Supplemental Notes

EE503 Week 14 Dr. Franzke

HW #13

① Handout - Least squares

② Leon-Garcia 6.71 - 6.72

Topics: Markov Chains and Estimation.

① Mean-Square Estimation (MSE)

$$\hat{\theta}_{n\times 1}^{ms} = E[\theta_{n\times 1} | X_n]$$

Orthogonality Principle: $E_{\theta x}[(\theta - \hat{\theta}_n^{ms})\, g^T(X_n)] = 0$ for any $g(X_n)$

② Simple Least Squares: $\hat{y} = \hat{a}_n + \hat{b}_n \cdot x$

$$\therefore \hat{a}^{LS} = \bar{y}_n - \hat{b}^{LS}\bar{x}_n$$

$$\hat{b}^{LS} = s_{xy}/s_{xx}^2$$

③ Ordinary least squares (OLS): $Y = X\beta + \varepsilon$.

$$\therefore \hat{\beta}^{LS} = (X^TX)^{-1}X^TY \sim N(\beta, \sigma^2(X^TX)^{-1})$$

Optimal Estimators

Recall: ① Maximum Likelihood (ML)

$$\hat{\theta}_n^{ML} = \arg\max_\theta g(x|\theta)$$

- need closed form pdf g
- Consistency: $\hat{\theta}_n^{ML} \xrightarrow{P} \theta$ (smoothness condition)
- Asymptotic normality $\sqrt{n}(\hat{\theta}_n^{ML} - \theta) \xrightarrow{d} W \sim N(0, \frac{1}{I(\theta)})$ for Fisher Information I (Cramer-Rao)
- Invariance principle $\widehat{g(\theta)}^{ML} = g(\hat{\theta}^{ML})$

- Applications: - EM algorithm for missing data
  - Viterbi algorithm (dynamic programming on "hidden" Markov chain)

② Maximum A Posteriori (MAP)

$$\hat{\theta}_n^{MAP} = \arg\max_\theta f(\theta|x) \quad \text{if } \theta \sim h(\theta) \text{ and } f(\theta|x) \propto h(\theta) \cdot g(x|\theta)$$

($h(\theta)$: prior; $g(x|\theta)$: likelihood)

- MAP = MLE if $h(\theta) \sim$ uniform.
- Bayes minimization of squared error loss $\longrightarrow E[\theta|X]$ is optimal estimator (MSE result)
- Gibbs Sampler (MCMC)
- Subjectivity: who gives prior $h(\theta)$?

③ Mean square estimation (MSE)

$$\hat{\theta}_n^{ms} = E[\theta|X] \text{ for random } \theta \text{ (in general)}$$

focus of these notes.

Data set $\mathcal{X}_n = \{X_1, X_2, \ldots, X_n\}$ for random vectors X_k

Estimator $\underset{k\times 1}{\hat{\theta}_n} = g(\mathcal{X}_n)$

$= g(X_1, \ldots, X_n)$ $\therefore \hat{\theta}_n$ is a statistic.

Random Parameter $\underset{k\times 1}{\theta} \sim f(\theta_1, \ldots, \theta_k)$

generalizes deterministic case

Assume: Data $\mathcal{X}_n$ depends on θ

$\therefore$ there exists a joint pdf: $f(\theta, \mathcal{X}_n) = f(\theta, X_1, \ldots, X_n)$

★ **Thrm:** $\boxed{\underset{k\times 1}{\hat{\theta}_n^{ms}} = E[\underset{k\times 1}{\theta} | \mathcal{X}_n]}$ if $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ and θ obey joint pdf $f(\theta, \mathcal{X}_n)$

Prf: $0 \le \underset{1\times 1}{MSE}[\hat{\theta}_n^{ms}]$

$= E_{\theta\mathcal{X}}[\underset{1\times k}{(\theta - \hat{\theta}_n^{ms})^T} \underset{k\times 1}{(\theta - \hat{\theta}_n^{ms})}]$

where $\theta\mathcal{X} \longleftrightarrow f(\theta, \mathcal{X}_n) = f(\theta_1, \ldots, \theta_n, X_1, \ldots X_n)$

total exp. $= E_{\mathcal{X}}[E[(\theta - \hat{\theta}_n^{ms})^T(\theta - \hat{\theta}_n^{ms}) | \mathcal{X}_n]]$.

