Normal Equations & Geo Interpretations

1、Simple linear regression model

Data (x_i, Y_i), $i = 1, \dots, n$

model : $Y_i = \beta_0 + \beta_1 x_i + e_i$

random $\mathbb{E}[e_i] = 0$

dependent var./observation

intercept

indep. var./predictor

noise : measurement noise, biological variation

$\theta = [\beta_0, \beta_1]^T$ is the param vector/weights

$\mathbb{E}[Y_i] = \beta_0 + \beta_1 x_i$ = lin. comb. of unknows β's, w/ known coeff $(1, x_i)$.

$$\underset{\sim}{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} \quad X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}_{n\times 2}, \quad \underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}_{2\times 1}, \quad \underset{\sim}{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}$$

$$X = \Big[\; \underbrace{\begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}}_{\underset{\sim}{\mathbb{1}}} \;\Big|\; \underbrace{\begin{matrix} x_1 \\ \vdots \\ x_n \end{matrix}}_{\underset{\sim}{x}} \;\Big]$$

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{e}$$ "Matrix-vector form".

$\uparrow$ data matrix

$$E[\underset{\sim}{Y}] = X\underset{\sim}{\beta} = [\underset{\sim}{\mathbb{1}} \;\; \underset{\sim}{x}] \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \beta_0 \underset{\sim}{\mathbb{1}} + \beta_1 \underset{\sim}{x}$$

2. (Multiple) Linear regression model

$$y_i = \sum_{j=1}^{p} x_{ij} \beta_j + e_i , \quad i=1,\dots,n.$$

$$\underset{\sim}{Y}_{n\times 1} = X_{n\times p} \underset{\sim}{\beta}_{p\times 1} + \underset{\sim}{e}_{n\times 1}$$ vector of rand. elements.

$$\mathrm{VarCov}(u) = \mathbb{E}\left[(u-\mathbb{E}u)(u-\mathbb{E}u)^T\right]$$

$n\times 1$ $1\times n$

Common Assumptions:

$$\mathbb{E}[\underset{\sim}{e}] = \underset{\sim}{0}$$

$$\mathrm{Var}(e_i) = \sigma^2 , \quad i=1,\dots,n.$$

$$\mathbb{E}[e_i^2] = \mathrm{Var}(e_i) = \sigma^2$$

$$\mathrm{Cov}(e_i, e_{i'}) = \mathbb{E}[e_i e_{i'}] = 0 , \quad \forall i \neq i'$$

$$\mathrm{VarCov}(\underset{\sim}{e}) = \sigma^2 \mathbb{I} = \mathbb{E}[ee^T]$$

$$\mathrm{VarCov}(\underset{\sim}{Y}) = \mathrm{VarCov}(\underset{\sim}{e}) = \sigma^2 \mathbb{I}$$

Ex : (Regression in narrow sense) $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$, $i = 1, \dots, 50$.

Y_i: grade

x_{i1}: time spent on hw

x_{i2}: time spent on review

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{50} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{50,1} & x_{50,2} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_{50} \end{bmatrix}$$

Ex : [Analysis of Variance (ANOVA)] $Y_{ij} = \mu + \alpha_i + e_{ij}$, $i = 1, 2, 3$; $j = 1, 2$.

$$\underset{\sim}{Y} = \begin{pmatrix} y_{11} \\ y_{12} \\ \hline y_{21} \\ y_{22} \\ \hline y_{31} \\ y_{32} \end{pmatrix} \qquad X = \begin{array}{c} \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \\ \begin{array}{cccc} \mu & \alpha_1 & \alpha_2 & \alpha_3 \end{array} \end{array} \qquad \underset{\sim}{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$$

rank$(X) = 3$

μ = "overall effect"

α_i = "effect of i^{th} treatment"

$\hat{\theta}_1 = (\widehat{\alpha_1 - \alpha_2})$?

$\hat{\theta}_2 = (\widehat{\alpha_1 - (\alpha_2 + \alpha_3)/2})$

"Linear (in)dependent", "vector space", "basis", "column/row rank".

