beta_2 = \frac{z_2}{r_{22}}.$$

- Plug β_2 into the first equation:

$$r_{11}\beta_1 + r_{12}\beta_2 = z_1 \quad \Rightarrow \quad \beta_1 = \frac{z_1 - r_{12}\beta_2}{r_{11}}.$$

- This is called **backward substitution**
  - No matrix inversion. Just division and subtraction.

Numeric example of backward substitution

- Let $\boldsymbol{Y} = \begin{bmatrix} 3 & 2 & 0 \end{bmatrix}^\top$.
- Compute

$$\boldsymbol{z} = \boldsymbol{Q}^\top \boldsymbol{Y} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}.$$

- Now solve $\boldsymbol{R}\hat{\boldsymbol{\beta}} = \boldsymbol{z}$:

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}.$$

- Bottom equation: $\hat{\beta}_2 = 2$.
- Top equation: $\hat{\beta}_1 + \hat{\beta}_2 = 3 \Rightarrow \hat{\beta}_1 = 1$.

Hypothesis testing: Basic principle (1)

- **Procedure**
  - STEP 1: Assume the null hypothesis $H_0 : \theta \in \Theta_0$ ($\Leftrightarrow H_1 : \theta \in \Theta_1$, with $\Theta = \Theta_0 \cup \Theta_1$)
  - STEP 2: Calculate the distribution of test statsitics under the null (known as **reference distribution**)
  - STEP 3: If the observed statistics is too extreme compared to the reference distribution, we reject the null hypothesis
    - Two decisions: (i) reject null or (ii) accept null

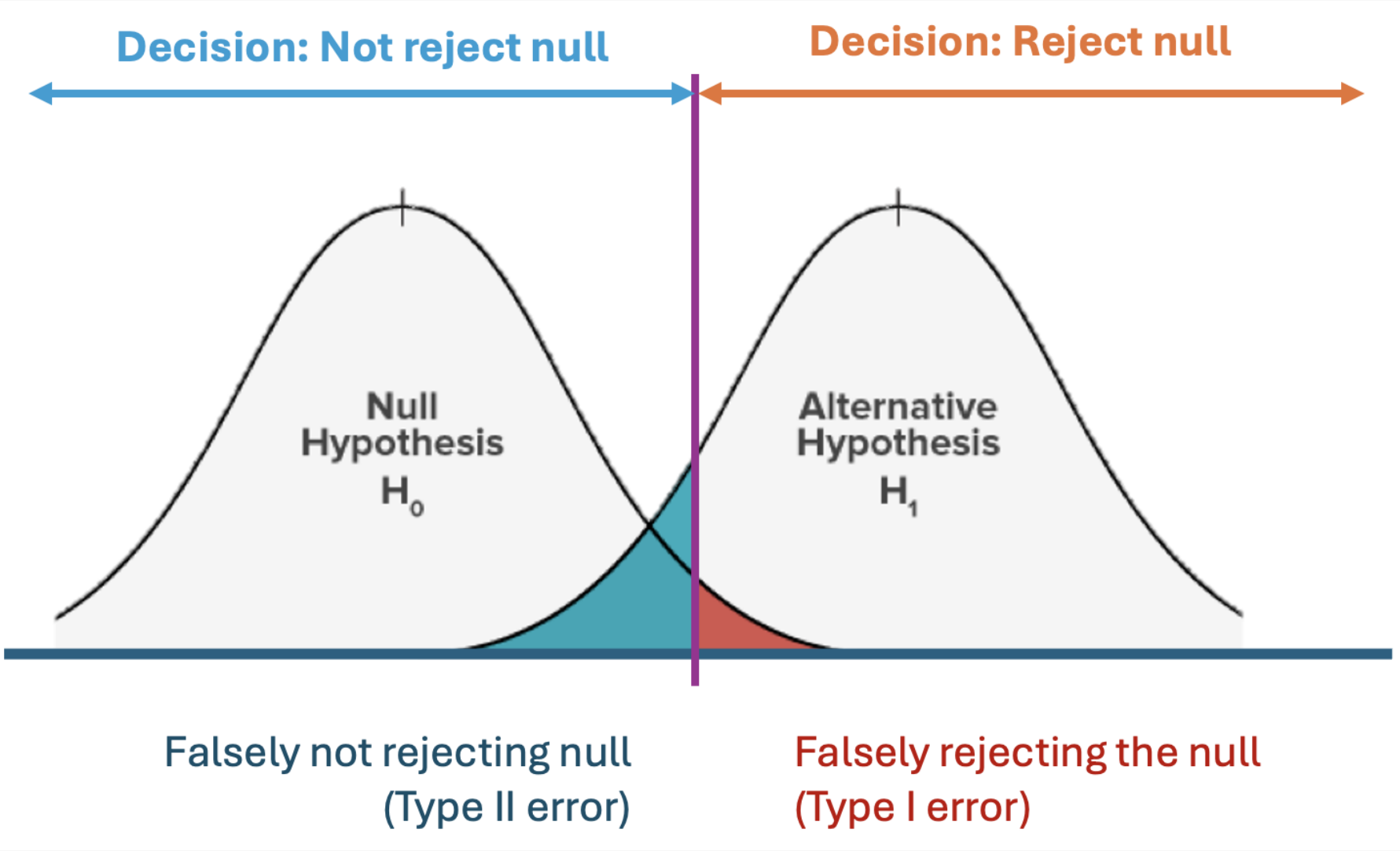
- Importantly, we need to know the distribution of test statistics
  - Thus, we typically either assume the assumption on our data or use asymptotic approach

