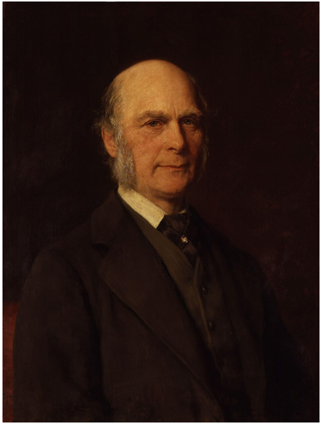
BIOST/STAT 533, Sp 2024
Theory of Linear Models

Richard Guo

Lecture # 1: §1–2

'Linear regression' and 'linear model' are used interchangeably.

☞ What is 'regression'?



- Francis Galton (1822–1911), Darwin's nephew, controversial guy

Galton's 'regression to the mean'

Francis Galton's example.

```
> GaltonFamilies
   family father mother midparentHeight children childNum gender childHeight
1     001   78.5   67.0           75.43        4        1   male        73.2
2     001   78.5   67.0           75.43        4        2 female        69.2
3     001   78.5   67.0           75.43        4        3 female        69.0
4     001   78.5   67.0           75.43        4        4 female        69.0
5     002   75.5   66.5           73.66        4        1   male        73.5
6     002   75.5   66.5           73.66        4        2   male        72.5
7     002   75.5   66.5           73.66        4        3 female        65.5
8     002   75.5   66.5           73.66        4        4 female        65.5
9     003   75.0   64.0           72.06        2        1   male        71.0
10    003   75.0   64.0           72.06        2        2 female        68.0
...
```

Galton's 'regression to the mean'

$$x_i = \texttt{midparentHeight}_i = (\texttt{father} + 1.08\,\texttt{mother})/2$$
$$y_i = \texttt{childHeight}_i.$$

- Galton determines that the 'best fitted line' can be written as

$$\frac{y - \bar{y}}{\hat{\sigma}_y} = \hat{\rho}\,\frac{x - \bar{x}}{\hat{\sigma}_x},$$

where $|\hat{\rho}| < 1$ — "regression to the mean / mediocre".

Galton's 'regression to the mean'

Ordinary least squares (OLS)

The best fitted line $y = \hat{\alpha} + \hat{\beta}x$ is defined to be

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_i (y_i - \alpha - \beta x_i)^2.$$

- Gauss and Legendre
- Not $\sum_i |y_i - \alpha - \beta x_i|$

Normal equations:

$$\sum_i y_i - \hat{\alpha} - \hat{\beta}x_i = 0 \implies \bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = 0$$

$$\sum_i x_i(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \implies \overline{xy} - \hat{\alpha}\bar{x} - \hat{\beta}\overline{x^2} = 0.$$

- goes through data center

Subtracting 1st eqn multiplied by $\bar{x}$, we get

$$(\overline{x^2} - \bar{x}^2)\hat{\beta} = \overline{xy} - \bar{x}\bar{y}$$

Ordinary least squares (OLS) for univariate X

Univariate OLS

$$\boxed{\hat{\beta} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} = \frac{\hat{\rho}_{xy}\hat{\sigma}_x\hat{\sigma}_y}{\hat{\sigma}_x^2} = \frac{\hat{\rho}_{xy}\hat{\sigma}_y}{\hat{\sigma}_x}.}$$

The fitted line is

$$\begin{aligned} y = \hat{\alpha} + \hat{\beta}x &= (\bar{y} - \hat{\beta}\bar{x}) + \hat{\beta}x \\ y - \bar{y} &= \hat{\beta}(x - \bar{x}) \\ y - \bar{y} &= \frac{\hat{\rho}_{xy}\hat{\sigma}_y}{\hat{\sigma}_x}(x - \bar{x}) \end{aligned}$$

$$\boxed{\frac{y - \bar{y}}{\hat{\sigma}_y} = \hat{\rho}_{xy}\,\frac{x - \bar{x}}{\hat{\sigma}_x}.}$$

- $\hat{\rho}_{xy} = 0.32 < 1$ — 'regression to the mean' by Galton, "the average regression of the offspring is a constant fraction of their respective mid-parental deviations"
  - Does it make sense?

Without intercept

$$\hat{\beta} = \arg\min_b \sum_i (y_i - b\,x_i)^2$$

From normal equation

$$\sum_i x_i(y_i - \hat{\beta} x_i) = 0,$$

we get

$$\boxed{\hat{\beta} = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{\langle x, y \rangle}{\langle x, x \rangle}.}$$

- Will be used a lot later!

Multiple linear regression

Multiple covariates

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{pmatrix} = (X_1, \dots, X_p)$$

- row x_i, column X_j

and single outcome

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Jargon

X_j	Y
Regressor	Response
Covariate	Outcome
Feature	Label
Predictor	
Explanatory variable	
Independent variable	Dependent variable

OLS

Find the best fitted line

$$y_i = \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{i1} = x_i^\intercal \hat{\beta}$$

for

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

$$\hat{\beta} = \arg\min_b \sum_{i=1}^{n} (y_i - b^\intercal x_i)^2 = \arg\min_b \| Y - Xb \|^2.$$

Normal equation

$$\sum_i (y_i - x_i^\intercal \hat{\beta}) x_i = \mathbf{0} \iff X^\intercal (Y - X\hat{\beta}) = 0 \iff X^\intercal Y = X^\intercal X \hat{\beta}.$$

OLS

If $X^\intercal X$ is invertible,

$$\boxed{\hat{\beta} = (X^\intercal X)^{-1} X^\intercal Y.} \iff \boxed{\hat{\beta} = \left(\sum_i x_i x_i^\intercal\right)^{-1} \left(\sum_i y_i x_i\right).}$$

$X^\intercal X$ invertible requires that for any $\mathbf{0} \neq a \in \mathbb{R}^p$,

$$a^\intercal (X^\intercal X) a = \|Xa\|^2 \neq 0 \iff Xa \neq 0,$$

i.e., $X = (X_1, \ldots, X_p)$ are linearly independent.

☞ Throughout, we assume

Condition: Column vectors of X are linearly independent. ▶ Must have $n \geq p$ (why?)

BIOST/STAT 533, Sp 2024
Theory of Linear Models

Richard Guo

Lecture # 2: Review of linear algebra
Appendix A

Recall

- Univariate OLS $y = \alpha + \beta x$:

$$y - \bar{y} = \widehat{\beta}(x - \bar{x}), \quad \widehat{\beta} = \frac{\widehat{\sigma}_{xy}}{\widehat{\sigma}_x^2}.$$

- Univariate OLS $y = \beta x$:

$$\widehat{\beta} = \frac{\langle x, y \rangle}{\langle x, x \rangle}.$$

- Multivariate OLS $y = \beta^\intercal x$:

$$X^\intercal (Y - X\widehat{\beta}) = 0, \quad \widehat{\beta} = (X^\intercal X)^{-1} X^\intercal Y.$$

Vectors

For $x, y \in \mathbb{R}^n$,

- $\langle x, y \rangle = x^\intercal y = \sum_i x_i y_i$
- $\|x\|^2 = \langle x, x \rangle$
- Cauchy-Schwartz $|\langle x, y \rangle| \le \|x\| \|y\|$.
  ☞ '$=$' holds iff $ax = by$ for some scalar a, b.
- Triangle $\|x + y\| \le \|x\| + \|y\|$
  ▶ Follows from above
- Orthogonal $x \perp y$: $x^\intercal y = 0$
- $\widehat{\rho}_{xy} = \cos \angle(x - \bar{x}, y - \bar{y}) = \frac{\langle x - \bar{x}, y - \bar{y} \rangle}{\|x - \bar{x}\| \|y - \bar{y}\|}$
  ▶ when achieves ± 1?

Matrix, row space, column space

$$A = (a_{ij})_{n\times m} = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nm} \end{pmatrix} = \begin{pmatrix} a_1^\intercal \\ \vdots \\ a_n^\intercal \end{pmatrix} = (A_1, \dots, A_m).$$

- Column space

$$\mathcal{C}(A) = \{\alpha_1 A_1 + \dots \alpha_m A_m : \alpha_1, \dots, \alpha_m \in \mathbb{R}\} = \{A\alpha : \alpha \in \mathbb{R}^m\}.$$

- Row space

$$\mathcal{R}(A) = \{r_1 a_1 + \dots + r_n a_n : r_1, \dots, r_n \in \mathbb{R}\} = \{A^\intercal r : r \in \mathbb{R}^n\}$$

- $\mathcal{C}(A) = \mathcal{R}(A^\intercal)$

Matrix as a linear map

A matrix $A \in \mathbb{R}^{p \times r}$ is a linear map from $\mathbb{R}^r$ to $\mathbb{R}^p$: $x \mapsto Ax$ for $x \in \mathbb{R}^r$.

- Rotation by θ counterclockwise: $A = \begin{bmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{bmatrix}$
- Reflection: $A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ and $A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$
- Scale by 2 in all directions: $A = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

Matrix multiplication, rank

- For $A : n \times m$, $B : m \times l$,

$$AB = A(B_1, \ldots, B_l) = (AB_1, \ldots, AB_l) \text{ has columns in } \mathcal{C}(A).$$

☞ Right multiply: cols!

$$AB = \begin{pmatrix} a_1^\intercal \\ \vdots \\ a_n^\intercal \end{pmatrix} B = \begin{pmatrix} a_1^\intercal B \\ \vdots \\ a_n^\intercal B \end{pmatrix} \text{ has rows in } \mathcal{R}(B) = \mathcal{C}(B^T).$$

☞ Left multiply: rows!

- A set of vectors $A_1, \ldots, A_m \in \mathbb{R}^n$ are linearly independent if

$$\alpha_1 A_1 + \cdots + \alpha_m A_m = 0 \iff \alpha = \mathbf{0}.$$

$$\text{rank}(A) := \text{rank}(A_1, \ldots, A_m) := \text{maximal } \# \text{ of linearly independent vectors}$$

- $\text{rank}(AB) \leq \min(\text{rank}\, A, \text{rank}\, B)$ (why?)

Matrix multiplication, rank

- If $A \in \mathbb{R}^{n \times m}$ has rank k, then we can write $A = \underbrace{B}_{n \times k} \underbrace{C}_{k \times m}$ (why?)

  ☞ Take B to be k linearly independent cols of A...

Orthogonal matrix

Definition $A \in \mathbb{R}^{n \times n}$ is orthogonal if A has orthonormal columns, i.e.,

$$\langle A_i, A_j \rangle = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}.$$

$$A \text{ is orthogonal} \iff A^\intercal A = I_n$$

$$\iff A \text{ has orthonormal rows}$$

(why?)

(Need A be symmetric?)

Inverse of a matrix

Definition For $A \in \mathbb{R}^{n \times n}$, A is invertible if there is $B \in \mathbb{R}^{n \times n}$ such that

$$AB = I_n$$

and $A^{-1} := B$.

- If so, $B^{-1} = A$. (why?)
- $(AB)^{-1} = B^{-1}A^{-1}$.

Eigendecomposition of a real, symmetric matrix

$A \in \mathbb{R}^{n\times n}$ has eigenvalue λ with eigenvector $x \in \mathbb{R}^n$ if

$$Ax = \lambda x, \quad x \neq \mathbf{0}.$$

- This must mean $(A - \lambda I)$ has rank $< n$. (why?)
- Characteristic-poly$(\lambda) := \det(\lambda I - A) = 0$.

If A is **symmetric**, it must admit

$$(\lambda_1, u_1), (\lambda_2, u_2), \ldots, (\lambda_n, u_n)$$

with $Au_i = \lambda_i u_i$, $\lambda_i \in \mathbb{R}$, $u_i \in \mathbb{R}^n$.

☞ Further, $\{u_i\}$ can be chosen such that they are **orthonormal**. (why?)

- For $\lambda_i \neq \lambda_j$, $u_i \perp u_j$; if $\lambda_i = \lambda_j$, orthogonalize.

Eigendecomposition of a real, symmetric matrix

Let $U = (u_1, \dots, u_n)$, then

$$AU = U \operatorname{diag}(\lambda_1, \dots, \lambda_n),$$

so

$$\boxed{A = U \operatorname{diag}(\lambda_1, \dots, \lambda_n) U^\intercal = \sum_i \lambda_i u_i u_i^\intercal.}$$

- $\operatorname{rank} A = \sum_i \mathbb{I}_{\lambda_i \neq 0}$
- A invertible $\iff \lambda_i \neq 0,\ i = 1, \dots, n$.
- If A is invertible, $A^{-1} = U \operatorname{diag}(\lambda_1^{-1}, \dots, \lambda_n^{-1}) U^\intercal$.
- Spectral function

$$A^k := \underbrace{A \dots A}_{k} = U \operatorname{diag}(\lambda_1^k, \dots, \lambda_n^k) U^\intercal.$$

- $\operatorname{Tr} A = \sum_i \lambda_i$ (why?)

Quadratic form

A quadratic form in $x \in \mathbb{R}^n$ is

$$\sum_{ij} a_{ij} x_i x_j = x^\mathsf{T} A x,$$

where WLOG we can assume $A \in \mathbb{R}^{n \times n}$ is symmetric. (why?)

For a symmetric A,

- $A \succeq 0$ (positive semidefinite): $x^\mathsf{T} A x \geq 0$ for every x
- $A \succ 0$ (positive definite): $x^\mathsf{T} A x > 0$ for every $x \neq 0$

☞ $\{A : A \succeq 0\}$ is a cone: For $a, a' > 0$, $aA + a'A' \succeq 0$ if $A, A' \succeq 0$.

Theorem A is psd iff every $\lambda_i(A) \geq 0$; A is pd iff every $\lambda_i(A) > 0$.

- Eigendecomposition $A = U \operatorname{diag}(\lambda_1, \ldots, \lambda_n) U^\mathsf{T}$

- For $A \succeq 0$, define

$$A^{1/2} := U \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) U^\mathsf{T}.$$

Rayleigh quotient

Theorem Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of a real, symmetric matrix A.

1. The optimization problem

$$\max_x x^\top A x, \text{s.t. } \|x\| = 1$$

has maximum λ_1, which is achieved by $\pm u_1$.

2. The optimization problem

$$\max_x x^\top A x, \text{s.t. } \|x\| = 1,\ x \perp u_1$$

has maximum λ_2, which is achieved by $\pm u_2$.

3. ...

Rayleigh quotient

For a real, symmetric matrix $A \in \mathbb{R}^{n \times n}$ and any $x \neq \mathbf{0}$,

$$\lambda_{\min}(A) \leq \frac{x^\top A x}{x^\top x} \leq \lambda_{\max}(A).$$

☞ All the diagonal elements of A are bounded between $\lambda_{\min}$ and $\lambda_{\max}$ (why?)

Trace

Trace of a square matrix is the sum of diagonal elements.

1. $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ (why?)

   ☞ Useful for changing dimension, e.g., for vectors $v_1, v_2 \in \mathbb{R}^n$,

$$\langle v_1, v_2 \rangle = v_1^\intercal v_2 = \mathrm{Tr}(v_1^\intercal v_2) = \mathrm{Tr}(v_2 v_1^\intercal) = \mathrm{Tr}(v_1 v_2^\intercal)$$

2. For a real, symmetric matrix A, $\mathrm{Tr}(A) = \sum_i \lambda_i$. (why?)

Projection matrix (important!)

Definition Matrix $H \in \mathbb{R}^{n \times n}$ is a projection matrix if it is (why?)

1. symmetric: $H = H^\intercal$
2. idempotent: $H^2 = H$. i.e., $HHx = Hx$ for any $x \in \mathbb{R}^n$

Theorem For a projection matrix H,

1. its eigenvalues are either 0 or 1, (why?)
2. $$\text{rank}(H) = \text{Tr}(H)$$

Singular Value Decomposition (SVD)

Any $n \times m$ matrix X can be written as

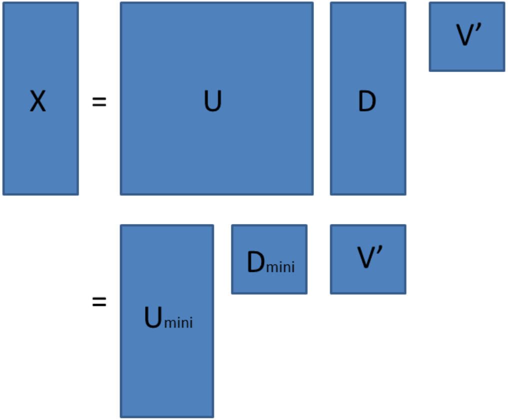
$$X = UDV^\intercal,$$

where

1. $U : n \times n$, orthogonal
2. $V : m \times m$, orthogonal
3. $D : n \times m$, 'diagonal': $D_{ii} \geq 0$ for $i \leq \min(m, n)$

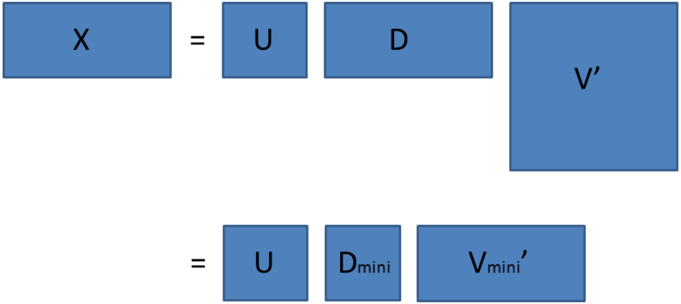
Full and mini SVDs: Tall matrix

U: $n \times n$, orthogonal: $U^\intercal U = UU^\intercal = I_n$
U_{mini}: $n \times m$ with orthogonal columns: $U_{\text{mini}}^\intercal U_{\text{mini}} = I_m$ but $U_{\text{mini}} U_{\text{mini}}^\intercal \neq I_n$

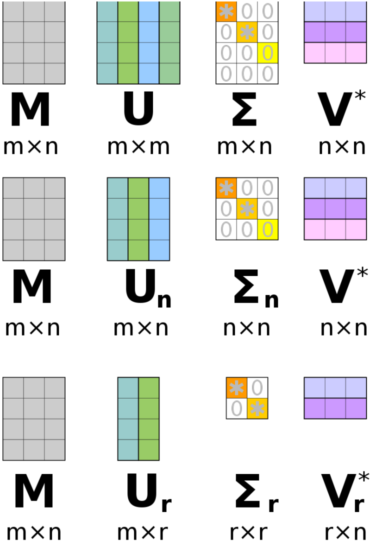


Full and mini SVDs: Wide matrix

V: $m \times m$, orthogonal: $V^\intercal V = VV^\intercal = I_m$
V_{mini}: $n \times m$ with orthogonal columns: $V_{\text{mini}}^\intercal V_{\text{mini}} = I_n$ but $V_{\text{mini}} V_{\text{mini}}^\intercal \neq I_m$



Mini SVD: rank-r matrix



Relation to eigendecomposition

Given SVD $X = UDV^\intercal$,

$$XX^\intercal = UDV^\intercal VDU^\intercal = UD^2U^\intercal$$

$$X^\intercal X = VDU^\intercal UDV^\intercal = VD^2V^\intercal$$

- The left singular vectors U of X are eigenvectors of $XX^\intercal$
- The right singular vectors V of X are eigenvectors of $X^\intercal X$
- The eigenvalues of $X^\intercal X$ and $XX^\intercal$ are squares of singular values of X

▶ When X is symmetric, SVD = eigendecomposition up to signs:

$$X = U\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)U^\intercal = U\,\mathrm{diag}(|\lambda_1|, \ldots, |\lambda_n|)V^T,$$

where $V_i = \mathrm{sign}(\lambda_i)U_i$.

