Financial Econometrics

Multivariate Regression

Tim C.C. Hung 洪志清

March 7^{th}, 2022

Multivariate Regression

- There are a few things we have not covered in univariate regression.
  - The general form of OLS.
  - Prove of BLUE and BUE of OLS under Gauss-Markov.
  - Why $\frac{1}{n-2}\sum_{i=1}^{n}\widehat{u_i}^2$ is an unbiased estimator for σ_u^2
  - How to deal with heteroskedasticity when we do not have A5.
  - Why sample mean is also a least square estimator.
  - Relationship between multiple independent (explanatory) variables.
  - Hypothesis testing of multiple coefficient estimates (and the entire model).
- We will first review for some matrix properties and show you the general multivariate regression derivations.

- A matrix is an array of numbers. It is usually denoted by an upper-case alphabet in boldface (e.g. $\mathbf{A}$), and its $(i,j)^{th}$ element (the element at the i^{th} row and j^{th} column) is denoted by the corresponding lower-case alphabet with subscripts ij (e.g., a_{ij}).

- The following is an example of a $m \times n$ matrix:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Review of Matrix Property

- An $n \times 1$ $(1 \times n)$ matrix is an n-dimensional column (row) vector.
- A matrix is square if its number of rows equals the number of columns.
- A matrix is said to be diagonal if its off-diagonal elements (i.e., a_{ij}, for $i \neq j$) are all zeros and at least one of its diagonal elements is non-zero, i.e., $a_{ii} \neq 0$ for some i.
- A diagonal matrix whose diagonal elements are all ones is an identity matrix, denoted as $\mathbf{I}$. We also write the $n \times n$ identity matrix as $\mathbf{I}_n$.

Matrix Operation

- Two matrices $\mathbf{A}_{mn}$ and $\mathbf{B}_{mn}$ are said to be the same if 1) they have the same number of rows (m) and same number of columns (n); and 2) $a_{ij} = b_{ij}$ for all i and j.

- **Matrix Addition**: defined only for two matrices of the same size (same m and n). $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$, $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$

- **Transpose**: the transpose of $\mathbf{A}$, denoted as $\mathbf{A}'$, is a matrix whose $(i, j)^{th}$ element is the $(j, i)^{th}$ element of $\mathbf{A}$. For $\mathbf{A}_{mn}$, its transpose $\mathbf{A}'$, has n rows and m columns, i.e. $\mathbf{A}'$ is a $n \times m$ matrix.

Matrix Operation

- **Scalar Multiplication**: $c\mathbf{A}_{mn}$ changes all elements in the $\mathbf{A}_{mn}$ matrix from a_{ij} to $c * a_{ij}$ for all (i, j).

- **Matrix Multiplication**: $\mathbf{AB}$ is only defined for matrix $\mathbf{A}$ and $\mathbf{B}$ when the number of columns of $\mathbf{A}$ is the same as the number of rows of $\mathbf{B}$.

- Therefore, $\mathbf{AB} \neq \mathbf{BA}$. $(\mathbf{BA}$ may not even be well defined.)

- Specifically, when $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times p$, their product, $\mathbf{C} = \mathbf{AB}$, is a $m \times p$ matrix whose $(i, j)^{th}$ element is

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

Matrix Operation

- **Matrix Multiplication**: *Rules*
  - Associative: $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
  - Distributive: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
  - $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$
  - For a $m \times n$ matrix $\mathbf{A}$, $\mathbf{I}_m\mathbf{A} = \mathbf{A}\mathbf{I}_n = \mathbf{A}$
  - A squared matrix $\mathbf{A}$ is idempotent if $\mathbf{AA} = \mathbf{A}$.

