References

Introduction to Linear Regression Analysis
- Douglas C. Montgomery
Applied Regression Analysis : a research tool
- John O. Rawlings
Brockwell, Peter J. - Intro. to TS and Forecasting

Midsem 30M Endsem 50M Project 20M

Regression Problem

Predicting dependent variable on the basis of independent variable(s).

- might be a categorical value / real value (of both the independent / dependent variable)

Polynomial regression also falls in the category of linear-regression.

Definition Conditional expectation of Y given X, is known as regression of Y on X. It can be denoted as $\hat{y} = g(x, \beta) = E[Y|X = x]$

→ $g(x, \beta)$ can be a line, curve, plane, surface ...
→ x can be stochastic / non-stochastic
→ Y is always stochastic or a random variable.

Associated statistical problems with regression

- Interval Estimation
- Testing of Hypothesis
- Estimability of certain combination of regressors
  Admissibility

[All combination of regressors won't be estimated. given data, some combination might]

- Contrast
- Necessity of building regression : ANOVA
- Classification (if categorical prediction)

Associated mathematical problems with regression

- Notion of error
- How error space would interact with prediction space
- optimal subspace of regressors
- Reducing bias and variance of regression coeff.
- Rate of reduction of error if we incorporate more regressors

Vector Representation

1. Simple Linear Regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

y_i: dependent; x_i: independent; ϵ_i: random error

$$\text{Data} = \{(y_i, x_i) \quad i = 1, 2, \ldots, n\}$$

$$\underset{\sim}{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = [\underset{\sim}{1} \; \underset{\sim}{x}] = \left[\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \right] \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad \underset{\sim}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$\underset{(n\times 1)}{\underset{\sim}{Y}} = \underset{(n\times 2)(2\times 1)}{X\underset{\sim}{\beta}} + \underset{(n\times 1)}{\underset{\sim}{\epsilon}}$$

Problem in $\mathbb{R}^n$.

Mathematics on $\mathbb{R}^n$

Vector Space $(V, +, \cdot)$ where $+ : V \times V \to V$; $\cdot : F \times V \to V$

$V = \mathbb{R}^n$; $F = \mathbb{R}$

$\mathbb{P}^n$: $\{ \sum_{i=0}^{n} a_i x^i \mid a_i \in \mathbb{R}, x \in \mathbb{R} \}$

$\underset{\sim}{a} = \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix}$ $\underset{\sim}{x} = \begin{pmatrix} 1 \\ x \\ \vdots \\ x^n \end{pmatrix}$ $\mathbb{P}_n = \{ \underset{\sim}{a}^T \underset{\sim}{x} \mid \underset{\sim}{a} \in \mathbb{R}^n ; x \in \mathbb{R} \}$

Since a polynomial can be identified by one variable irrespective of the domain, it still is a linear regression problem.
All the coefficient are identified uniquely by an $n \times 1$ vector ($\underset{\sim}{a}$ in the above case)

Subspace $S \subseteq V$ and it satisfies all the properties of V , then S is called subspace of V.

- check if $0 \in S$
- check if $x + a.y \in S$?

Example $V = \mathbb{R}^2$

- $S_1 = \{ (x, 0), x \in \mathbb{R} \}$
- $S_2 = \{ (0, y), y \in \mathbb{R} \}$
- $S_3 = \{ (x, mx), x \in \mathbb{R} \}$

generalizing, $S = \{ \underset{\sim}{a}^T \underset{\sim}{x} = a_1 x_1 + a_2 x_2 = 0 \mid a_i \in \mathbb{R}, x_i \in \mathbb{R}, i = 1, 2 \}$

$S \subseteq V$ and S is subspace.

- For $V = \mathbb{R}^3$, $S = \{ \underset{\sim}{a}^T \underset{\sim}{x} = a_1 x_1 + a_2 x_2 + a_3 x_3 = 0 \mid a_i \in \mathbb{R}, x_i \in \mathbb{R}, i = 1, 2, 3 \}$
- $V = \mathbb{P}_n$ $S = \mathbb{P}_k$, $k \leq n$ S is a subspace of $\mathbb{P}_n$.

0 element is always a subspace (trivial subspace)

- For $V = \mathbb{P}_n$, would $S = \{k\text{-degree polynomial}\}$ be a subspace?

Not necessarily. (only 0 elt. would be a subspace)

ex. $v_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$ and $v_2 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$, then $v_1 + v_2 = \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}$ is not a 2-degree polynomial. (Not closed under addition)

given $\underset{\sim}{x}$, we predict $\hat{y}$ if we know $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$$

$$= \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \underset{\sim}{\epsilon}$$

$$\hat{y} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \hat{\beta}_0 + \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \hat{\beta}_1 = \underset{\sim}{1}\hat{\beta}_0 + \underset{\sim}{x}\hat{\beta}_1$$

$\hat{y}$: prediction

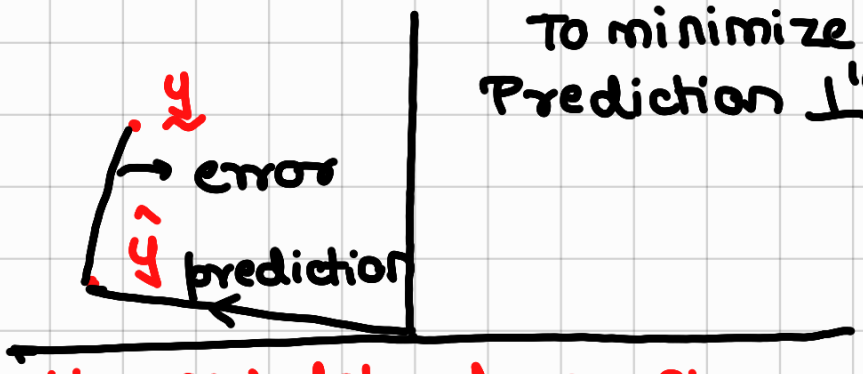
$\underset{\sim}{1}$, $\underset{\sim}{x}$: from $\mathbb{R}^n$

$\underset{\sim}{x}$: given $\underset{\sim}{x}$

We have 2 vectors $\underset{\sim}{1}$ and $\underset{\sim}{x}$. But their linear combination will not ever span full $\mathbb{R}^n$ space.

