Lecture 7-
Matrix Approach to Linear Regression & Multiple Linear Regression

周在莹

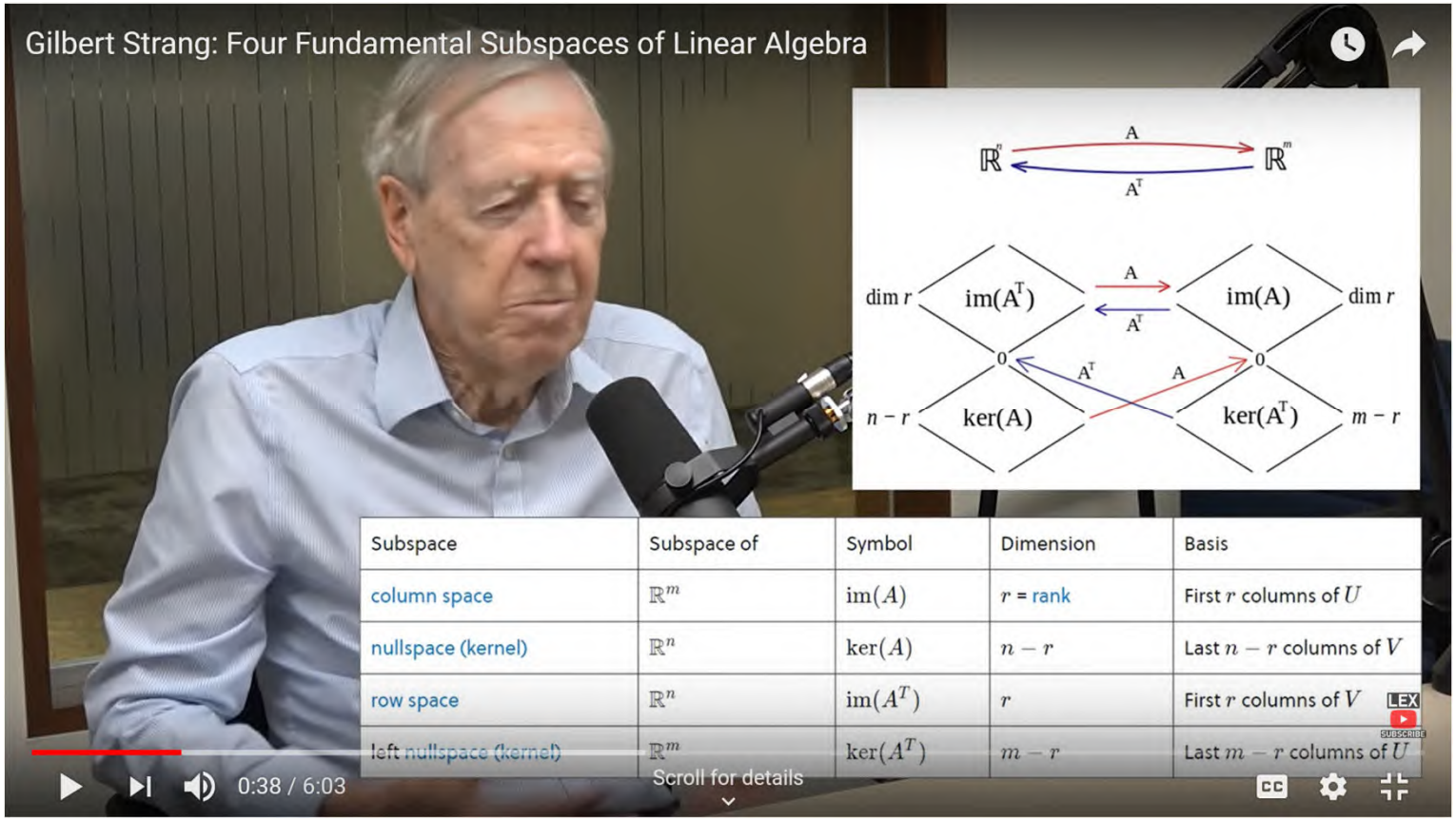
清华大学统计与数据科学系

zyzhou@mail.tsinghua.edu.cn



Warm up Before Class

- https://www.bilibili.com/video/BV1zx411g7gq?p=10



Some Properties of (Multivariate) Normal Distribution

(N1) Linear transformation of normal is still normal

$$Z \sim N(\mu, \Sigma) \Rightarrow AZ + c \sim N(A\mu + c, A\Sigma A^T).$$

($n \times 1$ vector; $n \times n$ matrix)

$E^*[(AZ+c)(AZ+c)^T] = E^*[(AZ)(AZ)^T] = E^*[AZZ^TA^T]$

$\mathrm{cov}(Z) = \begin{bmatrix} \mathrm{cov}(Z_1) & \mathrm{cov}(Z_1, Z_2) \\ T & \mathrm{cov}(Z_2) \end{bmatrix}$

(N2) When 1st and 2nd moments are given, the normal distribution is specified.

i.e., mean vector & variance-covariance matrix.

(N3) $Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim$ normal, and Z_1, Z_2 uncorrelated (i.e., $cov(Z_1, Z_2) = 0$)

By (N3) $\Rightarrow Z_1, Z_2$ independent

normal: $W = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} Z$, $\mathrm{cov}(W_1, W_2) = E^*(W_1 W_2^T) = E^*(A_1 Z Z^T A_2^T) = A_1 E^*(ZZ^T) A_2^T$

2 can be generalized to k

(N4) $Z \sim N(\mu, \Sigma)$, $W_1 = A_1 Z$, $W_2 = A_2 Z$

2 can be generalized to k & $A_i \Sigma A_j^T = 0$, $i \neq j$

By (N3) (N4) $\Rightarrow W_1, W_2$ are independent iff $A_1 \Sigma A_2^T = 0$.

If $\Sigma = \sigma^2 I$, then $A_1 \Sigma A_2^T = 0 \Leftrightarrow A_1 A_2^T = 0$

(N5) $Z \sim N(\mu, \Sigma)$, $W_1 = A_1 Z$, $W_2 = A_2 Z$, …, $W_k = A_k Z$, and $cov(W_i, W_j) = 0$ for $1 \leq i < j \leq k$, $\Rightarrow W_1^T W_1, W_2^T W_2, \ldots, W_k^T W_k$ are mutually independent.

length^2 of W_i

(N6) Z: an $n \times 1$ random vector and $Z \sim N(\mu, \Sigma)$, then

useful for the independence between sums of squares

$\Sigma^{-1} = (\Sigma^{-1/2})^T (\Sigma^{-1/2})$

- if Σ is non-singular, $(Z-\mu)^T \Sigma^{-1} (Z-\mu) \sim \chi_n^2$

$\Sigma^{-1/2}(Z-\mu) \sim N(0, I)$ ← standardization

- if Σ is singular and has rank r ($< n$), → The possible vectors of Z only occupy an r-dim subspace of $\mathbb{R}^n$

  let Σ^- be a generalized inverse of Σ (i.e., $\Sigma \Sigma^- \Sigma = \Sigma$), then

not unique

$$(Z-\mu)^T \Sigma^- (Z-\mu) \sim \chi_r^2$$



Theorems Used Today

- Suppose $U_{m\times 1} \sim N(\mu, \Sigma)$. Let $V_{d\times 1} = c + D_{d\times m}U$. Then,
$$V \sim N(c + D\mu, D\Sigma D^T)$$

- A_n is idempotent $\Leftrightarrow$ $rank(A) + rank(I - A) = n$

- $X \sim N(\mu, \Sigma) \Rightarrow E(X^T AX) = tr(A\Sigma) + \mu^T A\mu$



本节揭晓：

- 如何用矩阵和向量的形式来描述回归模型？
  - 相应的参数估计和分布函数又该怎么写？
- 线性回归的几何直观是什么？
- 你知道调整的 R^2 吗？好好的为什么要“调整”？



学习目标（精简版）

- 掌握
  - ✓ 熟练掌握用矩阵及向量描述回归数据、模型、参数估计、相应分布
  - ✓ LS的几何推导
  - ✓ 回归参数的解释
- 理解
  - ✓ Gauss-Markov 定理的条件和结论
  - ✓ MLR 模型相应的 ANOVA
  - ✓ 从 R^2 到调整的 R^2
- 了解
  - ✓ 线性变换四个基本子空间理论
  - ✓ 帽子（投影）矩阵的相关延拓



Topic 1:
Matrix Approach to Linear Regression



Outline

Simple Linear Regression (SLR) in another Perspective

- 1. Linear Regression in Matrix Form
- 2. OLS
  - 2.1 Analytic approach
  - 2.2 Geometric approach
- 3. Sampling Distribution



The Model in Scalar Form

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
  - The $\varepsilon_i's$ are independent Normally distributed random variables with mean 0 and variance σ^2
- Consider writing out the observations:

$$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1$$
$$Y_2 = \beta_0 + \beta_1 X_2 + \varepsilon_2$$
$$\vdots$$
$$Y_n = \beta_0 + \beta_1 X_n + \varepsilon_n$$

