Review Section (Midterm)

OLS / MLE

Kentaro Nakamura

GOV 2003

March 23th, 2026

Today's Agenda: Review

- OLS
  - Projection Matrix
  - Annihilator Matrix
  - FWL Theorem
  - Variance of OLS
  - Asymptotics of OLS

- MLE
  - Review of Theories
  - Logistic regression
  - Bradley-Terry models

OLS and Orthogonal Projection

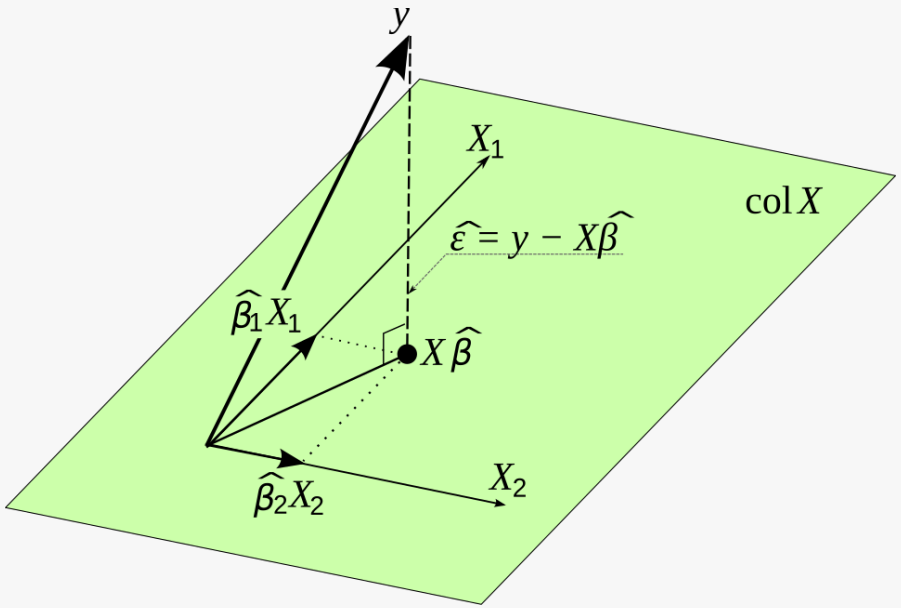
- Consider the linear model $\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon$. Recall that OLS estimator of β is given by

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

- As $\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon$,

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta} = \underbrace{\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top}_{:=\boldsymbol{P}_X} \boldsymbol{Y}$$

- Now, notice that $\boldsymbol{P}_X^2 = \boldsymbol{P}_X$ and $\boldsymbol{P}^\top = \boldsymbol{P}$ (Prove this!), and thus $\boldsymbol{P}_X$ is an orthogonal projection matrix



Annihilator Matrix

Definition (Annihilator Matrix)

$\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is called **annihilator matrix**.

- Notice that

$$\boldsymbol{M}\boldsymbol{X} = (\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = 0$$

- Annihilator matrix gives you residual since

$$\boldsymbol{M}\boldsymbol{Y} = (\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$$

  - Annihilator matrix is a projection onto orthogomal complement of the column space of $\boldsymbol{X}$

- We can also show that annihilator matrix is symmetric ($\boldsymbol{M}^\top = \boldsymbol{M}$) and idempotent $\boldsymbol{M}^2 = \boldsymbol{M}$
  - That is, annihilator matrix is a projection onto orthogonal complement of the column space

Residual and Annihilator Matrix

- Residual is a projection of $\boldsymbol{Y}$ onto $S^{\perp}(\boldsymbol{X})$ (i.e., orthogonal complement of the column space)
- Orthogonality 1:

$$\hat{\boldsymbol{Y}}^{\top}\hat{\epsilon} = (\boldsymbol{P}_X\boldsymbol{Y})^{\top}\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{Y}^{\top}\boldsymbol{P}_X^{\top}(\boldsymbol{I}-\boldsymbol{P}_X)\boldsymbol{Y} = 0$$