- **Determinant**:
  Given a square matrix $\mathbf{A}_n$, let $\mathbf{A}_{ij}$ denote the sub-matrix obtained from $\mathbf{A}$ by deleting its i^{th} row and j^{th} column. The determinant of $\mathbf{A}$ is:

$$det(\mathbf{A}) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \; det(\mathbf{A}_{ij})$$

  for any $j = 1, 2, \cdots, n$

- Example: $det(\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}) = (-1)^2 * 1 * 4 + (-1)^3 * 2 * 3 = 1 * 4 - 2 * 3 = -2$

- **Determinant**: *Rules*
  - The determinant of a scalar is itself.
  - A square matrix with non-zero determinant is said to be *nonsingular*; otherwise, it is *singular*.
  - $det(\mathbf{A}) = det(\mathbf{A}')$
  - $det(c\mathbf{A}) = c^n det(\mathbf{A})$
  - $det(\mathbf{AB}) = det(\mathbf{BA}) = det(\mathbf{A}) * det(\mathbf{B})$
  - $det(\mathbf{I}) = 1$ for all size of $\mathbf{I}$

- **Trace**:
  The *trace* of a square matrix is the sum of its diagonal elements:

$$trace(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$$

- *Rules*:
  - $trace(\mathbf{A}) = trace(\mathbf{A}')$
  - $trace(c\mathbf{A} + d\mathbf{B}) = c\ trace(\mathbf{A}) + d\ trace(\mathbf{B})$
  - $trace(\mathbf{I}_n) = n$
  - $trace(\mathbf{AB}) = trace(\mathbf{BA})$ ***Important Lemma!

Matrix Operation

- **Inverse**: a nonsingular matrix $\mathbf{A}$ possesses a unique inverse $\mathbf{A}^{-1}$ in the sense that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$

- Given a invertible $\mathbf{A}$, its inverse

$$\mathbf{A}^{-1} = \frac{1}{det(\mathbf{A}^{-1})}\mathbf{F}'$$

where $\mathbf{F}$ is the matrix of cofactors, i.e., the $(i,j)^{th}$ element of $\mathbf{F}$ is the cofactor: $(-1)^{i+j}det(\mathbf{A}_{ij})$. The matrix $\mathbf{F}'$ is known as the *adjoint* of $\mathbf{A}$.

- Example: for a 2×2 matrix $\mathbf{A}$, $\mathbf{A}^{-1} = \frac{1}{a_{11}a_{22}-a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$

- **Inverse**: *Rules*
  - Matrix inversion and transposition can be interchanged, i.e. $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$
  - For nonsingular $\mathbf{A}$ and $\mathbf{B}$, $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
  - For a diagonal matrix $\mathbf{A}$,
    $\mathbf{A}^{-1}$ is also diagonal with the diagonal elements a_{ii}^{-1}.
  - $\mathbf{I}^{-1} = \mathbf{I}$ for all size of $\mathbf{I}$

- **Linear Dependence**:
  the vectors $\mathbf{z}_1, \cdots, \mathbf{z}_n$ are said to be linearly independent if the only solution to $c_1\mathbf{z}_1 + c_2\mathbf{z}_2 + \cdots + c_n\mathbf{z}_n = 0$ is the trivial solution: $c_1 = \cdots = c_n = 0$. Otherwise, *they are linearly dependent.*

- **Rank**:
  the *column (row) rank* of a matrix $\mathbf{A}$ is the maximum number of linearly independent *column (row) vectors* of $\mathbf{A}$.

- Assume for $\mathbf{A}_{n\times k}$, $n < k$, and that $\mathbf{A}$ has $r < n$ linearly independent rows. Row vectors can be written as $\mathbf{a}_i = q_{i1}\mathbf{a}_1 + q_{i2}\mathbf{a}_2 + \cdots + q_{ir}\mathbf{a}_r$, with the j^{th} element, $a_{ij} = q_{i1}a_{1j} + q_{i2}a_{2j} + \cdots + q_{ir}a_{rj}$.
  We can then see that the column rank is also r!

- **Lemma**: the column rank and row rank of a matrix are equal.

