Section: Module 3

OLS

Kentaro Nakamura

GOV 2003

February 13th, 2026

Today's agenda

- Today, we will cover
  - Review of Linear Algebra
    - Matrix derivative
  - OLS
    - Difference between Estimand and Estimator
    - FWL Theorem
    - Gauss Markov Theorem

- Logistics
  - Problem set 1 is due on next Wednesday

- Recommended Textbook:
  - Hansen (2022) Econometrics, Ch2-7, Ch9

Matrix Derivative / Gradients (1)

Definition (The derivative of matrix by a scalar)

Let $A = (a_{ij})_{ij}$ be an $m \times n$ matrix and let each a_{ij} be a function of x. Then, we define the derivative of A with respect to the variable x as follows:

$$\frac{\partial A(x)}{\partial x} = \left(\frac{\partial a_{ij}(x)}{\partial x}\right)_{ij}$$

Example (The derivative of matrix by a scalar)

Let $A(x) = \begin{bmatrix} x^2 + 2x & x^3 - 1 \\ x^3 - x^2 + 1 & x^2 + x \end{bmatrix}$. Then,

$$\frac{\partial A(x)}{\partial x} = \begin{bmatrix} 2x + 2 & 3x^2 \\ 3x^2 - 2x & 2x + 1 \end{bmatrix}$$

Matrix Derivative / Gradients (2)

Definition (The derivative of function by a matrix)

Let $f(X)$ be a function of an $m \times n$ matrix X ($f : \mathbb{R}^{m \times n} \mapsto \mathbb{R}$). Then, the derivative of $f(X)$ with respect to the matrix X is

$$\frac{\partial f(X)}{\partial X} = \left(\frac{\partial f(X)}{\partial x_{ij}} \right)_{ij}$$

Example (The derivative of function by a matrix)

Let $X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{bmatrix}$ and $f(X) = x_{11}^2 x_{23} + 2x_{12}x_{13} - 5x_{21}x_{22}^2$.

Then,

$$\frac{\partial f(X)}{\partial X} = \begin{bmatrix} 2x_{11}x_{23} & 2x_{13} & 2x_{12} \\ -5x_{22}^2 & -10x_{21}x_{22} & x_{11}^2 \end{bmatrix}$$

Matrix Derivative / Gradients (3)

Definition (The derivative of function by a vector)

Let $f(x)$ be a function of vector $x \in \mathbb{R}^m$ ($f : \mathbb{R}^m \mapsto \mathbb{R}$). Then, we define the first and second derivatives of $f(x)$ as follows:

$$\frac{\partial f(x)}{\partial x} = \left[\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \cdots, \frac{\partial f(x)}{\partial x_m}\right]^\top$$

$$\frac{\partial f(x)}{\partial x \partial x^\top} = \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right)_{ij}$$

Example (The derivative of function by a vector)

Let $x = [x_1, x_2]^\top$ and consider $f(x) = x_1 x_2^3 + x_1^2 x_2^4$. Then,

$$\frac{\partial f(x)}{\partial x} = \begin{bmatrix} x_2^3 + 2x_1 x_2^4 \\ 3x_1 x_2^2 + 4x_1^2 x_2^3 \end{bmatrix}$$

$$\frac{\partial f(x)}{\partial x \partial x^\top} = \begin{bmatrix} 2x_2^4 & 3x_2^2 + 8x_1 x_2^3 \\ 3x_2^2 + 8x_1 x_2^3 & 6x_1 x_2 + 12x_1^2 x_2^2 \end{bmatrix}$$

Some useful formulas

- **Rule 1**: If a is not a function of x,

$$\frac{\partial a^\top x}{\partial x} = a^\top$$

- **Rule 2**: If A and b are not functions of x,

$$\frac{\partial b^\top A x}{\partial x} = A^\top b$$

- **Rule 3**: If A is not a function of x,

$$\frac{\partial x^\top A x}{\partial x} = (A + A^\top)^\top x$$

  - If A is symmetric, $A + A^\top = 2A$

OLS at the population: Estimand

- Consider $\arg\min_\beta \mathbb{E}[(Y_i - \boldsymbol{X}_i^\top \beta)^2]$
- Then, the objective function is written as

$$\mathbb{E}[(Y_i - \boldsymbol{X}^\top \beta)^2] = \mathbb{E}[Y_i^2] - 2 \cdot \beta^\top \mathbb{E}[\boldsymbol{X}_i Y_i] + \beta^\top \mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^\top]\beta$$

- By taking the derivative with respect to β, we get

$$\begin{aligned} &- 2 \cdot \mathbb{E}[\boldsymbol{X}_i Y_i] + 2\mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^\top]\beta = 0 \\ &\Rightarrow \beta = (\mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^\top])^{-1}\mathbb{E}[\boldsymbol{X}_i Y_i] \end{aligned}$$