The Model in Matrix Form

- $$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The Model in Matrix Form

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

- Vector of responses

$$\boldsymbol{Y}_{n\times 1} = (Y_1, Y_2, \cdots, Y_n)^T$$

- Design Matrix

$$\boldsymbol{X}_{n\times 2} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

- Vector of parameters (coefficients)

$$\boldsymbol{\beta}_{2\times 1} = (\beta_0, \beta_1)^T$$

- Vector of error terms

$$\boldsymbol{\varepsilon}_{n\times 1} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$$



Variance-Covariance Matrix

$$\sigma^2(\boldsymbol{\varepsilon})_{n\times n} = Cov\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{bmatrix} = \sigma^2\boldsymbol{I}_{n\times n}$$

- The error terms are independent

  ⇒ the covariance between any two error terms is zero

- The error terms have common variance

  ⇒ the diagonal are equal to σ^2



Distributional Model

- The covariance matrices of $\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$ are the same, because the design matrix $\boldsymbol{X}$ is fixed

$$\sigma^2(\boldsymbol{Y})_{n\times n} = Cov\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \sigma^2\boldsymbol{I}_{n\times n}$$

- The distributional model:

$$\boldsymbol{Y}\sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}_n)$$



Parametric Estimation: Least Squares

- Analytic approach:

$$Q(\beta) = \|Y - X\beta\|^2 = (Y - X\beta)^T(Y - X\beta)$$

$$\hat{\beta} = \underset{\beta \in R^2}{\operatorname{argmin}}\, Q(\beta),$$

(Normal Equation)

$$\frac{\partial Q}{\partial \beta} = -2X^T(Y - X\beta) = 0, \qquad \frac{\partial^2 Q}{\partial \beta^2} = 2X^T X$$

- Under the condition $rank(X) = 2$, the minimizer of $Q(\beta)$ exists and is unique:

$$b = \hat{\beta} = (X^T X)^{-1} X^T Y$$



Least Squares Estimation

- Vector of predicted/fitted responses

$$\hat{Y} = Xb = X(X^T X)^{-1} X^T Y$$

$$\hat{Y} = HY$$

- Hat matrix (Projection matrix)

$$H = X(X^T X)^{-1} X^T$$

- Vector of residuals

$$e = Y - \hat{Y} = Y - HY = (I - H)Y$$



Column Space

$$\begin{bmatrix} X_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

- $\hat{Y} = Xb$

$$\begin{bmatrix} \hat{Y} \end{bmatrix} = \begin{bmatrix} X_1 & X_2 & \cdots & X_p \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} = b_1 \begin{bmatrix} X_1 \end{bmatrix} + b_2 \begin{bmatrix} X_2 \end{bmatrix} + \cdots b_p \begin{bmatrix} X_p \end{bmatrix} = \sum_{j=1}^{p} b_j X_j$$

  - $\hat{Y}$ is expressed as a linear combination of the columns X_j

- Column space:

  $C(X)$ = linear space spanned by column vectors of $X = Im(X)$

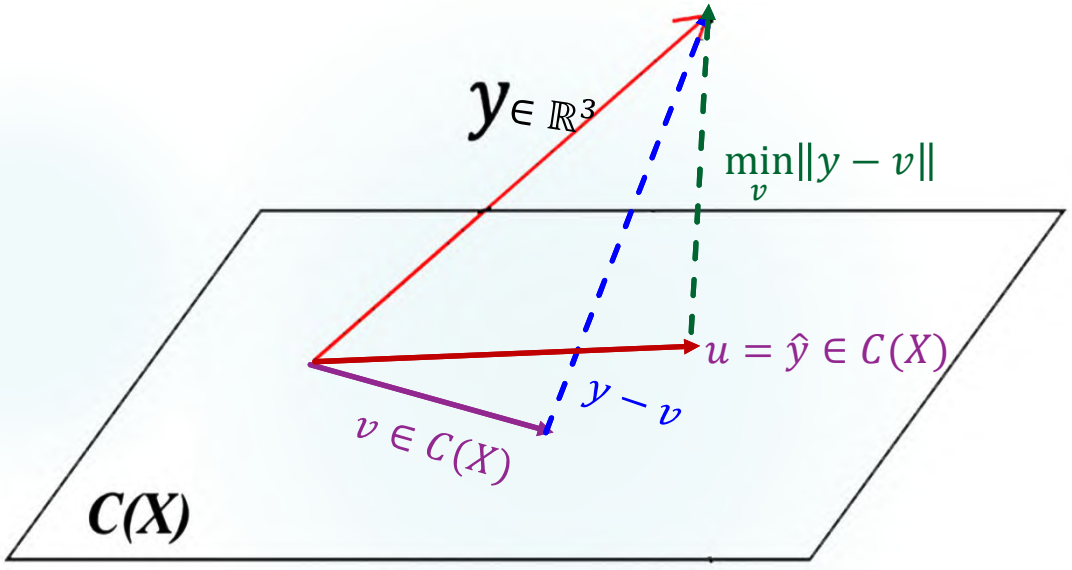
  - $\hat{Y} \in C(X)$

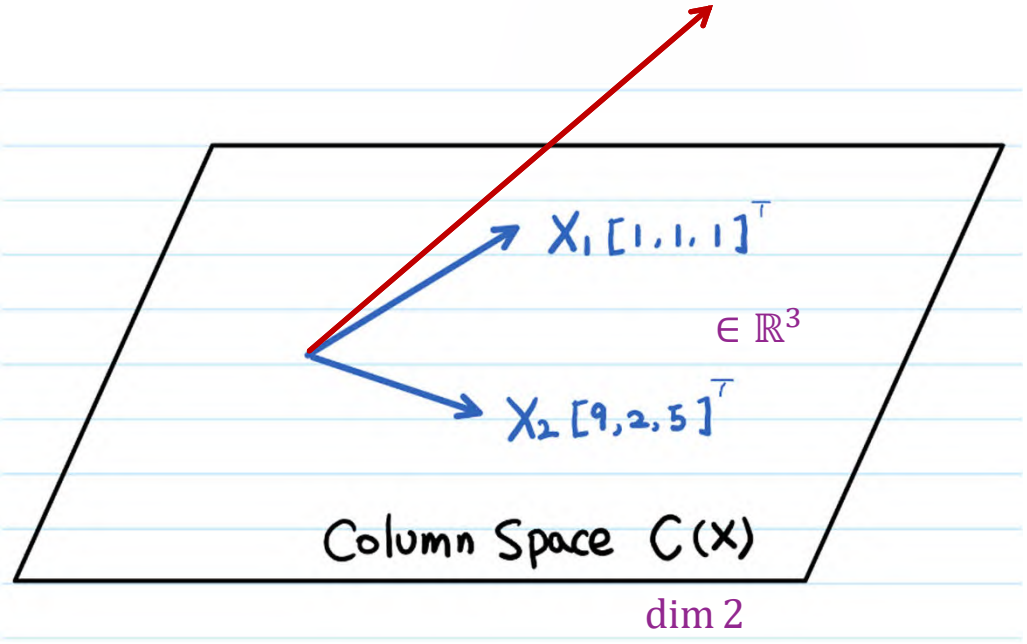


Projective Geometric Approach for LSE

- Column space:

$C(X)$ = linear space spanned by column vectors of $X = Im(X)$





单选题 1分

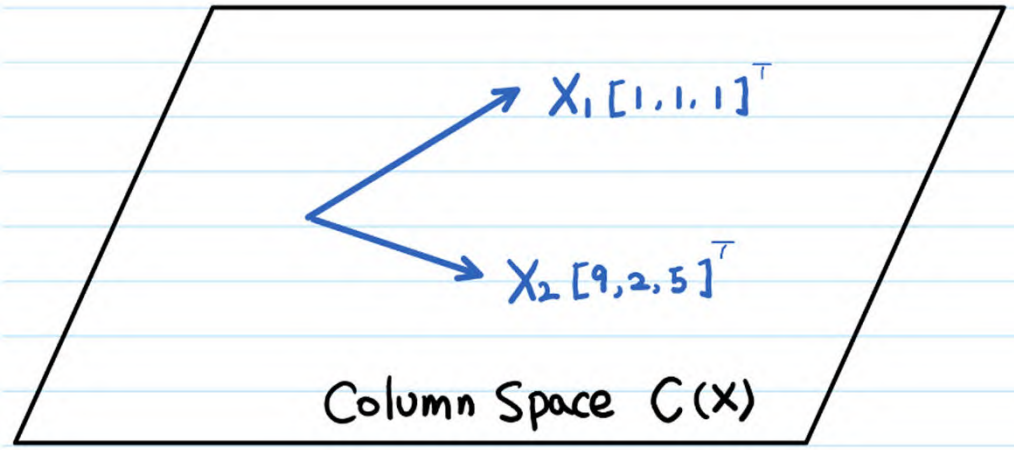
设置

17

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$X_{n\times p}$ 是从哪个空间（定义域）作用到哪个空间（值域）？