linear $= E_{\mathcal{X}}[E[\theta^T\theta | \mathcal{X}_n] + E[\hat{\theta}_n^{ms\,T}\hat{\theta}_n^{ms} | \mathcal{X}_n]$

$- E[\theta^T\hat{\theta}_n^{ms} | \mathcal{X}_n] - E[\hat{\theta}_n^{ms}\theta | \mathcal{X}_n]]$.

$$= E_x\Big[E[\theta^T\theta \mid \chi_n] + \hat{\theta}_n^{ms\,T}\hat{\theta}_n^{ms} - E[\theta^T \mid \chi_n]\cdot\hat{\theta}_n^{ms} - \hat{\theta}_n^{ms\,T} E[\theta \mid \chi_n]\Big].$$

since $\hat{\theta}_n^{ms} = g(\chi_n)$ and $E[g(x)z \mid x] = g(x)\cdot E[z \mid x]$

comp. the square $= E_x\Big[E[\theta^T\theta \mid \chi_n] - E[\theta^T \mid \chi_n]\cdot E[\theta \mid \chi_n] + \big(\hat{\theta}_n^{ms} - E[\theta \mid \chi_n]\big)^T\big(\hat{\theta}_n^{ms} - E[\theta \mid \chi_n]\big)\Big]$

since $\big(\hat{\theta}_n^{ms} - E[\theta \mid \chi_n]\big)^T\big(\hat{\theta}_n^{ms} - E[\theta \mid \chi_n]\big)$

$= \hat{\theta}_n^{ms\,T}\hat{\theta}_n^{ms} - \hat{\theta}_n^{ms\,T}E[\theta \mid \chi_n] - E[\theta^T \mid \chi_n]\,\hat{\theta}_n^{ms} + E[\theta^T \mid \chi_n]\cdot E[\theta \mid \chi_n]$

linear $= E_x\Big[E[\theta^T\theta \mid \chi_n] - E[\theta^T \mid \chi_n] E[\theta \mid \chi_n]\Big] + E_x\Big[\big(\hat{\theta}_n^{ms} - E[\theta \mid \chi_n]\big)^T\big(\hat{\theta}_n^{ms} - E[\theta \mid \chi_n]\big)\Big]$

↑ only this term depends on $\hat{\theta}_n^{ms}$

$\therefore\ \hat{\theta}_n^{ms}$ minimizes $MSE[\hat{\theta}_n^{ms}] \geq 0$ if it minimizes $E_x\Big[\big(\hat{\theta}_n^{ms} - E[\theta \mid \chi_n]\big)^T\big(\hat{\theta}_n^{ms} - E[\theta \mid \chi_n]\big)\Big]$

≥ 0.

Latter occurs iff $\hat{\theta}_n^{ms} = E[\theta \mid \chi_n]$.

QED

Same result if minimize $E_{\theta x}\big[(\theta - \hat{\theta}_n^{ms})^T W (\theta - \hat{\theta}_n^{ms})\big]$.

weight matrix $W = W^T$ is positive semi-definite

Unbiased: $E_x[\hat{\theta}_n^{ms}] \overset{\text{thm}}{=} E_x[E[\theta|X_n]] \overset{\text{total exp.}}{=} E_\theta[\theta] \quad \forall n.$

★ **Thrm:** (Orthogonality Principle) MSE error is orthogonal to data (functions of data): $\mathcal{E}^{ms} \perp f(X)$.

$$E_{\theta x}[(\theta - \hat{\theta}_n^{ms}) g^T(X_n)] = 0$$

for ANY function $g(X_n) = g(X_1, \ldots, X_n)$

- applications of orthogonal projections.
  - Wiener-filter (Wiener-Hopf equation) EE562
  - Yule-Walker equations EE583
  - Kalman-filter (linear Gauss-Markov model)