Pseudoinverse

For $n \times m$ matrix A with rank r, its SVD can be written as

$$A = U \operatorname{diag}(d_1, \ldots, d_r) V^\intercal,$$

from which the pseudoinverse is defined to be the $m \times n$ matrix

$$A^\dagger := V \operatorname{diag}(d_1^{-1}, \ldots, d_r^{-1}) U^\intercal$$

In terms of the full SVD

$$A = U \begin{pmatrix} D^* & 0 \\ 0 & 0 \end{pmatrix} V^\intercal, \quad D^* = \operatorname{diag}(d_1, \ldots, d_r) > 0,$$

we have

$$A^\dagger = V \begin{pmatrix} D^{*-1} & 0 \\ 0 & 0 \end{pmatrix} U^\intercal,$$

i.e., **inverting what can be inverted and leave zeros alone**.

- When A is invertible, $A^\dagger = A^{-1}$.

Vector calculus

For $f : \mathbb{R}^p \to \mathbb{R}$,

$$\partial f(x)/\partial x = (\partial f(x)/\partial x_1, \ldots, \partial f(x)/\partial x_p)^\mathsf{T}.$$

Hence,

$$\partial a^\mathsf{T} x/\partial x = a, \quad \partial x^\mathsf{T} A x/\partial x = 2Ax.$$

- For $f(x) = (f_1(x), \ldots, f_q(x))^\mathsf{T} : \mathbb{R}^p \to \mathbb{R}^q$,

$$\partial f(x)/\partial x = (\partial f_1(x)/\partial x, \ldots, \partial f_q(x)/\partial x) \in \mathbb{R}^{p \times q}.$$

Hence, for $x \in \mathbb{R}^p$ and $B \in \mathbb{R}^{q \times p}$,

$$\frac{\partial Bx}{\partial x} = B^\mathsf{T}.$$

▶ See Appendix A.2

BIOST/STAT 533, Sp 2024
Theory of Linear Models

Richard Guo

Lecture # 3: OLS with multiple covariates
§3

Recall: OLS with a single covariate

- Univariate OLS $y = \alpha + \beta x$:

$$y - \bar{y} = \widehat{\beta}(x - \bar{x}), \quad \widehat{\beta} = \frac{\widehat{\sigma}_{xy}}{\widehat{\sigma}_x^2}(x - \bar{x}) = \frac{\widehat{\rho}_{xy}\widehat{\sigma}_y}{\widehat{\sigma}_x}(x - \bar{x}).$$

- Univariate OLS $y = \beta x$:

$$\widehat{\beta} = \frac{\langle x, y \rangle}{\langle x, x \rangle}.$$

OLS

Find the best fitted line

$$y_i = \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{i1} = x_i^\intercal \widehat{\beta}$$

for

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

$$\boxed{\widehat{\beta} = \arg\min_b \sum_{i=1}^n (y_i - b^\intercal x_i)^2 = \arg\min_b \|Y - Xb\|^2.}$$

Normal equation (why?)

$$\sum_i (y_i - x_i^\intercal \widehat{\beta}) x_i = \mathbf{0} \iff X^\intercal (Y - X\widehat{\beta}) = 0 \iff X^\intercal Y = X^\intercal X \widehat{\beta}.$$

OLS

If $X^\intercal X$ is invertible,

$$\boxed{\widehat{\beta} = (X^\intercal X)^{-1} X^\intercal Y.} \iff \boxed{\widehat{\beta} = \left(\sum_i x_i x_i^\intercal\right)^{-1} \left(\sum_i y_i x_i\right).}$$

$X^\intercal X$ invertible requires that for any $\mathbf{0} \neq a \in \mathbb{R}^p$,

$$a^\intercal (X^\intercal X) a = \|Xa\|^2 \neq 0 \iff Xa \neq 0,$$

i.e., $X = (X_1, \ldots, X_p)$ are linearly independent.

☞ In this course (unless stated otherwise), we assume

Condition Column vectors of X are linearly independent. $\implies n \geq p$.

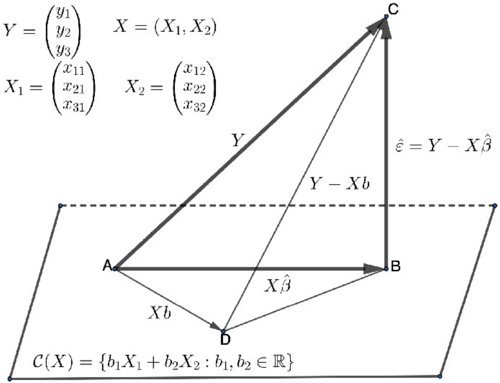
☞ Fill X with iid normal draws (random design). Then the above is satisfied with prob. 1. (why?)

Geometry

Because $\mathcal{C}(X) = \{Xb : b \in \mathbb{R}^p\}$, the least squares finds

$$\min \|Y - Xb\|^2 \iff \min_{\widehat{Y} \in \mathcal{C}(X)} \|Y - \widehat{Y}\|^2,$$

where $\widehat{Y} = X\widehat{\beta} = X(X^\intercal X)^{-1}X^\intercal Y$ is the vector of fitted values.



Geometry

Orthogonal projection Projection matrix / Hat matrix $\boxed{H = X(X^\intercal X)^{-1}X^\intercal}$.

$$Y = HY + (I - H)Y = \widehat{Y} + \widehat{\varepsilon},$$

where the residual vector $\widehat{\varepsilon} = Y - \widehat{Y}$ satisfies

$$\boxed{X^\intercal \widehat{\varepsilon} = \begin{pmatrix} X_1^\intercal \widehat{\varepsilon} \\ \vdots \\ X_p^\intercal \widehat{\varepsilon} \end{pmatrix} = \mathbf{0}}$$

$$\iff X^\intercal (Y - X\widehat{\beta}) = \mathbf{0}. \quad \text{normal equation}$$

- Implications

1. $\widehat{\varepsilon} \perp v$ for any $v \in \mathcal{C}(X) \iff 0 = \langle Xb, \widehat{\varepsilon} \rangle = b^\intercal X^\intercal \widehat{\varepsilon}$ for any $b \in \mathbb{R}^p$.
2. If X contains a column of 1's (intercept), then $1^\intercal \widehat{\varepsilon} = 0$.

Geometry

- Pythagorean Theorem

$$\boxed{\|Y\|^2 = \|\widehat{Y}\|^2 + \|\widehat{\varepsilon}\|^2}$$

- OLS is the best fitted line

For any $b \in \mathbb{R}^n$, (why?)

$$\|Y - Xb\|^2 = \|Y - X\widehat{\beta}\|^2 + \|X(\widehat{\beta} - b)\|^2$$

and hence

$$\|Y - Xb\|^2 \geq \|Y - X\widehat{\beta}\|^2,$$

with equality iff $b = \widehat{\beta}$.

Projection matrix H

In HW, we have verified that $H = X(X^\intercal X)^{-1}X^\intercal$ is symmetric and idempotent — indeed, a projection matrix.

- We know (What are the eigenvalues?)

$$\boxed{\text{Tr}(H) = \text{rank}(H) = p.}$$

Theorem $H = X(X^\intercal X)^{-1}X^\intercal$ satisfies

1. $Hv = v \iff v \in \mathcal{C}(X)$ (why?)
2. $Hw = 0 \iff w \perp \mathcal{C}(X)$ (why?)

☞ If X contains a column of 1's, then

$$\boxed{H\mathbf{1}_n = \mathbf{1}_n \implies \text{Every row of } H \text{ sums to } 1}$$

Examples

- Example m treated, n controls

$$X = \begin{pmatrix} 1_m & 1_m \\ 1_n & 0_n \end{pmatrix}.$$

$$H = \begin{pmatrix} m^{-1} 1_m 1_m^\top & 0 \\ 0 & n^{-1} 1_n 1_n^\top \end{pmatrix}$$

- Example J treatment levels, level j has n_j units

$$X = \mathrm{diag}(1_{n_1}, \ldots, 1_{n_J})$$

(What is H?)

BIOST/STAT 533, Sp 2024
Theory of Linear Models

Richard Guo

Lecture # 4: Gauss–Markov model and theorem
§4

Recall: OLS

- $H = X(X^\top X)^{-1}X^\top$ is the projection onto $\mathcal{C}(X)$.
- Orthogonal decomposition

$$Y = \underbrace{\widehat{Y}}_{HY} + \underbrace{\widehat{\varepsilon}}_{(I_n - H)Y}, \quad \|Y\|^2 = \|\widehat{Y}\| + \|\widehat{\varepsilon}\|^2.$$

- To make sure that $X^\top X$ is invertible, we shall assume throughout

(☞ Or equivalently, to make sure that $\widehat{Y} = X\widehat{\beta}$ for a uniquely defined $\widehat{\beta}$)

Assumption $X \in \mathbb{R}^{n \times p}$ has linearly independent columns.

Gauss–Markov model

GM The data generating process obeys

$$Y = X\beta + \varepsilon,$$

where

1. X is fixed and has linearly independent columns,
2. $\mathbb{E}\,\varepsilon = \mathbf{0}$, $\operatorname{cov}\varepsilon = \sigma^2 I_n$.

The unknown parameters are (β, σ^2).

- No distributional nor independence assumption on ε — only the first two moments of the random vector are concerned.
- X fixed not essential — if random, we can condition on X. (why?)

Mean and covariance of OLS

- Under **GM**, the OLS satisfies

$$\mathbb{E}\,\widehat{\beta} = \beta, \quad \mathrm{cov}\,\widehat{\beta} = \sigma^2(X^\intercal X)^{-1}.$$

- What if X is not fixed?

Mean and covariance of $(\widehat{Y}, \widehat{\varepsilon})$

- Recall that H (onto $\mathcal{C}(X)$) and $I_n - H$ (onto $\mathcal{C}(X)^{\perp}$) are both projection matrices satisfying

$$HX = X, \quad (I - H)X = 0, \quad H(I_n - H) = (I_n - H)H = 0.$$

With

$$\begin{pmatrix} \widehat{Y} \\ \widehat{\varepsilon} \end{pmatrix} = \begin{pmatrix} H \\ I_n - H \end{pmatrix} Y = \begin{pmatrix} H \\ I_n - H \end{pmatrix} (X\beta + \varepsilon),$$

under **GM** we have

$$\mathbb{E} \begin{pmatrix} \widehat{Y} \\ \widehat{\varepsilon} \end{pmatrix} = \begin{pmatrix} X\beta \\ 0 \end{pmatrix},$$

$$\operatorname{cov} \begin{pmatrix} \widehat{Y} \\ \widehat{\varepsilon} \end{pmatrix} = \sigma^2 \begin{pmatrix} H & 0 \\ 0 & I_n - H \end{pmatrix}.$$

☞ For any i, j,

$$\operatorname{cov}(\widehat{Y}_i, \widehat{Y}_j) = \sigma^2 h_{ij}, \quad \operatorname{cov}(\widehat{\varepsilon}_i, \widehat{\varepsilon}_j) = \sigma^2(\mathbb{I}\{i = j\} - h_{ij}), \quad \operatorname{cov}(Y_i, \widehat{\varepsilon}_j) = 0.$$

Estimating σ^2

It is natural to estimate σ^2 based on the **residual sum of squares**

$$\boxed{\mathrm{RSS} := \sum_{i=1}^{n} \widehat{\varepsilon}_i^2}$$

We have

$$\mathbb{E}\,\mathrm{RSS} = \mathbb{E}\sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \sigma^2 \sum_i (1 - h_{ii}) = \sigma^2(n - \mathrm{Tr}(H)) = \sigma^2(n-p) \quad \textit{(why?)}$$

Theorem $\widehat{\sigma}^2 := \mathrm{RSS}/(n-p)$ is an unbiased estimator of σ^2 under **GM**.

Gauss–Markov theorem

- Question unanswered so far — why should we focus on OLS?
- The next theorem establishes that OLS $\widehat{\beta}$ is the Best Linear Unbiased Estimator (BLUE) for β under **GM**.

☞ Recall that '$\succeq$' is positive semidefinite order. For real, symmetric A, B,

$$A \succeq B \iff A - B \succeq \mathbf{0} \iff c^{\mathsf{T}}(A - B)c \geq 0 \text{ for every } c.$$

- Natural notion for comparing covariances.

Gauss–Markov Theorem. Under **GM**, let $\widetilde{\beta}$ be any linear, unbiased estimator of β in the sense that

1. $\widetilde{\beta} = AY$ for some $A \in \mathbb{R}^{p \times n}$ that does not depend on Y, (linear in what?)
2. $\mathbb{E}\,\widetilde{\beta} = \beta$ for every β.

Then the OLS $\widehat{\beta}$ satisfies

$$\operatorname{cov}\widetilde{\beta} \succeq \operatorname{cov}\widehat{\beta}.$$

- $\implies \operatorname{var}\widetilde{\beta}_j \geq \operatorname{var}\widehat{\beta}_j$ (why?)

Proof.

1. OLS is linear,
2. unbiased.
3. Covariance comparison.

BLUE, necessarily good?

1. OLS is BLUE under **GM**, which is a restrictive model. In particular, it assumes homoskedasticity $\operatorname{var} \varepsilon_i^2 \equiv \sigma^2$. ▶ homo-*skedastikos* (Greek, disperse)

   Under heteroskedasticity $\operatorname{cov}(\varepsilon) = \Sigma$, it makes more sense to weigh observations inverse proportionally to Σ:

   $$\text{Generalized least squares (GLS): } \widehat{\beta}_\Sigma = (X^\intercal \Sigma^{-1} X)^{-1} X^\intercal \Sigma^{-1} Y.$$

   - GLS is also linear and unbiased. (why?)

2. Unbiased estimator is important in classic statistics (e.g., U-stat).
   In terms of estimation error, it can be worse than a biased estimator when p is large.

BIOST/STAT 533, Sp 2024
Theory of Linear Models

Richard Guo

Lecture # 5: Normal linear model: inference and prediction
§5, Appendix B

Recall: Gauss–Markov model

GM The data generating process obeys

$$Y = X\beta + \varepsilon,$$

where

1. X is fixed and has linearly independent columns,
2. $\mathbb{E}\,\varepsilon = \mathbf{0}$, $\operatorname{cov}\varepsilon = \sigma^2 I_n$.

The unknown parameters are (β, σ^2).

Theorem Under **GM**, OLS is **BLUE**.

Gauss–Markov–Normal model

GM-$\mathcal{N}$ The data generating process obeys

$$Y = X\beta + \varepsilon,$$

where

1. X is fixed and has linearly independent columns,
2. $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$.

The unknown parameters are (β, σ^2).

☞ Or equivalently,

$$\boxed{Y \sim \mathcal{N}(X\beta, \sigma^2 I_n).}$$

▶ **GM-$\mathcal{N}$** implies **GM**

Distributions, finite sample

Under **GM-$\mathcal{N}$**,

$$\begin{pmatrix} \widehat{\beta} \\ \widehat{\varepsilon} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta \\ \mathbf{0} \end{pmatrix}, \sigma^2 \begin{pmatrix} (X^\intercal X)^{-1} & 0 \\ 0 & I_n - H \end{pmatrix} \right),$$

and with $\widehat{\sigma}^2 = \|\widehat{\varepsilon}\|^2/(n-p)$,

$$\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}/(n-p).$$

It holds that

$$\widehat{\beta} \perp\!\!\!\perp \widehat{\varepsilon}, \quad \widehat{\beta} \perp\!\!\!\perp \widehat{\sigma}^2$$

▶ $\widehat{\sigma}^2$ is unbiased (why?)

Distributions, finite sample

Under **GM-$\mathcal{N}$**,

$$\begin{pmatrix} \widehat{Y} \\ \widehat{\varepsilon} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} X\beta \\ \mathbf{0} \end{pmatrix}, \sigma^2 \begin{pmatrix} H & 0 \\ 0 & I_n - H \end{pmatrix} \right).$$

- $\widehat{\varepsilon}$'s distribution is degenerate. (why?)