- Orthogonality 2:

$$\boldsymbol{X}_k^{\top}\hat{\epsilon} = (\boldsymbol{P}_X\boldsymbol{X}_k)^{\top}\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{X}_k^{\top}\boldsymbol{P}_X^{\top}(\boldsymbol{I}-\boldsymbol{P}_X)\boldsymbol{Y} = 0$$

  - $\boldsymbol{P}_X\boldsymbol{X}_k = \boldsymbol{X}_k$ because
    $\boldsymbol{P}_X\boldsymbol{X}_k = \boldsymbol{P}_X\boldsymbol{X}\boldsymbol{e}_k = \boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{e}_k = \boldsymbol{X}\boldsymbol{e}_k = \boldsymbol{X}_k$
- Orthogonality 3: $\boldsymbol{x}^{\top}\hat{\epsilon} = 0$ for any $\boldsymbol{x} \in S(\boldsymbol{X})$
  - Proof: Recall the definition of column space: $\boldsymbol{x} = \boldsymbol{X}a$ for some a.

$$\boldsymbol{x}^{\top}\hat{\epsilon} = (\boldsymbol{X}a)^{\top}\hat{\epsilon} = a^{\top}\boldsymbol{X}^{\top}\boldsymbol{M}\boldsymbol{Y} = a^{\top}\boldsymbol{X}^{\top}(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}) = 0$$

- Zero mean: $\sum_{i=1}^{n}\hat{\epsilon}_i = 0$:
  - As $\boldsymbol{X}$ includes intercept, there exists a such that $\boldsymbol{1} = \boldsymbol{X}a$. As a result, by the previous orthogonality 3, we have zero mean.

Unbiasedness of OLS

Theorem

Unbiasedness of OLS Estimator Assume the linearity: $\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon$ *with exogeneity* $\mathbb{E}[\epsilon \mid \boldsymbol{X}] = 0$. *Then,*

$$\mathbb{E}[\hat{\beta}] = \beta$$

- **Proof**

$$\begin{aligned}
\mathbb{E}[\hat{\beta}] &= \mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}] \\
&= \mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top (\boldsymbol{X}\beta + \epsilon)] \\
&= \mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\beta + (\boldsymbol{X}^\top \boldsymbol{X})^\top \boldsymbol{X}^\top \epsilon] \\
&= \beta + \mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \epsilon] \\
&= \beta + \mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \mathbb{E}[\epsilon \mid \boldsymbol{X}]] = \beta
\end{aligned}$$

Variance of OLS

- Now,
$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \epsilon$$
and thus
$$\mathbb{V}[\hat{\beta} \mid \mathbf{X}] = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{V}[\epsilon \mid \mathbf{X}] \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}$$

- Under homoskedasticity, $\mathbb{V}[\epsilon \mid \mathbf{X}] = \sigma^2 I$.

- Thus,
$$\mathbb{V}[\hat{\beta} \mid \mathbf{X}] = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

Variance estimator

- 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 biased.
  - 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

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

Partitioned Matrix

- Sometimes, partitioned matrix can help us formulate the proof / computation
  - Let $\boldsymbol{A}_{11} \in \mathbb{R}^{m \times m}$, $\boldsymbol{A}_{12} \in \mathbb{R}^{m \times n}$, $A_{21} \in \mathbb{R}^{n \times m}$, and $\boldsymbol{A}_{22} \in \mathbb{R}^{n \times n}$.
    Then, you can write matrix as $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}$
  - You can calculate the addition and multiplication with this partitioned matrix in the usual way

Theorem (Inverse of Partitioned Matrix)

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{A}_{11}^{-1} + \boldsymbol{B}_{12}\boldsymbol{B}_{22}^{-1}\boldsymbol{B}_{21} & -\boldsymbol{B}_{12}\boldsymbol{B}_{22}^{-1} \\ -\boldsymbol{B}_{22}^{-1}\boldsymbol{B}_{21} & \boldsymbol{B}_{22}^{-1} \end{bmatrix}$$

where $\boldsymbol{B}_{12} := \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$, $\boldsymbol{B}_{21} := \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}$, *and* $\boldsymbol{B}_{22} := \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$

- We can later use this to prove partitioned regression in a different way from the last time

Partitioned Regression / FWL Theorem

- In this proof, I derive FWL theorem using partitioned matrix.