（注意 X 的像空间 $Im(X) = C(X)$ ）



$\mathbb{R}^n$ 到 $\mathbb{R}^p$，右图是 $\mathbb{R}^3$ 到 $\mathbb{R}^2$

$\mathbb{R}^p$ 到 $\mathbb{R}^n$，右图是 $\mathbb{R}^2$ 到 $\mathbb{R}^3$

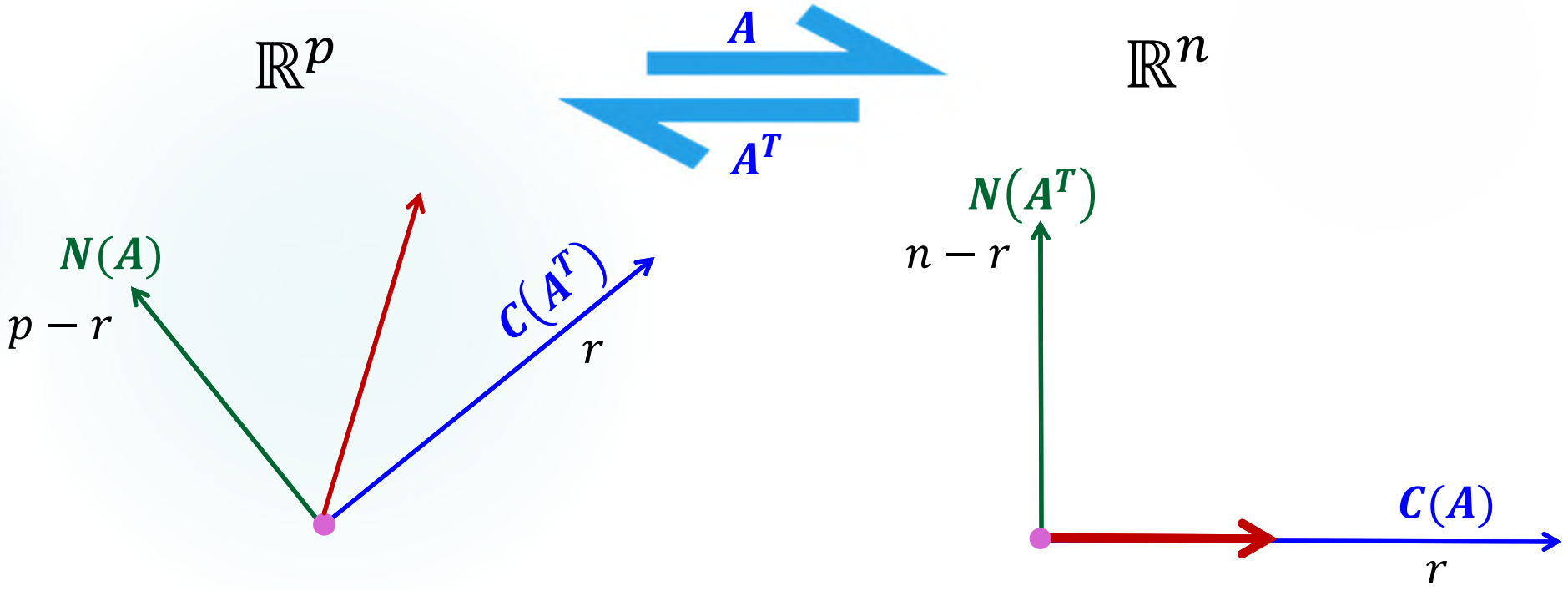
$\mathbb{R}^n$ 到 $\mathbb{R}^n$，右图是 $\mathbb{R}^3$ 到 $\mathbb{R}^3$

$\mathbb{R}^p$ 到 $\mathbb{R}^p$，右图是 $\mathbb{R}^2$ 到 $\mathbb{R}^2$

提交

Four Fundamental Subspaces (Gilbert Strang)

- At the heart of Linear Algebra. For any $A_{n\times p}$ with rank r



Alternative: Projective Geometry Approach for LSE

- Reformulate the problem

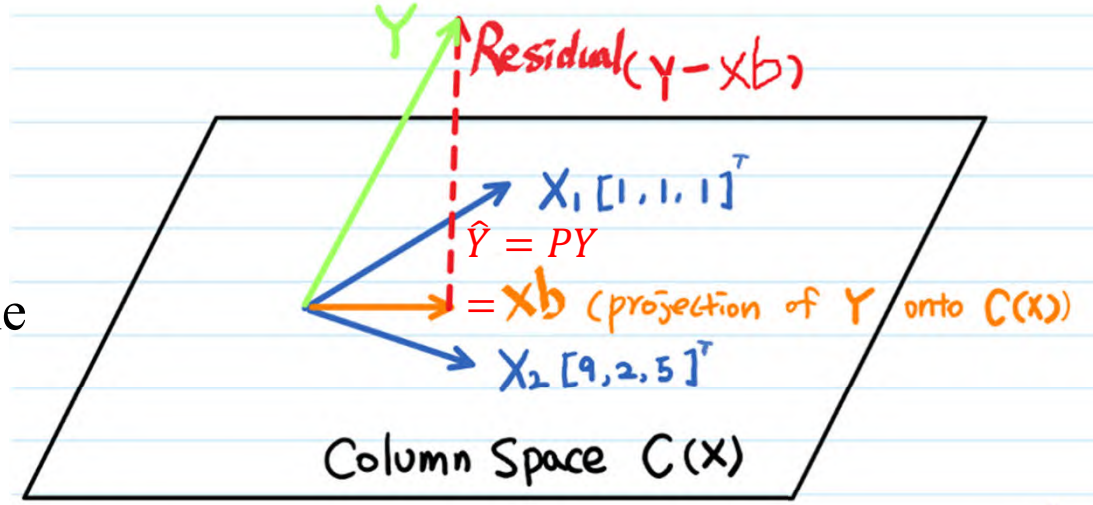
$$\min_{\beta \in R^2} \|Y - X\beta\|^2 = \min_{y = X\beta} \|Y - y\|^2$$

$$= \min_{y \in C(X)} \|Y - y\|^2$$

- The solution: the minimizer must be the orthogonal projection of Y onto the column space of X:

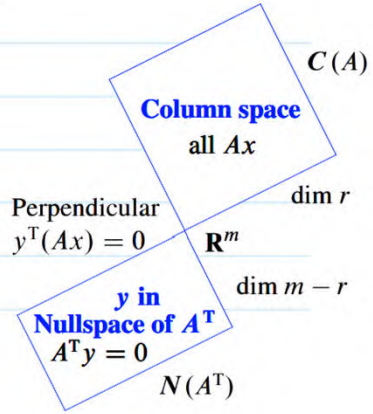
$$\hat{Y} = PY$$

where P is the orthogonal projection matrix/operator (Pythagoras)



$$X^T(y - X\hat{\beta}) = X^T \cdot e = 0$$

Also True : $X_1^T \cdot e = 0$, $X_2^T \cdot e = 0$



Projection Matrix and LS Estimates

- $X = \left[X_1 \middle| X_2 \middle| \cdots \middle| X_p \right]$

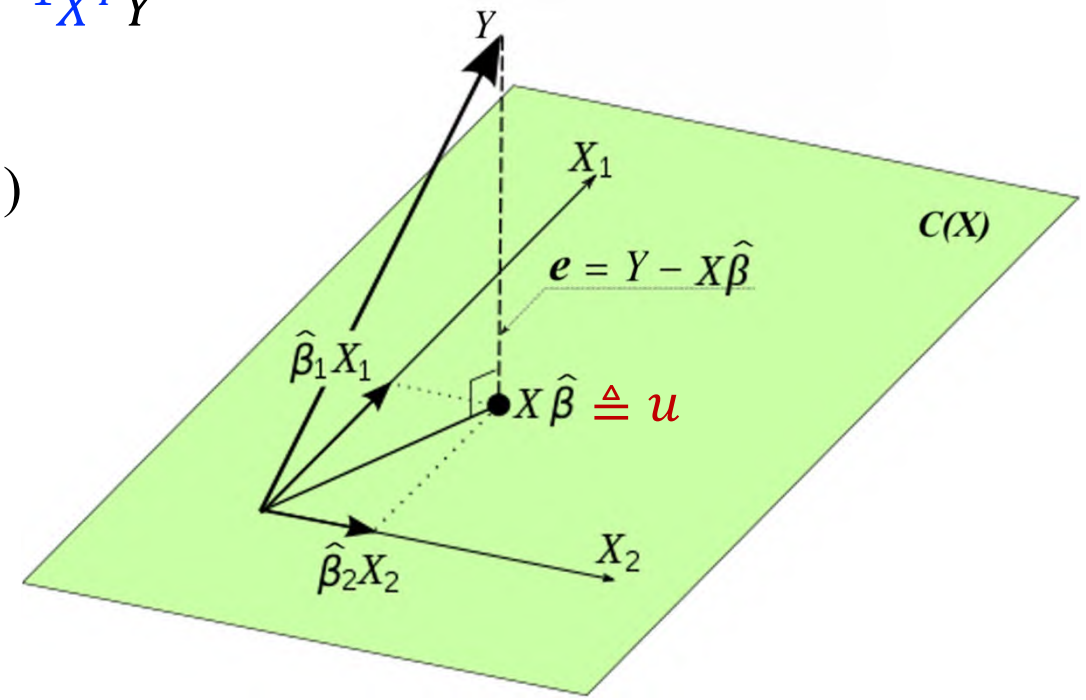
$X_j^T(Y - Xb) = 0$ for every j → $X^T(Y - Xb) = 0$ → $b = (X^TX)^{-1}X^TY$