Inference for scalar $c^\intercal \beta$

- **Pivot**: A real-valued quantity $f(\text{statistic}, \text{unknown})$ (called a 'root') whose distribution is known — a bridge for inference.

$$c^\intercal \widehat{\beta} \sim \mathcal{N}(c^\intercal \beta, \sigma^2 c^\intercal (X^\intercal X)^{-1} c)$$

Pivot for $c^\intercal \beta$. Under **GM-$\mathcal{N}$**,

$$T_c := \frac{c^\intercal \widehat{\beta} - c^\intercal \beta}{\sqrt{\widehat{\sigma}^2 c^\intercal (X^\intercal X)^{-1} c}} \sim t_{n-p}.$$

Finite-sample CI can be constructed from

$$\mathbb{P}\{|T_c| \le t_{1-\alpha/2, n-p}\} = 1 - \alpha.$$

Quadratic forms of MVN

Theorem B.10

1. If $Y \sim \mathcal{N}(\mu, \Sigma)$,

$$(Y-\mu)^\top \Sigma^\dagger (Y-\mu) \sim \chi^2_k, \quad k = \mathrm{rank}(\Sigma)$$

2. If $Y \sim \mathcal{N}(0, I_n)$ and H is a projection matrix of rank k, then

$$Y^\top H Y \sim \chi^2_k.$$

3. If $Y \sim \mathcal{N}(0, H)$ and H is a projection matrix of rank k, then

$$Y^\top Y \sim \chi^2_k.$$

Inference for vector $C\beta$

- For $C : l \times p$, consider inferring $C\beta \in \mathbb{R}^p$.

$$\boxed{C(\widehat{\beta} - \beta) \sim \mathcal{N}(\mathbf{0}, \sigma^2 C(X^\intercal X)^{-1} C^\intercal)}$$

☞

$$(C\widehat{\beta} - C\beta)^\intercal \left\{\sigma^2 C(X^\intercal X)^{-1} C^\intercal\right\}^{-1} (C\widehat{\beta} - C\beta) \sim \chi^2_l,$$

if we assume C has linearly independent rows. (why?)

Pivot for $C\beta$. Suppose $C \in \mathbb{R}^{l\times p}$ has linearly independent rows. Under **GM-$\mathcal{N}$**,

$$F_C := \frac{(C\widehat{\beta} - C\beta)^\intercal \left\{C(X^\intercal X)^{-1} C^\intercal\right\}^{-1} (C\widehat{\beta} - C\beta)}{l\widehat{\sigma}^2} \sim F_{l,n-p}.$$

Quadratic forms of random vectors

Theorem B.8 If a random vector Y has mean μ and covariance Σ, then

$$\mathbb{E}\, Y^\intercal A Y = \operatorname{Tr}(A\Sigma) + \mu^\intercal A \mu.$$

Related distributions

$$\mathrm{Ga}(k/2, 1/2) =_d \chi^2_k = \underbrace{N(0,1)^2 + \cdots + N(0,1)^2}_{k \text{ times}}$$

$$t_k = \frac{N(0,1)}{\sqrt{\chi^2_k/k}}$$

$$F_{k,l} = \frac{\chi^2_k/k}{\chi^2_l/l},$$

$$\text{so } t_l^2 = F_{1,l}$$

with appropriate independence between relevant random variables.

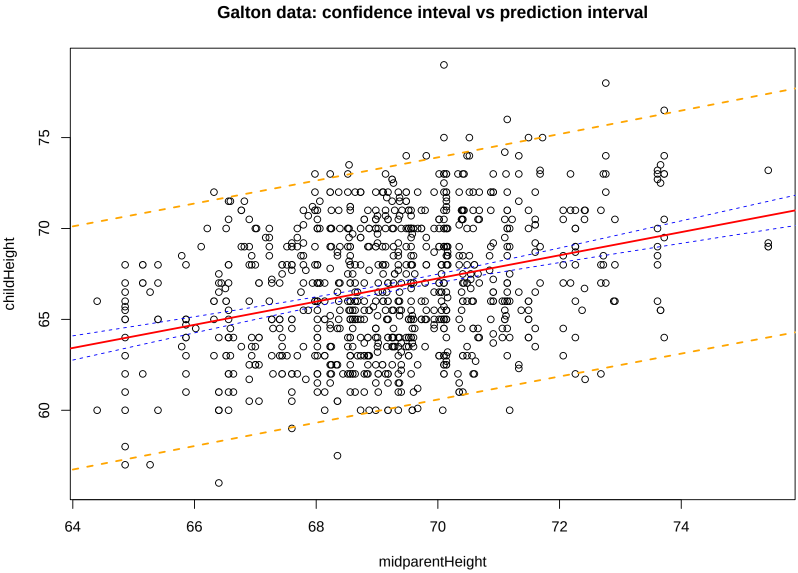
Prediction

Want to get a prediction interval on a new observation

$$y_{n+1} = x_{n+1}^{\mathsf{T}}\beta + \varepsilon_{n+1}, \quad \varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2).$$

Theorem Under **GM-$\mathcal{N}$**, we have the following pivot for prediction:

$$\frac{y_{n+1} - x_{n+1}^{\mathsf{T}}\widehat{\beta}}{\sqrt{\widehat{\sigma}^2 + \widehat{\sigma}^2 x_{n+1}^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}x_{n+1}}} \sim t_{n-p}.$$



R/week-2-Galton.R

BIOST/STAT 533, Sp 2024
Theory of Linear Models

Richard Guo

Lecture # 6: Asymptotic inference of OLS: heteroskedastic errors
§6, Appendix C.2

Recall: finite-sample inference under Normal linear model

GM-$\mathcal{N}$ The data generating process obeys

$$Y = X\beta + \varepsilon,$$

where

1. X is fixed and has linearly independent columns,
2. $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. i.e., $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$

The unknown parameters are (β, σ^2).

- $\widehat{\beta} - \beta \sim \mathcal{N}(0, \sigma^2(X^\intercal X)^{-1}) \quad \perp\!\!\!\perp \quad \widehat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}/(n-p)$.
- Constructing t or F pivots
- Efficiency: OLS $\widehat{\beta}$ is MLE and **BLUE**

Heteroskedastic linear model

Hetero The data generating process obeys

$$Y = X\beta + \varepsilon, \quad \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\intercal,$$

where

1. X is fixed and has linearly independent columns,
2. ε_i's are independent with $\mathbb{E}\,\varepsilon_i = 0$, $\operatorname{var}\varepsilon_i = \sigma_i^2$

The unknown parameters are $(\beta, \sigma_1^2, \ldots, \sigma_n^2)$.

☞ The errors might not be normal (though still cannot be arbitrary). (why?)

☞ Compared to **GM**, the errors are independent (though not iid).

Heteroskedastic linear model: a simulation

		$\widehat{\text{SE}}_{\text{R}}$	$\widehat{\text{SE}}_{\text{hccm}}$
Homoskedastic	normal		
	exp		
Heteroskedastic	normal		
	unif		

`R/HuberWhite.R`

Heteroskedastic linear model: a simulation

		$\widehat{\mathrm{SE}}_{\mathrm{R}}$	$\widehat{\mathrm{SE}}_{\mathrm{hccm}}$
Homoskedastic	normal	✓	✓
	exp	✓	✓
Heteroskedastic	normal	✗	✓
	unif	✗	✓

`R/HuberWhite.R`

OLS: asymptotic expansion

Hetero The data generating process obeys $Y = X\beta + \varepsilon$, $\quad \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\intercal$, where

1. X is fixed and has linearly independent columns,
2. ε_i's are independent with $\mathbb{E}\,\varepsilon_i = 0$, $\operatorname{var}\varepsilon_i = \sigma_i^2$

The unknown parameters are $(\beta, \sigma_1^2, \ldots, \sigma_n^2)$.

Lemma Under **Hetero**, we have $\widehat{\beta} - \beta = B_n^{-1}\xi_n$ with

$$B_n = n^{-1}\sum_{i=1}^n x_i x_i^\intercal, \quad \xi_n = n^{-1}\sum_{i=1}^n x_i\varepsilon_i$$

☞ $\mathbb{E}\,\widehat{\beta} = \beta$ (why?)

☞ When n is large, $\widehat{\beta} - \beta \approx B^{-1}\xi_n = B^{-1}\underbrace{(n^{-1}\sum_i x_i\varepsilon_i)}_{\text{avg of ind terms}}$.

OLS: consistency

Assumption 1. We have

$$B_n := n^{-1}\sum_{i=1}^{n} x_i x_i^{\mathsf{T}} \to B, \quad M_n := n^{-1}\sum_{i=1}^{n} \sigma_i^2 x_i x_i^{\mathsf{T}} \to M,$$

where B is invertible.

Theorem Under **Hetero** and **Assumption 1**, $\widehat{\beta}$ is consistent for β.

OLS: asymptotic normality

Assumption 1 (good limits). $B_n \to B$, $M_n \to M$, where B is invertible.

Assumption 2 (moment condition). For some $\delta > 0$ and $C > 0$, it holds that

$$d_{2+\delta,n} := n^{-1} \sum_{i=1}^{n} \|x_i\|^{2+\delta}\, \mathbb{E}\, |\varepsilon_i|^{2+\delta} < C \quad \text{for all } n.$$

Theorem Consider **Hetero** model. Under **Assumption 1** and **2**, we have

$$\sqrt{n}(\widehat{\beta} - \beta) \to_d \mathcal{N}(\mathbf{0}, B^{-1} M B^{-1}).$$

- Approximately,

$$\widehat{\beta} - \beta \overset{a}{\sim} \mathcal{N}(\mathbf{0}, n^{-1} B^{-1} M B^{-1}),$$

where $n^{-1} B^{-1} M B^{-1}$ is the **standard error** of $\widehat{\beta}$.

Lindeberg-Feller CLT

- Triangular array $(Z_{n,1}, \dots, Z_{n,k_n})$ with $k_n \to \infty$ as $n \to \infty$, e.g.,

$$\begin{array}{cccc} Z_{1,1} & & & \\ Z_{2,1} & Z_{2,2} & & \\ Z_{3,1} & Z_{3,2} & \cdots & Z_{3,3} \\ \vdots & & & \ddots \end{array}$$

Theorem For each n, let $Z_{n,1}, \dots, Z_{n,k_n}$ be independent random variables with finite variances such that

(LF-1) $\sum_{i=1}^{k_n} \mathbb{E}\left[\|Z_{n,i}\|^2 \mathbb{I}\{\|Z_{n,i}\| > c\}\right] \to 0$ for every $c > 0$,

(LF-2) $\sum_{i=1}^{k_n} \operatorname{cov} Z_{n,i} \to \Sigma$.

Then, $\sum_{i=1}^{k_n}(Z_{n,i} - \mathbb{E}\, Z_{n,i}) \to_d \mathcal{N}(\mathbf{0}, \Sigma)$. The result still holds if (LF-1) is replaced by

$$\text{(LF-1')} \quad \sum_{i=1}^{k_n} \mathbb{E}\, \|Z_{n,i}\|^{2+\delta} \to 0 \text{ for some } \delta > 0.$$

☞ See also Theorem 2.2 in Shorack (2017), Probability for Statisticians.

BIOST/STAT 533, Sp 2024
Theory of Linear Models

Richard Guo

Lecture # 7: Asymptotic inference of OLS: Eicker–Huber–White
§6.4–6.5

Recall: ASN under heteroskedastic errors

Hetero The data generating process obeys $Y = X\beta + \varepsilon, \quad \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, where

1. X is fixed and has linearly independent columns,
2. ε_i's are independent with $\mathbb{E}\,\varepsilon_i = 0$, $\operatorname{var}\varepsilon_i = \sigma_i^2$

The unknown parameters are $(\beta, \sigma_1^2, \ldots, \sigma_n^2)$.

☞ Is OLS still **BLUE**? (why?)

Theorem Consider **Hetero** model. Under

$$\text{(A1) (good limits)} \quad B_n := n^{-1}\sum_{i=1}^n x_i x_i^\top \to B \text{ (full rank)}, \quad M_n := n^{-1}\sum_{i=1}^n \sigma_i^2 x_i x_i^\top \to M$$

$$\text{(A2) (moment condition)} \quad d_{2+\delta,n} := n^{-1}\sum_{i=1}^n \|x_i\|^{2+\delta}\,\mathbb{E}\,|\varepsilon_i|^{2+\delta} < C \quad \text{for } \delta > 0,\ C > 0,$$

$$\sqrt{n}(\widehat{\beta} - \beta) \to_d \mathcal{N}(\mathbf{0}, B^{-1}MB^{-1}).$$

Estimating asymptotic covariance

The asymptotic covariance

$$\boxed{\Sigma = B^{-1} M B^{-1}.}$$

- B is the limit of B_n and can be naturally estimated by $B_n = n^{-1} X^\intercal X$.
- For $M = n^{-1} \sum_i \sigma_i^2 x_i x_i^\intercal = n^{-1} X^\intercal \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2) X$, an ideal (but infeasible) estimator is

$$\widetilde{M}_n := n^{-1} \sum_i \varepsilon_i^2 x_i x_i^\intercal.$$

  - unbiased (why?)

- Replace it with

$$\widehat{M}_n := n^{-1} \sum_i \widehat{\varepsilon}_i^2 x_i x_i^\intercal.$$

  - Recall: $\widehat{\varepsilon} = (I - H)Y$
  - To show its consistency, suffices to show $\widetilde{M}_n \to M$ and $\widehat{M}_n - \widetilde{M}_n \to_p \mathbf{0}$. (why?)

Consistency of $\widehat{M}_n$

Theorem Consider **Hetero** model. Suppose it holds that

$$\text{(A1) (good limits)} \quad B_n := n^{-1}\sum_{i=1}^{n} x_i x_i^{\mathsf{T}} \to B \text{ (full rank)}, \quad M_n := n^{-1}\sum_{i=1}^{n} \sigma_i^2 x_i x_i^{\mathsf{T}} \to M.$$

We have $\widehat{M}_n \to_p M$ if the following **(A3) (extra moment condition)** holds:

$$n^{-1}\sum_i \text{var}(\varepsilon_i^2) x_{i,j_1}^2 x_{i,j_2}^2, \quad n^{-1}\sum_i |x_{i,j_1} x_{i,j_2} x_{i,j_3} x_{i,j_4}|, \quad n^{-2}\sum_i \sigma_i^2 x_{i,j_1}^2 x_{i,j_2}^2 x_{i,j_3}^2$$

are bounded above by some constant C for all n and every $j_1, j_2, j_3, j_4 \in \{1, \ldots, p\}$.

☞ Then,

$$\boxed{\widehat{\Sigma}_{\text{EHW}} := B_n^{-1}\widehat{M}_n B_n^{-1} \to_p \Sigma = B^{-1} M B^{-1}.}$$

Consistency of $\widehat{M}_n$

Proof.

$$\widehat{M}_n - M = \widetilde{M}_n - M + \widehat{M}_n - \widetilde{M}_n$$

1. $\widetilde{M}_n \to_p M$.
2. $\widehat{M}_n - \widetilde{M}_n \to_p \mathbf{0}$.

Eicker–Huber–White

- Consistent estimator of asymptotic covariance

$$\widehat{\Sigma}_{\mathrm{EHW}} = \Bigg(\overbrace{n^{-1}\sum_i x_i x_i^{\mathsf{T}}}^{B_n}\Bigg)^{-1} \Bigg(\overbrace{n^{-1}\sum_i \widehat{\varepsilon}_i^2 x_i x_i^{\mathsf{T}}}^{\widehat{M}_n}\Bigg) \Bigg(\overbrace{n^{-1}\sum_i x_i x_i^{\mathsf{T}}}^{B_n}\Bigg)^{-1}$$

$$= n\,\underbrace{(X^{\mathsf{T}}X)^{-1}(X^{\mathsf{T}}\widehat{\Omega}X)(X^{\mathsf{T}}X)^{-1}}_{\widehat{V}_{\mathrm{EHW}}}, \qquad \widehat{\Omega} = \mathrm{diag}(\widehat{\varepsilon}_1^2, \ldots, \widehat{\varepsilon}_n^2).$$

- Convergence in distribution: $\widehat{\Sigma}_{\mathrm{EHW}}^{-1/2}\sqrt{n}(\widehat{\beta} - \beta) \to_d \mathcal{N}(0, I_p)$. (why?)
- Approximately,

$$\widehat{\beta} \overset{\mathrm{a}}{\sim} \mathcal{N}(\beta, \widehat{V}_{\mathrm{EHW}}).$$

☞ $\widehat{V}_{\mathrm{EHW}}$ yields robust/sandwich or HC (heteroskedasticity-consistent) standard errors:

$$\widehat{\mathrm{SE}}_{\mathrm{EHW}}(\beta_j) = \sqrt{(\widehat{V}_{\mathrm{EHW}})_{j,j}}, \quad j = 1, \ldots, p.$$

Eicker–Huber–White: HC variants

$$\widehat{\Sigma}_{\mathrm{EHW},k}/n = \widehat{V}_{\mathrm{EHW},k} = (X^\top X)^{-1} \left(X^\top \operatorname{diag}(\widehat{\varepsilon}^2_{1,k}, \ldots, \widehat{\varepsilon}^2_{n,k})X\right) (X^\top X)^{-1},$$

with

$$\widehat{\varepsilon}_{i,k} = \begin{cases} \widehat{\varepsilon}_i, & \text{HC0} \\ \widehat{\varepsilon}_i\sqrt{n/(n-p)}, & \text{HC1} \\ \widehat{\varepsilon}_i/\sqrt{1-h_{ii}}, & \text{HC2} \\ \widehat{\varepsilon}_i/(1-h_{ii}), & \text{HC3} \\ \widehat{\varepsilon}_i/(1-h_{ii})^{\min\{2, nh_{ii}/(2p)\}}, & \text{HC4} \end{cases}$$

- HC0: vanilla
- HC1: d.o.f. correction
- HC2: unbiased under homoskedasticity
- HC3: jackknife

☞ For practice, maybe consider HC2 or HC3.

Special case: homoskedastic

- When $\sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2$,

$$\sqrt{n}(\widehat{\beta} - \beta) \to_d \mathcal{N}(0, B^{-1}MB^{-1}) = \mathcal{N}(0, \sigma^2 B^{-1})$$

(why?)

Theorem Consider **Hetero** model. Suppose it holds that

$$\text{(A1) (good limits)} \quad B_n := n^{-1}\sum_{i=1}^{n} x_i x_i^{\mathsf{T}} \to B \text{ (full rank)}, \quad M_n := n^{-1}\sum_{i=1}^{n} \sigma_i^2 x_i x_i^{\mathsf{T}} \to M.$$

Further, if

$$\sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2 \text{ and } n^{-1}\sum_i \mathrm{var}(\varepsilon_i^2) \text{ is bounded},$$

then

$$\widehat{\sigma}^2 := \mathrm{RSS}/(n-p) \to_p \sigma^2.$$

- We already know that $\widehat{\sigma}^2$ is unbiased. (why?)