Prf:

$$E_{\theta x}[(\theta - \hat{\theta}_n^{ms}) g^T(X_n)]$$

$$\overset{\text{total exp.}}{=} E_x[E[(\theta - \hat{\theta}_n^{ms}) g^T(X_n) | X_n]]$$

$$= E_x[E[\theta - \hat{\theta}_n^{ms} | X_n] \cdot g(X_n)].$$

since $E[g(X)Y|X] = g(X) \cdot E[Y|X]$

$$\overset{\text{MSE thm.}}{=} E_x[E[\theta - E[\theta|X_n] | X_n] g^T(X_n)]$$

$$\overset{\text{linear}}{=} E_x[(E[\theta|X_n] - E[\theta|X_n] \cdot E[1|X_n]) g^T(X_n)].$$

since $E[\theta|X_n] = \phi(X_n)$ and $E[\phi(X_n)|X_n] = \phi(X_n) \cdot E[1|X_n]$

$$= E_x\left[\left(\underbrace{E[\theta|X_n] - E[\theta|X_n]}_{=0} \right) g^T(X_n) \right].$$

$$= E_x[\, 0 \; g^T(X_n)]$$

$$= 0$$

QED

- Hard to find closed form $E[\theta|X_n]$ in general
- $\hat{\theta}_n^{MS}$ is <u>not robust</u> to outliers or impulsive data (unlike conditional median)

Special (but common) case: Jointly-Gaussian $\underset{n\times 1}{X}$ and $\underset{p\times 1}{Y}$

$$E[\underset{n\times 1}{X}|Y] = \underset{n\times 1}{\mu_X} + \underset{n\times p}{K_{XY}} \underset{p\times p}{K_{YY}^{-1}} (\underset{p\times 1}{Y} - \mu_Y)$$

$$\underset{n\times n}{K_{X|Y}} = \underset{n\times n}{K_{XX}} - \underset{n\times p}{K_{XY}} \underset{p\times p}{K_{YY}^{-1}} \underset{p\times n}{K_{YX}}$$

<u>Warning</u>: Must <u>TEST</u> for normality (EE517)

<u>Ex</u>: Bayesian conjugacy for finding $E[\theta|X]$.
(proved in Week 5: "beta is conjugate to binomial")

<u>Thm</u>: If $\theta \sim h(\theta) = \text{Beta}(\alpha, \beta)$
$g(x|\theta) = \text{binomial}(n, x, \theta)$ ← Bernoulli trials

then $f(\theta|x) = \text{Beta}(\alpha + x, \beta + n - x)$

where θ estimates "success probability" p after n-trials and n successes (∴ $n-x$ failures)

Recall: Also proved $E[\theta] = \frac{\alpha}{\alpha+\beta}$ if $\theta \sim \text{Beta}(\alpha, \beta)$

$\therefore$ For squared-error loss and $f(\theta|x)$

$$\hat{\theta}_n^{ms} = E[\theta \mid X=x]$$

$$\overset{\text{conjugacy}}{=} \frac{\alpha'}{\alpha'+\beta'} \quad \text{since } f(\theta|x) = \text{Beta}(\alpha', \beta')$$

$$= \frac{\alpha+x}{\alpha+x+\beta+n-x} \quad \text{since } f(\theta|x) = \text{Beta}(\alpha+x, \beta+n-x)$$

$$= \boxed{\frac{\alpha+x}{\alpha+\beta+n}} \leftarrow \text{closed form answer.}$$

$$= \frac{\alpha}{\alpha+\beta+n} + \frac{x}{\alpha+\beta+n}$$

$$= \frac{\alpha}{\alpha+\beta+n} \cdot \frac{\alpha+\beta}{\alpha+\beta} + \frac{x}{\alpha+\beta+n} \cdot \frac{n}{n}$$

$$= \underbrace{\frac{\alpha+\beta}{\alpha+\beta+n}}_{\lambda_n} \cdot \frac{\alpha}{\alpha+\beta} + \underbrace{\frac{n}{\alpha+\beta+n}}_{1-\lambda_n} \cdot \frac{x}{n}$$

$$= \lambda_n \cdot \underbrace{\frac{\alpha}{\alpha+\beta}}_{E[\theta]} + (1-\lambda_n) \underbrace{\frac{x}{n}}_{\bar{X}_n = \hat{\theta}_n^{ml}} \quad \text{for convex coefficients } 0 \le \lambda_n \le 1$$

$$= \lambda_n \cdot E[\theta] + (1-\lambda_n)\bar{X}_n$$

$\longrightarrow \bar{X}_n$ as $n \longrightarrow \infty$ because $\lambda_n = \frac{\alpha+\beta}{\alpha+\beta+n} \downarrow 0$

$\therefore 1-\lambda_n \longrightarrow 1.$

$\therefore$ Effect $E[\theta]$ of subjective prior "washes out" for large n.