Linear Algebra Review

Hayes 2.3.2 ; Scheffe App. I

Given $\{\underset{\sim}{v}_1, \dots, \underset{\sim}{v}_n\}$. $\alpha_1 \underset{\sim}{v}_1 + \cdots + \alpha_n \underset{\sim}{v}_n = \underset{\sim}{0} \Rightarrow \begin{cases} \alpha_i = 0, \forall i : \text{"linearly independent"}. \\ \text{not all } \alpha_i \text{ are } 0 : \text{"linearly dependent"}. \end{cases}$

For "lin dependent" case, may write

$$\underset{\sim}{v}_1 = \beta_2 \underset{\sim}{v}_2 + \cdots + \beta_n \underset{\sim}{v}_n$$

Ex: $\underset{\sim}{v}_1 = [1\ 2\ 1]^T \quad \underset{\sim}{v}_2 = [1\ 0\ 1]^T$

$$\alpha_1 \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \underset{\sim}{0} \qquad \begin{cases} \alpha_1 + \alpha_2 = 0 \\ 2\alpha_1 + 0 = 0 \\ \alpha_1 + \alpha_2 = 0 \end{cases} \Rightarrow \begin{cases} \alpha_1 = 0 \\ \alpha_2 = 0 \end{cases} \Rightarrow \text{"lin. indep"}$$

Ex: $\underset{\sim}{v}_1 = [1\ 2\ 1]^T \qquad \underset{\sim}{v}_4 = [-2\ -4\ -2]^T$

$$\underset{\sim}{v}_4 = -2 \cdot \underset{\sim}{v}_1 \Rightarrow \text{"lin dependent"}$$

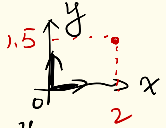
Ex: $v_1, v_2, v_3 = [0\ 1\ 0]^T$

$$v_1 = v_2 + 2v_3 \Rightarrow \text{"lin dep."}$$

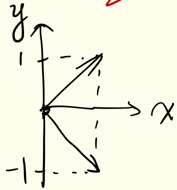
Def: <u>Vector space</u>: A set of all vectors that are lin. comb. of $\{\underset{\sim}{v_i}\}_{i=1}^{n}$, i.e., $V = \left\{ \underset{\sim}{v} = \sum_{i=1}^{n} \alpha_i \underset{\sim}{v_i} \,,\ \alpha_i \in \mathbb{R} \right\}$.

v_i's are said to <u>span</u> the vector space V, i.e., $V = \operatorname{span}\{\underset{\sim}{v_1}, \ldots, \underset{\sim}{v_n}\}$.

Ex: $V^{(1)} = \left\{ \alpha_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \alpha_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix},\ \alpha_i \in \mathbb{R} \right\}$

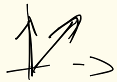


$V^{(2)} = \left\{ r_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + r_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix},\ r_i \in \mathbb{R} \right\}$



Def: A <u>basis</u> for V is a set of lin indep vectors that span V.

Ex: $\left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\}$ ✓ $\qquad \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$ ✓



$\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$ ✓ $\qquad \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\}$ ✗

Def: The <u>dimension</u> of vector space V is the # of vectors in any (a) basis of V. <u>Column/row rank</u>: dim of column/row vector space, respectively.

Ex: What's the rank of $V = \left\{ \underset{\sim}{v} = \alpha_1 \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + \alpha_3 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \alpha_i \in \mathbb{R} \right\}$?

Ans1: Basis $= \left\{ \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$ or

$\left\{ \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \right\}$ or $\left\{ \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$.

$\text{rank}(V) = 2$.

Ans2: $V = \left\{ \underset{\sim}{v} = \alpha_1 \left(\underset{\sim}{v_2} + 2\underset{\sim}{v_3} \right) + \alpha_2 \underset{\sim}{v_2} + \alpha_3 \underset{\sim}{v_3} \right.$

$\left. = (\alpha_1 + \alpha_2) \underset{\sim}{v_2} + (2\alpha_1 + \alpha_3) \underset{\sim}{v_3} \right\}$, $\underset{\sim}{v_2} \perp \underset{\sim}{v_3} \Rightarrow \text{rank}(V) = 2$.