Review: models

GM We have $Y = X\beta + \varepsilon$ with

1. X is fixed and has linearly independent columns,
2. $\mathbb{E}\,\varepsilon = \mathbf{0}$, $\operatorname{cov}\varepsilon = \sigma^2 I_n$.

The unknown parameters are (β, σ^2).

- Errors need not be independent.

GM-$\mathcal{N}$ We have $Y = X\beta + \varepsilon$ with

1. X is fixed and has linearly independent columns,
2. $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$.

The unknown parameters are (β, σ^2).

- Errors are iid.

Hetero We have $Y = X\beta + \varepsilon, \quad \varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\mathsf{T}$, where

1. X is fixed and has linearly independent columns,
2. ε_i's are independent with $\mathbb{E}\,\varepsilon_i = 0$, $\operatorname{var}\varepsilon_i = \sigma_i^2$

The unknown parameters are $(\beta, \sigma_1^2, \dots, \sigma_n^2)$.

- Errors are independent.

BIOST/STAT 533, Sp 2024
Theory of Linear Models

Richard Guo

Lecture # 8: Partial regression and Frisch–Waugh–Lovell Theorem
§7

Recall: ASN under heteroskedastic errors

Hetero The data generating process obeys $Y = X\beta + \varepsilon$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\intercal$, where

1. X is fixed and has linearly independent columns,
2. ε_i's are independent with $\mathbb{E}\,\varepsilon_i = 0$, $\operatorname{var}\varepsilon_i = \sigma_i^2$

The unknown parameters are $(\beta, \sigma_1^2, \ldots, \sigma_n^2)$.

☞ Is OLS still **BLUE**? (why?)

Theorem Consider **Hetero** model. Under

$$\text{(A1) (good limits)} \quad B_n := n^{-1}\sum_{i=1}^{n} x_i x_i^\intercal \to B \text{ (full rank)}, \quad M_n := n^{-1}\sum_{i=1}^{n} \sigma_i^2 x_i x_i^\intercal \to M$$

$$\text{(A2) (moment condition)} \quad d_{2+\delta,n} := n^{-1}\sum_{i=1}^{n} \|x_i\|^{2+\delta}\,\mathbb{E}\,|\varepsilon_i|^{2+\delta} < C \quad \text{for } \delta > 0,\ C > 0,$$

$$\sqrt{n}(\widehat{\beta} - \beta) \to_d \mathcal{N}(\mathbf{0}, B^{-1}MB^{-1}).$$

Recall: Consistency of $\widehat{\Sigma}_{\text{EHW}}$

$$
\begin{aligned}
\widehat{\Sigma}_{\text{EHW}} &= B_n^{-1}\widehat{M_n}B_n^{-1} \\
&= \left(n^{-1}\sum_i x_i x_i^\top\right)^{-1}\left(n^{-1}\sum_i \widehat{\varepsilon}_i^2 x_i x_i^\top\right)\left(n^{-1}\sum_i x_i x_i^\top\right)^{-1} \\
&= n\,\underbrace{(X^\top X)^{-1}(X^\top\widehat{\Omega}X)(X^\top X)^{-1}}_{\widehat{V}_{\text{EHW}}}, \qquad \widehat{\Omega} = \text{diag}(\widehat{\varepsilon}_1^2,\ldots,\widehat{\varepsilon}_n^2).
\end{aligned}
$$

We have $\widehat{\Sigma}_{\text{EHW}} \to_p \Sigma = B^{-1}MB^{-1}$ under (A1) (good limits) and (A3) (extra moment conditions):

$$
n^{-1}\sum_i \text{var}(\varepsilon_i^2)x_{i,j_1}^2 x_{i,j_2}^2 \le C, \quad n^{-1}\sum_i |x_{i,j_1}x_{i,j_2}x_{i,j_3}x_{i,j_4}| \le C, \quad n^{-2}\sum_i \sigma_i^2 x_{i,j_1}^2 x_{i,j_2}^2 x_{i,j_3}^2 \le C.
$$

Long and short regressions

Suppose $X : n \times (k + l)$ has linearly independent columns. Partition X and β into

$$X = (\underbrace{X_1}_{n \times k}, \underbrace{X_2}_{n \times l}), \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Long regression $Y \sim X_1 + X_2$

$$\begin{aligned} Y &= X\widehat{\beta} + \widehat{\varepsilon} \\ &= X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + \widehat{\varepsilon}. \end{aligned}$$

Short regression $Y \sim X_2$

$$Y = X_2\widetilde{\beta}_2 + \widetilde{\varepsilon}.$$

FWL theorem

Short regression $Y = \underbrace{X_2}_{n\times l}\widetilde{\beta}_2 + \widetilde{\varepsilon}$:

$$\widetilde{\beta}_2 = (X_2^\intercal X_2)^{-1} X_2^\intercal Y.$$

Long regression $Y = \underbrace{X_1}_{n\times k}\widehat{\beta}_1 + \underbrace{X_2}_{n\times l}\widehat{\beta}_2 + \widehat{\varepsilon}$

Frisch–Waugh–Lovell Theorem Suppose X has linearly independent columns. In the long regression, the OLS for β_2 has the following equivalent forms:

$$\begin{aligned}
\widehat{\beta}_2 &= [(X^\intercal X)^{-1} X^\intercal Y]_{(k+1):(k+l)} \\
&= \{X_2^\intercal (I_n - H_1) X_2\}^{-1} X_2^\intercal (I_n - H_1) Y, \quad \text{where } H_1 = X_1 (X_1^\intercal X_1)^{-1} X_1^\intercal \\
&= (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1} \widetilde{X}_2^\intercal Y, \quad \text{where } \widetilde{X}_2 = (I_n - H_1) X_2 \\
&= (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1} \widetilde{X}_2^\intercal \widetilde{Y}, \quad \text{where } \widetilde{Y} = (I_n - H_1) Y.
\end{aligned}$$

FWL theorem

Frisch–Waugh–Lovell Theorem Suppose X has linearly independent columns. In the long regression, the OLS for β_2 has the following equivalent forms:

$$
\begin{aligned}
\widehat{\beta}_2 &= [(X^\intercal X)^{-1} X^\intercal Y]_{(k+1):(k+l)} \\
&= \{X_2^\intercal (I_n - H_1) X_2\}^{-1} X_2^\intercal (I_n - H_1) Y, \quad \text{where } H_1 = X_1 (X_1^\intercal X_1)^{-1} X_1^\intercal \\
&= (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1} \widetilde{X}_2^\intercal Y, \quad \text{where } \widetilde{X}_2 = (I_n - H_1) X_2 \\
&= (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1} \widetilde{X}_2^\intercal \widetilde{Y}, \quad \text{where } \widetilde{Y} = (I_n - H_1) Y.
\end{aligned}
$$

- $\widetilde{X}_2$ is the residual matrix from columnwise OLS fit of X_2 on X_1.
- $\widetilde{Y}$ is the residual from OLS fit of Y on X_2.
- $\widehat{\beta}_2$ is the OLS from $\widetilde{Y} \sim \widetilde{X}_2$ (partial regression)
- $\widehat{\beta}_2$ is also the OLS from $Y \sim \widetilde{X}_2$ (no need to residualize Y).

☞ That is, short regression but with X_2 replaced by $\widetilde{X}_2$

Proof I: using inverse of 2×2 block matrix

Recall from HW 3,

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B\widetilde{D}^{-1}CA^{-1} & -A^{-1}B\widetilde{D}^{-1} \\ -\widetilde{D}^{-1}CA^{-1} & \widetilde{D}^{-1} \end{pmatrix},$$

where $\widetilde{D} = D - CA^{-1}B$ is the Schur complement of A.

Lemma We have

$$(X^\intercal X)^{-1} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},$$

where

$$\begin{aligned} S_{11} &= (X_1^\intercal X_1)^{-1} + (X_1^\intercal X_1)^{-1} X_1^\intercal X_2 (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1} X_2^\intercal X_1 (X_1^\intercal X_1)^{-1}, \\ S_{12} &= -(X_1^\intercal X_1)^{-1} X_1^\intercal X_2 (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1}, \\ S_{21} &= S_{12}^\intercal, \\ S_{22} &= (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1}. \end{aligned}$$

Proof II: using orthogonality

Properties

Lemma Let

$$\widetilde{H}_2 := \widetilde{X}_2(\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1}\widetilde{X}_2^\intercal.$$

We have

$$H_1\widetilde{H}_2 = \widetilde{H}_2 H_1 = 0, \quad H = H_1 + \widetilde{H}_2.$$

☞ $H \neq H_1 + H_2$ in general!

Corollary Long regression $Y \sim X$ and the partial regression $\widetilde{Y} \sim \widetilde{X}_2$ have the same residuals.

Corollary (under orthogonality) When $X_1^\intercal X_2 = 0$, i.e., $\mathcal{C}(X_1) \perp \mathcal{C}(X_2)$, we have

$$\widetilde{X}_2 = X_2,$$

$$\widehat{\beta}_2 \text{ from } Y \sim X_1 + X_2 = \widetilde{\beta}_2 \text{ from } Y \sim X_2.$$

Gram–Schmidt

- Projection of $V_2 \in \mathbb{R}^n$ on $V_1 \in \mathbb{R}^n$:

$$\widehat{\beta}_{V_2|V_1} V_1 = \underbrace{V_1(V_1^\mathsf{T} V_1)^{-1} V_1^\mathsf{T}}_{H_{V_1}} V_2$$

Gram–Schmidt orthogonalization: Sequentially orthogonalize $X = (X_1, \ldots, X_p)$ to orthogonal vectors $(U_1, \ldots, U_p)$ such that $\mathcal{C}(U_1, \ldots, U_m) = \mathcal{C}(X_1, \ldots, X_m)$ for $m = 1, \ldots, p$.

1. $X_1 = U_1$
2. $X_2 = \widehat{\beta}_{X_2|U_1} U_1 + U_2$ $\qquad U_2 \perp U_1$ (why?)
3. $X_3 = \widehat{\beta}_{X_3|U_1} U_1 + \widehat{\beta}_{X_3|U_2} U_2 + U_3$ $\qquad$ (why?)

   $\vdots$ $\qquad \mathcal{C}(X_1, X_2, X_3) = \mathcal{C}(U_1, U_2, U_3)$ (why?)

- $X_p = \sum_{j=1}^{p-1} \widehat{\beta}_{p|U_j} U_j + U_p$.

QR decomposition

Normalization

$$Q_j = U_j / \|U_j\|, \quad j = 1, \ldots, p.$$

QR decomposition

$$\begin{aligned}
X &= (X_1, \ldots, X_p) \\
&= (U_1, \ldots, U_p) \begin{pmatrix} 1 & \widehat{\beta}_{X_2|U_1} & \widehat{\beta}_{X_3|U_1} & \cdots & \widehat{\beta}_{X_p|U_1} \\ 0 & 1 & \widehat{\beta}_{X_3|U_2} & \cdots & \widehat{\beta}_{X_p|U_2} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \\
&= Q \operatorname{diag}(\|U_1\|, \ldots, \|U_p\|) \begin{pmatrix} 1 & \widehat{\beta}_{X_2|U_1} & \widehat{\beta}_{X_3|U_1} & \cdots & \widehat{\beta}_{X_p|U_1} \\ 0 & 1 & \widehat{\beta}_{X_3|U_2} & \cdots & \widehat{\beta}_{X_p|U_2} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \\
&= Q\,R.
\end{aligned}$$

Solving OLS

- Instead of inverting $X^\intercal X$ (numerically unstable), R solves OLS using $X = QR$:

$$\begin{aligned} X^\intercal X\widehat{\beta} &= X^\intercal Y \\ R^\intercal Q^\intercal QR\widehat{\beta} &= R^\intercal Q^\intercal Y \\ R^\intercal R\widehat{\beta} &= R^\intercal Q^\intercal Y \\ R\widehat{\beta} &= Q^\intercal Y, \end{aligned}$$

then backsolves $\widehat{\beta}$.

BIOST/STAT 533, Sp 2024
Theory of Linear Models

Richard Guo

Lecture # 9: Applications of FWL; ANOVA and Wald; Midterm Review
§8

Recall: FWL theorem

Short regression $Y = \underbrace{X_2}_{n\times l}\widetilde{\beta}_2 + \widetilde{\varepsilon}$

Long regression $Y = \underbrace{X_1}_{n\times k}\widehat{\beta}_1 + \underbrace{X_2}_{n\times l}\widehat{\beta}_2 + \widehat{\varepsilon}$

Frisch–Waugh–Lovell Theorem Suppose X has linearly independent columns. In the long regression, the OLS for β_2 has the following equivalent forms:

$$\begin{aligned}
\widehat{\beta}_2 &= [(X^\intercal X)^{-1}X^\intercal Y]_{(k+1):(k+l)} \\
&= \{X_2^\intercal(I_n - H_1)X_2\}^{-1}X_2^\intercal(I_n - H_1)Y, \quad \text{where } H_1 = X_1(X_1^\intercal X_1)^{-1}X_1^\intercal \\
&= (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1}\widetilde{X}_2^\intercal Y, \quad \text{where } \widetilde{X}_2 = (I_n - H_1)X_2 \\
&= (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1}\widetilde{X}_2^\intercal \widetilde{Y}, \quad \text{where } \widetilde{Y} = (I_n - H_1)Y.
\end{aligned}$$

Application of FWL: intercept and centering

- Coefficients in $Y \sim 1 + X$ can be obtained from $Y \sim (I_n - H_{\mathbf{1}})X$, where

$$H_{\mathbf{1}} = 1_n(1_n^\intercal 1_n)^{-1} 1_n = n^{-1} 1_n 1_n = \begin{pmatrix} n^{-1} & \dots & n^{-1} \\ \vdots & \dots & \vdots \\ n^{-1} & \dots & n^{-1} \end{pmatrix}.$$

- $H_{\mathbf{1}} y = \begin{pmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix}$.
- $(I_n - H_{\mathbf{1}})y = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}$.

☞ Centering

- So what is $(I_n - H_{\mathbf{1}})X$?

Application of FWL: intercept and centering

- $y^\intercal(I_n - H_{\mathbf{1}})y = [(I_n - H_{\mathbf{1}})y]^\intercal \, (I_n - H_{\mathbf{1}})y = \sum_i (y_i - \bar{y})^2 = (n-1)\widehat{\sigma}_y^2$.
- For $X : n \times p$,

$$X^\intercal(I_n - H_{\mathbf{1}})X = (n-1)\begin{pmatrix} \widehat{\sigma}_{11}^2 & \cdots & \widehat{\sigma}_{1p}^2 \\ \vdots & \ldots & \vdots \\ \widehat{\sigma}_{p1}^2 & \cdots & \widehat{\sigma}_{pp}^2 \end{pmatrix}.$$

☞ $X^\intercal(I_n - H_{\mathbf{1}})X/(n-1)$ is the sample covariance.

- Why divide by $n-1$?

Simpson's paradox: correlation and partial correlation

- (Marginal) correlation between $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$:

$$\widehat{\rho}_{xy} = \frac{\langle x - \bar{x}, y - \bar{y} \rangle}{\|x - \bar{x}\| \, \|y - \bar{y}\|} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}.$$

  - $\widehat{\rho}_{xy} \in [-1, 1]$ (why?)

- Partial correlation between x and y given $W \in \mathbb{R}^{n \times p}$:

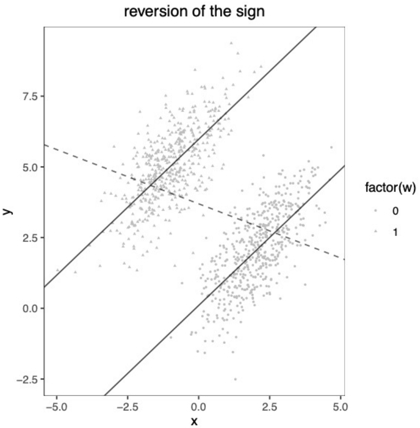
$$\widehat{\rho}_{xy|W} := \widehat{\rho}_{\widehat{\varepsilon}_{x|W}, \widehat{\varepsilon}_{y|W}},$$

where $\widehat{\varepsilon}_{x|W}, \widehat{\varepsilon}_{y|W}$ are respectively residuals from $x \sim 1 + W$ and $y \sim 1 + W$.

☞ This is the correlation between x and y while controlling for W, or after partialling out W.

Simpson's paradox: correlation and partial correlation

- Simpson's paradox: $\widehat{\rho}_{xy}$ and $\widehat{\rho}_{xy|W}$ can have opposite signs.



Hypothesis testing: Wald F-test

Consider a long regression

$$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon,$$

where $X_1 : n \times k$, $X_2 : n \times l$, $\beta_1 \in \mathbb{R}^k$ and $\beta_2 \in \mathbb{R}^l$.

- Want to test

$$\boxed{H_0 : \beta_2 = \mathbf{0}.}$$

Under **GM-$\mathcal{N}$** $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, we can use F-test with $C = (0_{l\times k}, I_l)$ so $C\beta = \beta_2$.