$\therefore \hat{\theta}_n^{ms}$ is asymptotically unbiased for θ. since

$$E[\bar{X}_n] = \frac{E[X]}{n} = \frac{np}{n} = p \quad \text{for binomial } n$$

$$- V[E[\theta|X]] = V\left[\frac{\alpha+X}{\alpha+\beta+n}\right]$$

$$= \frac{1}{(\alpha+\beta+n)^2} V[X]$$

$$= \frac{npq}{(\alpha+\beta+n)^2} \quad \text{since } X \sim \text{binomial}$$

$$\downarrow 0 \quad \text{as} \quad n \uparrow \infty$$

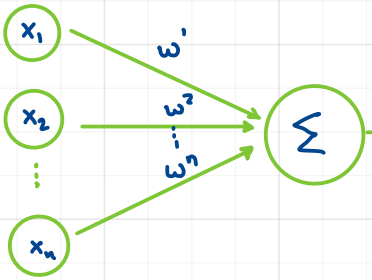
$$\therefore E[\theta|X] \xrightarrow{m} p$$

$\therefore$ Consistent: $E[\theta|X] \xrightarrow{\text{plim}} p$ by $m \longrightarrow p$

Adaptive MS Estimation (LMS) EE583

"Linear combiner" → adaptive noise suppression
→ adaptive equalizer and modems

Weights: $w \in \mathbb{R}^n$



desired output $d_j \in \mathbb{R}$

Learn MS optimal w^* from $(x_1, d_1), \ldots, (x_m, d_m)$

r.v. $\varepsilon_k = d_k - y_k = d_k - x_k^T w$

$\therefore$ MSE $= J_k(w) = E[\varepsilon_k^2]$

← unknown pdf $p(x)$

$\therefore \nabla_w J = 0 : \; w^* = R^{-1} P$

$= (E[X_k X_k^T])^{-1} E[d_k x_k^T]$

unknown pdfs

★

LMS idea: $E[\varepsilon_k^2] \approx \varepsilon_k^2$ (Widrow-Hopf, late 1950's)

single realization of r.v.

∴ **LMS algorithm**

LMS estimate

$$w_{k+1} = w_k - \mu \hat{\nabla}_w J \leftarrow -2\varepsilon_k x_k \text{ (derivative below)}$$

constant learning rate

general GRADIENT DESCENT algorithm

$$= \boxed{w_k + 2\mu \varepsilon_k x_k} \quad \star$$

↑ arguably most widely used EE algorithm.

if WSS process (wide-sense stationary) EE562

Not robust to outliers or impulsive ($S\alpha S$-type) noise

Robust LMS: $\varepsilon_k \longmapsto \text{sign}(\varepsilon_k)$ "signed LMS"

$S\alpha S$ noise: $\alpha = 2$ Gaussian; $\alpha = 1$ Cauchy

$\alpha \downarrow \Longrightarrow$ tail-thickness ↑ ($\alpha < 2 \longrightarrow$ infinite variance)

∴ LMAD: "least mean absolute deviation"

- LMS similar to CENTROID learning in clustering models (pattern recognition)
- LMS: constant learning rate μ (unlike most neural network algorithms) ↔ still converges if WSS signal

Fact: $w_k \xrightarrow{\text{LMS}} w^*$ if $\boxed{0 < \mu < \frac{1}{\lambda_{max}}}$

eigenvalues of R: $0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_{max}$

$$\hat{\nabla}_w J = \nabla_w \varepsilon_k^2 \qquad \text{since LMS idea } E[\varepsilon_k^2] = \varepsilon_k^2$$

$$= \left(\frac{\partial \varepsilon_k^2}{\partial w^1}, \cdots, \frac{\partial \varepsilon_k^2}{\partial w^n} \right)$$

$$= \left(\frac{\partial}{\partial w^1}(d_k - y_k)^2, \cdots, \frac{\partial}{\partial w^n}(d_k - y_k)^2 \right)$$