Hypothesis testing: Basic principle (2)

	Accept H_0	**Reject** H_0
H_0 **true**	Correct Decision	Type I Error
H_1 **true**	Type II Error	Correct Decision

- The probability of Type I error is called the **size** of the test
- The rejection probability under the alternative hypothesis is called the **power** of the test
  - I.e., power equals 1 minus the probability of a Type II error
- There is a trade-off between size and power
  - Understand this trade-off from the next page's illustration

Illustration: Type I & Type II error



Normal Distribution

Definition (Normal distribution)

A random variable Z is **normal** with mean μ and variance σ^2, written $Z \sim \mathcal{N}(\mu, \sigma^2)$, if it has density

$$f(z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right).$$

- If $\boldsymbol{Z}$ is a vector, the distribution is called **multivariate normal**.
  - For example,

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right)$$

    where ρ is the correlation
- Important properties:
  - In the case of multivariate normal, no correlation is equivalent to independence (not true for other cases)
  - **reproductive property**: If $\boldsymbol{Z} \sim \mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\boldsymbol{AZ} + \boldsymbol{a} \sim \mathrm{MVN}(\boldsymbol{A\mu} + \boldsymbol{a}, \boldsymbol{A\Sigma A}^\top)$$

Chi-squared distribution

Definition (Chi-squared distribution)

If $Z_1, \ldots, Z_\nu \overset{iid}{\sim} \mathcal{N}(0,1)$, then

$$\sum_{j=1}^{\nu} Z_j^2 \sim \chi^2_\nu,$$

where ν is called the **degrees of freedom**.

t-distribution and F-distribution

Definition (Student t distribution)

Let $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi^2_\nu$ be independent. Then

$$T := \frac{Z}{\sqrt{U/\nu}} \sim t_\nu.$$

Definition (F distribution)

Let $U \sim \chi^2_{d_1}$ and $V \sim \chi^2_{d_2}$ be independent. Then

$$F := \frac{U/d_1}{V/d_2} \sim F_{d_1,d_2}.$$

- We will derive t-test and F-test statistic for OLS

Distribution of OLS Coefficient and Residual

- **Setup**: $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \mid \boldsymbol{X} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$
- Recall that
$$\hat{\beta} - \beta = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon}$$
- On the other hand, annhilitaor matrix $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$ is orthogonal to $\boldsymbol{X}$ (i.e., $\boldsymbol{M}\boldsymbol{X} = 0$), and thus
$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{Y} = \boldsymbol{M}(\boldsymbol{X}\beta + \boldsymbol{\epsilon}) = \boldsymbol{M}\boldsymbol{\epsilon}$$
- Thus,
$$\begin{bmatrix} \hat{\beta} - \beta \\ \hat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \\ \boldsymbol{M} \end{bmatrix} \boldsymbol{\epsilon}$$
- Therefore, (try to show this part by yourself!)
$$\begin{bmatrix} \hat{\beta} - \beta \\ \hat{\boldsymbol{\epsilon}} \end{bmatrix} \mid \boldsymbol{X} \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} & 0 \\ 0 & \sigma^2 \boldsymbol{M} \end{bmatrix}\right)$$