Recall:

Pivot for $C\beta$. Suppose $C \in \mathbb{R}^{l\times p}$ has linearly independent rows. Under **GM-$\mathcal{N}$**,

$$F_C := \frac{(C\widehat{\beta} - C\beta)^\intercal \left\{C(X^\intercal X)^{-1}C^\intercal\right\}^{-1}(C\widehat{\beta} - C\beta)}{l\widehat{\sigma}^2} \sim F_{l,n-p}.$$

Hypothesis testing: Wald F-test

Also, recall:

Lemma We have

$$(X^\intercal X)^{-1} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},$$

where

$$\begin{aligned} S_{11} &= (X_1^\intercal X_1)^{-1} + (X_1^\intercal X_1)^{-1} X_1^\intercal X_2 (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1} X_2^\intercal X_1 (X_1^\intercal X_1)^{-1}, \\ S_{12} &= -(X_1^\intercal X_1)^{-1} X_1^\intercal X_2 (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1}, \\ S_{21} &= S_{12}^\intercal, \\ S_{22} &= (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1}. \end{aligned}$$

- F-test (aka Wald) under **GM-$\mathcal{N}$**:

$$\boxed{F_{\text{Wald}} = \frac{\widehat{\beta}_2^\intercal (S_{22})^{-1} \widehat{\beta}_2}{l \widehat{\sigma}^2} = \frac{\widehat{\beta}_2^\intercal \widetilde{X}_2^\intercal \widetilde{X}_2 \widehat{\beta}_2}{l \widehat{\sigma}^2} \sim F_{l, n-p}.}$$

Hypothesis testing: ANOVA

Long regression: $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ $\qquad X_1 : n \times k,\ X_2 : n \times l$

- Under $H_0 : \beta_2 = \mathbf{0}$, it is reduced to
Short regression: $Y = X_1\beta_1 + \varepsilon$.

☞ Under H_0, the two regressions should have 'similar' RSS's:

$$\mathrm{RSS}_{\mathsf{long}} = Y^\intercal(I_n - H)Y, \quad \mathrm{RSS}_{\mathsf{short}} = Y^\intercal(I_n - H_1)Y.$$

- $\mathrm{RSS}_{\mathsf{long}} \le \mathrm{RSS}_{\mathsf{short}}$ (why?)

- R. A. Fisher proposed the following ANOVA (Analysis of Variance) statistic:

$$\boxed{F_{\mathsf{ANOVA}} := \frac{(\mathrm{RSS}_{\mathsf{short}} - \mathrm{RSS}_{\mathsf{long}})/l}{\mathrm{RSS}_{\mathsf{long}}/(n-p)} = \frac{\mathrm{RSS}_{\mathsf{short}} - \mathrm{RSS}_{\mathsf{long}}}{l\,\widehat{\sigma}^2}.}$$

Equivalence: Wald and ANOVA

Theorem Suppose X has linearly independent columns. Consider testing $H_0 : \beta_2 = \mathbf{0}$ in

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon, \quad X_1 : n \times k,\, X_2 : n \times l$$

with

$$F_{\text{Wald}} = \frac{\widehat{\beta}_2^{\mathsf{T}} (S_{22})^{-1} \widehat{\beta}_2}{l \widehat{\sigma}^2} = \frac{\widehat{\beta}_2^{\mathsf{T}} \widetilde{X}_2^{\mathsf{T}} \widetilde{X}_2 \widehat{\beta}_2}{l \widehat{\sigma}^2}$$

and

$$F_{\text{ANOVA}} := \frac{(\mathrm{RSS}_{\text{short}} - \mathrm{RSS}_{\text{long}})/l}{\mathrm{RSS}_{\text{long}}/(n-p)} = \frac{\mathrm{RSS}_{\text{short}} - \mathrm{RSS}_{\text{long}}}{l\, \widehat{\sigma}^2}.$$

1. Under **GM-$\mathcal{N}$**, $F_{\text{ANOVA}} \sim F_{l,n-p}$ under H_0.
2. In fact, for any X, Y without assuming **GM-$\mathcal{N}$**, $F_{\text{Wald}} = F_{\text{ANOVA}}$ numerically.

Equivalence: Wald and ANOVA

Proof.

1. Under **GM-$\mathcal{N}$**, $F_{\text{ANOVA}} \sim F_{l,n-p}$ under H_0.
2. $F_{\text{Wald}} = F_{\text{ANOVA}}$ numerically.

Review: OLS

$$\widehat{\beta} = \arg\min_b \sum_{i=1}^n (y_i - b^\intercal x_i)^2 = \arg\min_b \|Y - Xb\|^2.$$

- Normal equation

$$\sum_i (y_i - x_i^\intercal \widehat{\beta}) x_i = \mathbf{0} \iff X^\intercal (Y - X\widehat{\beta}) = 0 \iff X^\intercal Y = X^\intercal X \widehat{\beta}.$$

- Projection, orthogonal decomposition

$$H = X(X^\intercal X)^{-1} X^\intercal, \quad \widehat{Y} = HY, \quad \widehat{\varepsilon} = (I_n - H)Y.$$

$$\|Y\|^2 = \|\widehat{Y}\|^2 + \|\widehat{\varepsilon}\|^2.$$

Review: Gauss–Markov

GM We have $Y = X\beta + \varepsilon$ with

1. X is fixed and has linearly independent columns,
2. $\mathbb{E}\,\varepsilon = \mathbf{0}$, $\operatorname{cov}\varepsilon = \sigma^2 I_n$.

The unknown parameters are (β, σ^2).

- Errors need not be independent.

Gauss–Markov Theorem. Under **GM**, let $\widetilde{\beta}$ be any linear, unbiased estimator of β in the sense that

1. $\widetilde{\beta} = AY$ for some $A \in \mathbb{R}^{p \times n}$ that does not depend on Y, (linear in what?)
2. $\mathbb{E}\,\widetilde{\beta} = \beta$ for every β.

Then the OLS $\widehat{\beta}$ satisfies

$$\operatorname{cov}\widetilde{\beta} \succeq \operatorname{cov}\widehat{\beta}.$$

Review: Gauss–Markov–Normal

GM-$\mathcal{N}$ We have $Y = X\beta + \varepsilon$ with

1. X is fixed and has linearly independent columns,
2. $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$.

The unknown parameters are (β, σ^2).

- Errors are iid.

- Inference:

$$T_c := \frac{c^\intercal \widehat{\beta} - c^\intercal \beta}{\sqrt{\widehat{\sigma}^2 c^\intercal (X^\intercal X)^{-1} c}} \sim t_{n-p}.$$

$$F_C := \frac{(C\widehat{\beta} - C\beta)^\intercal \left\{ C(X^\intercal X)^{-1} C^\intercal \right\}^{-1} (C\widehat{\beta} - C\beta)}{l\widehat{\sigma}^2} \sim F_{l,n-p}.$$

- Prediction:

$$\frac{y_{n+1} - x_{n+1}^\intercal \widehat{\beta}}{\sqrt{\widehat{\sigma}^2 + \widehat{\sigma}^2 x_{n+1}^\intercal (X^\intercal X)^{-1} x_{n+1}}} \sim t_{n-p}.$$

Review: heteroskedastic linear model

Hetero We have $Y = X\beta + \varepsilon, \quad \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\intercal$, where

1. X is fixed and has linearly independent columns,
2. ε_i's are independent with $\mathbb{E}\,\varepsilon_i = 0$, $\operatorname{var} \varepsilon_i = \sigma_i^2$

The unknown parameters are $(\beta, \sigma_1^2, \ldots, \sigma_n^2)$.

- Errors are independent.

Theorem Consider **Hetero** model. Under

$$\text{(A1) (good limits)} \quad B_n := n^{-1} \sum_{i=1}^n x_i x_i^\intercal \to B \text{ (full rank)}, \quad M_n := n^{-1} \sum_{i=1}^n \sigma_i^2 x_i x_i^\intercal \to M$$

$$\text{(A2) (moment condition)} \quad d_{2+\delta,n} := n^{-1} \sum_{i=1}^n \|x_i\|^{2+\delta}\, \mathbb{E}\,|\varepsilon_i|^{2+\delta} < C \quad \text{for } \delta > 0,\ C > 0,$$

$$\sqrt{n}(\widehat{\beta} - \beta) \to_d \mathcal{N}(\mathbf{0}, B^{-1} M B^{-1}).$$

Review: Eicker–Huber–White

Theorem Consider **Hetero** model. Suppose it holds that

$$\text{(A1) (good limits)} \quad B_n := n^{-1}\sum_{i=1}^{n} x_i x_i^\mathsf{T} \to B \text{ (full rank)}, \quad M_n := n^{-1}\sum_{i=1}^{n} \sigma_i^2 x_i x_i^\mathsf{T} \to M.$$

We have

$$\widehat{\Sigma}_n = B_n^{-1}\widehat{M}_m B_n^{-1} \to_p B^{-1} M B^{-1} = \Sigma$$

if the following **(A3) (extra moment condition)** holds:

$$n^{-1}\sum_i \text{var}(\varepsilon_i^2) x_{i,j_1}^2 x_{i,j_2}^2, \quad n^{-1}\sum_i x_{i,j_1} x_{i,j_2} x_{i,j_3} x_{i,j_4}, \quad n^{-2}\sum_i \sigma_i^2 x_{i,j_1}^2 x_{i,j_2}^2 x_{i,j_3}^2$$

are bounded above by some constant C for all n and every $j_1, j_2, j_3, j_4 \in \{1, \ldots, p\}$.

Review: Frisch–Waugh–Lovell

Long regression $Y = \underbrace{X_1}_{n\times k}\widehat{\beta}_1 + \underbrace{X_2}_{n\times l}\widehat{\beta}_2 + \widehat{\varepsilon}$

Frisch–Waugh–Lovell Theorem Suppose X has linearly independent columns. In the long regression, the OLS for β_2 has the following equivalent forms:

$$\begin{aligned}
\widehat{\beta}_2 &= [(X^\intercal X)^{-1}X^\intercal Y]_{(k+1):(k+l)} \\
&= \{X_2^\intercal(I_n - H_1)X_2\}^{-1}X_2^\intercal(I_n - H_1)Y, \quad \text{where } H_1 = X_1(X_1^\intercal X_1)^{-1}X_1^\intercal \\
&= (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1}\widetilde{X}_2^\intercal Y, \quad \text{where } \widetilde{X}_2 = (I_n - H_1)X_2 \\
&= (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1}\widetilde{X}_2^\intercal \widetilde{Y}, \quad \text{where } \widetilde{Y} = (I_n - H_1)Y.
\end{aligned}$$

- $\widetilde{X}_2$ is the residual matrix from columnwise OLS fit of X_2 on X_1.
- $\widetilde{Y}$ is the residual from OLS fit of Y on X_2.
- $\widehat{\beta}_2$ is the OLS from $\widetilde{Y} \sim \widetilde{X}_2$ (partial regression)
- $\widehat{\beta}_2$ is also the OLS from $Y \sim \widetilde{X}_2$ (no need to residualize Y).

Review: Gram–Schmidt and QR

Corollary (under orthogonality) When $X_1^\intercal X_2 = 0$, i.e., $\mathcal{C}(X_1) \perp \mathcal{C}(X_2)$, we have

$$\widetilde{X}_2 = X_2,$$

$$\widehat{\beta}_2 \text{ from } Y \sim X_1 + X_2 = \widetilde{\beta}_2 \text{ from } Y \sim X_2.$$

$$\begin{aligned}
X &= (X_1, \ldots, X_p) \\
&= (U_1, \ldots, U_p) \begin{pmatrix} 1 & \widehat{\beta}_{X_2|U_1} & \widehat{\beta}_{X_3|U_1} & \cdots & \widehat{\beta}_{X_p|U_1} \\ 0 & 1 & \widehat{\beta}_{X_3|U_2} & \cdots & \widehat{\beta}_{X_p|U_2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \\
&= Q \operatorname{diag}(\|U_1\|, \ldots, \|U_p\|) \begin{pmatrix} 1 & \widehat{\beta}_{X_2|U_1} & \widehat{\beta}_{X_3|U_1} & \cdots & \widehat{\beta}_{X_p|U_1} \\ 0 & 1 & \widehat{\beta}_{X_3|U_2} & \cdots & \widehat{\beta}_{X_p|U_2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} = Q\,R.
\end{aligned}$$

BIOST/STAT 533, Sp 2024
Theory of Linear Models

Richard Guo

Lecture # 10: More on ANOVA; Cochran's formula; Omitted variable bias
§9

Recall: Wald and ANOVA equivalence

Theorem Suppose X has linearly independent columns. Consider testing $H_0 : \beta_2 = \mathbf{0}$ in

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon, \quad X_1 : n \times k,\ X_2 : n \times l$$

with

$$F_{\text{Wald}} = \frac{\widehat{\beta}_2^{\mathsf{T}} (S_{22})^{-1} \widehat{\beta}_2}{l \widehat{\sigma}^2} = \frac{\widehat{\beta}_2^{\mathsf{T}} \widetilde{X}_2^{\mathsf{T}} \widetilde{X}_2 \widehat{\beta}_2}{l \widehat{\sigma}^2}$$

and

$$F_{\text{ANOVA}} := \frac{(\mathrm{RSS}_{\text{short}} - \mathrm{RSS}_{\text{long}})/l}{\mathrm{RSS}_{\text{long}}/(n-p)} = \frac{\mathrm{RSS}_{\text{short}} - \mathrm{RSS}_{\text{long}}}{l\,\widehat{\sigma}^2}.$$

1. Under **GM-$\mathcal{N}$**, $F_{\text{ANOVA}} \sim F_{l,n-p}$ under H_0.
2. In fact, for any X, Y without assuming **GM-$\mathcal{N}$**, $F_{\text{Wald}} = F_{\text{ANOVA}}$ numerically.

Jargon

Traditionally,

- ANOVA (analysis of variance) – regression with indicator variables.
- ANCOVA (analysis of covariance) – regression with indicator and quantitative variables.

One-way ANOVA: Example

A health researcher wishes to compare the effects of four anti-inflammatory drugs on arthritis patients. She takes a random sample of patients and divides them randomly into four groups, each of which receives one of the drugs.

One-way ANOVA: Example

- The type of drug is usually referred to as a factor or treatment.
- The four kinds of drug are referred to as levels of the factor.
- We can model this as follows: $Y_{ij} = \mu_i + \varepsilon_{ij}$ where ε_{ij} i.i.d. with mean zero and variance σ^2, and $i = 1, \ldots, I$, $j = 1, \ldots, J_i$.
- What does the design matrix look like?
- An alternative parametrization: $Y_{ij} = \alpha + \mu_i + \varepsilon_{ij}$; need an identifiability constraint.

An example

Suppose we have observations $\{\{y_{ij}\}_{j=1}^{J_i}\}_{i=1}^{I}$, $\mathbb{E}(y_{ij}) = \mu_i$, $\text{var}(y_{ij}) = \sigma^2$, and the y_{ij}'s are all independent. Let $n = \sum_{i=1}^{I} J_i$.

	Observations	Mean
Population 1	$y_{11}, \ldots, y_{1J_1}$	$\bar{y}_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$
Population I	$y_{I1}, \ldots, y_{IJ_I}$	$\bar{y}_{I\cdot}$

One-way ANOVA To test $H_0 : \mu_1 = \ldots = \mu_I$, we use the F-statistic:

$$F = \frac{(RSS_{H_0} - RSS)/(I-1)}{RSS/(n-I)} = \frac{\sum_i J_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2/(I-1)}{\sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2/(n-I)},$$

which has an $F_{I-1,n-I}$ distribution under H_0 if the errors are normally distributed, and approximately an F-distribution if the sample size is large. (why?)

A typical analysis

1. Test for overall model significance (i.e. $H_0 : \mu_1 = \ldots = \mu_I$).
2. If the model is significant overall, then test specific contrasts of interest.

- Since this analysis involves performing multiple tests, some method for multiple testing control must be applied, such as a Bonferroni correction.
  - Outside the scope of this course

Typical results table for one-way ANOVA

ANOVA table with $J_1 = \ldots = J_I$.

	D.F.	Sum of Squares	Mean Sum of Squares
Groups	$I-1$	$SSTrt = J\sum_i(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2$	$SSTrt/(I-1)$
Error	$I(J-1)$	$SSErr = \sum_i\sum_j(y_{ij} - \overline{y}_{i\cdot})^2$	$SSErr/(I(J-1))$
Total	$IJ-1$	$SSTot = \sum_i\sum_j(y_{ij} - \overline{y}_{\cdot\cdot})^2$	

Now we can see why it is called ANOVA.

F-test: $F = \frac{SSTrt/(I-1)}{SSErr/(I(J-1))}$

Two-way ANOVA with balanced design: Example

A health researcher wishes to compare the effects of I anti-inflammatory drugs (factor A), as well as J different dosages (factor B), on arthritis patients. In total, there are IJ different combinations of the levels. She randomly assigns K patients to each combination of levels; there are $n = IJK$ patients in total.

Two-way ANOVA with balanced design: Example

- We assume that $y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$ where the ε_{ijk} are i.i.d. with mean zero and variance σ^2 and where $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$.
- Tests:
  1. Test $H_0 : \mu_{ij} = \mu$ for all i, j.
  2. Test whether the factors interact: does the effect of factor A at level i depend on the level of factor B?

Cochran's formula

Recall that from FWL that for $X_1 : n \times k$, $X_2 : n \times l$, from

$$\begin{aligned} \text{long:} \quad & Y = X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + \widehat{\varepsilon}, \\ \text{short:} \quad & Y = X_2\widetilde{\beta}_2 + \widetilde{\varepsilon}, \end{aligned}$$

we generally expect $\widetilde{\beta}_2 \neq \widehat{\beta}_2$. (When are they equal?)

Cochran's formula The short regression coefficients can be written as

$$\widetilde{\beta}_2 = \widehat{\beta}_2 + \widehat{\delta}\,\widehat{\beta}_1,$$

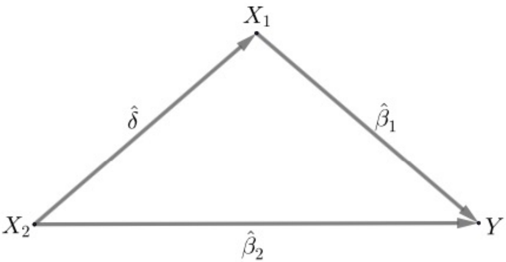
where $\delta : l \times k$ is from column-wise OLS

$$X_1 = X_2\widehat{\delta} + \widehat{U}.$$

SEM interpretation

Cochran's formula $\boxed{\widetilde{\beta}_2 = \widehat{\beta}_2 + \widehat{\delta}\,\widehat{\beta}_1}$ is a purely algebraic result that holds for OLS's

$$\begin{aligned} \text{long:} \quad & Y = X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + \widehat{\varepsilon}, \\ \text{short:} \quad & Y = X_2\widetilde{\beta}_2 + \widetilde{\varepsilon}, \\ \text{intermediate:} \quad & X_1 = X_2\widehat{\delta} + \widehat{U}. \end{aligned}$$



Proof.