$$= \left(-2\underbrace{(d_k - y_k)}_{\varepsilon_k} \frac{\partial y_k}{\partial w^1}, \cdots, -2\underbrace{(d_k - y_k)}_{\varepsilon_k} \frac{\partial y_k}{\partial w^n} \right)$$

$$= -2\varepsilon_k \left(\frac{\partial y_k}{\partial w^1}, \cdots, \frac{\partial y_k}{\partial w^n} \right)$$

$$= -2\varepsilon_k (x_k^1, \cdots, x_k^n) \qquad \text{since } y_k = x_k^T w = \sum_{i=1}^{n} x_k^i w^i$$

$$= -2\varepsilon_k x_k$$

QED.

$$\therefore \; w_{k+1} = w_k + 2\mu\varepsilon_k x_k \qquad \text{(LMS)}$$

Simple Linear Regression (OLS)

$$y_k = f(x_k)$$

$$= \underbrace{a}_{\text{constant}} + b x_k + \underbrace{\varepsilon_k}_{\text{noise}}$$

ε_k: modelling error

$$\triangleq \beta_0 + \beta_1 x_k + \varepsilon_k .$$

y_k is a (normal) r.v. because iid $\varepsilon_k \sim N(0, \sigma^2)$

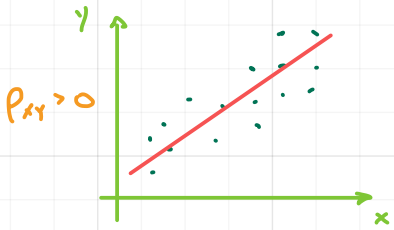
"TRUE" model: $y = a + bx$

Result: ★ $a_n^{LS} = \bar{y}_n - \hat{b}^{LS} \bar{x}_n$

$= \hat{\beta}_0$ unbiased and consistent.

★ $b_n^{LS} = \frac{S_{xy}}{S_x^2} = R_{xy} \cdot \frac{S_y}{S_x}$

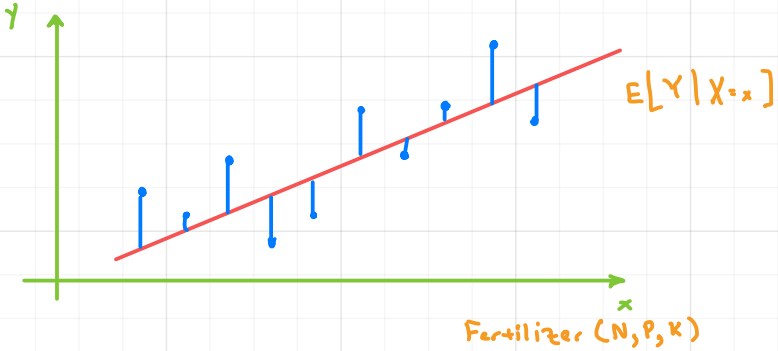
$= \hat{\beta}_1$ unbiased and consistent.



Least Squares

Model: $y_k = \beta_0 + \beta_1 x_k + \varepsilon_k$

Data: $(x_1, y_1), \ldots, (x_n, y_n)$; iid $\varepsilon_k \sim N(0, \sigma^2)$



Check: iid $\varepsilon_k \sim N(0, \sigma^2)$

Independence All ε_k independent

- "serial correlation" → time series (AR(1))
  Durbin-Watson test.

Homoscedasticity Constant σ_k^2

- pattern in residuals?

Normality $y_k - \hat{y}_k \sim N$

$K_{\varepsilon\varepsilon} = \sigma^2 I = \begin{pmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{pmatrix}$

else look at WLS.