Theorem (Partitioned Regression)

Consider the partition $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ *and* $\beta = (\beta_1, \beta_2)$*, and the regression model*

$$\boldsymbol{Y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \epsilon.$$

The least squred estimator $(\hat{\beta}_1, \hat{\beta}_2)$ *is written as*

$$\hat{\beta}_1 = (\boldsymbol{X}_1^\top \boldsymbol{M}_2 \boldsymbol{X}_1)^{-1}(\boldsymbol{X}_1^\top \boldsymbol{M}_2 \boldsymbol{Y})$$
$$\hat{\beta}_2 = (\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^{-1}(\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{Y})$$

where

$$\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{X}_1(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top$$
$$\boldsymbol{M}_2 = \boldsymbol{I}_n - \boldsymbol{X}_2(\boldsymbol{X}_2^\top \boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top$$

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,

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

Asymptotics of OLS: Consistency

- Now, notice that

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

- By law of large numbers,

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

- Similarly,

$$\frac{1}{n} \mathbf{X}^\top \boldsymbol{\epsilon} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \epsilon_i \xrightarrow{p} \mathbb{E}[\mathbf{X}_i^\top \epsilon_i] = 0$$

- Thus, by slutsky,

$$\hat{\beta} - \beta = \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} \right)^{-1} \frac{1}{n} \mathbf{X}^\top \boldsymbol{\epsilon} \xrightarrow{p} \mathbb{E}[\mathbf{X}_i \mathbf{X}_i^\top] \cdot 0 = 0$$

  which means that $\hat{\beta}$ is consistent.
  - If we use CLT for the second term, then we can derive the asymptotic normality.

MLE: Review of Basic Principle

- **Setting**: Observe data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ from a distribution with density (or pmf) $f_\theta(\boldsymbol{y})$, where $\theta \in \Theta$ is an unknown parameter.

- **Likelihood function**: Given observed data $\boldsymbol{y}$,

$$L(\theta) = L(\theta \mid \boldsymbol{y}) = f_\theta(\boldsymbol{y}) = \prod_{i=1}^{n} f_\theta(y_i) \quad \text{(under i.i.d. assumption)}$$

- **Likelihood principle**: The likelihood function contains all the information about the unknown parameter of interest.

- **MLE**: The maximum likelihood estimator is

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} L(\theta)$$

Score Function

Definition (Score Function)

The score function $S_n(\boldsymbol{Y}, \theta)$ (or simply $S_n(\theta)$) for the data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is defined as

$$S_n(\boldsymbol{Y}, \theta) = \frac{\partial \log f_\theta(\boldsymbol{Y})}{\partial \theta} = \frac{\partial \ell(\theta)}{\partial \theta}$$

where $\ell(\theta)$ is the log-likelihood function.

- **Property**: The expected score is zero. That is, $\mathbb{E}_\theta\left[S_n(\boldsymbol{Y}, \theta)\right] = 0$
  - $\mathbb{E}_\theta[S_n(\boldsymbol{Y}, \theta)] = \int \frac{f'_\theta(\boldsymbol{y})}{f_\theta(\boldsymbol{y})} f_\theta(\boldsymbol{y})\, d\boldsymbol{y} = \frac{\partial}{\partial \theta} \int f_\theta(\boldsymbol{y})\, d\boldsymbol{y} = \frac{\partial}{\partial \theta} 1 = 0$

Fisher Information

Definition (Fisher Information)

The Fisher information for data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is defined as

$$I(\theta) := \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \log f_\theta(\boldsymbol{Y})\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log f_\theta(\boldsymbol{Y})\right]$$

- Fisher information gives us the asymptotic variance under the correctly specified models

Theorem (Fisher's Theorem / MLE Asymptotic Normality)

Let $Y_1, \cdots, Y_n \sim f_{\theta_0}$. Under the regularity conditions, the MLE estimator $\hat{\theta}_n$ achieves the asymptotic normality

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$$

where $I(\theta_0)$ is the fisher information.