- BUT we cannot calculate this since we cannot calculate expectation.
  - The quantity of interest, which we cannot calculate from data, is called **estimand**

OLS at the sample: Estimator

- Recall that OLS solves the following optimization problem:

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

- Notice that

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

  and thus

$$\nabla_\beta ||\boldsymbol{Y} - \boldsymbol{X}\beta||_2^2 = -2\boldsymbol{X}^\top \boldsymbol{Y} + 2\boldsymbol{X}^\top \boldsymbol{X}\beta$$

- Therefore, $-\boldsymbol{X}^\top \boldsymbol{Y} + \boldsymbol{X}^\top \boldsymbol{X}\beta = 0$, which gives us

$$\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right)\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{Y}\right)$$

- This is calculatable from data $\Rightarrow$ called **estimator**

FWL theorem (1)

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

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

- Consider the OLS estimator of $\beta = (\beta_1, \beta_2)$, which is defined by

$$(\hat{\beta}_1, \hat{\beta}_2) = \arg\min_{\beta_1, \beta_2} \|\boldsymbol{Y} - \boldsymbol{X}_1\beta_1 - \boldsymbol{X}_2\beta_2\|_2^2$$

  Now, $\hat{\beta}_1$ is obtained by nested minimization

$$\hat{\beta}_1 = \arg\min_{\beta_1}\left(\min_{\beta_2}\|\boldsymbol{Y} - \boldsymbol{X}_1\beta_1 - \boldsymbol{X}_2\beta_2\|_2^2\right)$$

- Now, let's focus on the inner minimization problem. This is simply the OLS of $\boldsymbol{Y} - \boldsymbol{X}_1\beta_1$ on $\boldsymbol{X}_2$, which has the solution

$$\hat{\beta}_2 = (\boldsymbol{X}_2^\top \boldsymbol{X}_2)^{-1}(\boldsymbol{X}_2^\top(\boldsymbol{Y} - \boldsymbol{X}_1\beta_1))$$

- The residual $\boldsymbol{Y} - \boldsymbol{X}_1\beta_1 - \boldsymbol{X}_2\hat{\beta}_2$ is written as

$$\underbrace{(\boldsymbol{I}_n - \boldsymbol{X}_2(\boldsymbol{X}_2^\top \boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top)}_{\boldsymbol{M}_2}(\boldsymbol{Y} - \boldsymbol{X}_1\beta_1)$$

FWL theorem (2)

- Thus,

$$\begin{aligned}
&\arg\min_{\beta_1}\left(\min_{\beta_2}\|\boldsymbol{Y}-\boldsymbol{X}_1\beta_1-\boldsymbol{X}_2\beta_2\|_2^2\right)\\
&=\arg\min_{\beta_1}\left((\boldsymbol{M}_2(\boldsymbol{Y}-\boldsymbol{X}_1\beta_1))^\top(\boldsymbol{M}_2(\boldsymbol{Y}-\boldsymbol{X}_1\beta_1))\right)\\
&=\arg\min_{\beta_1}(\boldsymbol{Y}-\boldsymbol{X}_1\beta_1)^\top\boldsymbol{M}_2(\boldsymbol{Y}-\boldsymbol{X}_1\beta_1)
\end{aligned}$$

- Notice that this is the form of weighted least squares (WLS), which solution is written as

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

- This is equivalently expressed as

$$\begin{aligned}
\hat{\beta}_1&=(\boldsymbol{X}_1^\top\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}(\boldsymbol{X}_1^\top\boldsymbol{M}_2\boldsymbol{Y})\\
&=(\boldsymbol{X}_1^\top\boldsymbol{M}_2\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}(\boldsymbol{X}_1^\top\boldsymbol{M}_2\boldsymbol{M}_2\boldsymbol{Y}) \quad (\because \boldsymbol{M}_2^2=\boldsymbol{M}_2)\\
&=(\tilde{\boldsymbol{X}}_1^\top\tilde{\boldsymbol{X}}_1)^{-1}(\tilde{\boldsymbol{X}}_1^\top\tilde{\boldsymbol{e}})
\end{aligned}$$

where $\tilde{\boldsymbol{X}}_1=\boldsymbol{M}_2\boldsymbol{X}_1$ and $\tilde{\boldsymbol{e}}=\boldsymbol{M}_2\boldsymbol{Y}$.