Find: $\hat{\theta}_n^{ms} = \underset{2\times 1}{\hat{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$

Minimize J:

$$J = SSE = \sum_{k=1}^{n} \varepsilon_k^2 = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2$$

$$\overset{LS}{=} \sum_{k=1}^{n} (y_k - (\hat{\beta}_0 + \hat{\beta}_1 x_k))^2$$

$$\therefore \quad \odot = \nabla_\beta J = \begin{pmatrix} \partial J/\partial \beta_0 \\ \partial J/\partial \beta_1 \end{pmatrix}$$

$$= \begin{cases} \sum_{k=1}^{n} 2(y_k - (\hat{\beta}_0 + \hat{\beta}_1 x_k))(-1) \\ \sum_{k=1}^{n} 2(y_k - (\hat{\beta}_0 + \hat{\beta}_1 x_k))(-x_k) \end{cases}$$

Solve for $\hat{\beta}_0$ and $\hat{\beta}_1$ (positive definite Hessian matrix)

OLS Normal Equations (Ordinary Least Squares)

$$\underset{(k+1)\times 1}{\hat{\beta}^{LS}} = (X^T X)^{-1} X^T Y \quad \bigstar$$

Observation matrix

$$\underset{n\times 1}{Y} = \underbrace{\underset{n\times(k+1)}{X}\ \underset{(k+1)\times 1}{\beta}}_{n\times 1} + \underset{n\times 1}{\varepsilon}$$

$$y_k = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon_k$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}}_{X} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

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

- $\varepsilon \sim N(0, \sigma^2 I)$

∴ White and normal and homoscedastic (additive noise)

Recall:

(1) $\frac{d}{dm}\underbrace{(\underset{1\times n}{b^T}\ \underset{n\times 1}{m})}_{1\times 1} = \underset{n\times 1}{b}$ (linear form) $\quad \left(\frac{d}{dx} cx = c\right)$

(2) If $A = A^T$: $\frac{d}{dm}\underbrace{(\underset{1\times n}{m^T}\ \underset{n\times n}{A}\ \underset{n\times 1}{m})}_{1\times 1} = \underbrace{2\ \underset{n\times n}{A}\ \underset{n\times 1}{m}}_{n\times 1}$ (quadratic form) $\quad \left(\frac{d}{dx} x^2 = 2x\right)$

Minimize $(Y-\hat{Y})^T(Y-\hat{Y})$

$$J \triangleq SSE \quad \hat{Y} = \underset{1\times n}{(Y - X\beta)^T} \underset{n\times 1}{(Y - X\beta)} \quad \text{scalar.}$$

$$= Y^TY - 2Y^TX\beta + \beta^TX^TX\beta.$$

$$\therefore \nabla_\beta J = 0 = -2X^TY + 2X^TX\hat{\beta}.$$

$\longleftrightarrow$ "Normal Equations"

$$\boxed{(X^TX)\,\hat{\beta} = X^TY} \quad \star$$

$$\therefore \hat{\beta}^{LS} = (X^TX)^{-1}X^TY = C \cdot Y \quad \leftarrow \text{linear in } Y.$$

if $(X^TX)^{-1}$ exists, iff $\det(X^TX) \neq 0$

iff $\forall$ eigenvalues $\lambda > 0$

$$X^TX = \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum x_{i2}x_{i1} & & \vdots \\ \vdots & \vdots & \ddots & \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik}x_{i1} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}$$

$\therefore X^TX$ is real and symmetric. $(X^TX)^T = X^T(X^T)^T = X^TX.$

$\therefore$ for $\forall w \neq 0$

$$\underbrace{w^T (X^T X) w}_{1\times 1} \overset{\text{assoc.}}{=} (w^T X^T)(Xw)$$

$$= (Xw)^T (Xw)$$

$$= z^T \cdot z \qquad \text{if } z = Xw$$

$$\geq 0 \qquad \text{since } z \text{ real vector}$$

$\therefore X^T X \geq 0$ Positive Semi-definite

$\Longleftrightarrow$ all eigenvalues $\lambda \in \mathbb{R}^+ : \lambda_i \geq 0$

$\therefore \det(X^T X) = \prod_{i=1}^{k+1} \lambda_i \geq 0.$

- Problems with numerical stability if λ_i "small" (multicolinearity)
- Pseudo-inverse always exists and is unique

Thm: $Y = X\beta + \varepsilon$ and $\varepsilon \sim N(0, \sigma^2 I)$ and

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

$$\boxed{\hat{\beta}^{LS} \sim N(\beta, \sigma^2 (X^T X)^{-1})} \quad \star$$

Prf:

(1) **Normality**

$$\hat{\beta} = (X^TX)^{-1}X^TY = C \cdot Y$$ - linear function of Y

But $Y = X\beta + \varepsilon \sim N$ because $\varepsilon \sim N(0, \sigma^2 I)$