Extra: Derivation of OLS Variance Estimator

- So far, everything depends on the error variance $\sigma^2 = \mathbb{E}[\epsilon^2]$
  - We do not know the true error variance, so we need to propose the estimator $\hat{\sigma}^2$

- The natural estimator is to replace expectation with average; i.e.,

$$\hat{\sigma}^2 = \frac{1}{n}\hat{\boldsymbol{\epsilon}}^\top \hat{\boldsymbol{\epsilon}} = \frac{1}{n}(\boldsymbol{M}\boldsymbol{\epsilon})^\top \boldsymbol{M}\boldsymbol{\epsilon} = \frac{1}{n}\boldsymbol{\epsilon}^\top \boldsymbol{M}\boldsymbol{\epsilon}$$

- This is actually not unbiased.
  - To see that, use $\mathrm{tr}(\boldsymbol{\epsilon}^\top \boldsymbol{M}\boldsymbol{\epsilon}) = \mathrm{tr}(\boldsymbol{M}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top)$

$$\mathbb{E}[\hat{\sigma}^2 \mid \boldsymbol{X}] = \frac{1}{n}\mathrm{tr}(\mathbb{E}[\boldsymbol{M}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \mid \boldsymbol{X}]) = \frac{1}{n}\mathrm{tr}(\boldsymbol{M}\underbrace{\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \mid \boldsymbol{X}]}_{\sigma^2 I_n}) = \frac{n-p}{n}\sigma^2$$

  where the last line is by $\mathrm{tr}(\boldsymbol{M}) = n - p$

- Thus, the unbiased estimator of variance is given by

$$s^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \frac{1}{n-p}\boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}$$

Distribution of Variance Estimator

- Since $\boldsymbol{\epsilon} = \hat{\boldsymbol{\epsilon}} + \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$,

$$\begin{aligned}\boldsymbol{\epsilon}^\top \boldsymbol{\epsilon} &= \hat{\boldsymbol{\epsilon}}^\top \hat{\boldsymbol{\epsilon}} + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{X}^\top \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + 2(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{X}^\top \hat{\boldsymbol{\epsilon}} \\ &= \hat{\boldsymbol{\epsilon}}^\top \hat{\boldsymbol{\epsilon}} + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{X}^\top \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\end{aligned}$$

- $\frac{\boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}}{\sigma^2} \sim \chi^2_n$ because $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

- $\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{X}^\top \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} \sim \chi^2_p$ because $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \mid \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$

- Therefore,

$$\frac{\hat{\boldsymbol{\epsilon}}^\top \hat{\boldsymbol{\epsilon}}}{\sigma^2} = \frac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p}$$

t-test: Derivation

- **Setup**:
  - $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \mid X \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$
- **t-statistic** is

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}]_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}]_{jj}}} \Big/ \sqrt{\frac{(n-p)s^2}{\sigma^2}/(n-p)}$$

$$\sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-p}/(n-p)}} \sim t_{n-p}$$

- **Note**: Even though we start with conditional distribution $\boldsymbol{\epsilon} \mid \boldsymbol{X} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$, the resulting t-distribution does not depend on $\boldsymbol{X}$ (and also σ^2)
  - This means that the reference distribution does not depend on any other features of the data, except the degrees of freedom $n - p$
  - Such test statistic is called **pivotal**, meaning that it does not depend on unknowns.

F-test: Derivation

- **Setup**:
  - $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \mid X \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$
  - We are testing joint restriction $H_0 : \boldsymbol{A}\beta = \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{q \times p}$ with $\operatorname{rank}(\boldsymbol{A}) = q$ and $\beta \in \mathbb{R}^p$.

- As $\hat{\beta}$ is normal, under H_0,

$$\boldsymbol{A}\hat{\beta} - \boldsymbol{b} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{A}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{A}^\top)$$

  thus by normalization,

$$\frac{(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})^\top \{\boldsymbol{A}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{A}^\top\}^{-1} (\boldsymbol{A}\hat{\beta} - \boldsymbol{b})}{\sigma^2} \sim \chi^2_q$$

- You can easily derive denominator and see why the test statistic follows F-distribution using definition (try by yourself!)