3. Geometric interpretation of LS:

Problem Setup: $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{e}$, where $X \triangleq [\underset{\sim}{\xi}_1, \dots, \underset{\sim}{\xi}_p]$

Estimate $\underset{\sim}{\beta}$ such that $J(\underset{\sim}{\beta}) = \|\underset{\sim}{Y} - X\underset{\sim}{\beta}\|^2$ is minimized.

$$J(\underset{\sim}{\beta}) = \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$$

This is called "least-squares".

Claims: 1. $\hat{\underset{\sim}{\beta}}$ always exists, but

2. not always unique.

Goal Estimate β, σ^2; "error (variance) of estimators,

e.g., $Var(\hat\beta_0)$, $var(\hat\beta_1)$, ... ; $VarCov(\hat{\underset{\sim}{\beta}})$, $VarCov(\hat{\underset{\sim}{y}})$.

Method 1: $$\frac{\partial J}{\partial \beta_k} = \sum_{i=1}^{n} 2\left(Y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right) \underbrace{\frac{\partial}{\partial \beta_k}\left(-\left(\cdots + x_{ik}\beta_k + \cdots\right)\right)}_{-x_{ik}}$$

$$\boxed{J(\beta) = \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2}$$

$$= \left. 0 \right|_{\beta_j = \hat\beta_j}, \quad k = 1, \dots, p$$

$$\Longleftrightarrow \sum_i Y_i x_{ik} = \sum_i \sum_j x_{ij}\hat\beta_j x_{ik} \Longleftrightarrow \boxed{X^T \underset{\sim}{Y} = X^T X \hat{\underset{\sim}{\beta}}}$$ Normal Equation (NE)

where $$X^T \underset{\sim}{Y} = \left[\sum_{i=1}^{n} x_{ik} Y_i\right]_{p\times 1}, \quad X^T X = \left[\sum_{i=1}^{n} x_{ij} x_{ik}\right]_{p\times p} \quad \hat{\underset{\sim}{\beta}} = (X^T X)^{-1} X^T \underset{\sim}{Y}$$

$$X^T X \hat{\underset{\sim}{\beta}} = \left[\sum_{j=1}^{p}\left(\sum_{i=1}^{n} x_{ij} x_{ik}\right)\hat\beta_j\right]_{p\times 1}$$

Method 2 : $\nabla_{\underset{\sim}{\beta}} J(\underset{\sim}{\beta}) = \Big|_{\underset{\sim}{\beta} = \hat{\underset{\sim}{\beta}}} 0$

$$\boxed{J(\underset{\sim}{\beta}) = \|\underset{\sim}{Y} - X\underset{\sim}{\beta}\|^2}$$

$$\nabla_{\underset{\sim}{\beta}} J(\underset{\sim}{\beta}) = 2\left[-X^T(\underset{\sim}{Y} - X\underset{\sim}{\beta})\right] = \Big|_{\underset{\sim}{\beta} = \hat{\underset{\sim}{\beta}}} 0$$

$$\boxed{X^T \underset{\sim}{Y} = X^T X \hat{\underset{\sim}{\beta}}}$$

L.S. procedure: Find a vector in $\mathcal{C}(X)$ which is as close as possible to $\underset{\sim}{Y}$.

Claim: If $(\underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}}) \perp \mathcal{C}(X)$, then $\hat{\beta}$ solves N.E.