Omitted-variable bias

Omitted-variable bias $\widetilde{\beta}_2 - \widehat{\beta}_2 = \widehat{\delta}\,\widehat{\beta}_1.$

- So $\widetilde{\beta}_2 = \widehat{\beta}_2$ if either
  1. $\widehat{\delta} = 0 \iff X_1 \perp X_2$, **or**
  2. $\widehat{\beta}_1 = 0 \iff Y \perp \widetilde{X}_1 = (I - H_2)X_1.$

Example: confounding bias

- z_i: treatment (1: treated; 0: control)
- x_i: observed baseline covariates

$$\text{OLS:} \quad y_i = \tilde{\beta}_0 + \tilde{\beta}_1 z_i + \tilde{\beta}_2^{\mathsf{T}} x_i + \tilde{\varepsilon}_i.$$

▶ But confounder u_i may be unobserved. The ideal (but infeasible) OLS is

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 z_i + \hat{\beta}_2^{\mathsf{T}} x_i + \hat{\beta}_3^{\mathsf{T}} u_i + \hat{\varepsilon}_i.$$

Cochran's formula:

$$\begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} + \hat{\beta}_3 \begin{pmatrix} \hat{\delta}_0 \\ \hat{\delta}_1 \\ \hat{\delta}_2 \end{pmatrix} \implies \tilde{\beta}_1 - \hat{\beta}_1 = \hat{\beta}_3 \hat{\delta}_1,$$

where $(\hat{\delta}_0, \hat{\delta}_1, \hat{\delta}_2)$ comes from $u_i \sim 1 + z_i + x_i$.

BIOST/STAT 533, Sp 2024
Theory of Linear Models

Richard Guo

Lecture # 11: R^2, leverage scores and LOO
§10, §11

- Now, let us turn to the **art** part of linear models.

Multiple correlation coefficient R^2

Consider $Y \sim 1 + X$, where $X : n \times (p-1)$ and $(1_n, X)$ has linearly independent columns.
Recall variance decomposition:

$$\underbrace{\sum_i (y_i - \bar{y})^2}_{\text{total var}} = \underbrace{\sum_i (\widehat{y}_i - \bar{y})^2}_{\text{var explained}} + \underbrace{\sum_i (y_i - \widehat{y}_i)^2}_{\text{var unexplained (RSS)}} . \quad \textit{(why?)}$$

Multiple correlation coefficient

$$R^2 = (\text{var explained }\%) = \frac{\sum_i (\widehat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}.$$

☞

$$\mathrm{RSS} = (1 - R^2) \sum_i (y_i - \bar{y})^2.$$

R^2 in equivalent forms

1. $$R^2 = \widehat{\rho}^2_{y,\widehat{y}}.$$

2. Relation to ANOVA.

$$Y = 1_n \widehat{\beta}_0 + X\widehat{\beta} + \widehat{\varepsilon}$$
$$Y = 1_n \widetilde{\beta}_0 + \widetilde{\varepsilon}.$$

☞

$$R^2 = \frac{\mathrm{RSS}_{\mathsf{short}} - \mathrm{RSS}_{\mathsf{long}}}{\mathrm{RSS}_{\mathsf{short}}}.$$

Compare this with

$$F_{\mathsf{ANOVA}} = \frac{(\mathrm{RSS}_{\mathsf{short}} - \mathrm{RSS}_{\mathsf{long}})/(p-1)}{\mathrm{RSS}_{\mathsf{long}}/(n-p)}.$$

Distribution of R^2

We have

$$F_{\text{ANOVA}} = \frac{R^2}{1-R^2} \times \frac{n-p}{p-1}.$$

Null distribution of R^2. Assume a **GM-$\mathcal{N}$** model $Y = 1_n\beta_0 + X\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Suppose $(1_n, X)$ has linearly independent columns. Then, under

$$H_0 : \beta = 0 \iff X \text{ explains no variance of } Y \text{ in population},$$

we have

$$R^2 \sim \text{Beta}\left(\frac{p-1}{2}, \frac{n-p}{2}\right).$$

☞ $\mathbb{E}\, R^2 = (p-1)/(n-1)$ under the null.

Leverage

The **leverage** of observation i is

$$h_{ii} = (H)_{ii} = x_i^\intercal (X^\intercal X)^{-1} x_i.$$

- Recall that
$$\sum_i h_{ii} = \mathrm{Tr}(H) = \mathrm{rank}(H) = n - p.$$
- It holds that $0 \leq h_{ii} \leq 1$. (why?)

☞ Leverage only concerns X (not Y)!

Leverage as a measure of ...

1. **Sensitivity**.

$$\partial \widehat{y}_i / \partial y_i = h_{ii} \qquad (why?)$$

Also, under **GM**, $\text{var}(\widehat{y}_i) = \sigma^2 h_{ii}$. *(why?)*

2. **Outlier**. Suppose $X = (1_n, X_2)$ and let $H_1 = n^{-1} 1_n 1_n^\top$.

   Let $S := (n-1)^{-1} \sum_i (x_{2,i} - \bar{x}_2)(x_{2,i} - \bar{x}_2)^\top$ be sample covariance of X_2.

   - Consider D_i^2 that measures the Mahalanobis distance between x_{i2} and $\bar{x}_2$:

$$D_i^2 := (x_{i2} - \bar{x}_2)^T S^{-1} (x_{i2} - \bar{x}_2).$$

   - Theorem 11.2

$$\boxed{h_{ii} = \frac{1}{n} + \frac{D_i^2}{n-1}.}$$

Leverage and leave-one-out (LOO) formulae

- Consider OLS from deleting the i-th observation

$$\widehat{\beta}_{[-i]} := (X_{[-i]}^{\intercal} X_{[-i]})^{-1} X_{[-i]}^{\intercal} Y_{[-i]}, \quad i = 1, \ldots, n.$$

☞ Basic idea: If i is not an outlier, result should not change much upon deleting i.

LOO formula When $h_{ii} \neq 1$,

$$\widehat{\beta}_{[-i]} = \widehat{\beta} - (1 - h_{ii})^{-1} (X^{\intercal} X)^{-1} x_i \widehat{\varepsilon}_i.$$

(What if $h_{ii} = 1$?)

Predicted residual

- Recall that residual

$$\widehat{\varepsilon}_i = y_i - \widehat{y}_i = y_i - x_i^\top \widehat{\beta} = [(I_n - H)Y]_i$$

$$\text{Under } \textbf{GM}, \quad \operatorname{var} \widehat{\varepsilon}_i = \sigma^2(1 - h_{ii}) \quad (\textit{why?})$$

- We use LOO to define the **predicted residual**

$$\boxed{\widehat{\varepsilon}_{[-i]} := y_i - x_i^\top \widehat{\beta}_{[-i]}.}$$

Theorem We have

$$\widehat{\varepsilon}_{[-i]} = \widehat{\varepsilon}_i / (1 - h_{ii}).$$

Under **GM**,

$$\operatorname{var} \widehat{\varepsilon}_{[-i]} = \sigma^2 / (1 - h_{ii}).$$

Standardized residual and studentized residual

- Under **GM-$\mathcal{N}$**, $\widehat{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2(1 - h_{ii}))$. This motivates the **standardized residual**

$$\text{standr}_i := \frac{\widehat{\varepsilon}_i}{\sqrt{\widehat{\sigma}^2(1 - h_{ii})}}$$

(What is its distribution?)

- Under **GM-$\mathcal{N}$**, $\widehat{\varepsilon}_{[-i]} \sim \mathcal{N}(0, \sigma^2/(1 - h_{ii}))$ and we define

$$\text{studr}_i := \frac{\widehat{\varepsilon}_{[-i]}}{\sqrt{\widehat{\sigma}^2_{[-i]} \,/\, (1 - h_{ii})}}.$$

☞ Under **GM-$\mathcal{N}$**, $\text{studr}_i \sim t_{n-1-p}$. (why?)

Cook's distance

- A related measure is **Cook's distance**

$$\boxed{\text{cook}_i := \frac{\|X^\intercal(\widehat{\beta} - \widehat{\beta}_{[-i]})\|^2}{p\widehat{\sigma}^2}.}$$

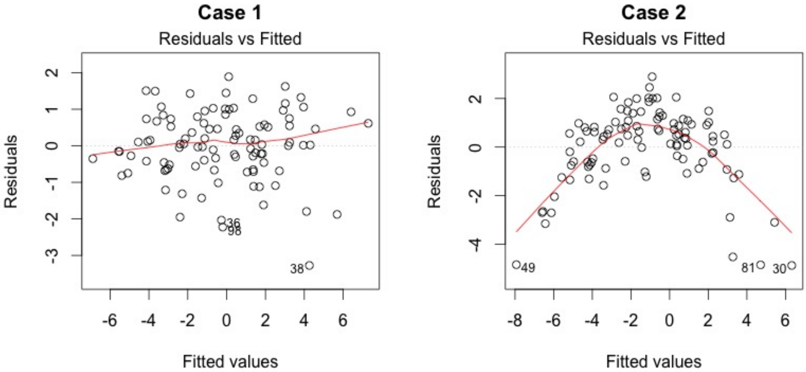
- Cook's distance is related to the standardized residual via

$$\text{cook}_i = \text{standr}_i^2 \times \frac{h_{ii}}{p(1 - h_{ii})}.$$

lm() diagnostic plots in R

1. Residuals vs Fitted: $\text{studr}_i \sim \widehat{y}_i$.

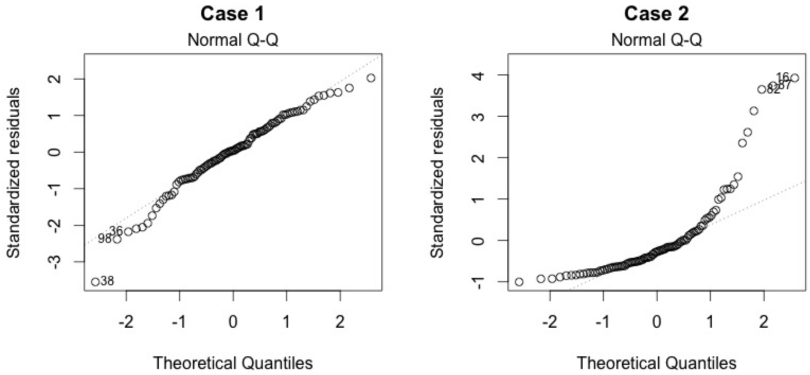
```
> plot(lm(y ~ X))
```



https://library.virginia.edu/data/articles/diagnostic-plots

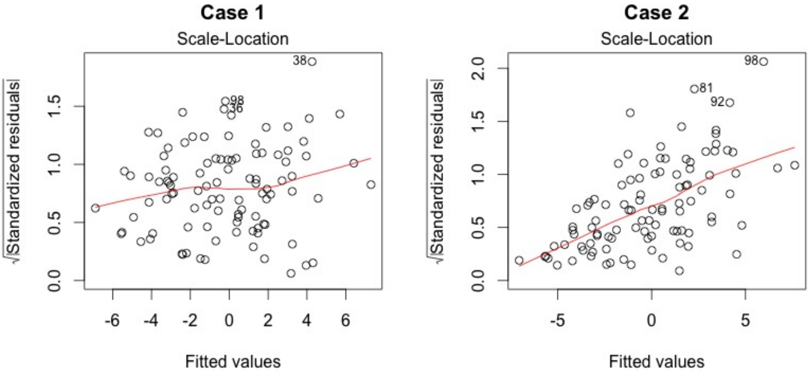
`lm()` diagnostic plots in R

2. Normal QQ plot: sample quantiles of $\text{studr}_i \sim$ quantiles of $\mathcal{N}(0,1)$.



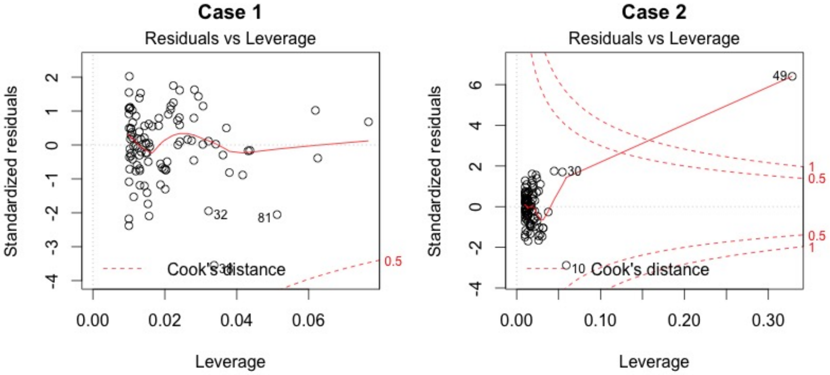
lm() diagnostic plots in R

3. Location-Scale plot: $\sqrt{|\text{studr}_i|} \sim \widehat{y}_i$.



`lm()` diagnostic plots in R

4. Residuals vs Leverage: $\text{studr}_i \sim h_{ii}$.



BIOST/STAT 533, Sp 2024
Theory of Linear Models

Richard Guo

Lecture # 12: Population OLS, misspecified linear model
§12

Population least squares

- ☞ Consider random variable Y and random vector $X \in \mathbb{R}^p$.
  - ☞ In this lecture, X and Y are no longer the data of n rows!
    - ☞ X is not fixed but also random now!

Theorem For any measurable, real-valued function $f(X)$ of X, we have bias-variance decomposition

$$\mathbb{E}(Y - f(X))^2 = \mathbb{E}\{\mathbb{E}[Y \mid X] - f(X)\}^2 + \mathbb{E}\,\mathrm{var}[Y \mid X].$$

Further, we have

$$\mathbb{E}[Y \mid X] = \arg\min_f \mathbb{E}(Y - f(X))^2,$$

where the minimization is over all square integrable, measurable function of X.

Population OLS

Now consider **linear functions** of X: $f(X) = \beta^\intercal X, \quad \beta \in \mathbb{R}^p$.

- Population OLS:

$$\boxed{\beta = \arg\min_b \mathbb{E}(Y - b^\intercal X)^2 = \arg\min_b \mathbb{E}([\mathbb{E}[Y \mid X] - b^\intercal X)^2.} \qquad (why?)$$

☞ What is the interpretation of $\beta^\intercal X$?

Theorem $\beta = (\mathbb{E}\, XX^\intercal)^{-1}\, \mathbb{E}[XY]$ when $\mathbb{E}\, XX^\intercal$ is invertible.

- For $X \in \mathbb{R}$, univariate population OLS: $(\alpha, \beta) = \arg\min_{a,b} \mathbb{E}(Y - a - bX)^2$.

$$\boxed{\alpha = \mathbb{E}\, Y - \beta\, \mathbb{E}\, X, \quad \beta = \frac{\mathrm{cov}(X, Y)}{\mathrm{var}\, X} = \rho_{XY} \sqrt{\frac{\mathrm{var}\, Y}{\mathrm{var}\, X}}.}$$

☞ $\rho_{XY} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}}$.

Population FWL

Suppose Y is a random variable and $X \in \mathbb{R}^{p-1}$ a random vector. Consider a population OLS

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_{p-1} + \varepsilon.$$

Also consider the following population OLS's:

☞ These equations *define* the residuals.

$$\begin{aligned} X_k &= \gamma_0 + \gamma_1 X_1 + \ldots \gamma_{k-1} X_{k-1} + \gamma_{k+1} X_{k+1} + \cdots + \gamma_{p-1} X_{p-1} + \widetilde{X}_k, \\ Y &= \delta_0 + \delta_1 X_1 + \ldots \delta_{k-1} X_{k-1} + \delta_{k+1} X_{k+1} + \cdots + \delta_{p-1} X_{p-1} + \widetilde{Y}, \\ \widetilde{Y} &= \widetilde{\beta}_k \widetilde{X}_k + \widetilde{\varepsilon}. \end{aligned}$$

Population FWL Theroem

1. $\beta_k = \widetilde{\beta}_k = \operatorname{cov}(\widetilde{X}_k, \widetilde{Y}) / \operatorname{var} \widetilde{X}_k = \operatorname{cov}(\widetilde{X}_k, Y) / \operatorname{var} \widetilde{X}_k$.
2. $\widetilde{\varepsilon} = \varepsilon$ almost surely.

Population R^2

- Nonparametric R^2: For $f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}(Y - f(X))^2$,

$$\boxed{\operatorname{var} Y = \mathbb{E}(Y - \mathbb{E}\, Y)^2 = \mathbb{E}(Y - f^*(X))^2 + \operatorname{var} f^*(X)}, \quad (why?)$$

and

$$R^2_{\mathcal{F}} = \frac{\operatorname{var} f^*(X)}{\operatorname{var} Y} \in [0, 1].$$

- When $\mathcal{F}$ is the set of linear functions of $(1, X)$,

$$\boxed{R^2 = \frac{\Sigma_{Y,X} \Sigma_{X,X}^{-1} \Sigma_{X,Y}}{\operatorname{var} Y},}$$

where

$$\operatorname{cov} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Sigma_{X,X} & \Sigma_{X,Y} \\ \Sigma_{Y,X} & \operatorname{var} Y \end{pmatrix}.$$

OLS inference when linear model is misspecified

Let (x_i, y_i) be iid copies of (X, Y). We do not assume a linear model for $Y \sim X$ holds in the population (data generating mechanism).

☞ Let $\widehat{\beta}$ be OLS from $(x_i, y_i) : i = 1, \ldots, n$.

Let β be the population OLS:

▶ Not assuming a linear model!

$$\boxed{Y = \beta^\intercal X + \varepsilon.}$$

Theorem Let $(x_i, y_i)_{i=1}^n$ be iid copies of (X, Y).