$PY = u = Xb$

$= X(X^TX)^{-1}X^TY$

$P = X(X^TX)^{-1}X^T$

- $P = H = X(X^TX)^{-1}X^T$
  - Symmetric and idempotent ($P^2 = P$)
- $b = \hat{\beta} = (X^TX)^{-1}X^TY$
- Predicted responses: $\hat{Y} = HY$
- Residuals: $e = Y - \hat{Y} = (I - H)Y$
- Orthogonal: $e^T\hat{Y} = 0$



Sampling Distribution of b

- **Theorem**: Suppose $U_{m\times 1} \sim N(\mu, \Sigma)$. Let $V_{d\times 1} = c + D_{d\times m}U$. Then,
$$V \sim N(c + D\mu, D\Sigma D^T)$$

- Recall $Y \sim N(X\beta, \sigma^2 I_{n\times n})$, and $b = (X^T X)^{-1} X^T Y$. Applying the theorem,
$$b \sim N\left(\beta, \sigma^2 (X^T X)^{-1}\right)$$

- The estimated covariance matrix of b is
$$s^2(b) = s^2 (X^T X)^{-1}$$

where
$$s^2 = \frac{e^T e}{n-2} = \frac{Y^T (I - H) Y}{n-2}$$



Background Reading

- We will use this framework to do multiple regression $\Longrightarrow$ Now we can have more than one explanatory variable
- Adding another explanatory variable is to add another column in the design matrix X



Topic 2:

Multiple

Linear Regression



Outline

- 1. Building the Model
  - 1.1 Data and notation
  - 1.2 Model, parameters and special cases
- 2. Estimation
  - 2.1 OLS
  - 2.2 The Gauss-Markov theorem
  - 2.3 Concerning H
- 3. ANOVA
  - 3.1 Elements of the F-test
  - 3.2 R^2 & Adjusted R^2

Recall notes from simple linear regression regarding differences



Data for Multiple Regression

- Cases still denoted by $i = 1$ to n

$$\{(Y_i;\ X_{i,1}, X_{i,2}, \ldots, X_{i,p-1})\}_{1 \le i \le n}$$

  - $Y_i =$ response variable for the i^{th} case
  - $X_{i,1}, X_{i,2}, \ldots, X_{i,p-1}$ are the $p-1$ explanatory (or predictor) variables for the i^{th} case

$$X = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} \cdots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots \ \vdots & \vdots \\ 1 & X_{n-1,1} & X_{n-1,2} \cdots & X_{n-1,p-1} \\ 1 & X_{n,1} & X_{n,2} \cdots & X_{n,p-1} \end{bmatrix} = \left[\begin{array}{c|c|c|c} 1 & & & \\ \vdots & X_1 & \cdots & X_{p-1} \\ 1 & & & \end{array}\right] = \begin{bmatrix} \frac{X_{(1)}}{} \\ \frac{X_{(2)}}{} \\ \frac{\cdots}{} \\ \frac{X_{(n)}}{} \end{bmatrix}$$

Multiple Regression Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

- β_0 is the intercept
- $\beta_1, \beta_2, \ldots, \beta_{p-1}$ are the regression coefficients for the explanatory variables
  - Notice switch from slope to regression coefficient. More on this soon
- $X_{i,k}$ is the value of the k^{th} explanatory variable for the i^{th} case
- $\varepsilon_i' s$ are independent Normally distributed random errors with mean 0 and variance σ^2



Multiple Regression Parameters

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

- β_0, the intercept
- $\beta_1, \beta_2, \ldots, \beta_{p-1}$, the regression coefficients for the explanatory variables
  - β_1 indicates the change in the mean response $E\{Y\}$ per unit increase in X_1 when $X_2, \ldots, X_{p-1}$ are held constant
  - They are also called partial regression coefficients because they reflect the partial effect of one predictor variable when the other predictor variables are included in the model and are held constant
- σ^2, the variance of the error term



Some Special Cases

- Polynomial Model of order $p-1$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_{p-1} X_i^{p-1} + \varepsilon_i$$

- Interactions between explanatory variables can be represented as products of X's and included in the model (crossed terms)
- Some explanatory variables X can be indicator or dummy variables taking values of 0 and 1 or other distinct numbers



An Example: Children's Height

- The response variable is Height in metric of standard units (Y), and two explanatory variables Age in Months X_1 and Gender X_2



An Example: Children's Height

- The response variable is Height in metric of standard units (Y), and two explanatory variables Age in Months X_1 and Gender X_2
- Consider the following model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

where X_2 is a dummy variable, that is, $X_2 = 0$ if the child is a girl; and $X_2 = 1$ if the child is a boy

  - If Child i is a girl:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i, \ (X_{i2} = 0)$$

  - If Child i is a boy:

$$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{i1} + \varepsilon_i, \ (X_{i2} = 1)$$

- One model represents two different regression lines of Y versus X_1 for girls and boys, respectively



Model in Matrix Form

- Response vector: $Y = (Y_1, Y_2, \cdots, Y_n)^T$
- Design matrix: $$X = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{pmatrix}$$
- Regression coefficients: $\beta = \left(\beta_0, \beta_1, \cdots, \beta_{p-1}\right)^T$
- Error terms: $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$
- The model:

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$$

$$\varepsilon \sim N(0, \sigma^2 I_n),\ Y \sim N(X\beta, \sigma^2 I_n)$$



Least Squares Estimation

- Analytic approach:

$$Q(\beta) = \|Y - X\beta\|^2 = (Y - X\beta)^T(Y - X\beta)$$

$$\hat{\beta} = \underset{\beta \in R^2}{\operatorname{argmin}}\, Q(\beta), \frac{\partial Q}{\partial \beta} = -2X^T(Y - X\beta) = 0 \quad \textbf{(Normal Equation)}$$

$$\frac{\partial^2 Q}{\partial \beta^2} = 2X^T X$$

- Under the condition $rank(X) = p$, the minimizer of $Q(\beta)$ exists and is unique:

$$b = \hat{\beta} = (X^T X)^{-1} X^T Y$$



Least Squares Estimation

- Vector of predicted/fitted responses

$$\hat{Y} = Xb = X(X^T X)^{-1} X^T Y$$

$$\hat{Y} = HY$$

- Hat matrix (Projection matrix)

$$H = X(X^T X)^{-1} X^T$$

- Vector of residuals

$$e = Y - \hat{Y} = Y - HY = (I - H)Y$$



Projective Geometry Approach for LSE

- Reformulate the problem

$$\min_{\beta \in R^p} \|Y - X\beta\|^2 = \min_{v = X\beta} \|Y - v\|^2$$

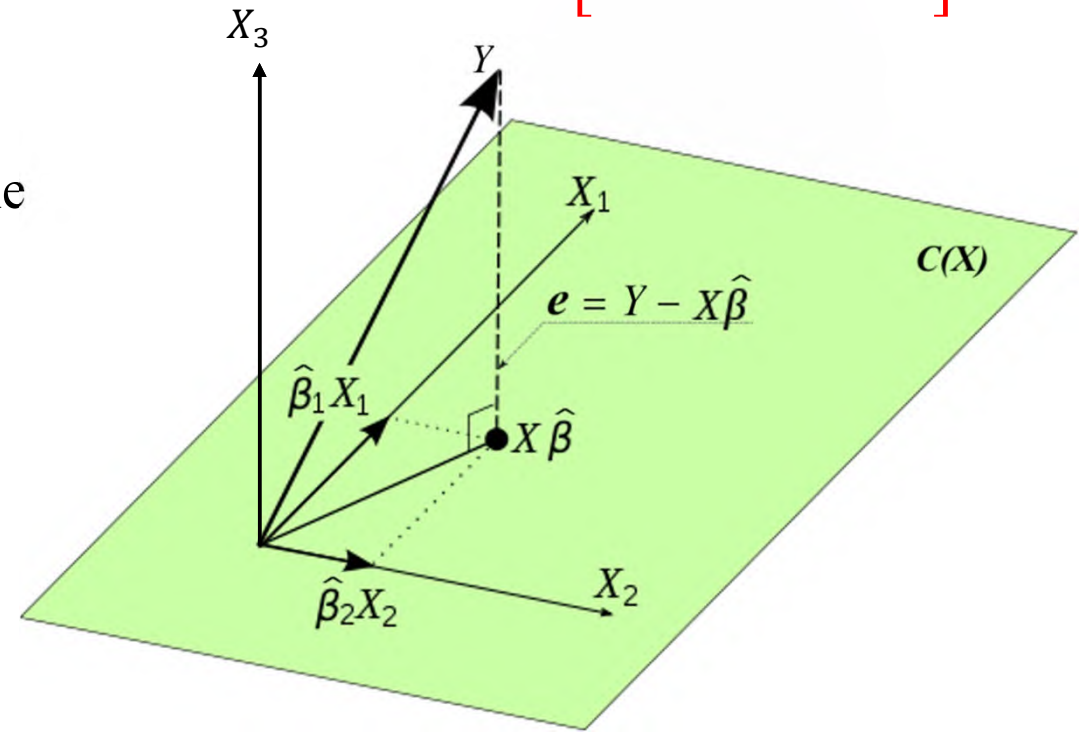
$$= \min_{v \in C(X)} \|Y - v\|^2$$

- The solution: the minimizer must be the orthogonal projection of Y onto the column space of X:

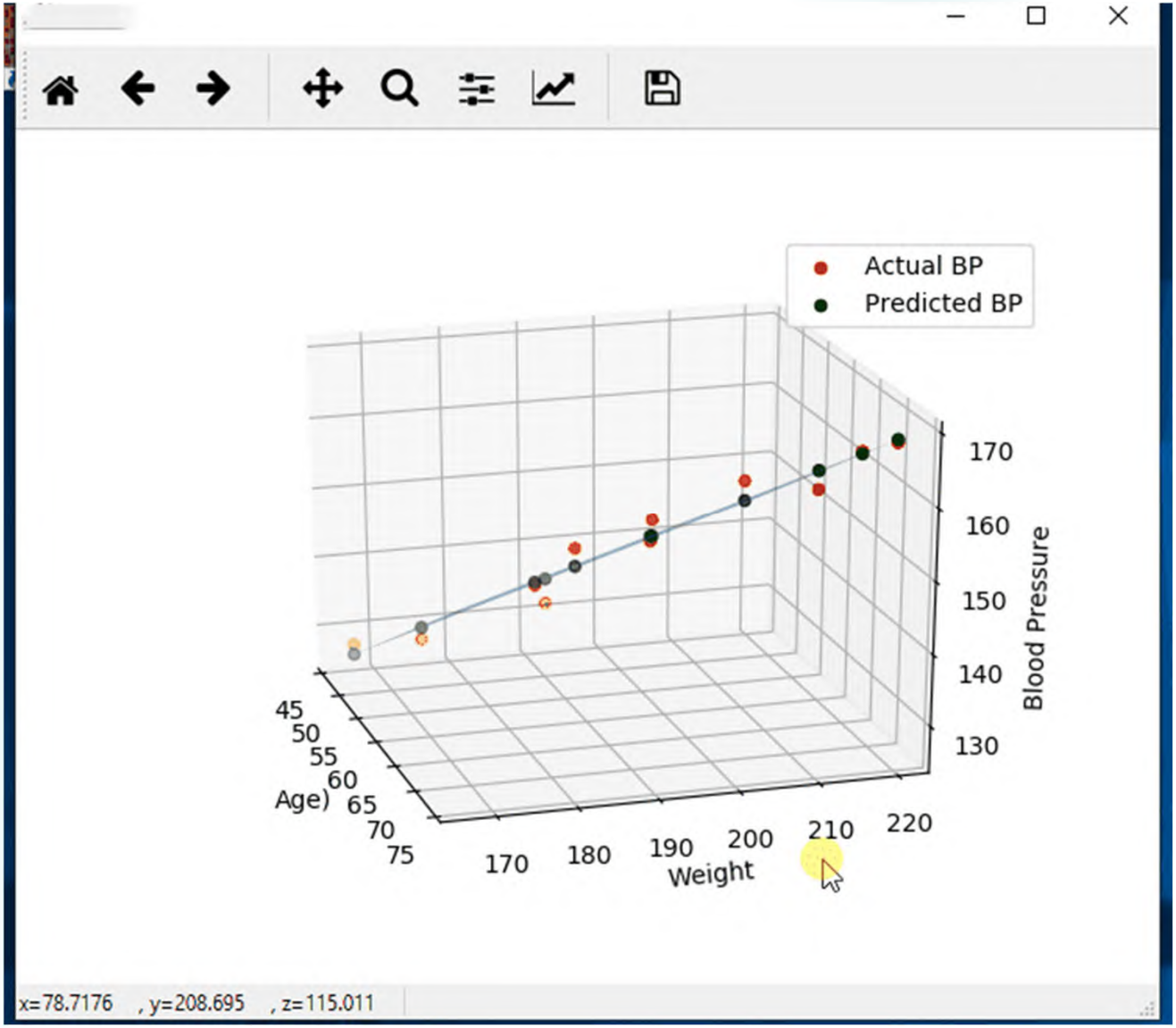
$$\hat{Y} = PY$$

where P is the orthogonal projection matrix/operator (Pythagoras)

$$\begin{bmatrix} X_1 & X_2 & \cdots & X_p \end{bmatrix}$$



Visualization



Projection Matrix and LS Estimates

- It can be shown that P is symmetric and idempotent ($P^2 = P$)
- It turns out that

$$P = H = X(X^TX)^{-1}X^T$$

- From

$$PY = X(X^TX)^{-1}X^TY = Xb$$
$$b = \hat{\beta} = (X^TX)^{-1}X^TY$$

- Predicted responses: $\hat{Y} = HY$
- Residuals: $e = Y - \hat{Y} = (I - H)Y$
- Residuals and predicted responses are orthogonal: $e^T\hat{Y} = 0$
- Degree of Freedom of residuals

$$df_E = rank(I - H) = n - p$$



Estimation of σ^2

- Sum of Squared Errors (Residuals)- SSE:

$$SSE = e^T e = (Y - Xb)^T (Y - Xb) = Y^T (I - H) Y$$

- df of SSE:

$$df_E = df \text{ of } e = n - p$$

- Mean of Squared Errors- MSE:

$$MSE = \frac{SSE}{df_E} = \frac{Y^T (I - H) Y}{n - p}$$

- Estimator of σ^2

$$s^2 = MSE$$

$$s = \sqrt{MSE}$$



Sampling Distribution of b

- Recall $Y \sim N(X\beta, \sigma^2 I_n)$, and $b = (X^T X)^{-1} X^T Y$. Applying the theorem,

$$b \sim N(\beta, \sigma^2 (X^T X)^{-1})$$

  - Mean: $E(b) = \beta$, unbiased
  - Covariance matrix: $Cov(b) = \sigma^2 (X^T X)^{-1}$, optimal in some sense

- The estimated covariance matrix of b is

$$s^2(b) = s^2 (X^T X)^{-1}$$

where

$$s^2 = \frac{e^T e}{n - p} = \frac{Y^T (I - H) Y}{n - p}$$



Gauss-Markov Theorem

- Five Gauss Markov assumptions/conditions:
  - **Linearity**: the parameters are themselves linear
  - **Randomness**: the data must have been randomly sampled from the population
  - **Non-Collinearity**: the regressors aren't perfectly correlated with each other
  - **Exogeneity**: the regressors aren't correlated with the error term $E(\varepsilon|X) = 0$
  - **Homoscedasticity**: no matter what the values of the regressors might be, the error of the variance is constant

- Then the ordinary least squares estimate for regression coefficients gives you the ***best linear unbiased estimate* (*BLUE*)** possible



Proof, Illustration, MSE

- $E(\hat{\beta}) = E(DY + b) = E(DX\beta + D\varepsilon) + \beta = \beta$
  - $\therefore DX = 0$

- $Var(\hat{\beta}) = Var(DY + b)$
  - $= Var(D\varepsilon + (X^TX)^{-1}X^T\varepsilon)$
  - $= \sigma^2\{D + (X^TX)^{-1}X^T\}\{D + (X^TX)^{-1}X^T\}^T$
  - $= \sigma^2\{DD^T + (X^TX)^{-1} + 2DX(X^TX)^{-1}\}$
  - $= \sigma^2\{DD^T + (X^TX)^{-1}\}$
  - $\geq \sigma^2(X^TX)^{-1} = Var(b)$