Matrix Operation

- **Rank**: *Rules*
  - $rank(\mathbf{A}) = rank(\mathbf{A}')$
  - for two $n \times k$ matrices $\mathbf{A}$ and $\mathbf{B}$, $rank(\mathbf{A}+\mathbf{B}) \leq rank(\mathbf{A}) + rank(\mathbf{B})$
  - for $\mathbf{A}_{n\times k}$ and $\mathbf{B}_{k\times m}$,
    $rank(\mathbf{A}) + rank(\mathbf{B}) - k \leq rank(\mathbf{AB}) \leq min[rank(\mathbf{A}), rank(\mathbf{B})]$
  - for a nonsingular matrix $\mathbf{A}$,
    $rank(\mathbf{AB}) \leq rank(\mathbf{B}) = rank(\mathbf{A}^{-1}\mathbf{AB}) \leq rank(\mathbf{AB})$
    $\Rightarrow rank(\mathbf{AB}) = rank(\mathbf{B})$
    similarly, $rank(\mathbf{BC}) = rank(\mathbf{C}'\mathbf{B}') = rank(\mathbf{B}') = rank(\mathbf{B})$
- **Lemma**: let $\mathbf{A}_{n\times n}$ and $\mathbf{C}_{k\times k}$ be nonsingular matrices.
  then for any $n \times k$ matrix $\mathbf{B}$, $rank(\mathbf{B}) = rank(\mathbf{AB}) = rank(\mathbf{BC})$

- **Matrix Derivative**: scalar-valued function
  Say, we have a function $y = f(\mathbf{x}) = f(x_1, x_2, \cdots, x_n)$

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \partial y/\partial x_1 \\ \partial y/\partial x_2 \\ \vdots \\ \partial y/\partial x_n \end{bmatrix}$$

- A second derivatives matrix or Hessian is computed as:

$$\begin{bmatrix} \partial^2 y/\partial x_1 \partial x_1 & \partial^2 y/\partial x_1 \partial x_2 & \cdots & \partial^2 y/\partial x_1 \partial x_n \\ \partial^2 y/\partial x_2 \partial x_1 & \partial^2 y/\partial x_2 \partial x_2 & \cdots & \partial^2 y/\partial x_2 \partial x_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2 y/\partial x_n \partial x_1 & \partial^2 y/\partial x_n \partial x_2 & \cdots & \partial^2 y/\partial x_n \partial x_n \end{bmatrix}$$

- **Matrix Derivative**:
  Now, suppose we have a linear combination $y = \mathbf{a}'\mathbf{x} = \mathbf{x}'\mathbf{a} = \sum_i a_i x_i$

- We can easily see that

$$\frac{\partial y}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

- Similarly, in a set of linear function $\mathbf{Y} = \mathbf{A}\mathbf{x}$, $\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}'$

- **Matrix Derivative**:
  Lastly, for a quadratic form with a symmetric $\mathbf{A}$,
  $\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_i \sum_j x_i x_j a_{ij}$,

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$$

- For example, $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix}$, $\mathbf{x}'\mathbf{A}\mathbf{x} = x_1^2 + 4x_2^2 + 6x_1x_2$

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \begin{bmatrix} 2x_1 + 6x_2 \\ 6x_1 + 8x_2 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2\mathbf{A}\mathbf{x}$$

- If $\mathbf{A}$ is not symmetric, then $\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}')\mathbf{x}$

Matrix - Skipped Concepts

- Things you don't need to worry about in this course:
  - Kronecker product
  - Orthogonalization
  - Eigen Value and Eigen Vector

Multivariate Regression

National Taiwan University

- Suppose we have k regressors $(X_1, X_2, \cdots, X_k)$ and n observations, the regression function / Data Generating Process (DGP) is as the following:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

  for $i = 1, 2, \cdots, n$

- Let's group things into matrices!

$$\mathbf{X}_{n\times(k+1)} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix} \quad \boldsymbol{\beta}_{(k+1)\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

Multivariate Regression in Matrix

- $\mathbf{X}$ is a $n \times (k+1)$ matrix of independent variables.
- $\boldsymbol{\beta}$ is a $(k+1) \times 1$ column vector of coefficients.
- $\mathbf{X}\boldsymbol{\beta}$ will be of $n \times 1$ dimension.

$$\mathbf{X}\boldsymbol{\beta} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \cdots + \beta_k \mathbf{x}_k$$

  where x_i, $i = 1, \cdots, k$ are also $(n \times 1)$ column vectors.
- We can compactly write the linear model as the following:

$$\mathbf{Y}_{(n\times 1)} = \mathbf{X}\boldsymbol{\beta}_{(n\times 1)} + \mathbf{u}_{(n\times 1)}$$