1. $\widehat{\beta} \to_p \beta$.
2. $\sqrt{n}(\widehat{\beta} - \beta) \to_d \mathcal{N}(0, \Sigma)$, where

$$\Sigma = B^{-1} M B^{-1}, \quad B = \mathbb{E}\, XX^\intercal, \quad M = \mathbb{E}(\varepsilon^2 XX^\intercal).$$

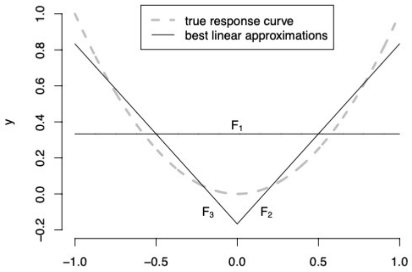
3. Eicker-Huber-White $\Sigma_{\text{EHW}} \to_p \Sigma$ if $\mathbb{E}\,\|X\|^4 < \infty$ and $\mathbb{E}\, Y^4 < \infty$.

Best linear approximation

The target of OLS $\widehat{\beta}$ is the

- the correct β if linear model holds;
- when linear model does not hold, the population OLS $\beta = \arg\min_b \mathbb{E}(Y - b^\intercal X)^2$;
- however, β not only depends on $\mathbb{E}[Y \mid X]$, but also the distribution of X

$$\beta = \arg\min_b \mathbb{E}\{\mathbb{E}[Y \mid X] - b^\intercal X\}^2.$$



BIOST/STAT 533, Sp 2024
Theory of Linear Models

Richard Guo

Lecture # 13: Overfitting, bias-variance trade-off, model selection
§13

Caution of R^2

$$R^2 = (\text{var explained \%}) = \frac{\sum_i (\widehat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}.$$

Recall that

Null distribution of R^2. Assume a **GM-$\mathcal{N}$** model $Y = 1_n\beta_0 + X\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Suppose $(1_n, X)$ has linearly independent columns. Then, under

$$H_0 : \beta = 0 \iff X \text{ explains no variance of } Y \text{ in population},$$

we have

$$R^2 \sim \text{Beta}\left(\frac{p-1}{2}, \frac{n-p}{2}\right).$$

☞ $\mathbb{E}\, R^2 = (p-1)/(n-1)$ under the **null**.

What happens when $p/n \to \gamma$?

Variance inflation factor

Let $X : n \times (p-1)$ be a fixed design matrix. Suppose

$$y_i = f(x_i) + \varepsilon_i,$$

where ε_i's are uncorrelated with mean zero and variance σ^2.

Short regression: $Y = \widetilde{\alpha} + \widetilde{\beta}_j X_j + \widetilde{\varepsilon}$
Long regression: $Y = \widehat{\alpha} + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_{p-1} X_{p-1} + \widehat{\varepsilon}$.

Theorem We have

$$\operatorname{var} \widehat{\beta}_j = \operatorname{var} \widetilde{\beta}_j \times \underbrace{\frac{1}{1 - R_j^2}}_{\text{VIF}},$$

where R_j^2 is the R^2 from OLS $X_j \sim 1 + X_{-j}$.

Bias-variance trade-off

Suppose the true data generating mechanism is

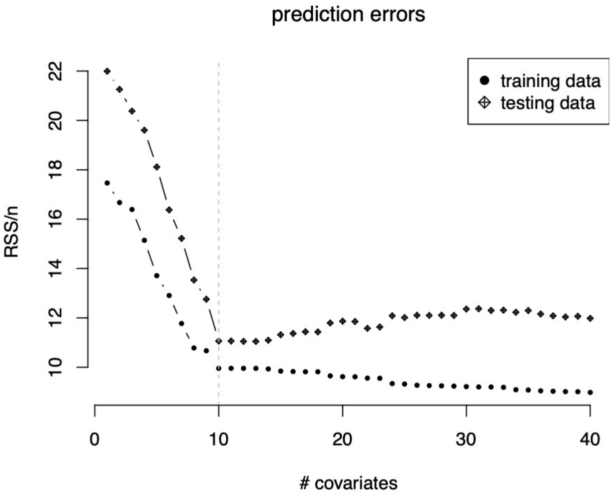
$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_s x_{i,s} + \varepsilon_i$$

with s non-null covariates.

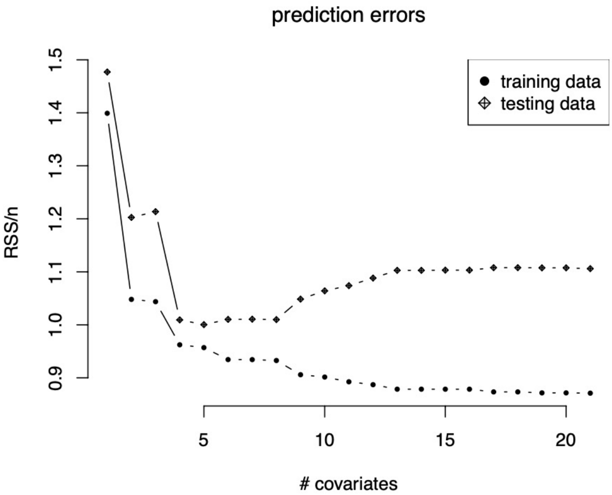
☞ Consider running OLS fitting Y on $X_1, \ldots, X_p$.

- What happens when $p < s$?
- What happens when $p > s$?

Typical trade-off: under linear model



Typical trade-off: under a non-linear model



Adjusted R^2

Recall that

$$\underbrace{\sum_i (y_i - \bar{y})^2}_{\text{total var}} = \underbrace{\sum_i (\widehat{y}_i - \bar{y})^2}_{\text{var explained}} + \underbrace{\sum_i (y_i - \widehat{y}_i)^2}_{\text{var unexplained (RSS)}} .$$

- $$1 - R^2 = \frac{\sum_i (y_i - \widehat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = \frac{\|(I-H)Y\|^2}{\|(I-H_1)Y\|^2}$$

To account for model complexity, define adjusted R^2 as

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p} = 1 - \frac{\|(I-H)Y\|^2/(n-p)}{\|(I-H_1)Y\|^2/(n-1)} = 1 - \frac{\widehat{\sigma}^2}{\widehat{\sigma}^2_Y}.$$

Mallow's C_p

$$C_p := \|Y - \widehat{Y}\|^2 + 2p\sigma^2.$$

- Infeasible but unbiased for MSPE over the same X but new Y

Akaike's information criterion (AIC)

Consider the more general set up: $Y_i \sim f(y), i = 1, \ldots, n$ but a parametric model $Y_i \sim f(y; \theta)$ is fitted over an Euclidean model space Θ.

$$\text{AIC} = -2 \sum_{i=1}^{n} \log f(Y_i; \widehat{\theta}) + 2\text{dim}(\Theta)$$

- For model selection, AIC attempts to estimate prediction error $-2E_f\{\log f(Y_{n+1}; \widehat{\theta}) \mid \widehat{\theta}\}$, where the expectation is taken over a new observation Y_{n+1}.
- $2\,\text{dim}(\Theta)$ is a correction term

Bayesian information criterion (BIC)

BIC is motivated by Bayesian perspective on model selection and is defined as

$$\mathrm{BIC}(\Theta) = -2\sum_{i=1}^{n} \log f(Y_i;\widehat{\theta}) + \dim(\Theta)\log n$$

- Intuition: suppose $\{\Theta_1, \ldots, \Theta_m\}$ is a collection of model spaces. If we assign a uniform prior on the model spaces, $P(\Theta_k) = \frac{1}{m}$ for all k. Then as $n \to \infty$, the posterior probability for a model is approximately $P(\Theta_k \mid \mathrm{Data}) \propto e^{-\mathrm{BIC}(\Theta_k)/2}$.
- Compared with AIC, BIC puts a larger penalty on model complexity and thus selects a smaller model.
- Rule of thumb: AIC is more suitable for prediction and BIC is more suitable for selecting the "correct" model.

AIC and BIC

Under **GM-$\mathcal{N}$**,

$$\mathrm{AIC} = n \log \frac{\mathrm{RSS}}{n} + 2p$$
$$\mathrm{BIC} = n \log \frac{\mathrm{RSS}}{n} + p \log n.$$

(why?)

- Shao (1997):
  - If the linear model is correctly specified, BIC can consistently select the true model.
  - Even when the linear model is misspecified, AIC can select the model that minimizes the prediction error.

Cross-validation and its approximation

- We can use K-fold CV to select covariates.
- When $K = n$, we can use LOO formula to approximate the actual CV.

Recall the **LOO predicted residual**:

$$\boxed{\hat{\varepsilon}_{[-i]} := y_i - x_i^\top \hat{\beta}_{[-i]} = \frac{\hat{\varepsilon}_i}{1 - h_{ii}}.}$$

- Define the predicted residual error sum of squares $\boxed{\mathrm{PRESS} := \sum_i \hat{\varepsilon}_{[-i]}^2.}$

Replacing $h_{ii} \approx p/n$ (why?),

$$\mathrm{PRESS} \approx \mathrm{GCV} := \sum_i \frac{\varepsilon_i^2}{(1 - p/n)^2} = (1 - p/n)^{-2} \times \mathrm{RSS}.$$

- When $p/n \approx 0$, $\log \mathrm{GCV}$ is approximately equivalent to AIC. (why?)

Algorithms for model selection

- Best subset selection
- Forward stepwise
- Backward stepwise

Best subset selection

- We have p possible predictors and we want to know which to use in our model.
- We could consider every possible model (there are 2^p of them) and select the one with smallest cross-validation error.
- If $p = 3$ there are $2^3 = 8$ possible models.
- If $p = 6$ there are $2^6 = 64$ possible models.
- If $p = 250$ there are $2^{250} \approx 10^{80}$ possible models.
- Obviously we need a more efficient alternative.

Forward stepwise selection

1. Fit p univariate regression models – one with each predictor – and select the predictor corresponding to the most significant model (largest F-stat, or equivalently reduces the RSS the most).
2. Then fit $p-1$ models containing the predictor that we just selected and each of the $p-1$ other predictors. Select the predictor corresponding to the most significant model.
3. Now we have selected 2 predictors. Fit the $p-2$ models containing these 2 predictors, and each of the $p-2$ other predictors. Select the predictor corresponding to the most significant model.
4. And so on....

This procedure will result in $p+1$ distinct models, containing between 0 and p predictors.

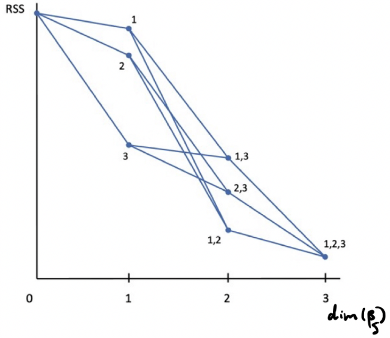
Backward stepwise selection

Just like forward stepwise, but we instead start with the model containing all of the features and remove features one-at-a-time.

Pros and cons of stepwise selection

- Backward and forward stepwise selection are much more efficient than best subset selection... they require looking at $p + (p-1) + \ldots$ (on the order of p^2) models, rather than 2^p models!
- However, backward stepwise and forward stepwise will give us different answers!
- They will not give us the "best" model for a fixed number of predictors.

Example



For 0,1,2,3 regressors,

- The best subset algorithm selects $\emptyset, \{3\}, \{1,2\}, \{1,2,3\}$
- The forward stepwise algorithm selects $\emptyset, \{3\}, \{2,3\}, \{1,2,3\}$
- The backward stepwise algorithm selects $\emptyset, \{2\}, \{1,2\}, \{1,2,3\}$

Model selection

- With the path plot, we can then select a single model by using one of the quantitative criteria introduced above.
- This can be combined with model diagnostics.

BIOST/STAT 533, Sp 2024
Theory of Linear Models

Richard Guo

Lecture # 14: Generalized and weighted least squares; Transformations in OLS
§19, §16

Recall: Gauss-Markov

GM The data generating process obeys $Y = X\beta + \varepsilon$,

1. X is fixed and has linearly independent columns,
2. $\mathbb{E}\,\varepsilon = \mathbf{0}$, $\operatorname{cov}\varepsilon = \sigma^2 I_n$.

The unknown parameters are (β, σ^2).

Gauss–Markov Theorem. Under **GM**, let $\widetilde{\beta}$ be any linear, unbiased estimator of β in the sense that

1. $\widetilde{\beta} = AY$ for some $A \in \mathbb{R}^{p\times n}$ that does not depend on Y,
2. $\mathbb{E}\,\widetilde{\beta} = \beta$ for every β.

Then the OLS $\widehat{\beta}$ satisfies

$$\operatorname{cov}\widetilde{\beta} \succeq \operatorname{cov}\widehat{\beta}.$$

▶ OLS is **BLUE** under **GM**.

Gauss-Markov Generalized model

GM-Generalized (aka. Aitkin model) The data generating process obeys $Y = X\beta + \varepsilon$,

1. X is fixed and has linearly independent columns,
2. $\mathbb{E}\,\varepsilon = \mathbf{0}$, $\operatorname{cov}\varepsilon = \sigma^2\Sigma$.

The unknown parameters are (β, σ^2); Σ is a known positive definite matrix.

Special cases:

- $\Sigma = I_n$ ▶ **GM**
- $\Sigma = \operatorname{diag}(w_1^{-1}, \ldots, w_n^{-1})$ ▶ weighted least squares
- $\Sigma = \operatorname{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_K)$ ▶ clusters

☞ Note that Σ is known under **GM-Generalized**.
Hence, $\Sigma = \operatorname{diag}(w_1^{-1}, \ldots, w_n^{-1})$ is different from **Hetero**.

Recall: Heteroskedastic linear model

Hetero The data generating process obeys

$$Y = X\beta + \varepsilon, \quad \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\mathsf{T},$$

where

1. X is fixed and has linearly independent columns,
2. ε_i's are independent with $\mathbb{E}\,\varepsilon_i = 0$, $\operatorname{var}\varepsilon_i = \sigma_i^2$

The unknown parameters are $(\beta, \sigma_1^2, \ldots, \sigma_n^2)$.

OLS, GLS and BLUE

GM-Generalized (aka. Aitkin model) The data generating process obeys $Y = X\beta + \varepsilon$,

1. X is fixed and has linearly independent columns,
2. $\mathbb{E}\,\varepsilon = \mathbf{0}$, $\operatorname{cov}\varepsilon = \sigma^2\Sigma$.

The unknown parameters are (β, σ^2); Σ is a known positive definite matrix.

- OLS is unbiased but not **BLUE** under **GM-Generalized**. (why?)

Theorem Under **GM-Generalized**, the generalized least squares (GLS) is **BLUE**:

$$\widehat{\beta}_\Sigma := (X^\intercal\Sigma^{-1}X)^{-1}X^\intercal\Sigma^{-1}Y.$$

Further, we have

$$\mathbb{E}\,\widehat{\beta}_\Sigma = \beta, \quad \operatorname{cov}\widehat{\beta}_\Sigma = \sigma^2(X^\intercal\Sigma^{-1}X)^{-1}.$$

OLS, GLS and BLUE

☞ From comparing OLS and GLS,

$$(X^\intercal \Sigma^{-1} X)^{-1} \preceq (X^\intercal X)^{-1} X^\intercal \Sigma X (X^\intercal X)^{-1}.$$

☞ What happens if using GLS $\widehat{\beta}_\Omega$ (for some covariance Ω) under **GM-Generalized** with covariance Σ?

Weighted least squares

When $\Sigma = \text{diag}(w_1^{-1}, \dots, w_n^{-1})$, the weighted least squares (WLS) is

$$\widehat{\beta}_w = \widehat{\beta}_\Sigma = (X^\intercal \Sigma^{-1} X)^{-1} X^\intercal \Sigma^{-1} Y$$
$$= \left(\sum_i w_i x_i x_i^\intercal \right)^{-1} \sum_i w_i x_i y_i.$$

- Under **GM-Generalized**,

$$\text{cov}\, \widehat{\beta}_w = \sigma^2 \left(\sum_i w_i x_i x_i^\intercal \right)^{-1}.$$

- Under **Hetero**, we have Eicker-Huber-White estimator for the asymptotic covariance of $\widehat{\beta}_w$:

$$\widehat{\Sigma}_{\text{EHW},w} = \left(n^{-1} \sum_i w_i x_i x_i^\intercal \right)^{-1} \left(n^{-1} \sum_i w_i^2 \widehat{\varepsilon}_{w,i}^2 x_i x_i^\intercal \right) \left(n^{-1} \sum_i w_i x_i x_i^\intercal \right)^{-1} \quad \textit{(why?)}$$

Weighted least squares: Two-stage under heteroskedasticity

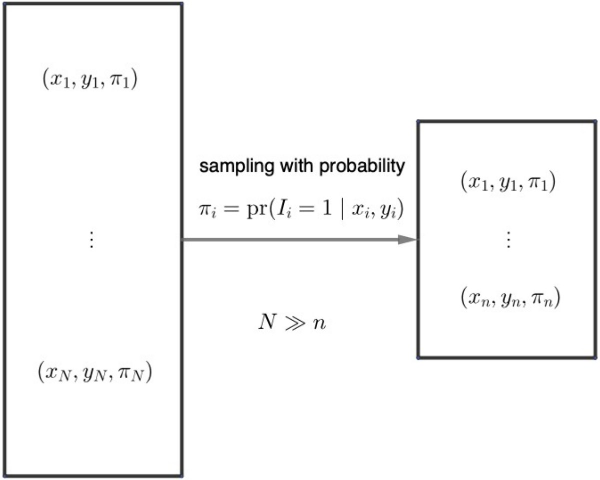
Consider **Hetero** model where $\sigma_1^2, \ldots, \sigma_n^2$ are unknown.

- OLS is consistent, but not efficient.
- WLS (with $w_i = \sigma_i^{-2}$) is consistent and efficient — but the true weights are unknown!

Two-stage method:

1. Use OLS to estimate β and get residuals $\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n$.
2. Use $\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n$ to estimate a postulated model of $\sigma_i^2 = \sigma^2(x_i; \theta)$.
   - E.g., fit a linear model $\log(\widehat{\varepsilon}_i^2) \sim X$ and exponentiate.
3. Fit WLS $\widehat{\beta}_{\widehat{w}}$ with $\widehat{w}_i = \sigma^{-2}(x_i; \widehat{\theta})$, i=1,. . . ,n.
4. Inference with Eicker-Huber-White covariance for $\widehat{\beta}_{\widehat{w}}$.