Example: Logistic Regression (1)

- The likelihood function is

$$L(\beta) = \prod_{i=1}^{n} \pi(\boldsymbol{X}_i)^{Y_i} (1 - \pi(\boldsymbol{X}_i))^{1-Y_i}$$

  and thus the logistic regression is

$$\ell(\beta) = \sum_{i=1}^{n} \Big(Y_i \log \pi(\boldsymbol{X}_i) + (1 - Y_i) \log(1 - \pi(\boldsymbol{X}_i)) \Big)$$

- The score function is

$$\frac{\partial}{\partial \beta} \ell(\beta) = \sum_{i=1}^{n} \boldsymbol{X}_i (Y_i - \pi(\boldsymbol{X}_i))$$

Example: Logistic Regression (1)

- The Hessian is

$$\frac{\partial^2}{\partial \beta^2} \ell(\beta) = -\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top \pi(\boldsymbol{X}_i)(1 - \pi(\boldsymbol{X}_i))$$

- By information equality,

$$I_n(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \beta^2} \ell(\beta)\right] = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top \pi(\boldsymbol{X}_i)(1 - \pi(\boldsymbol{X}_i))$$

- Note that this assumes fixed design (i.e., $\boldsymbol{X}_i$ is non-random.)

Example 2: Bradley Terry Model (1)

- The model is

$$\Pr(j \text{ beats } j') = \frac{e^{\beta_j}}{e^{\beta_j} + e^{\beta_{j'}}},$$

- The likelihood is thus given by

$$L(\beta) = \prod_{1 \le j < j' \le J} \left(\frac{e^{\beta_j}}{e^{\beta_j} + e^{\beta_{j'}}} \right)^{W_{jj'}} \left(\frac{e^{\beta_{j'}}}{e^{\beta_j} + e^{\beta_{j'}}} \right)^{W_{j'j}}$$

- And thus the log-likelihood is

$$\ell(\beta) = \sum_{1 \le j < j' \le J} \Big[W_{jj'}\beta_j + W_{j'j}\beta_{j'} - (W_{jj'} + W_{j'j}) \log(\theta_j + \theta_{j'}) \Big].$$

where $\theta_j := e^{\beta_j}$

Example 2: Bradley-Terry Model (2)

- The score is

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{j' \neq j} \left[W_{jj'} - (W_{jj'} + W_{j'j}) \frac{\theta_j}{\theta_j + \theta_{j'}} \right].$$

- The Hessian is, for $j \neq k$,

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = (W_{jk} + W_{kj}) \frac{\theta_j \theta_k}{(\theta_j + \theta_k)^2},$$

and for $j = k$,

$$\frac{\partial^2 \ell}{\partial \beta_j^2} = -\sum_{j' \neq j} (W_{jj'} + W_{j'j}) \frac{\theta_j \theta_{j'}}{(\theta_j + \theta_{j'})^2}.$$

Example 2: Bradley-Terry Model (3)

- By information equality,

$$I_n(\beta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\beta\partial\beta^\top}\ell(\beta)\right].$$

- Thus, for $j \neq k$,

$$I_n(\beta)_{jk} = -(W_{jk} + W_{kj})\frac{\theta_j\theta_k}{(\theta_j+\theta_k)^2},$$

and

$$I_n(\beta)_{jj} = \sum_{j'\neq j}(W_{jj'} + W_{j'j})\frac{\theta_j\theta_{j'}}{(\theta_j+\theta_{j'})^2}.$$