Proof: $(\underset{\sim}{Y} - \underset{\sim}{\hat{Y}}) \perp \mathcal{C}(X)$

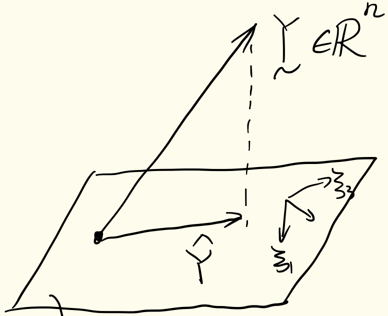
$\Longleftrightarrow (\underset{\sim}{Y} - \underset{\sim}{\hat{Y}}) \perp X\underset{\sim}{b}, \ \forall \underset{\sim}{b} \in \mathbb{R}^p$

$\Longleftrightarrow \underset{\sim}{\xi}_j^T(\underset{\sim}{Y} - \underset{\sim}{\hat{Y}}) = 0, \ j = 1, \dots, p$

$\Longleftrightarrow [\underset{\sim}{\xi}_1, \dots, \underset{\sim}{\xi}_p]^T(\underset{\sim}{Y} - X\hat{\beta}) = \underset{\sim}{0}$

$\Longleftrightarrow X^T\underset{\sim}{Y} = X^TX\underset{\sim}{\hat{\beta}}$

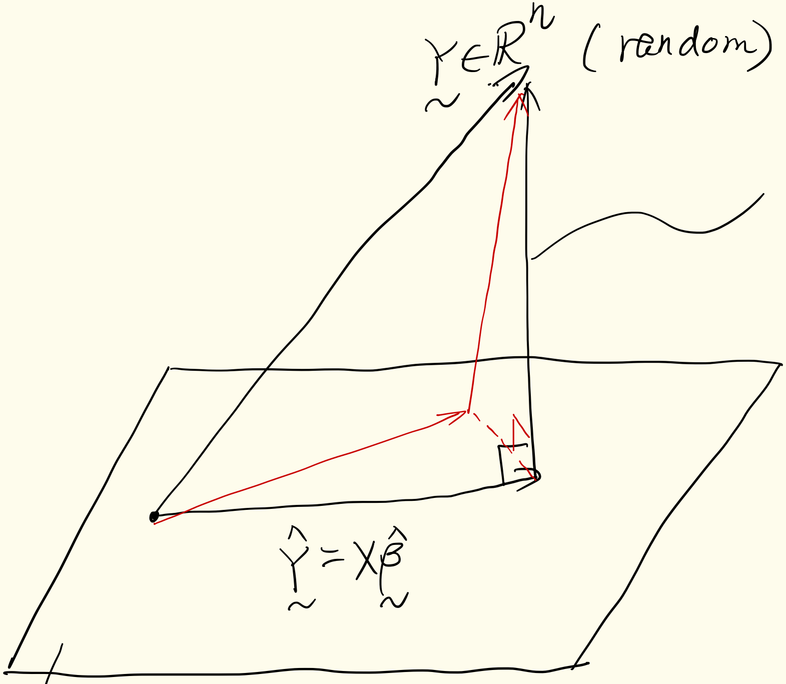
Regardless the rank of X^TX, there's a soln to the problem.



$\mathcal{C}(X) = \{X\underset{\sim}{b}, \ \underset{\sim}{b} \in \mathbb{R}^p\}$

$X \overset{\text{def}}{=} [\underset{\sim}{\xi}_1, \underset{\sim}{\xi}_2, \dots, \underset{\sim}{\xi}_p]$

$\dim(\mathcal{C}(X)) = r \le p$



$$\text{length}^2 = \|Y - \hat{Y}\|^2 = \|Y - X\hat{\beta}\|^2 = \sum \Big(Y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j\Big)^2$$

(random)

Column space of X (fixed), namely,

$$\mathcal{C}(X) \subset \mathbb{R}^n$$

$$\overset{\text{def}}{=} \{X\underset{\sim}{b},\ \underset{\sim}{b} \in \mathbb{R}^p\}$$

$\underline{\text{If } \operatorname{rank}(X) \triangleq r = p}$ ① $\hat{\underset{\sim}{\beta}} = (X^T X)^{-1} X^T \underset{\sim}{Y}$ is unique soln.