Weighted least squares: Survey sampling



Weighted least squares: Survey sampling

- Ideal estimator

$$\widehat{\beta}_{\text{ideal}} = \left(\sum_{i=1}^{N} x_i x_i^{\mathsf{T}} \right)^{-1} \sum_{i=1}^{N} x_i y_i .$$

- Sampling probability

$$I_i = \mathbb{I}\{\text{unit } i \text{ is included in the sample}\}, \quad \pi_i = \mathbb{P}(I_i = 1 \mid X_i, y_i).$$

- Horvitz and Thompson (1952) inverse probability weighting (IPW)

$$\widehat{\beta}_{\text{IPW}} := \left(\sum_{i=1}^{N} \frac{I_i}{\pi_i} x_i x_i^{\mathsf{T}} \right)^{-1} \sum_{i=1}^{N} \frac{I_i}{\pi_i} x_i y_i = \left(\sum_{j=1}^{n} \pi_j^{-1} x_j x_j^{\mathsf{T}} \right)^{-1} \sum_{j=1}^{n} \pi_j^{-1} x_j y_j .$$

☞ $\mathbb{E}[\frac{I_i}{\pi_i} \mid x_i, y_i] = 1.$

- Some tricks of the trade: transformations of outcome and covariates.

Transform of the outcome: $\log$

For $y_i > 0$,

$$\log y_i = x_i^\top \beta + \varepsilon_i.$$

- Would do you interpret it?
- When $y_i \sim \mathcal{N}(\mu_i, \sigma^2 \mu_i^2)$, where

$$sd(y_i) \propto \mathbb{E}\, y_i,$$

then it is a good idea to take $\log$ transform:

$$\log y_i - \log \mu_i \approx (y_i - \mu_i)/\mu_i \sim \mathcal{N}(0, \sigma^2). \quad \textit{(why?)}$$

☞ e.g., y_i is the time that runner i takes to finish distance μ_i

- When $y \geq 0$, $\log(y_i + 1)$ is used frequently. ☞ e.g., gene expression

Transform of the outcome: Box–Cox

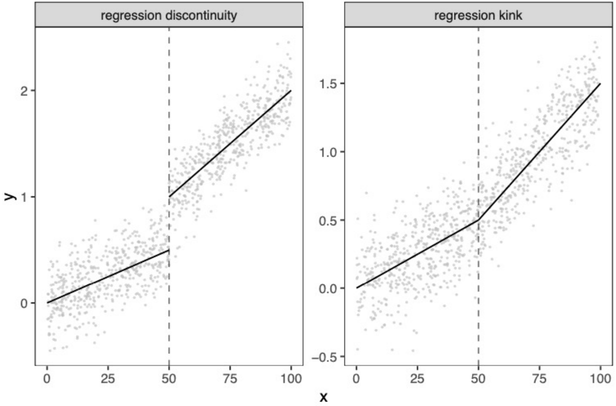
George Box and Sir David Cox (1964) consider a family of transformations on y:

$$g_\lambda(y) = \begin{cases} (y^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \log y, & \lambda = 0 \end{cases}.$$

- $Y_\lambda = (g_\lambda(y_1), \dots g_\lambda(y_n))^\intercal \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ yields likelihood $L(\beta, \sigma^2, \lambda; Y)$.
- Draw profile log-likelihood $l_p(\lambda) = \log L(\widehat{\beta}(\lambda), \widehat{\sigma}^2(\lambda), \lambda; Y)$ and construct 95% CI for λ around the maximizer $\widehat{\lambda}$ based on

$$2(l_P(\widehat{\lambda}) - l_P(\lambda)) \to_d \chi^2(1).$$

Transform of the covariates: regression discontinuity and kink



Transform of the covariates: regression discontinuity and kink

- Testing H_0: no treatment effect boils down to testing

$$H_0 : \text{regression is continuous (i.e. a kink) at } x = c. \iff \beta_3 = 0 \text{ in}$$

$$y_i = \begin{cases} \beta_1 + \beta_2(x_i - c) + \varepsilon_i, & x_i \leq c \\ (\beta_1 + \beta_3) + (\beta_2 + \beta_4)(x_i - c) + \varepsilon_i, & x_i > c. \end{cases}$$

- This piecewise linear model can be parameterized as a linear model

$$\boxed{y_i = \beta_1 + \beta_2(x_i - c) + \beta_3 \, \mathbb{I}(x_i > c) + \beta_4 \, (x_i - c) \, \mathbb{I}(x_i > c) + \varepsilon_i}$$

- Similarly, we can test H_0' : no kink using $R_c(x_i) := \max(0, x_i - c)$:

$$\boxed{y_i = \beta_1 + \beta_2 R_c(x_i) + \beta_3(x_i - c) + \varepsilon_i,}$$

where $H_0' \iff \beta_2 = 0$.

BIOST/STAT 533, Sp 2024
Theory of Linear Models

Richard Guo

Lecture # 15: Final Review

Final Exam

Monday June 3: 2:30 – 4:20 PM, this classroom. Open notes / books. No electronics. Covers the whole course.

Covered: Before midterm

1. Linear algebra: column space, orthogonal matrix, eigendecomposition, projection
2. OLS: algebra and geometry
3. **GM** model, RSS, $\widehat{\sigma}^2$, Gauss-Markov theorem
4. **GM-$\mathcal{N}$** model, pivotal t and F inference
5. **Hetero** model, consistency and asymptotic normality of $\widehat{\beta}$
6. Eicker-Huber-White covariance estimation
7. Long and short regressions; Frisch–Waugh–Lovell theorem; QR decomposition

Covered: After midterm

1. ANOVA (and its equivalence to Wald test); one-way and two-way ANOVA; ANOVA with parameters under constraints; degrees of freedom
2. Orthogonal decomposition of RSS and variance; R^2
3. Cochran's formula; omitted variable bias
4. Leverage; leave-one-out; predicted residuals; diagnostic plots
5. Misspecified linear model and its interpretation; population OLS; inference for the population OLS
6. Overfitting; bias-variance tradeoff; mean squared prediction error; Mallow's C_p; AIC and BIC; model selection
7. Generalized least squares; weighted least squares

Gauss–Markov

GM We have $Y = X\beta + \varepsilon$ with

1. X is fixed and has linearly independent columns,
2. $\mathbb{E}\,\varepsilon = \mathbf{0}$, $\operatorname{cov}\varepsilon = \sigma^2 I_n$.

The unknown parameters are (β, σ^2). ☞ Errors need not be independent.

- OLS is **BLUE** under **GM**.

GM-Generalized (aka. Aitkin model) We have $Y = X\beta + \varepsilon$, where

1. X is fixed and has linearly independent columns,
2. $\mathbb{E}\,\varepsilon = \mathbf{0}$, $\operatorname{cov}\varepsilon = \sigma^2\Sigma$.

The unknown parameters are (β, σ^2); Σ is a known positive definite matrix.

- GLS is **BLUE** under **GM-Generalized**.

Gauss–Markov–Normal

GM-$\mathcal{N}$ We have $Y = X\beta + \varepsilon$ with

1. X is fixed and has linearly independent columns,
2. $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$.

The unknown parameters are (β, σ^2).

- Errors are iid.

- Inference:

$$T_c := \frac{c^\intercal \widehat{\beta} - c^\intercal \beta}{\sqrt{\widehat{\sigma}^2 c^\intercal (X^\intercal X)^{-1} c}} \sim t_{n-p}.$$

$$F_C := \frac{(C\widehat{\beta} - C\beta)^\intercal \left\{C(X^\intercal X)^{-1} C^\intercal\right\}^{-1} (C\widehat{\beta} - C\beta)}{l\widehat{\sigma}^2} \sim F_{l,n-p}.$$

- Prediction:

$$\frac{y_{n+1} - x_{n+1}^\intercal \widehat{\beta}}{\sqrt{\widehat{\sigma}^2 + \widehat{\sigma}^2 x_{n+1}^\intercal (X^\intercal X)^{-1} x_{n+1}}} \sim t_{n-p}.$$

Gauss–Markov–Normal

- Question: For $Y = X\beta + \varepsilon$, Under **GM-$\mathcal{N}$**, how would you test $\beta_1 = 0$?

Lemma We have

$$(X^\intercal X)^{-1} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},$$

where

$$\begin{aligned}
S_{11} &= (X_1^\intercal X_1)^{-1} + (X_1^\intercal X_1)^{-1} X_1^\intercal X_2 (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1} X_2^\intercal X_1 (X_1^\intercal X_1)^{-1}, \\
S_{12} &= -(X_1^\intercal X_1)^{-1} X_1^\intercal X_2 (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1}, \\
S_{21} &= S_{12}^\intercal, \\
S_{22} &= (\widetilde{X}_2^\intercal \widetilde{X}_2)^{-1}.
\end{aligned}$$

Heteroskedastic linear model

Hetero We have $Y = X\beta + \varepsilon, \quad \varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\intercal$, where

1. X is fixed and has linearly independent columns,
2. ε_i's are independent with $\mathbb{E}\,\varepsilon_i = 0$, $\operatorname{var}\varepsilon_i = \sigma_i^2$

The unknown parameters are $(\beta, \sigma_1^2, \dots, \sigma_n^2)$.

- Errors are independent.

Theorem Consider **Hetero** model. Under

$$\text{(A1) (good limits)} \quad B_n := n^{-1}\sum_{i=1}^n x_i x_i^\intercal \to B \text{ (full rank)}, \quad M_n := n^{-1}\sum_{i=1}^n \sigma_i^2 x_i x_i^\intercal \to M$$

$$\text{(A2) (moment condition)} \quad d_{2+\delta,n} := n^{-1}\sum_{i=1}^n \|x_i\|^{2+\delta}\,\mathbb{E}\,|\varepsilon_i|^{2+\delta} < C \quad \text{for } \delta > 0,\ C > 0,$$

$$\sqrt{n}(\widehat{\beta} - \beta) \to_d \mathcal{N}(\mathbf{0}, B^{-1}MB^{-1}).$$

Eicker–Huber–White

Theorem Consider **Hetero** model. Suppose it holds that

$$\text{(A1) (good limits)} \quad B_n := n^{-1}\sum_{i=1}^{n} x_i x_i^{\mathsf{T}} \to B \text{ (full rank)}, \quad M_n := n^{-1}\sum_{i=1}^{n} \sigma_i^2 x_i x_i^{\mathsf{T}} \to M.$$

We have

$$\widehat{\Sigma}_n = B_n^{-1}\widehat{M}_n B_n^{-1} \to_p B^{-1} M B^{-1} = \Sigma$$

if the following **(A3) (extra moment condition)** holds:

$$n^{-1}\sum_i \operatorname{var}(\varepsilon_i^2) x_{i,j_1}^2 x_{i,j_2}^2, \quad n^{-1}\sum_i x_{i,j_1} x_{i,j_2} x_{i,j_3} x_{i,j_4}, \quad n^{-2}\sum_i \sigma_i^2 x_{i,j_1}^2 x_{i,j_2}^2 x_{i,j_3}^2$$

are bounded above by some constant C for all n and every $j_1, j_2, j_3, j_4 \in \{1, \ldots, p\}$.

Eicker–Huber–White

Logic of Eicker–Huber–White:

1. Write $\widehat{\beta}$ as a function of the random response
2. Derive the covariance of $\widehat{\beta}$ — sandwich form
3. Estimate each piece

Review: Gram–Schmidt and QR

Corollary (under orthogonality) When $X_1^\top X_2 = 0$, i.e., $\mathcal{C}(X_1) \perp \mathcal{C}(X_2)$, we have

$$\widetilde{X}_2 = X_2,$$

$$\widehat{\beta}_2 \text{ from } Y \sim X_1 + X_2 = \widetilde{\beta}_2 \text{ from } Y \sim X_2.$$

$$
\begin{aligned}
X &= (X_1, \ldots, X_p) \\
&= (U_1, \ldots, U_p) \begin{pmatrix} 1 & \widehat{\beta}_{X_2|U_1} & \widehat{\beta}_{X_3|U_1} & \cdots & \widehat{\beta}_{X_p|U_1} \\ 0 & 1 & \widehat{\beta}_{X_3|U_2} & \cdots & \widehat{\beta}_{X_p|U_2} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \\
&= Q\,\mathrm{diag}(\|U_1\|, \ldots, \|U_p\|) \begin{pmatrix} 1 & \widehat{\beta}_{X_2|U_1} & \widehat{\beta}_{X_3|U_1} & \cdots & \widehat{\beta}_{X_p|U_1} \\ 0 & 1 & \widehat{\beta}_{X_3|U_2} & \cdots & \widehat{\beta}_{X_p|U_2} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix} = Q\,R.
\end{aligned}
$$

OLS and population OLS

OLS As an algebraic operation on data $X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^n$,

$$\widehat{\beta} = \arg\min_b \|Y - Xb\|^2 = (X^\intercal X)^{-1} X^\intercal Y.$$

We have orthogonal decomposition

☞ Orthogonal in what sense?

$$Y = \widehat{Y} + \widehat{\varepsilon} = HY + (I - H)Y.$$

Population OLS As an approximation to $P(X, Y)$ of random $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$,

$$\beta = \arg\min_b \mathbb{E}\,\|Y - X^\intercal b\|^2 = \mathbb{E}(XX^\intercal)^{-1}\,\mathbb{E}[XY]$$

☞ How do we interpret β?

☞ Orthogonal in what sense?

We have orthogonal decomposition

$$Y = \beta^\intercal X + \varepsilon.$$

OLS and population OLS: FWL theorems

OLS FWL Suppose X has linearly independent columns. Consider long regression

$$Y = X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + \widehat{\varepsilon}.$$

- $\widehat{\beta}_2$ from equals OLS from short regression $Y \sim \widetilde{X}_2$ and $\widetilde{Y} \sim \widetilde{X}_2$.
- $\widehat{\varepsilon}$ equals residuals from short regression $\widetilde{Y} \sim \widetilde{X}_2$.

☞ How do you partial out X_1?

Population OLS FWL Consider a population OLS

☞ What does this equation mean?

$$Y = X_1^\mathsf{T}\beta_1 + X_2^\mathsf{T}\beta_2 + \varepsilon.$$

- β equals population OLS $Y \sim \widetilde{X}_2$ and $\widetilde{Y} \sim \widetilde{X}_2$.
- ε equals the residual from $\widetilde{Y} \sim \widetilde{X}_2$ almost surely.

☞ How do you partial out X_1?

OLS and population OLS: FWL theorems

- Question: Consider a fixed design $X : n \times p$ with linearly independent columns. Let $Z : n \times (p-1)$ be X without the last column.

For a random response vector $Y \in \mathbb{R}^n$, let

$$\widehat{\beta}_X := \arg\min_b \|Y - Xb\|^2, \quad \widehat{\beta}_Z := \arg\min_{b\text{:last entry of } b \text{ is zero}} \|Y - Xb\|^2.$$

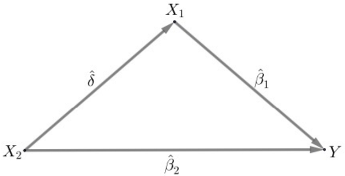
1. What are $\widehat{\beta}_X$ and $\widehat{\beta}_Z$ in closed form?
2. Which gives a higher R^2?
3. For fitted values, is it true that $\|\widehat{Y}_Z\| \leq \|\widehat{Y}_X\|$?
4. Is it true that $\mathrm{var}(\widehat{\beta}_Z)_1 \leq \mathrm{var}(\widehat{\beta}_X)_1$?

FWL theorem and Cochran's formula

- **FWL**: Get long regression coefficients from a partialled-out short regression.

Cochran's (omitted variable bias) formula

$$\widetilde{\beta}_2 = \widehat{\beta}_2 + \widehat{\delta}\,\widehat{\beta}_1$$



$$\begin{aligned} \text{long:} \quad & Y = X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + \widehat{\varepsilon}, \\ \text{short:} \quad & Y = X_2\widetilde{\beta}_2 + \widetilde{\varepsilon}, \\ \text{intermediate:} \quad & X_1 = X_2\widehat{\delta} + \widehat{U}. \end{aligned}$$

FWL theorem and Cochran's formula

- Question: When do you have $\widehat{\beta}_2 = \widetilde{\beta}_2$?

Multiple correlation coefficient R^2

Consider $Y \sim 1 + X$, where $X : n \times (p-1)$ and $(1_n, X)$ has linearly independent columns. Recall variance decomposition:

$$\underbrace{\sum_i (y_i - \bar{y})^2}_{\text{total var}} = \underbrace{\sum_i (\widehat{y}_i - \bar{y})^2}_{\text{var explained}} + \underbrace{\sum_i (y_i - \widehat{y}_i)^2}_{\text{var unexplained (RSS)}} . \quad (why?)$$

Multiple correlation coefficient

$$R^2 = (\text{var explained }\%) = \frac{\sum_i (\widehat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}.$$

☞

$$\mathrm{RSS} = (1 - R^2) \sum_i (y_i - \bar{y})^2.$$

Leverage and Leave-one-out

- Leverage of i-th observation:

$$h_{ii} = (X(X^\intercal X)^{-1}X^\intercal)_{ii} = x_i^\intercal (X^\intercal X)^{-1} x_i.$$

- LOO predicted residual:

$$\widehat{\varepsilon}_{[-i]} = \widehat{\varepsilon}_i/(1 - h_{ii}).$$

- Question: Under **GM**,

$$\operatorname{var} \widehat{\varepsilon}_i = ? \qquad \operatorname{var} \widehat{\varepsilon}_{[-i]} = ?$$

Model selection

- Mallow's $C_p = \|Y - \widehat{Y}\|^2 + 2p\sigma^2$

  ☞ unbiased estimate of MSPE (mean squared prediction error)

- $$\text{AIC} = n \log \frac{\text{RSS}}{n} + 2p$$

  ☞ For selecting model with small prediction error

- $$\text{BIC} = n \log \frac{\text{RSS}}{n} + p \log n$$

  ☞ For selecting the true model, when the linear model holds