Matrix OLS

Readings

Linear algebra

Matrix probability theory

Matrix statistics

Matrix OLS

Readings

Readings

Appendix E

- Skip Theorem E.3's proof, but re-read the proof we did in class
- Skip Section E.3
- Skip "Wald Statistics for Testing Multiple Hypothesis"

Linear algebra

Overview

- Matrix and vector operations
- Invertibility
  - linear independence
  - rank
  - positive (semi-)definiteness
- Trace
- Partitioned matrix multiplication

Equality

Two matrices are equal if all their elements are equal, i.e.

$$A = [a_{ij}] = B = [b_{ij}]$$

if and only if

$$a_{ij} = b_{ij} \text{ for all } i, j$$

Inequality

- $A > B$ if $A - B$ is positive definite
- $A \geq B$ if $A - B$ is positive _semi_definite

Useful for defining *efficiency* for estimators that are vectors.

Partitioned matrix multiplication

Let A be an $n \times k$ matrix and let B be an $n \times m$ matrix, i.e.

$$\mathbf{A} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

Then $A'B = \sum_{i=1}^{n} a_i' b_i$

PMM: Try it!

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

1. Compute $A'A$ directly
2. Compute $A'A$ using PMM

PMM: Try it (2)!

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$$

1. Compute $A'A$ directly
2. Compute $A'A$ using PMM

Trace

The trace of a square matrix is the sum of its diagonal elements,

$$tr(A) = \sum_i a_{ii}$$

1. Show that $tr(A) = tr(A')$. *whiteboard:sketch4.3*
2. Trace allows you to work with matrices as though they are numbers:
   - $tr(AB) = tr(BA)$

Matrix probability theory

Probability theory

Matrices give us economical notation for expectation and variances of collections of random variables.

Random vectors

A k-dimensional **random vector** X is a column vector of random variables $(X_1, \cdots, X_k)$,

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}.$$

Mean

The mean of a random vector $\mathbf{X}$ is a vector:

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_k) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} = \mu_{\mathbf{X}}.$$

Variance (definition)

The variance of a $k-$dimensional vector X is defined as

$$Var(\mathbf{X}) = E\left((\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))'\right)$$

What is the dimension of $Var(X)$?

Variance (looks)

The variance of a random vector $\mathbf{X}$ is a $k \times k$ matrix

$$Var\,(\mathbf{X}) = \begin{bmatrix} V(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_k) \\ Cov(X_1, X_2) & V(X_2) & \cdots & Cov(X_2, X_k) \\ \vdots & & & \\ Cov(X_1, X_k) & Cov(X_2, X_k) & \cdots & Var(X_k) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \cdots & \sigma_{1,k} \\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} & \cdots & \sigma_{2,k} \\ \vdots & & & & \\ \sigma_{1,k} & \sigma_{2,k} & \cdots & \cdots & \sigma_k^2 \end{bmatrix} = \Sigma_X$$

Example

Roll and flip:

- Let X_1 be the outcome of a dice roll, and let X_2 be the outcome of a coin flip.
- $X = (X_1, X_2)$,

then

- $\mu_1 = 7/2,\ \mu_2 = 1/2$
- $\sigma_1^2 = 35/12,\ \sigma_2^2 = 1/4,\ \sigma_{1,2} = 0$

so that

$$\mu = \begin{bmatrix} 7/2 \\ 1/2 \end{bmatrix},\ \Sigma_X = \begin{bmatrix} 35/12 & 0 \\ 0 & 1/4 \end{bmatrix}$$

Linear combinations

Consider

- a $k-$dimensional random vector $\mathbf{X}$ with
  - mean μ and
  - variance Σ.
- a (conformable) matrix $A \in \mathbb{R}^{m \times k}$
- a $k \times 1$ vector b

For the linear combination $Z = AX + b$, we have:

- $E(Z) = AE(X) + b$
  - proof: element by element
- $Var(Z) = A\Sigma A'$
  - proof: *whiteboard*

Matrix statistics

Intro

Matrix notation is extremely useful for linear regression and related methods. As an intermediate step, we look at estimating the *scalar* mean μ of a random variable X

$$E(X) = \mu$$

using a random sample of observations $X_n = (X_1, \cdots, X_n)$.