Asymptotic Inference

- However, the assumption that error is distributed according to normal distribution is a strong assumption
  - Both t-test and F-test depends on this assumption
  - We want to relax it

- Instead, we want to resort to **asymptotic statistics**
  - **Idea**: Assume we have infinite amount of data (i.e., $N \to \infty$)
  - Use tools in asymptotic statistics to derive the distribution at the limiting (called **limiting distribution**)

Tools in Asymptotic Statistics

- **Law of Large Numbers (LLN)**: If $X_1, \cdots, X_n$ are i.i.d.,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}[X_i]$$

- **Central Limit Theorem (CLT)**: If $X_1, \cdots, X_n$ are i.i.d.,

$$\sqrt{n}(\bar{X} - \mathbb{E}[X_i]) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}[X_i])$$

- **Slutsky's Lemma**: If $X_n \xrightarrow{d} X$ for some random variable X and $Y_n \xrightarrow{p} c$ for some constant c,

$$X_n Y_n \xrightarrow{d} cX$$

Model-based asymptotic inference (1)

- Notice that by law of large numbers, the first term converges in probability to

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^\top \xrightarrow{p} \mathbb{E}[\mathbf{X}^\top\mathbf{X}]$$

- On the other hand, because each $\mathbf{X}_i\epsilon_i$ are i.i.d., by central limit theorem, the second term converges in distribution to

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i = \sqrt{n}\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i}_{\text{Form of Avg!}} \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[(\mathbf{X}^\top\boldsymbol{\epsilon}(\mathbf{X}^\top\boldsymbol{\epsilon})^\top)])$$

- Note that mean of normal distribution here is 0 because $\mathbb{E}[\mathbf{X}_i\epsilon_i] = 0$

Model-based asymptotic inference (2)

- Therefore,

$$\sqrt{n}(\hat{\beta} - \beta) = \sqrt{n}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{\epsilon}$$

$$\rightarrow \mathcal{N}\left(0, \underbrace{\mathbb{E}[\mathbf{X}^\top \mathbf{X}]^{-1}\mathbb{E}[\mathbf{X}^\top \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \mathbf{X}]\mathbb{E}[\mathbf{X}^\top \mathbf{X}]^{-1}}_{\text{Asymptotic Variance}}\right)$$

- **Homoskedasticity**: $\mathbb{V}[\epsilon_i \mid \mathbf{X}_i] = \sigma^2$
  - This gives $\mathbb{V}[\epsilon_i \mid \mathbf{X}_i] = \mathbb{E}[\epsilon_i^2 \mid \mathbf{X}_i] - \underbrace{\mathbb{E}[\epsilon_i \mid \mathbf{X}_i]^2}_{=0} = \mathbb{E}[\epsilon_i^2 \mid \mathbf{X}_i]$
  - NOTE: Homoskedasticity is usually not plausible

$$\mathbb{E}[\mathbf{X}^\top \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \mathbf{X}] = \mathbb{E}[\mathbb{E}[\mathbf{X}^\top \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \mathbf{X} \mid \mathbf{X}]] \quad (\because \text{Law of Iterated Expectation})$$

$$= \mathbb{E}[\mathbf{X}^\top \underbrace{\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \mid \mathbf{X}]}_{=\sigma^2 I_n} \mathbf{X}] = \sigma^2 \mathbb{E}[\mathbf{X}^\top \mathbf{X}]$$

and thus the asymptotic variance is simplified to

$$\mathbb{E}[\mathbf{X}^\top \mathbf{X}]^{-1}\mathbb{E}[\mathbf{X}^\top \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \mathbf{X}]\mathbb{E}[\mathbf{X}^\top \mathbf{X}]^{-1} = \sigma^2 \mathbb{E}[\mathbf{X}^\top \mathbf{X}]^{-1}$$

Wald Test Statistic

- **Null hypothesis**: $H_0 : \boldsymbol{A}\beta = \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{q \times p}$ is full row rank
  - Any linear restriction can be written as $a_1\beta_1 + \cdots + a_p\beta_p = b$
  - Can test any number of linearly independent restrictions (joint test)

- From asymptotic normality, assume

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma})$$

  Then under H_0,

$$\sqrt{n}(\boldsymbol{A}\hat{\beta} - \boldsymbol{b}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top).$$

- **Wald statistic** (quadratic form):

$$\begin{aligned} W &:= (\boldsymbol{A}\hat{\beta} - \boldsymbol{b})^\top \left\{ \boldsymbol{A}\widehat{\mathbb{V}}(\hat{\beta})\boldsymbol{A}^\top \right\}^{-1} (\boldsymbol{A}\hat{\beta} - \boldsymbol{b}) \\ &= \left[\sqrt{n}(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})\right]^\top \left\{ \boldsymbol{A}\widehat{\boldsymbol{\Sigma}}\boldsymbol{A}^\top \right\}^{-1} \left[\sqrt{n}(\boldsymbol{A}\hat{\beta} - \boldsymbol{b})\right], \qquad \widehat{\boldsymbol{\Sigma}} := n\,\widehat{\mathbb{V}}(\hat{\beta}). \end{aligned}$$

- Under H_0, $W \xrightarrow{d} \chi^2_{\mathrm{rank}(\boldsymbol{A})}$.
- If W is larger than the critical value, reject H_0.