$\therefore \hat{\beta} \sim N$ QED (1)

(2) **Unbiasedness**

$$E[\hat{\beta}] \overset{LS}{=} E[(X^TX)^{-1}X^TY]$$

$$\overset{\text{def}Y}{=} E[(X^TX)^{-1}X^T(X\beta+\varepsilon)].$$

$$= E[(X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\varepsilon]$$

$$\overset{\text{linear}}{=} E[\underbrace{(X^TX)^{-1}(X^TX)}_{=I}\beta] + E[(X^TX)^{-1}X^T\varepsilon].$$

$$= E[\beta] + E[(X^TX)^{-1}X^T\varepsilon].$$

$$= \beta + E[(X^TX)^{-1}X^T\varepsilon]$$ since β not random

$$= \beta + (X^TX)^{-1}X^T E[\varepsilon]$$ since X not random

$$= \beta + \underline{0}$$ since $\varepsilon \sim N(\underline{0}, \sigma^2 I)$

$$= \beta$$

$\therefore \hat{\beta}$ is an unbiased estimator for β.

QED (2)

(3) **Covariance** (Consistency w/ Slutsky's Theorem)

$$K_{\hat\beta\hat\beta} = E\left[(\hat\beta - E[\hat\beta])(\hat\beta - E[\hat\beta])^T\right].$$

First, find K_{YY}:

$$K_{YY} = E\left[(Y - E[Y])(Y - E[Y])^T\right].$$

$$= E\left[(Y - X\beta)(Y - X\beta)^T\right] \quad \text{since } E[Y] = E[X\beta + \varepsilon] = X\beta + E[\varepsilon] = X\beta.$$

$$= E\left[\varepsilon\varepsilon^T\right] \quad \text{since } \varepsilon = Y - X\beta.$$

$$= E\left[(\varepsilon - \Theta)(\varepsilon - \Theta)^T\right]. \quad \text{by Hypo.}$$

$$= E\left[(\varepsilon - E[\varepsilon])(\varepsilon - E[\varepsilon])^T\right] \quad \text{since } \varepsilon \sim N(\Theta, \sigma^2 I)$$

$$\overset{\text{def}}{=} K_{\varepsilon\varepsilon}$$

$$\overset{\text{hypo.}}{=} \sigma^2 I \quad \text{since } \varepsilon \sim N(\Theta, \sigma^2 I)$$

$$\therefore K_{YY} = \sigma^2 I$$

$$K_{\hat\beta\hat\beta} = E\left[(\hat\beta - E[\hat\beta])(\hat\beta - E[\hat\beta])^T\right]$$

$$\overset{(3)}{=} E\left[(\hat\beta - \beta)(\hat\beta - \beta)^T\right] \quad \text{since } E[\hat\beta] = \beta. \text{ unbiased}$$

$$\overset{LS}{=} E\left[\left((X^TX)^{-1}X^TY - \beta\right)\left((X^TX)^{-1}X^TY - \beta\right)^T\right].$$

$$\overset{\text{"trick"}}{=} E\left[\left((X^TX)^{-1}X^TY - \underbrace{(X^TX)^{-1}X^TX}_{=I}\beta\right)\left((X^TX)^{-1}X^TY - \underbrace{(X^TX)^{-1}X^TX}_{=I}\beta\right)^T\right] \quad \text{since } \hat\beta = (X^TX)^{-1}X^TY.$$

dist

$$= E\left[\left((X^TX)^{-1}X^T(Y-\beta)\right)\left((X^TX)^{-1}X^T(Y-\beta)\right)^T\right]$$

$$= (X^TX)^{-1}X^T \cdot E\left[(Y-X\beta)(Y-X\beta)^T\right](X^TX)^{-1}X^T)^T$$

T

$$= (X^TX)X^T E\left[(Y-X\beta)(Y-X\beta)^T\right] \times (X^TX)^{-1}$$

since X is not random

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

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

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

$= X \cdot (X^TX)^{-1}$

$$= (X^TX)^{-1}X^T E\left[(Y-E[Y])(Y-E[Y])^T\right] \cdot X(X^TX)^{-1}$$

since $E[Y] = E[X\beta+\varepsilon] = X\beta$.