$$\mathbb{E}[\hat{\underset{\sim}{\beta}}] = \mathbb{E}[(X^T X)^{-1} X^T \underset{\sim}{Y}] = (X^T X)^{-1} X^T X \underset{\sim}{\beta} = \underset{\sim}{\beta} \quad \text{(unbiased)}$$

$$\operatorname{VarCov}(\hat{\underset{\sim}{\beta}}) = (X^T X)^{-1} X^T \underbrace{\operatorname{VarCov}(Y)}_{\sigma^2 \mathbb{I}} X (X^T X)^{-1}$$

Note: $\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$

$\operatorname{VarCov}(\underset{\sim}{a}^T \underset{\sim}{w}) = \underset{\sim}{a}^T \operatorname{VarCov}(\underset{\sim}{w}) \underset{\sim}{a}$

$$= \sigma^2 (X^T X)^{-1}$$

② $\hat{\underset{\sim}{Y}} = X \hat{\underset{\sim}{\beta}} = \underbrace{X (X^T X)^{-1} X^T}_{} \underset{\sim}{Y} = H \underset{\sim}{Y}$

H : "hat" matrix, "orthogonal projector". $H^n = H$. Why?

$$\operatorname{VarCov}(\hat{\underset{\sim}{Y}}) = H \operatorname{Cov}(\underset{\sim}{Y}) H^T = \sigma^2 H H^T = \sigma^2 H^2 = \sigma^2 H = \sigma^2 \cdot X (X^T X)^{-1} X^T$$

Ex: $X = \begin{bmatrix} 2 & 1 \\ 1 & 2 \\ 1 & 1 \end{bmatrix}$ $\quad Y = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ $\quad \boxed{Y = X\beta + e}$

$$\hat{\beta} = (X^T X)^{-1} X^T Y = \left(\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \\ 1 & 1 \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 6 & 5 \\ 5 & 6 \end{bmatrix}^{-1} \begin{bmatrix} 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 6 & -5 \\ -5 & 6 \end{bmatrix} \cdot \frac{1}{11} \cdot \begin{bmatrix} 4 \\ 4 \end{bmatrix}$$

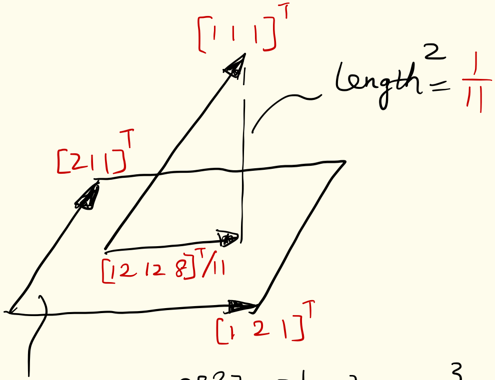
$$= \frac{4}{11} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\mathrm{VarCov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} = \frac{\sigma^2}{11} \begin{bmatrix} 6 & -5 \\ -5 & 6 \end{bmatrix}$$

$$H = X(X^T X)^{-1} X^T = \begin{bmatrix} 2 & 1 \\ 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 6 & -5 \\ -5 & 6 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \end{bmatrix} / 11 = \frac{1}{11} \begin{bmatrix} 10 & -1 & 3 \\ -1 & 10 & 3 \\ 3 & 3 & 2 \end{bmatrix}$$

$$\hat{Y} = HY = \frac{1}{11} \begin{bmatrix} 12 \\ 12 \\ 8 \end{bmatrix} \neq Y$$

$$\text{LS err: } \|Y - \hat{Y}\|^2 = \frac{1}{11^2} \left\| \begin{bmatrix} 12-11 \\ 12-11 \\ 8-11 \end{bmatrix} \right\|^2 = \frac{1}{11^2}(1+1+9) = \frac{1}{11}$$



$$\mathcal{C}(X) = \mathrm{span}\left\{ \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \right\} \subset \mathbb{R}^3$$

If $\text{rank}(X) \triangleq r < p$ (Not full column rank), what does $\hat{\beta}$ look like?

Ans: $r < p \iff X$ has a non-trivial null space.