- We can also look at individual level, where $\mathbf{x}_i'$ is the i^{th} row of $\mathbf{X}$:

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

Multivariate Regression in Matrix

- Let $\widehat{\boldsymbol{\beta}}$ be the matrix of estimated regression coefficients and $\widehat{\mathbf{Y}}$ be the vector of fitted values:

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \\ \widehat{\beta_2} \\ \vdots \\ \widehat{\beta_k} \end{bmatrix} \qquad \widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

- It might be helpful to see this again more written out:

$$\widehat{\mathbf{Y}} = \begin{bmatrix} \widehat{Y_1} \\ \widehat{Y_2} \\ \vdots \\ \widehat{Y_n} \end{bmatrix} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 1\widehat{\beta_0} + X_{11}\widehat{\beta_1} + \cdots + +X_{k1}\widehat{\beta_k} \\ 1\widehat{\beta_0} + X_{12}\widehat{\beta_1} + \cdots + +X_{k2}\widehat{\beta_k} \\ \vdots \\ 1\widehat{\beta_0} + X_{1n}\widehat{\beta_1} + \cdots + +X_{kn}\widehat{\beta_k} \end{bmatrix}$$

Residual in Matrix Form

- We can easily write the residuals in matrix form:

$$\widehat{\mathbf{u}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$$

- Our goal, again, is to minimize the sum of the squared residuals:

$$\sum_{i=1}^{n} \widehat{u_i}^2 = \widehat{\mathbf{u}}'\widehat{\mathbf{u}} = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

$$= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}$$

OLS in Matrix Form

National Taiwan University

- Goal: minimize the sum of the squared residuals.
- Take (matrix) derivatives, set equal to 0.
- Resulting first order conditions:

$$-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = 0$$

- Rearranging:

$$\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$$

- Assume that $\mathbf{X}'\mathbf{X}$ is nonsigular and invertible, $\Leftarrow$ **A3!**

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- Pronunciation: ex prime ex inverse ex prime y

Intuition for the OLS in Matrix Form

National Taiwan University

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- What's the intuition here?
  - Numerator $\mathbf{X}'\mathbf{Y}$: is approximately composed of the covariances between the columns of $\mathbf{X}$ and $\mathbf{Y}$
  - Denominator $\mathbf{X}'\mathbf{X}$ is approximately composed of the sample variances and covariances of variables within $\mathbf{X}$
- Thus, we have something like:

$$\widehat{\boldsymbol{\beta}} \approx (\text{Variance of } \mathbf{X})^{-1}(\text{Covariance between } \mathbf{X} \text{ and } \mathbf{Y})$$

$\Rightarrow$ an analogous to the simple linear regression case!

- Check the univariate regression on board: $\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{x} \end{bmatrix}$ and $\boldsymbol{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix}$!

CLRM Assumptions in Matrix Form

National Taiwan University

1. Linearity: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$

2. Randomness: $(y_i, \mathbf{x_i'})$ are IID samples from the populaiton.

3. No Perfect Collinearity: $\mathbf{X}$ is an $n \times (k+1)$ matrix with rank $k+1$
   $\rightarrow$ If $< k+1$, $\mathbf{X}'\mathbf{X}$ will not be invertible!

4. Zero Conditional Error: $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$

5. Homoskedasticity: $Var(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$

6. Normality: $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_n)$

Unbiasedness of $\widehat{\boldsymbol{\beta}}$

- Again, with CLRM assumptions 1-4:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad \text{(linear form and no collinearity)} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}
\end{aligned}$$

$$\begin{aligned}
E[\widehat{\boldsymbol{\beta}}|\mathbf{X}] &= E[\boldsymbol{\beta}|\mathbf{X}] + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{u}|\mathbf{X}] \qquad \text{(zero conditional error)} \\
&= \boldsymbol{\beta}
\end{aligned}$$

CLRM Assumption 5

- What does $Var(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$ mean?
- $\mathbf{I}_n$ is the $n \times n$ identity matrix, σ_u^2 is a scalar.
- Visually:

$$Var(\mathbf{u}) = \sigma_u^2 \mathbf{I}_n = \begin{bmatrix} \sigma_u^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_u^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_u^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_u^2 \end{bmatrix}$$