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

lemma

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

$$= \sigma^2 \cdot (X^TX)^{-1} \underbrace{X^TX}_{=I} (X^TX)^{-1}$$

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

QED

$\therefore \hat{\beta} \xrightarrow{P} \beta$ consistent. by "m-p-d" (Slutsky)

Note: asymptotic normality (CLT) if only iid ε_k and $E[\varepsilon_k] = 0$

Thm: $\hat{\beta}^{LS} = \hat{\beta}^{ML}$ $\therefore$ invarance principle holds.

$$Y \sim N(X\beta, \sigma^2 I)$$

Prf:

$$L(\beta, \sigma^2) \overset{N}{=} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\tfrac{1}{2}(Y-X\beta)^T K_{YY}^{-1} (Y-X\beta)\right]. \quad K_{YY} = \sigma^2 I.$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}(Y-X\beta)^T I (Y-X\beta)\right]$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}(Y-X\beta)^T (Y-X\beta)\right].$$

$\therefore$ Maximize L $\longleftrightarrow$ minimize exponent.

$\longleftrightarrow$ minimize J

where $J = \varepsilon^T\varepsilon = (Y-X\beta)^T(Y-X\beta)$

$\therefore$ gives $\hat{\beta}^{LS}$

QED.

Similarly,

Thm: $\hat{\sigma}^{2\,ML} = \dfrac{(Y-X\hat{\beta})^T(Y-X\hat{\beta})}{n}$

Weighted Least Squares (WLS)

Positive definite $W \gg 0$ often diagonal $W = \begin{pmatrix} w_{11} & & 0 \\ & w_{22} & \\ 0 & & \ddots \\ & & & w_{nn} \end{pmatrix}$

$\therefore$ W has a "square root" Q: $W = Q \cdot Q^T$
- find with Cholesky Decomposition, $O(n^3)$
- Q lower triangular

W can correct over-weighted errors if errors are uncorrelated (white) but Heteroscedastic

$$\hat{\beta}^{WLS} = (X^T W X)^{-1} X^T W Y \qquad (W = I)$$

$$= (X^T K_{\varepsilon\varepsilon}^{-1} X)^{-1} X^T K_{\varepsilon\varepsilon}^{-1} Y \qquad \text{if } W = K_{\varepsilon\varepsilon}^{-1}$$

where $K_{\varepsilon\varepsilon}^{-1} = \begin{pmatrix} 1/\sigma_1^2 & & 0 \\ & 1/\sigma_2^2 & \\ 0 & & \ddots \\ & & & 1/\sigma_n^2 \end{pmatrix}$ since $K_{\varepsilon\varepsilon}$ diagonal (uncorrelated errors)

Generalized Least Squares (GLS)

- Apply if errors heteroscedastic (non-constant variance)

  or correlated ($K_{\varepsilon\varepsilon} \neq$ diagonal)

Use WLS and positive-definite $K_{\varepsilon\varepsilon}^{-1}$ or estimate

$$\boxed{W = K_{\varepsilon\varepsilon}^{-1}} = (QQ^T)^{-1} \quad \text{since } K_{\varepsilon\varepsilon} = QQ^T$$

$\therefore$ Scale with Q^{-1} to <u>decouple</u>

Set $\varepsilon^* = Q^{-1}\varepsilon$

$\therefore K_{\varepsilon^*\varepsilon^*} = Q^{-1} K_{\varepsilon\varepsilon} Q = Q^{-1}(QQ^T)(Q^{-1})^T = I$

$\therefore \varepsilon^*$ white ($\therefore$ Gauss-Markov holds.) BLUE.

$\therefore J^{GLS} = (Y - X\beta)^T K_{\varepsilon\varepsilon}^{-1} (Y - X\beta)$ ← Malhalanobis distance

$= (\underbrace{Q^{-1}Y}_{Y^*} - (\underbrace{Q^{-1}X}_{X^*})\beta)^T (Q^{-1}Y - (Q^{-1}X)\beta)$

$= (Y^* - X^*\beta)^T (Y^* - X^*\beta)$

$\therefore$ OLS again with scaling $Y^* = Q^{-1}Y$ and $X^* = Q^{-1}X$.