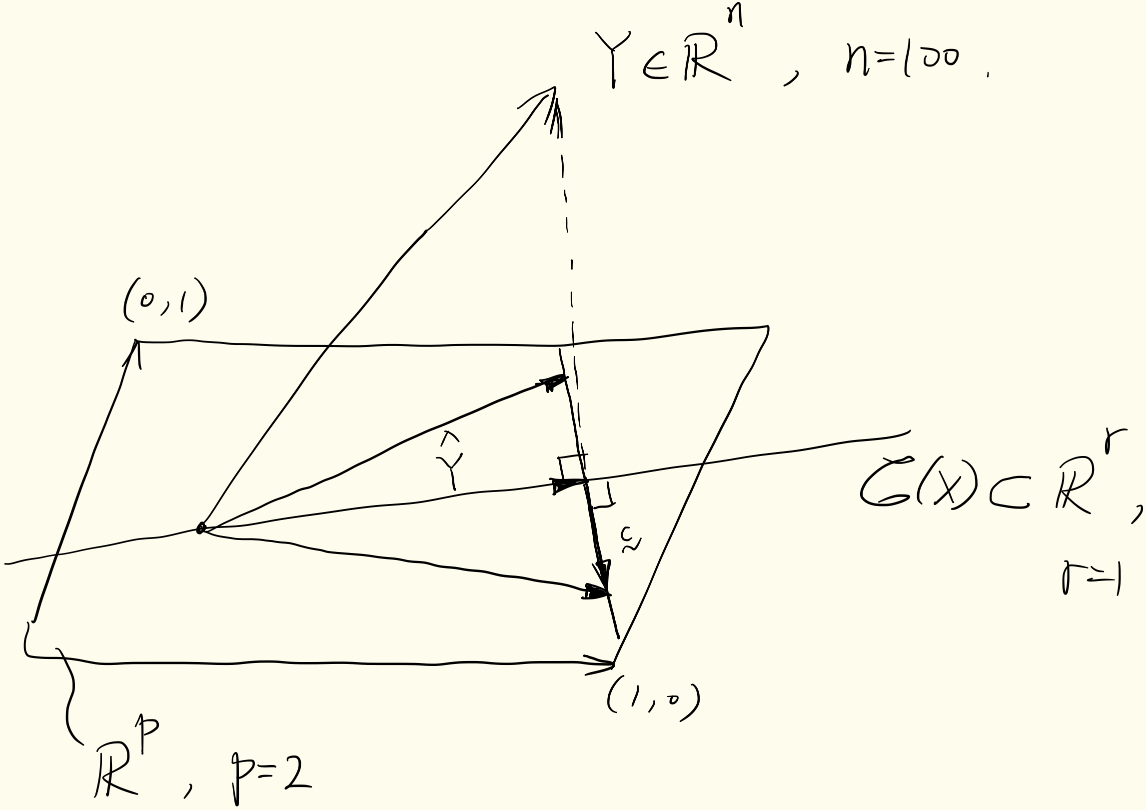
$\iff \underset{\sim}{c} \in \mathbb{R}^p$ s.t. $X\underset{\sim}{c} = \underset{\sim}{0} = \sum_{i=1}^{p} c_i \underset{\sim}{\xi}_i$. Then

$\exists\ \underset{\sim}{c} \neq \underset{\sim}{0}$

$$X\hat{\underset{\sim}{\beta}} = X\hat{\underset{\sim}{\beta}} + X\underset{\sim}{c} = X\underbrace{(\hat{\underset{\sim}{\beta}} + \underset{\sim}{c})}_{\tilde{\beta}}$$

$$\tilde{\beta} = \hat{\beta} + k\underset{\sim}{c}, \quad k \in \mathbb{R} \text{ are all LS estimates.}$$

However, $\hat{y}$ is unique: $\hat{y} = X(\hat{\beta} + k\underset{\sim}{c}) = X\hat{\beta} + k\underbrace{X\underset{\sim}{c}}_{\underset{\sim}{0}} = X\hat{\beta}$.



Y ∈ R^n , n=100 .
(0,1)
Ŷ
ε
G(X) ⊂ R^r ,
r=1
(1,0)
R^p , p=2