Unbiasedness of OLS Estimator

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**

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

Gauss-Markov Theorem

Theorem (Gauss-Markov Theorem)

Suppose that (i) $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$ *(linearity), (ii)* $\mathbb{E}[\boldsymbol{\epsilon} \mid \boldsymbol{X}] = 0$ *(exogeneity), and (iii)* $\mathbb{V}[\boldsymbol{\epsilon} \mid \boldsymbol{X}] = \sigma^2 I_n$ *(homoskedasticity). Then, OLS estimator* $\hat{\beta}$ *is Best Linear Unbiased Estimator in the sense that*

$$\mathbb{V}[\tilde{\beta}] \geq \mathbb{V}[\hat{\beta}]$$

for any $\tilde{\beta}$ *satisfying (i)* $\tilde{\beta} = \boldsymbol{A}(\boldsymbol{X})\boldsymbol{Y}$ *(linear) and (ii)* $\mathbb{E}[\tilde{\beta} \mid \boldsymbol{X}] = \beta$ *(unbiasedness).*

- But homoskedasticity is rarely true!

Gauss-Markov Theorem: Proof (1)

- Now, $\tilde{\beta}$ satisfies $\tilde{\beta} = \boldsymbol{A}(\boldsymbol{X})\boldsymbol{Y}$. So,

$$\begin{aligned}
\tilde{\beta} &= \boldsymbol{A}(\boldsymbol{X})\boldsymbol{Y} = \boldsymbol{A}(\boldsymbol{X})\boldsymbol{Y} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} + \underbrace{(\boldsymbol{A}(\boldsymbol{X}) - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)}_{:=\boldsymbol{B}(\boldsymbol{X})}\boldsymbol{Y} \\
&= \{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{B}(\boldsymbol{X})\}\boldsymbol{Y} \\
&= \{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{B}(\boldsymbol{X})\}(\boldsymbol{X}\beta + \epsilon) \\
&= \beta + \{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{B}(\boldsymbol{X})\}\epsilon + \boldsymbol{B}(\boldsymbol{X})\boldsymbol{X}\beta
\end{aligned}$$

- Now, we assume unbiasedness, which means $\mathbb{E}[\tilde{\beta} \mid \boldsymbol{X}] = \mathbb{E}[\hat{\beta} \mid \boldsymbol{X}] = \beta$. Thus, for any β,

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

- As a result, we have

$$\tilde{\beta} = \beta + \{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{B}(\boldsymbol{X})\}\epsilon$$

Gauss-Markov Theorem: Proof (2)

- Then, recall that for a vector $\boldsymbol{X} \in \mathbb{R}^k$, the covariance matrix $\mathbb{V}[\boldsymbol{X}]$ is defined as

$$\mathbb{V}[\boldsymbol{X}] = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^\top]$$

  where the (i, j)-th entry is covariance $\mathrm{Cov}(X_i, X_j)$.

- Therefore,

$$\begin{aligned}
\mathbb{V}[\tilde{\beta} \mid \boldsymbol{X}] &= \mathbb{V}[\beta + \{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{B}(\boldsymbol{X})\}\epsilon \mid \boldsymbol{X}] \\
&= \{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{B}(\boldsymbol{X})\}\mathbb{V}[\epsilon \mid \boldsymbol{X}]\{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{B}(\boldsymbol{X})\}^\top \\
&= \sigma^2\{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{B}(\boldsymbol{X})\}\{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{B}(\boldsymbol{X})\}^\top \\
&= \sigma^2\{(\boldsymbol{X}^\top \boldsymbol{X})^{-1} + \boldsymbol{B}(\boldsymbol{X})\boldsymbol{B}(\boldsymbol{X})^\top\}^\top \\
&\geq \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \mathbb{V}[\hat{\beta} \mid \boldsymbol{X}]
\end{aligned}$$

- As we consider the variance (which is the diagonal element of covariance matrix), this shows the claim of Gauss-Markov Theorem.
  - This is because the diagonal element of semi-positive definite matrix is non-negative (see Problem Set 0)

Appendix: 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

Appendix: 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

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

Appendix: Proof: Partitioned Regression / FWL Theorem (1)

- Now, recall that $\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$.
- Using partitioned matrix notation,

$$\boldsymbol{X}^\top \boldsymbol{Y} = \begin{bmatrix} \boldsymbol{X}_1^\top \boldsymbol{Y} \\ \boldsymbol{X}_2^\top \boldsymbol{Y} \end{bmatrix}$$

- On the other hand,

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1^\top \\ \boldsymbol{X}_2^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1^\top \boldsymbol{X}_1 & \boldsymbol{X}_1^\top \boldsymbol{X}_2 \\ \boldsymbol{X}_2^\top \boldsymbol{X}_1 & \boldsymbol{X}_2^\top \boldsymbol{X}_2 \end{bmatrix}$$