- In other words, $var(u_i) = \sigma_u^2$ for all i (constant variance) $cov(u_i, u_j) = 0$ for all $i \neq j$ (implied by IID)

Conditional Variance of $\widehat{\boldsymbol{\beta}}$

National Taiwan University

- A quick note: For a linear transformation of matrices: $\mathbf{Au}+\mathbf{B}$, $Var(\mathbf{Au}+\mathbf{B}) = \mathbf{A}Var(\mathbf{u})\mathbf{A}'$

- Now, with $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$,

$$\begin{aligned}
Var[\widehat{\boldsymbol{\beta}}|\mathbf{X}] &= Var[\boldsymbol{\beta}|\mathbf{X}] + Var[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] && \text{(no covariance term)}\\
&= Var[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] && \text{(Var(scalar)=0)}\\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var[\mathbf{u}|\mathbf{X}]((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')'\\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var[\mathbf{u}|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma_u^2\mathbf{I}_n\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} && \text{(homoskedasticity)}\\
&= \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\\
&= \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

- This is a $(k+1)\times(k+1)$ variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$.

BLUE of $\widehat{\beta}$

National Taiwan University

- Our purpose is to show that given CLRM assumptions 1-5, $\widehat{\beta}$ has the minimum variance among all unbiased linear estimators.

- Suppose we have some unbiased linear estimator $\tilde{\boldsymbol{\beta}} = \mathbf{A}'\mathbf{Y}$.
For $\tilde{\boldsymbol{\beta}}$ to be an unbiased estimator, We need $\mathbf{A}'\mathbf{X} = \mathbf{I}_{(k+1)}$ so that
$E[\tilde{\boldsymbol{\beta}}|\mathbf{X}] = \mathbf{A}'E[\mathbf{Y}|\mathbf{X}] = \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathbf{A}'E[\mathbf{u}|\mathbf{X}] = \mathbf{A}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$

- Under such circumstance, $Var[\tilde{\boldsymbol{\beta}}|\mathbf{X}] = Var[\mathbf{A}'\mathbf{u}|\mathbf{X}] = \sigma_u^2\mathbf{A}'\mathbf{A}$

- Our goal is to show that $\sigma_u^2\mathbf{A}'\mathbf{A} \geq \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}$

- Now assume some $\mathbf{C} = \mathbf{A} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$

BLUE of $\widehat{\boldsymbol{\beta}}$

National Taiwan University

- We can first show that
$\mathbf{X}'\mathbf{C} = \mathbf{X}'\mathbf{A} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{I}_{(k+1)} - \mathbf{I}_{(k+1)} = \mathbf{0}$

- Then,

$$\begin{aligned}
\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1} &= (\mathbf{C} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})'(\mathbf{C} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) - (\mathbf{X}'\mathbf{X})^{-1} \\
&= \mathbf{C}'\mathbf{C} + \mathbf{C}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C} \\
&+ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1} \\
&= \mathbf{C}'\mathbf{C} + \mathbf{0} + \mathbf{0} + (\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1} \\
&= \mathbf{C}'\mathbf{C} \geq \mathbf{0}
\end{aligned}$$

- The matrix $\mathbf{C}'\mathbf{C}$ is positive semi-definite. Therefore, $\sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}$ is the minimum variance of all linear unbiased estimators for $\boldsymbol{\beta}$.

BUE of $\widehat{\beta}$

- To go from BLUE to BUE, we need to relax the set of candidates from linear unbiased estimators to all unbiased estimators.
- We will not cover the details in this course!
- But the intuition is:
  with CLRM assumption 6, we know the exact distribution of $\mathbf{u}$.
- With the distribution known, we may use another estimator, the Maximum Likelihood Estimator (MLE).
- In a more advanced econometric course, you will learn that MLE is BUE and that with CLRM assumption 6, the MLE closed form solution coincides with OLS estimator, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.
- Lastly, Hansen, B. (2022). A Modern Gauss-Markov Theorem. *Econometrica, forthcoming.* finds that A1-A5 $\Rightarrow$ OLS is BUE!

Estimating σ_u^2

國立臺灣大學 National Taiwan University

- Similar to the univariate version, our goal is again to infer the variance of the error term from the residual!