Random sample

We treat the random sample as an $n-$dimensional random vector X_n

1. What is the mean, $E(X_n)$?
2. What is the variance, $Var(X_n)$?

The sample mean

Can we express the sample mean

$$\bar{X}_n = \frac{1}{n} \sum_i X_i$$

as a linear combination of the random sample X_n?

Unbiasedness

- The sample mean is

$$\bar{X}_n = \frac{1}{n}\iota_n' X_n$$

- The expectation of the sample mean is

$$\begin{aligned} E(\bar{X}_n) &= E(\frac{1}{n}\iota_n' X_n) \\ &= \frac{1}{n}\iota_n' E(X_n) \\ &= \frac{1}{n}\iota_n' \iota_n \mu \\ &= \frac{1}{n} n\mu \\ &= \mu. \end{aligned}$$

The sample mean is unbiased.

Variance

The variance of

$$\bar{X}_n = \frac{1}{n}\iota_n' X_n$$

is

whiteboard

Matrix OLS

Intro

You import all your linear regression model from your previous courses:

- interpretation
- estimation
- inference
- ...

We will do a quick review, with matrix tools in hand.

Model

The **linear model** relates - a **dependent variable** y_i, to - a set of **explanatory variables** $x_{i1}, \cdots, x_{ik}$ for a randomly selected unit (individual/country/firm) i

Example 4.3. Student i's college GPA (y_i) may be related to her high school GPA (x_{i1}), number of lectures skipped (x_{i2}), and her ACT score (x_{i3}).

Model (2)

The model consists of

1. The population regression *equation*:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i$$

2. A set of assumptions

Model (3)

For the college GPA example,

- What is in u_i?
- What is the interpretation of the regression coefficient β_2 on "lectures skipped"?

Equation

1. Gather the explanatory variables in a row vector of dimension $k+1$,
$$X_i = \begin{bmatrix} 1 & x_{i1} & \cdots & x_{ik} \end{bmatrix}$$
2. Gather the regression coefficients in a column vector of dimension $k+1$,
$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

The population regression equation can then be written as

$$y_i = X_i\beta + u_i$$

Equation (2)

We obtain a random sample on (y_i, X_i) and create

- the random vector

$$y = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}'$$

- matrix of explanatory variables

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

- vector of error terms u

Equation (3)

The population regression model can now be written in matrix form, for all the observations in our random sample jointly:

$$y = X\beta + u$$

Approximate solutions

The linear model therefore looks like an approximate linear system of equations,

$$y = X\beta + u$$
$$b \approx Ax$$

In the current application:

- we know y
- we know X
- we do not know β

We know how to obtain an approximate solution: **projection**

Approximate solution (2)

The **ordinary least squares** estimator for β is

$$\hat{\beta} = (X'X)^{-1}X'y$$

Properties

1. What do we need for the OLS estimator to be well-defined?
2. What do we need for the OLS estimator to have good statistical properties?

Model (4)

Assumption E.

1. The population regression equation can be written as $y = X\beta + u$
2. The matrix X has full column rank, $rank(X) = k + 1$
3. Zero conditional mean: $E(u|X) = O_{n\times 1}$

Mean / unbiasedness

The mean of $\hat{\beta}$ is

$$E(\hat{\beta}) = E(E(\hat{\beta} \Big| X))$$

Conditional expectation is worked out on *whiteboard*

Homoskedasticity

Until now, we did not use the fact that our observations constitute a random sample.

Assumption E.4

The conditional variance of u given X is

$$Var(u|X) = \sigma^2 I_n$$

Variance

The variance of $\hat{\beta}$ is...

whiteboard

Efficiency

Two estimators for β, $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ with variances V_1 and V_2

- $\hat{\beta}^{(1)}$ is **efficient relatively to** $\hat{\beta}^{(1)}$ iff $V_2 - V_1$ is psd
  - for any linear combination $a'\beta$, ...
- An estimator is **efficient** if it is efficient relative to all estimators (in a class)

A proof of the efficiency of the OLS estimator in the class of linear, unbiased estimators: *whiteboard*

Probability limit

1. Indirect: Unbiasedness + Chebyshev
2. Direct: Closed-form expression + LLN
3. Properties of consistency

Asymptotic distribution

Asymptotic normality: *whiteboard*