Appendix: Proof: Partitioned Regression / FWL Theorem (2)

- Now, applying the result of inverse for the partitioned matrix, we get

$$\left(\begin{bmatrix} \boldsymbol{X}_1^\top \boldsymbol{X}_1 & \boldsymbol{X}_1^\top \boldsymbol{X}_2 \\ \boldsymbol{X}_2^\top \boldsymbol{X}_1 & \boldsymbol{X}_2^\top \boldsymbol{X}_2 \end{bmatrix}\right)^{-1} = \begin{bmatrix} \boldsymbol{X}_{11}^{-1} + \boldsymbol{Z}_{12}\boldsymbol{Z}_{22}^{-1}\boldsymbol{Z}_{21} & -\boldsymbol{Z}_{12}\boldsymbol{Z}_{22}^{-1} \\ -\boldsymbol{Z}_{22}^{-1}\boldsymbol{Z}_{21} & \boldsymbol{Z}_{22}^{-1} \end{bmatrix}$$

where $\boldsymbol{X}_{ij} = \boldsymbol{X}_i^\top \boldsymbol{X}_j$, and

$$\begin{aligned} \boldsymbol{Z}_{12} &= \boldsymbol{X}_{11}^{-1}\boldsymbol{X}_{12} \\ \boldsymbol{Z}_{21} &= \boldsymbol{X}_{21}\boldsymbol{X}_{11}^{-1} \\ \boldsymbol{Z}_{22} &= \boldsymbol{X}_{22} - \boldsymbol{X}_{21}\boldsymbol{X}_{11}^{-1}\boldsymbol{X}_{12} \end{aligned}$$

Appendix: Proof: Partitioned Regression / FWL Theorem (3)

- As a result,

$$\begin{aligned}\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix} &= \begin{bmatrix}\boldsymbol{X}_{11}^{-1}+\boldsymbol{Z}_{12}\boldsymbol{Z}_{22}^{-1}\boldsymbol{Z}_{21} & -\boldsymbol{Z}_{12}\boldsymbol{Z}_{22}^{-1}\\ -\boldsymbol{Z}_{22}^{-1}\boldsymbol{Z}_{21} & \boldsymbol{Z}_{22}^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}_1^\top\boldsymbol{Y}\\ \boldsymbol{X}_2^\top\boldsymbol{Y}\end{bmatrix}\\ &= \begin{bmatrix}\left(\boldsymbol{X}_{11}^{-1}+\boldsymbol{Z}_{12}\boldsymbol{Z}_{22}^{-1}\boldsymbol{Z}_{21}\right)\boldsymbol{X}_1^\top\boldsymbol{Y}-\boldsymbol{Z}_{12}\boldsymbol{Z}_{22}^{-1}\boldsymbol{X}_2^\top\boldsymbol{Y}\\ -\boldsymbol{Z}_{22}^{-1}\boldsymbol{Z}_{21}\boldsymbol{X}_1^\top\boldsymbol{Y}+\boldsymbol{Z}_{22}^{-1}\boldsymbol{X}_2^\top\boldsymbol{Y}\end{bmatrix}.\end{aligned}$$

- Therefore,

$$\hat{\beta}_2 = \boldsymbol{Z}_{22}^{-1}\left(\boldsymbol{X}_2^\top\boldsymbol{Y}-\boldsymbol{Z}_{21}\boldsymbol{X}_1^\top\boldsymbol{Y}\right) = \boldsymbol{Z}_{22}^{-1}\left(\boldsymbol{X}_2^\top\boldsymbol{Y}-\boldsymbol{X}_{21}\boldsymbol{X}_{11}^{-1}\boldsymbol{X}_1^\top\boldsymbol{Y}\right)$$

Appendix: Proof: Partitioned Regression / FWL Theorem (4)

- Recall

$$\begin{aligned}\boldsymbol{Z}_{22} &= \boldsymbol{X}_{22} - \boldsymbol{X}_{21}\boldsymbol{X}_{11}^{-1}\boldsymbol{X}_{12} = \boldsymbol{X}_2^\top \boldsymbol{X}_2 - \boldsymbol{X}_2^\top \boldsymbol{X}_1 (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2 \\ &= \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2.\end{aligned}$$

- Also,

$$\boldsymbol{X}_2^\top \boldsymbol{Y} - \boldsymbol{X}_{21}\boldsymbol{X}_{11}^{-1}\boldsymbol{X}_1^\top \boldsymbol{Y} = \boldsymbol{X}_2^\top \left(\boldsymbol{I}_n - \boldsymbol{X}_1 (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top\right) \boldsymbol{Y} = \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{Y}.$$

- Hence,

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

- By symmetry (swap the roles of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$),

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