- Let's first introduce another matrix $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
  $\mathbf{M}$ is called the residual matrix or the orthogonal projection matrix.
  E.g. $\mathbf{M}\mathbf{Y} = \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{u}}$.

- For some $\tilde{\mathbf{Y}}$, the output of $\mathbf{M}\tilde{\mathbf{Y}}$ is the residual of regressing $\tilde{\mathbf{Y}}$ on $\mathbf{X}$!

- Important properties of $\mathbf{M}$:
  - $\mathbf{M}$ is a symmetric square, i.e. $\mathbf{M} = \mathbf{M}'$
  - $\mathbf{M}$ is idempotent. i.e. $\mathbf{M}\mathbf{M} = \mathbf{M}$
  - As $\mathbf{M}\mathbf{Y} = \widehat{\mathbf{u}}$, $\mathbf{M}\mathbf{u} = \widehat{\mathbf{u}}$

- Thus, $Var[\widehat{\mathbf{u}}|\mathbf{X}] = Var[\mathbf{M}\mathbf{u}|\mathbf{X}]$!

- Now, let's calculate $MSD(\widehat{u}) = \frac{1}{n}\sum_{i=1}^{n} \widehat{u_i}^2$.

$$
\begin{aligned}
MSD(\widehat{u}) &= \frac{1}{n}\sum_{i=1}^{n} \widehat{u_i}^2 \\
&= \frac{1}{n}\widehat{\mathbf{u}}'\widehat{\mathbf{u}} \\
&= \frac{1}{n}\mathbf{u}'\mathbf{M}'\mathbf{M}\mathbf{u} \\
&= \frac{1}{n}\mathbf{u}'\mathbf{M}\mathbf{u} \\
&= \frac{1}{n}trace(\mathbf{u}'\mathbf{M}\mathbf{u}) && \text{(because it's a scalar)} \\
&= \frac{1}{n}trace(\mathbf{M}\mathbf{u}\mathbf{u}') && \text{(trace property)}
\end{aligned}
$$

Estimating σ_u^2

国立臺灣大學 National Taiwan University

$$
\begin{aligned}
E[MSD(\widehat{u})|\mathbf{X}] &= \frac{1}{n} trace(E[\mathbf{M}\mathbf{u}\mathbf{u}'|\mathbf{X}]) = \frac{1}{n} trace(\mathbf{M}E[\mathbf{u}\mathbf{u}'|\mathbf{X}]) \\
&= \frac{1}{n} trace(\mathbf{M}\sigma_u^2 \mathbf{I}_n) \qquad \text{(homoskedasticity)} \\
&= \frac{1}{n}\sigma_u^2 \; trace(\mathbf{M}) = \frac{1}{n}\sigma_u^2 \; trace(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\
&= \frac{1}{n}\sigma_u^2 \; [trace(\mathbf{I}_n) - trace(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')] \\
&= \frac{1}{n}\sigma_u^2 \; [trace(\mathbf{I}_n) - trace((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X})] \\
&= \frac{1}{n}\sigma_u^2 \; [trace(\mathbf{I}_n) - trace(\mathbf{I}_{k+1})] = \frac{n-(k+1)}{n}\sigma_u^2
\end{aligned}
$$

- Therefore, an unbiased estimator for σ_u^2 is: $s^2 = \frac{1}{n-(k+1)} \sum_{i=1}^{n} \widehat{u_i}^2$
- We can now estimate $Var(\widehat{\beta}|\mathbf{X}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$

- CLRM assumption 5 assumes for Homoskedasticity, or IID, in the error term. All the u_i are drawn from the exact same distribution and are drawn independently. Therefore, they should have the same variance, σ_u^2.
- Now, let's still assume for independence but relax the assumption that they all have the same variance. For each $u_1, u_2, \cdots, u_n$, the corresponding variance is: $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$.
- The variance-covariance matrix goes from

$$\sigma_u^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \quad \text{to} \quad \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Heteroskedasticity

- What are the consequences?

- First, OLS is not BLUE anymore, but OLS estimates are still unbiased.
$\rightarrow$ To resume BLUE, we need GLS (Generalized Least Square)!