投票 最多可选1项

设置

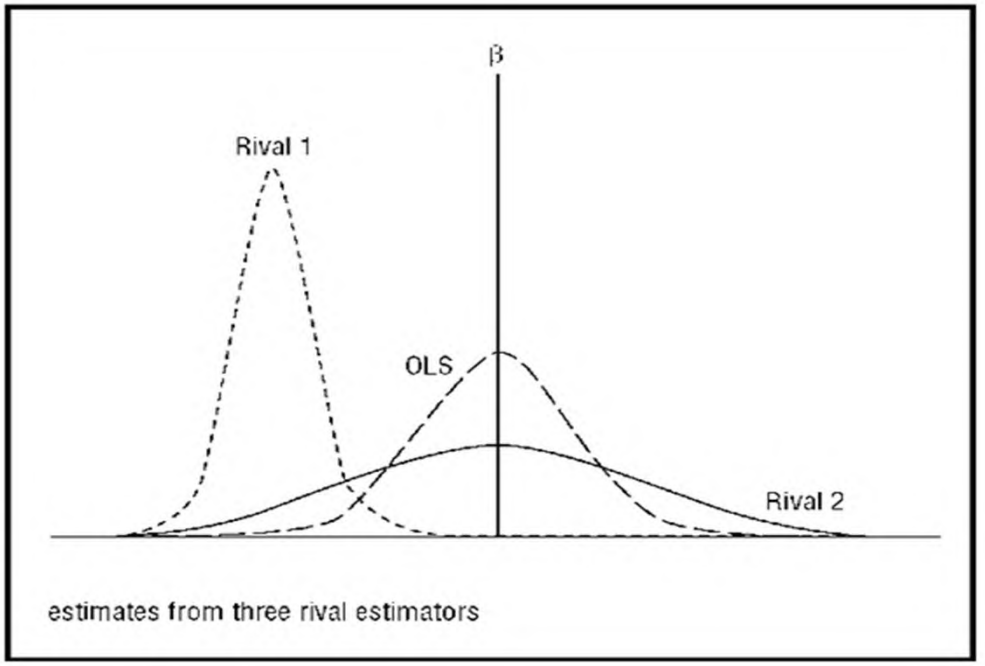
41

以下三个估计，大家觉得实际应用中你会选用谁？

A Rival 1

B Rival 2

C OLS



提交

Concerning H

- Recall that

$$P = H = X(X^T X)^{-1} X^T$$

  - The projection matrix/ hat matrix/ influence matrix

- $\hat{Y} = HY$
  - It maps the vector of response values to the vector of fitted/predicted values
  - It describes the influence each response value has on each fitted value
- $h_{ii} = \frac{\partial \hat{Y}_i}{\partial Y_i} = \partial \frac{\sum_{j=1}^{n} h_{ij} Y_j}{\partial Y_i}$
  - The diagonal elements are the leverages, which describe the influence each response value has on the fitted value for that same observation

  - $H = H^2 \Rightarrow h_{ii} = h_{ii}^2 + \sum_{i \neq j} h_{ij}^2 \geq 0$
  - $h_{ii} \geq h_{ii}^2 \Rightarrow h_{ii} \leq 1$



Why the Distance?

- $X_{(i)} \triangleq (1, X_{i1}, X_{i2}, \ldots, X_{i,p-1})$ for $i = 1, \ldots, n$

  $H = X(X^TX)^{-1}X^T \Rightarrow h_{ii} = X_{(i)}(X^TX)^{-1}X_{(i)}^T;\ \ h_{ij} = X_{(i)}(X^TX)^{-1}X_{(j)}^T$

- E.g., for the simple linear regression,

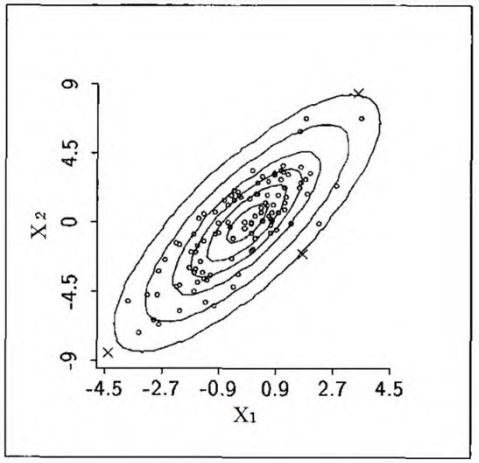
$$h_{ii} = [1 \quad X_{1i}](X^TX)^{-1}\begin{bmatrix}1\\X_{1i}\end{bmatrix}$$

$$X^TX = \begin{bmatrix} n & \sum X_{1j} \\ \sum X_{1j} & \sum X_{1j}^2 \end{bmatrix},\ (X^TX)^{-1} = \frac{1}{n\sum X_{1j}^2 - (\sum X_{1j})^2}\begin{bmatrix} \sum X_{1j}^2 & -\sum X_{1j} \\ -\sum X_{1j} & n \end{bmatrix}$$

$$\Rightarrow\ h_{ii} = \frac{\sum X_{1j}^2 + nX_{1i}^2 - 2X_{1i}\sum X_{1j}}{n\sum X_{1j}^2 - (\sum X_{1j})^2} = \frac{\left(\sum X_{1j}^2 - n\bar{X}^2\right) + n(X_{1i} - \bar{X})^2}{n(\sum X_{1j}^2 - n\bar{X}^2)} = \frac{1}{n} + \frac{(X_{1i} - \bar{X})^2}{\sum_{j=1}^{n}(X_{1j} - \bar{X})^2}$$

- To measure the distance between $X_{(i)}$ and $\bar{X}$ in terms of the density of the sample. All points on the same contour have the same leverage
  - (Recall the Mahalanobis Distance $D(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{y})}$)

Write $X_{i.} = (1; \boldsymbol{x_i})$, then $D^2(\boldsymbol{x_i}, \bar{\boldsymbol{x}}) = (n-1)(h_{ii} - 1/n)$



Notes on $df_E = rank(I - H) = n - p$

- df of SS = the rank of a quadratic form $x^T Ax = rank(A)$
- For any idempotent matrix A, $rank(A) = tr(A) = \sum \lambda_i$, here only 0 or 1

- **Theorem**: A_n is idempotent $\Leftrightarrow$ $rank(A) + rank(I - A) = n$

- With $X_{n,p}$ and idempotent H, $I - H$ is also idempotent, that is, $(I - H)^2 = I - H$,

$$tr(H) = tr\left(X\left(X^T X\right)^{-1} X^T\right) = tr\left(I_p\right) = p \qquad \sum h_{ii} = p$$

- For residuals $e = (I - H)Y, Q = e^T e = Y^T (I - H)Y$

$$df(Q) = rank(I - H) = tr(I - H) = n - p$$



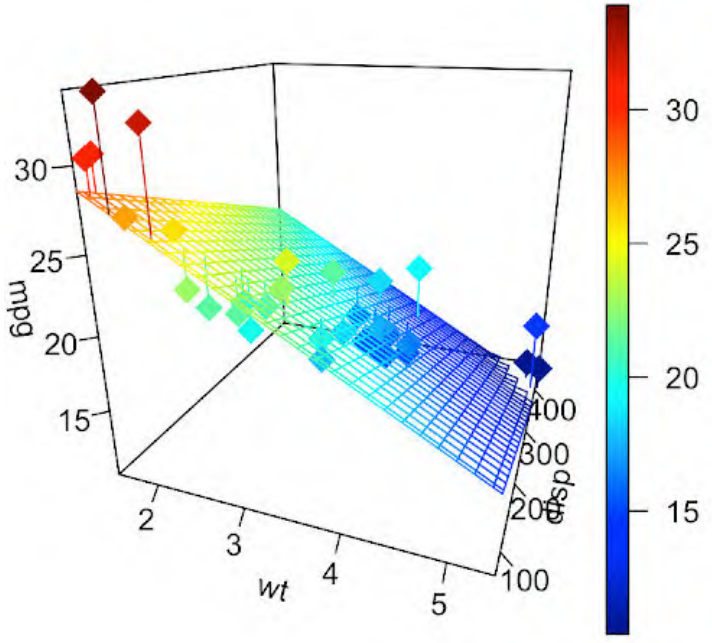
Covariance Matrix of Residual $e = (I - H)Y$

- Eigenvalues of $I - H$ are either 1 or 0
- Covariance matrix of residuals e

$$Cov(e) = Cov\big((I - H)Y\big)$$

$$= (I - H)\sigma^2 I(I - H)^T = \sigma^2(I - H)$$

$$\Rightarrow \begin{cases} Var(e_i) = \sigma^2(1 - h_{ii}) \\ Cov(e_i, e_j) = -\sigma^2 h_{ij} \end{cases}$$



单选题 1分

设置

51

提醒：随机误差($\boldsymbol{\varepsilon}$)不是残差 ($\boldsymbol{e}$)！由

$$\boldsymbol{Var(e_i) = \sigma^2(1 - h_{ii})},$$

在随机误差同方差假设满足的条件下，残差的方差有什么规律？由此可知线性回归中，$\boldsymbol{X}$ 位于中间或者两端的拟合能力是否有差别？

两端强

中间强

一样强

无法比较

提交

Does Dependency Affect a Normality Test?

$$Cov(e) = \sigma^2(I - H), \qquad Cov(e_i, e_j) = -\sigma^2 h_{ij}$$

- $\sum_{i=1}^{n} e_i = 0 \Rightarrow 0 = \sum_{i=1}^{n} Var(e_i) + \sum\sum_{i \neq j} Cov(e_i, e_j) = \sigma^2(n - tr(H)) + n(n-1)\overline{Cov(e_i, e_j)}$
  - $\Rightarrow \overline{Var(e_i)} = \sigma^2(1 - p/n),\ \overline{Cov(e_i, e_j)} = -\frac{\sigma^2(1-p/n)}{n-1}$
  - $\Rightarrow \overline{Cor(e_i, e_j)} \approx -\frac{1}{n-1} \approx 0$
  - Don't make much difference in practice, as long as p/n is small
- From the LM test, the Jarque–Bera test uses skewness S & kurtosis K

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right) \overset{A}{\sim} \chi_2^2$$

- Jarque, C. & Bera, A. (1980) Efficient tests for normality homoscedasticity and serial independence of regression residuals, Econometric Letters, 6, 255–259.
- Jarque, C. & Bera, A. (1987) A test for normality of observations and regression residuals, International Statistical Review, 55, 163–172
- Thorsten Thadewald & Herbert Büning (2007) Jarque–Bera Test and its Competitors for Testing Normality – A Power Comparison, Journal of Applied Statistics, 34:1, 87-105.
- Zacharias Psaradakis & Marián Vávra (2020) Normality tests for dependent data: large-sample and bootstrap approaches, Communications in Statistics - Simulation and Computation, 49:2, 283-304.