Appendix: OLS as Best Linear Approximation

- Linear regression is a **best linear approximation** of conditional expectation function $m(\mathbf{X}_i) = \mathbb{E}[Y_i \mid \mathbf{X}_i]$
- To see this, consider minimizing the mean squared loss between $m(\mathbf{X}_i)$ and $\mathbf{X}_i^\top \beta$

$$d(\beta) = \mathbb{E}[(m(\mathbf{X}_i) - \mathbf{X}_i^\top \beta)^2]$$

- Notice that this minimization problem is just replacing Y_i with $m(\mathbf{X}_i)$ for OLS. Thus, the best linear approximation $\tilde{\beta}$ is given by

$$\begin{aligned}
\tilde{\beta} &= \left(\mathbb{E}[\mathbf{X}_i \mathbf{X}_i^\top]\right)^{-1} \mathbb{E}[\mathbf{X}_i m(\mathbf{X}_i)] \\
&= \left(\mathbb{E}[\mathbf{X}_i \mathbf{X}_i^\top]\right)^{-1} \mathbb{E}[\mathbf{X}_i \mathbb{E}[Y_i \mid \mathbf{X}_i]] \\
&= \left(\mathbb{E}[\mathbf{X}_i \mathbf{X}_i^\top]\right)^{-1} \mathbb{E}[\mathbb{E}[\mathbf{X}_i Y_i \mid \mathbf{X}_i]] \\
&= \left(\mathbb{E}[\mathbf{X}_i \mathbf{X}_i^\top]\right)^{-1} \mathbb{E}[\mathbf{X}_i Y_i] = \beta
\end{aligned}$$

- OLS at the population (covered last week) is equal to the best linear approximation of $m(\mathbf{X}_i) = \mathbb{E}[Y_i \mid \mathbf{X}_i]$

Appendix: Sherman–Morrison Formula

- We derive the OLS estimate without the ith observation
- To do this, it is useful to know the following theorem.

Theorem (Sherman–Morrison formula Formula)

Suppose $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ *be an invertible square matrix and* $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ *be vectors. If* $\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^\top$ *is invertible, then*

$$(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^\top)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top\boldsymbol{A}^{-1}}{1 + \boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}}$$

- See next page for the proof.

Appendix: Sherman–Morrison Formula: Proof

$$\left(\boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top\boldsymbol{A}^{-1}}{1+\boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}}\right)(\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}^\top)$$

$$= \boldsymbol{A}^{-1}\boldsymbol{A} + \boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{A}}{1+\boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top}{1+\boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}}$$

$$= \boldsymbol{I} + \boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top}{1+\boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\left(\boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}\right)\boldsymbol{v}^\top}{1+\boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}}$$

$$= \boldsymbol{I} + \boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\left(1+\boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}\right)\boldsymbol{v}^\top}{1+\boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}}$$

$$= \boldsymbol{I} + \boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top - \boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top$$

$$= \boldsymbol{I}.$$

Appendix: OLS estimates without ith observation (1)

- Now, notice that $\boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)}$ ($\boldsymbol{X}^\top \boldsymbol{X}$ removing ith observations) is written as

$$\boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)} = \boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{x}_{(i)} \boldsymbol{x}_{(i)}^\top$$

- As a result, the OLS estimates without ith observation is given by

$$\hat{\beta}_{(i)} = (\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{x}_{(i)} \boldsymbol{x}_{(i)}^\top)^{-1} (\boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{x}_{(i)} \boldsymbol{y}_{(i)}^\top)$$

- Using Sherman-Morrison formula, we obtain

$$(\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{x}_{(i)} \boldsymbol{x}_{(i)}^\top)^{-1} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)} \boldsymbol{x}_{(i)}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}}{1 - \boldsymbol{x}_{(i)} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)}^\top}$$