- Second, we cannot estimate $Var(\widehat{\beta}|\mathbf{X}) = s^2(\mathbf{X'X})^{-1}$ anymore.

- Assume the Var-Cov matrix of the error term is denoted as Ω.

$$
\begin{aligned}
Var[\widehat{\beta}|\mathbf{X}] &= (\mathbf{X'X})^{-1}\mathbf{X'}Var[\mathbf{u}|\mathbf{X}]\mathbf{X}(\mathbf{X'X})^{-1} \\
&= (\mathbf{X'X})^{-1}\mathbf{X'}\Omega\mathbf{X}(\mathbf{X'X})^{-1} \\
&= (\mathbf{X'X})^{-1}\left(\sum_{i=1}^{n}\mathbf{X_i}\mathbf{X_i}'u_i^2\right)(\mathbf{X'X})^{-1}
\end{aligned}
$$

- If we can observe u_i, we could estimate $Var[\widehat{\beta}|\mathbf{X}]$ as above.

- Since we cannot observe u_i, we could only approximate using the residual terms $\widehat{u}_i$.

$$Var^{robust}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] = \frac{n}{n-(k+1)}(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n}\mathbf{X_i}\mathbf{X_i}'\widehat{u}_i{}^2\right)(\mathbf{X}'\mathbf{X})^{-1}$$

- This is the heteroskedasticity-consistent or heteroskedasticity-robust variance-covariance matrix estimator. It is also sometimes called the robust covariance matrix estimator.

- People also use the term, the White robust covariance matrix estimator, giving reference to White (1980) which first introduces this concept.

- In Stata, you can simply implement this using the command:
  reg Y X, r

Partition (Frisch—Waugh—Lovell Theorem)

National Taiwan University

- It is equally important to study the cross-relationship between different independent variables.

- Now, let's partition $\mathbf{X}$ into $[\mathbf{X_1}\ \mathbf{X_2}]$, and $\boldsymbol{\beta}$ into $[\boldsymbol{\beta_1}\ \boldsymbol{\beta_2}]$

- We can then rewrite $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ as

$$\mathbf{Y} = \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \mathbf{u}$$

- The solution to $[\widehat{\boldsymbol{\beta_1}}\ \widehat{\boldsymbol{\beta_2}}]$ is

$$\widehat{\boldsymbol{\beta_1}} = (\mathbf{X_1'M_2X_1})^{-1}\mathbf{X_1'M_2Y}$$
$$\widehat{\boldsymbol{\beta_2}} = (\mathbf{X_2'M_1X_2})^{-1}\mathbf{X_2'M_1Y}$$

where $\mathbf{M_i} = \mathbf{I}_n - \mathbf{X_i}(\mathbf{X_i'X_i})^{-1}\mathbf{X_i'}$, $i = 1, 2$

- Quick demonstration:

$$
\begin{aligned}
\mathbf{Y} &= \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \mathbf{u} \\
\mathbf{M_1Y} &= \mathbf{M_1X_1}\boldsymbol{\beta_1} + \mathbf{M_1X_2}\boldsymbol{\beta_2} + \mathbf{M_1u} \\
\mathbf{M_1Y} &= \mathbf{M_1X_2}\boldsymbol{\beta_2} + \mathbf{M_1u} \\
\tilde{\mathbf{Y}} &\equiv \tilde{\mathbf{X}}\boldsymbol{\beta_2} + \tilde{\mathbf{u}}
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\boldsymbol{\beta_2}} &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}} \\
&= (\mathbf{X_2'M_1'M_1X_2})^{-1}\mathbf{X_2'M_1'M_1Y} \\
&= (\mathbf{X_2'M_1X_2})^{-1}\mathbf{X_2'M_1Y}
\end{aligned}
$$

- Important takeaway: To obtain $\widehat{\boldsymbol{\beta_2}}$
    1. Regress $\mathbf{Y}$ on $\mathbf{X_1}$ and obtain residuals $\tilde{\mathbf{u_1}}$
    2. Regress $\mathbf{X_2}$ on $\mathbf{X_1}$ and obtain residuals $\tilde{\mathbf{X_2}}$
    3. Regress $\tilde{\mathbf{u_1}}$ on $\tilde{\mathbf{X_2}}$ and obtain $\widehat{\boldsymbol{\beta_2}}$ as well as residuals $\hat{\mathbf{u}}$
- What if $\mathbf{X_1}$ and $\mathbf{X_2}$ are independent/orthogonal to each other?