ANOVA

- Sources of variation include
  - Model (SAS) or Regression (KNNL)
  - Error (SAS, KNNL) or Residual (R)
  - Total
- SS and df add/decompose as before
- $SSM + SSE\ (RSS\ in\ R) = SST$
- $df_M + df_E = df_T$



Sum of Squares

- Sum of Squares due to Model- SSM:

$$SSM = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = (\hat{Y} - \bar{Y}1_n)^T (\hat{Y} - \bar{Y}1_n), \qquad 1_n = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

- Sum of Squares due to Error- SSE:

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = (Y - \hat{Y})^T (Y - \hat{Y})$$

- Sum of Squares in Total- SST:

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = (Y - \bar{Y}1_n)^T (Y - \bar{Y}1_n)$$



Sum of Squares and Mean Squares

- Can show that

$$(Y-\bar{Y}1_n)^T(Y-\bar{Y}1_n)=\left(Y-\hat{Y}+\hat{Y}-\bar{Y}1_n\right)^T\left(Y-\hat{Y}+\hat{Y}-\bar{Y}1_n\right)$$
$$=\left(\hat{Y}-\bar{Y}1_n\right)^T\left(\hat{Y}-\bar{Y}1_n\right)+\left(Y-\hat{Y}\right)^T\left(Y-\hat{Y}\right)$$

- That is,

$$SST=SSM+SSE$$

- Can show that

$$df_T=n-1;\ df_M=p-1;\ df_E=n-p$$

- Mean Squares:

$$MST=\frac{SST}{df_T};MSM=\frac{SSM}{df_M};MSE=\frac{SSE}{df_E}$$



Notes on df of SST

- $\because \bar{Y} = \frac{1}{n} 1_n^T Y = \frac{1}{n} Y^T 1_n$
- $SST = (Y - \bar{Y}1_n)^T (Y - \bar{Y}1_n) = Y^T Y - 2Y^T \bar{Y} 1_n + 1_n^T 1_n \bar{Y}^2$
- $-2Y^T 1_n \bar{Y} + 1_n^T 1_n \bar{Y}^2 = -\frac{2}{n} Y^T 1_n 1_n^T Y + n\bar{Y}^2$

$$= -\frac{2}{n} Y^T 1_n 1_n^T Y + \frac{1}{n} Y^T 1_n 1_n^T Y = -\frac{1}{n} Y^T J Y$$

- $\therefore SST = Y^T \left(I - \frac{1}{n} J \right) Y, \quad J = \begin{pmatrix} 1 & \dots & 1 \\ \dots & \dots & \dots \\ 1 & \dots & 1 \end{pmatrix}$
- Notice that $\frac{1}{n} J$ is idempotent, $\text{rank}(J) = 1$, so $\text{rank}\left(I - \frac{1}{n} J \right) = n - 1$



ANOVA Table

Source	SS	df	MS	F
Model	$SSM = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$	$df_M = p - 1$	$MSM = \frac{SSM}{df_M}$	$\frac{MSM}{MSE}$
Error	$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	$df_E = n - p$	$MSE = \frac{SSE}{df_E}$	
Total	$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$	$df_T = n - 1$	$MST = \frac{SST}{df_T}$	



Expected Mean Squares

- Formula for expected MS

$$MSE = \frac{Y^T(I-H)Y}{n-p} = \frac{\varepsilon^T(I-H)\varepsilon}{n-p}$$

$$E(MSE) = \frac{E\left(\varepsilon^T(I-H)\varepsilon\right)}{n-p} = \frac{tr\left[(I-H)\left(\sigma^2 I\right)\right]}{n-p} = \sigma^2$$

- Can show that

$$E(MSM) = \sigma^2 + \frac{1}{p-1}(X\beta)^T\left(I-\frac{1}{n}J\right)X\beta$$

- Under good design $rank(X) = p$,
  - $(X\beta)^T\left(I-\frac{1}{n}J\right)X\beta = 0$ if and only if $\beta_1 = \beta_2 = \cdots \beta_{p-1} = 0$
  - $(X\beta)^T\left(I-\frac{1}{n}J\right)X\beta > 0$ when at least one $\beta_i \neq 0$



ANOVA F-Test

- Hypotheses:
  - $H_0: \beta_1 = \beta_2 = \cdots \beta_{p-1} = 0$
  - $H_1: \beta_k \neq 0$ for at least one k in $1, 2, \cdots, p-1$
- Test statistic:

$$F^* = \frac{MSM}{MSE}$$

- Sampling distribution under H_0:

$$F^* \sim F_{p-1,n-p}$$

- Decision rule at α
  - Reject H_0 if the calculated $F_0 > F_{p-1,n-p,\alpha}$
  - Reject H_0 if the P-value $P(F^* > F_0 | H_0) < \alpha$



Interpret Test Results

- Reject H_0:

  There exists evidence suggesting that one or more of the explanatory variables in the linear model is potentially useful for predicting (explaining) the response variable.

- Fail to reject H_0:

  There does not exist evidence to conclude that any of the explanatory variables can help model/predict/explain the response variable using the linear model



Notes on $E(MSM)$

Expectation of quadratic forms

- **Theorem**: $X \sim N(\mu, \Sigma) \Rightarrow E(X^T AX) = tr(A\Sigma) + \mu^T A\mu$
- $SSE = Y^T(I-H)Y$, $SSM = Y^T\left(H - \frac{1}{n}J\right)Y$, $E(MSM) = \frac{1}{p-1}E\left[Y^T\left(H - \frac{1}{n}J\right)Y\right]$
- $\because Y \sim N(X\beta, \sigma^2 I_n)$

$$E\left[Y^T\left(H - \frac{1}{n}J\right)Y\right] = (X\beta)^T\left(H - \frac{1}{n}J\right)(X\beta) + \sigma^2 tr\left(H - \frac{1}{n}J\right)$$

- $tr\left(H - \frac{1}{n}J\right) = tr(H) - tr\left(\frac{1}{n}J\right) = rank(H) - rank\left(\frac{1}{n}J\right) = p - 1$
- $(X\beta)^T\left(H - \frac{1}{n}J\right)(X\beta) = \beta^T\left(X^T HX - \frac{1}{n}X^T JX\right)\beta$

$$= \beta^T\left(X^T X - \frac{1}{n}X^T JX\right)\beta = (X\beta)^T\left(I - \frac{1}{n}J\right)X\beta$$

- Since $\left(I - \frac{1}{n}J\right)$ is idempotent, thus semidefinite positive

Notes on the Proof of the Previous Theorem

- **Theorem**: $X \sim N(\mu, \Sigma) \Rightarrow E(X^T AX) = tr(A\Sigma) + \mu^T A\mu$

- Proof: (The trace trick!)
  - $E\left(X^T AX\right) = E(X-\mu)^T A(X-\mu) + \mu^T A\mu$

$$= tr\{E[A(X-\mu)(X-\mu)^T]\} + \mu^T A\mu$$

$$= tr\{AE[(X-\mu)(X-\mu)^T]\} + \mu^T A\mu$$

$$= tr(A\Sigma) + \mu^T A\mu$$



Coefficient of Multiple Determination R^2

- Correlation between responses Y and predicted responses $\hat{Y}$

$$r = \frac{\sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum(Y_i - \bar{Y})^2}\sqrt{\sum(\hat{Y}_i - \bar{Y})^2}} = \frac{\sqrt{\sum(\hat{Y}_i - \bar{Y})^2}}{\sqrt{\sum(Y_i - \bar{Y})^2}}$$