Appendix: OLS estimates without ith observation (2)

- By multiplying $\boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{x}_{(i)} \boldsymbol{y}_{(i)}^\top$, we obtain

$$
\begin{aligned}
\hat{\beta}_{(i)} &= \left(\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{x}_{(i)} \boldsymbol{x}_{(i)}^\top \right)^{-1} \left(\boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{x}_{(i)} y_{(i)} \right) \\
&= \left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)} \boldsymbol{x}_{(i)}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}}{1 - \boldsymbol{x}_{(i)}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)}} \right] \\
&\qquad \times \left(\boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{x}_{(i)} y_{(i)} \right) \\
&= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)} y_{(i)} \\
&\qquad + \frac{(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)} \boldsymbol{x}_{(i)}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}}{1 - \boldsymbol{x}_{(i)}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)}} \\
&\qquad - \frac{(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)} \boldsymbol{x}_{(i)}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)} y_{(i)}}{1 - \boldsymbol{x}_{(i)}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)}}
\end{aligned}
$$

Appendix: OLS estimates without ith observation (3)

- Thus,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{(i)} &= \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)} y_{(i)} \\
&\qquad + \frac{(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)} \boldsymbol{x}_{(i)}^\top \hat{\boldsymbol{\beta}}}{1 - \ell_i} - \frac{(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)} \ell_i y_{(i)}}{1 - \ell_i} \\
&= \hat{\boldsymbol{\beta}} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)} \left[\frac{\boldsymbol{x}_{(i)}^\top \hat{\boldsymbol{\beta}} - \ell_i y_{(i)}}{1 - \ell_i} - y_{(i)} \right] \\
&= \hat{\boldsymbol{\beta}} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)} \left[\frac{\boldsymbol{x}_{(i)}^\top \hat{\boldsymbol{\beta}} - y_{(i)}}{1 - \ell_i} \right] \\
&= \hat{\boldsymbol{\beta}} - \frac{(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)} \hat{\varepsilon}_{(i)}}{1 - \ell_i},
\end{aligned}$$

where $\ell_i \equiv \boldsymbol{x}_{(i)}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_{(i)}$ and $\hat{\varepsilon}_{(i)} \equiv y_{(i)} - \boldsymbol{x}_{(i)}^\top \hat{\boldsymbol{\beta}}$.

Appendix: Fixed effect (1)

- Consider the fixed effect model:

$$\boldsymbol{Y} = \boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

  where $\boldsymbol{D}$ is a $n \times G$ matrix for dummy variable for G groups

- One approach to estimate β in this context is to regress the regression model above
  - This approach is called the **least squares dummy variable (LSDV)**
- However, this is not scalable when G is large
  - Especially problematic when applied to panel data, where you want to use both time and unit fixed effects
- In such case, we can apply **within-transformation** approach to estimate it
  - This is actually well-explained from FWL theorem

Appendix: Fixed effect (2)

- Consider the partitioned regressor $(\boldsymbol{D}, \boldsymbol{X})$. Then, the regression coefficient β is obtained by

$$\begin{aligned}\hat{\beta} &= (\boldsymbol{X}^\top \{\boldsymbol{I} - \boldsymbol{P_D}\}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \{\boldsymbol{I} - \boldsymbol{P_D}\}\boldsymbol{Y} \\ &= (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{Y}}\end{aligned}$$

where $\boldsymbol{P_D} = \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1}\boldsymbol{D}^\top$, $\tilde{\boldsymbol{X}} = \{\boldsymbol{I} - \boldsymbol{P_D}\}\boldsymbol{X}$, and $\tilde{\boldsymbol{Y}} = \{\boldsymbol{I} - \boldsymbol{P_D}\}\boldsymbol{Y}$

- Now, notice that

$$\tilde{\boldsymbol{Y}} = \{\boldsymbol{I} - \boldsymbol{P_D}\}\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{P_D}\boldsymbol{Y}$$

  and you see that $\boldsymbol{P_D}\boldsymbol{Y}$ is the regression of $\boldsymbol{Y}$ on $\boldsymbol{D}$, which gives us the average of $\boldsymbol{Y}$ on each group

- Now, $\tilde{\boldsymbol{Y}}$ is the demaned outcome (demaned by group averages)
  - You can also think $\tilde{\boldsymbol{X}}$ is also the demeaned regressor
  - This is basically the within-transformation