Hypothesis Test: Multivariate

National Taiwan University

- $\frac{\widehat{\beta_1}-E[\widehat{\beta_1}]}{\sqrt{Var[\widehat{\beta_1}]}} \sim N(0,1)$ under CLT.

- Thus, under some null hypothesis about β_1, $\widehat{\beta_1}$ can be tested using the t-statistics. And the 95% confidence interval is: $\widehat{\beta_1} \pm 1.96 \times SE[\widehat{\beta_1}]$

- This is also the case for $\beta_2, \cdots, \beta_k$!

- However, we might want to test more than one coefficient at the same time as well!

- For example, for $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + u_i$ we might be interested in testing the following hypotheses:
    - $\beta_1 = \beta_2$
    - $2\beta_3 + 3\beta_4 = 0$
    - $\beta_1 = 10$

Hypothesis Test: Multivariate

- The hypotheses are called **linear hypotheses**:

- To test linear hypotheses **jointly**, let's form the a matrix of constraints:

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$$

- Let's first rewrite the constraints above as:
  - $\beta_1 - \beta_2 = 0$
  - $2\beta_3 + 3\beta_4 = 0$
  - $\beta_1 - 10 = 0$

- Each entry of $\mathbf{R}$ is a coefficient for $\boldsymbol{\beta}$ and each row is a constraint we want to test. $\mathbf{q}$ is a vector of scalars.

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 2 & 3 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 10 \end{bmatrix}$$

Hypothesis Test: Multivariate

- Now, suppose we want to test j hypotheses jointly:

$$H_0 : \mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$$
$$H_1 : \mathbf{R}\boldsymbol{\beta} - \mathbf{q} \neq \mathbf{0}$$

- Now Assumes that $\mathbf{m}_{(j\times 1)} = \mathbf{R}\boldsymbol{\beta} - \mathbf{q}$
- Asymptotically, $W \equiv \mathbf{m}'(Var[\mathbf{m}|\mathbf{X}])^{-1}\mathbf{m} \sim \chi^2(j)$
- Lastly, $F_{j,n-k-1} = \frac{W/j}{s^2/\sigma_u^2} = \frac{W}{j}\frac{\sigma_u^2}{s^2}$
- W is called a Wald Statistic, and the F test is called a Wald test.
- Under homoskedasticity, $F_{j,n-k-1} = \frac{(\mathbf{R}\widehat{\boldsymbol{\beta}}-\mathbf{q})'\{\mathbf{R}[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}'\}^{-1}(\mathbf{R}\widehat{\boldsymbol{\beta}}-\mathbf{q})}{j}$

Hypothesis Test: Multivariate

National Taiwan University

- Under heteroskedasticity,
  $F_{j,n-k-1} = \frac{(\mathbf{R}\widehat{\boldsymbol{\beta}}-\mathbf{q})'\{\mathbf{R}[s^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}'\}^{-1}(\mathbf{R}\widehat{\boldsymbol{\beta}}-\mathbf{q})}{j}$,
  where Ω is to be estimated.

- If we only consider one constraint and if we only test one β coefficient, the F-statistic will nest down to a t-statistic.

- Under homoskedasticity, there is a special solution for F:

$$F = \frac{(R^2_{\text{unrestricted}} - R^2_{\text{restricted}})/j}{(1 - R^2_{\text{unrestricted}})/(n - k_{\text{unrestricted}} - 1)}$$

  where $R^2_{\text{restricted}}$ is the R^2 when we restrict the testing model under the null hypotheses.

Model Specification

- Problem of potential omitted variable.
  $\rightarrow$ Check the white board!

- Is adding control variables always a good idea?
  $\rightarrow$ Think about *collinearity* and *bad controls*.
  We'll talk more about bad controls under the potential outcome framework!

- Problem of measurement bias or error.
  $\rightarrow$ Check the white board!

- Interpretation of R^2
  $\rightarrow$ Is a higher R^2 always better?