- CMD R^2:

$$R^2 = r^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{SSM}{SST}$$

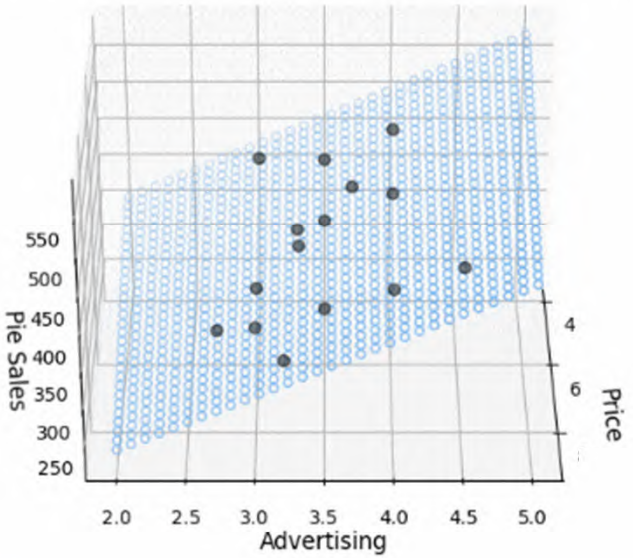
- Interpretation:

  R^2 gives the proportion of variation in the response variable, which can be explained by the model or all the explanatory variables in the model

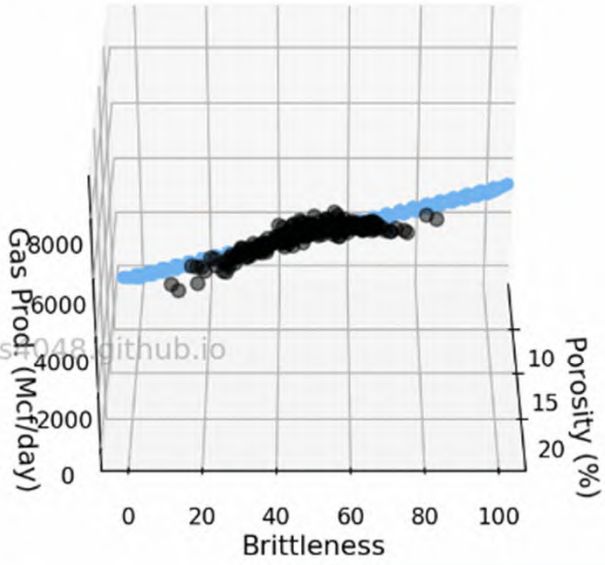


Illustration

Multi-Linear Regression Model Visualization ($R^2 = 0.52$)

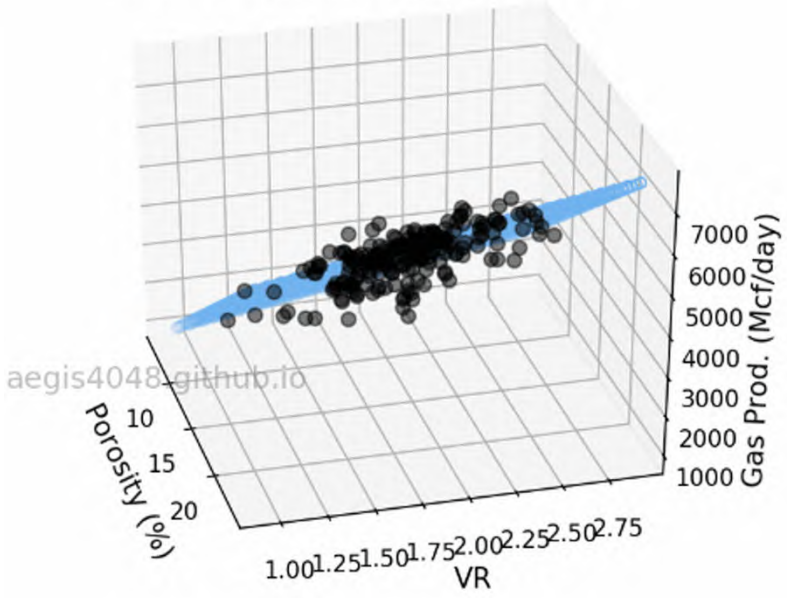


Porosity and Brittleness, $R^2 = 0.93$



aegis4048.github.io

Porosity and VR, $R^2 = 0.79$



aegis4048.github.io



单选题 1分

设置

65

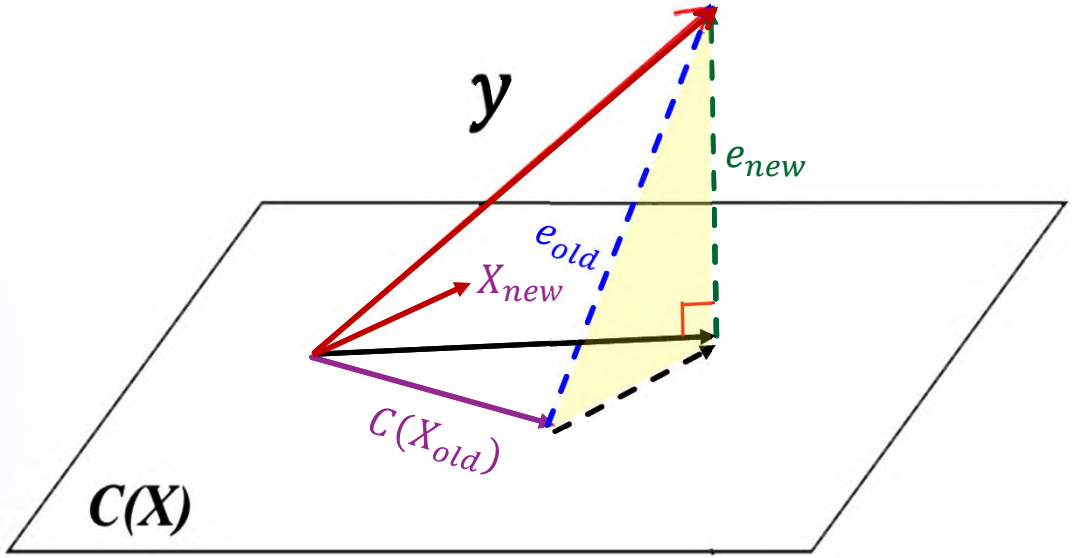
模型中增加一个新的解释变量，且它不能被之前的解释变量线性表出，则模型的SSE会?

下降

上升

不变

不一定



提交

R^2 & Adjusted R^2

- In terms of SSE:

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-p}$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \text{proportion not explained}$$

- Relation to F-test statistics:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-p}{p-1}$$

- Adjusted CMD R_a^2:

$$R_a^2 = 1 - \frac{MSE}{MST} = 1 - \frac{n-1}{n-p} \cdot \frac{SSE}{SST}$$

- When adding more explanatory variables, R^2 always increases, but R_a^2 can decrease



第7讲 矩阵方法以及多重回归

- 7.1 矩阵视角下的简单线性回归
  - 7.1.1 矩阵形式的回归模型
  - 7.1.2 最小二乘估计
    - 7.1.2.1 解析法
    - 7.1.2.2 几何法
  - 7.1.3 抽样分布
- 7.2 多重线性回归分析
  - 7.2.1 建立模型
    - 7.2.1.1 数据、记号
    - 7.2.1.2 模型、参数及其特例
  - 7.2.2 参数估计
    - 7.2.2.1 OLS
    - 7.2.2.2 Gauss-Markov定理
    - 7.2.2.3 与 H 相关
  - 7.2.3 ANOVA
    - 7.2.3.1 F检验的各个要素
    - 7.2.3.2 R^2、调整的 R^2



Objectives

- Know how to write data and model in the matrix form.
- Remember the formulae for the least squares estimation.
- Know the four fundamental subspaces for a linear transformation.
- Understand the projective geometry approach for LSE.
- Be able to interpret the coefficients of a multiple regression model.
- Understand what the scope of the model is in the multiple regression model.
- Know the sampling distributions of b and e.
- Know the Gauss-Markov Theorem for the LSE.
- Understand the importance and related concepts of the hat matrix.
- State the ANOVA F-test for the whole model.
- Understand the calculation and interpretation of R^2 in a multiple regression setting.
- Understand the calculation and use of adjusted R^2 in a multiple regression setting.



Background Reading

- KNNL 5.1 to 5.7



Homework

- Prob-1: KNNL 5.4, KNNL 5.12
- Prob-2: KNNL 5.23. (For part *c*, find the hat matrix *H* and its rank, and verify that *H* is idempotent.)
- Bonus question: A student shared this interview question with me. Try to answer it yourself.

2）一列随机变量 X_i 期望为 0，方差为 1，两两之间相关系数为 p。

Y=a1X1+a2X2+...+anXn。用矩阵形式求 E(Y) Var(Y)。求 p 的最小值



下节预告：

- 还记得建立 SLR模型的流程吗？比较而言，建立MLR模型的难点在于？一般如何应对？
- 单个变量的显著性和整个回归模型的显著性有什么联系？
- 关于SLR的统计推断，你可以用矩阵形式重述出来吗？
- …
- 下一次课，我们一起找寻这些问题的答案啊~~~