($n \leq 1$??)

Entire regression problem is that you will try to predict $\hat{y}$ but would not be able to span whole $\mathbb{R}^n$ space.



To minimize error, Prediction $\perp^{ar}$ error.

Recap of linear independence, Span and Basis & dimensions.

Orthogonal Vectors

vectors $\underset{\sim}{u}, \underset{\sim}{v} \in \mathbb{R}^n$ are said to be orthogonal if
$\langle \underset{\sim}{v}, \underset{\sim}{u} \rangle = \underset{\sim}{u}^T \underset{\sim}{v} = \sum_{i=1}^{n} u_i v_i = 0$

functions $u(t) : [a,b] \longrightarrow \mathbb{R}$ and $v(t) : [a,b] \longrightarrow \mathbb{R}$ are said to be orthogonal functions if
$\int_a^b u(t)v(t)\,dt = 0$

Orthogonal Complement

Projection Matrix

Idempotent Matrix

A projection matrix is an idempotent matrix (Prove)
An idempotent matrix has eigenvalues 0 or 1 (Prove)
If $\{v_1, v_2, \ldots, v_k\}$ is an orthonormal basis the subspace $S \subseteq V$ then the orthogonal projection matrix of S is $P_S = \sum_{i=1}^{k} v_i v_i^T$

Column Space

Properties: 1> $C(A:B) = C(A) + C(B)$
2> $C(AB) \subseteq C(A)$
3> $\dim(C(A)) = \text{Rank}(A)$
4> $C(AA^T) = C(A) \Longrightarrow \text{Rank}(AA^T) = \text{Rank}(A)$ [Prove it]

Quadratic Form

$\underset{\sim}{v}^T A \underset{\sim}{v} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} v_i^T v_j$

Definition of Metric

(Σ should be symm. and psd)

Statistics on $\mathbb{R}^n$

$X = (x_1, x_2, \ldots, x_n)^T$ a random vector with finite expectation for each component

Existence of Expectation

$E(X)$ exists $\Rightarrow \int |x| f(x)\,dx < \infty$, or,
$\sum_x |x| f(x) < \infty$

Dispersion matrix

$$D(X) = ((\mathrm{Cov}(X_i, X_j)))_{n\times n}$$
$$= E[(X - E(X))(X - E(X))^T] = \mathrm{Cov}(X, X)$$
$$= \Sigma, \text{ or,}$$
$$D(X) = \Sigma_X = ((\sigma_{ij}))_{n\times n} = ((\mathrm{Cov}(X_i, X_j)))$$

Note $\mathrm{Cov}(U_p, V_q) = ((\mathrm{Cov}(U_i, V_j)))_{p\times q}$

$E(X+b) = E(X) + b$

$D(X+b) = D(X)$

$\mathrm{Cov}(X+b, Y+c) = \mathrm{Cov}(X, Y)$

$$\mathrm{Cov}(X_i, X_j) = E(X_iX_j) - E(X_i)E(X_j)$$
$$= E(X_jX_i) - E(X_j)E(X_i) = \mathrm{Cov}(X_j, X_i)$$

→ commutative in nature

Hence, $\Sigma_X^T = \Sigma_X$

Important Results

$X \leftarrow$ random vector with n-components

$E(X) = \mu$; $D(X) = \Sigma$

1) $E(I^TX) = I^T\mu$, $I \in \mathbb{R}^n \leftarrow$ constant vector

2) $D(I^TX) = I^T\Sigma I$

3) $E(AX) = A\mu$, $A \in \mathbb{R}^{p\times n} \leftarrow$ constant matrix

4) $D(AX) = A\Sigma A^T$ and $\mathrm{Cov}(AX, BX) = A\Sigma B^T$

5) If $\mathrm{Cov}(U_p, V_q) = \Gamma$ then $\mathrm{Cov}(AU, BV) = A\Gamma B^T$

5) ⇒ 4) ⇒ 2) ; 3) ⇒ 1)

$E(\underset{\sim}{l}^T\underset{\sim}{X}) = E\sum_{i=1}^{n} l_i X_i = \sum l_i E(X_i) = \underset{\sim}{l}^T\underset{\sim}{\mu}$

$D(\underset{\sim}{l}^T\underset{\sim}{X}) = \text{var}(\underset{\sim}{l}^T\underset{\sim}{X}) = \text{cov}(\underset{\sim}{l}^T\underset{\sim}{X}, \underset{\sim}{l}^T\underset{\sim}{X})$

$= \sum_i \sum_j l_i \text{Cov}(x_i, x_j) l_j$

$= \underset{\sim}{l}^T \Sigma \underset{\sim}{l}$

$D(X) = \Sigma$ is symmetric

$D(X) = \Sigma$ is positive semi-definite

To show $D(\underset{\sim}{X}) = \Sigma$ is positive semi-definite.

$Y = \underset{\sim}{l}^T\underset{\sim}{X} \in \mathbb{R} \quad \underset{\sim}{l} \neq 0$

We know $\text{var}(Y) \geq 0$

$\text{Var}(Y) = \text{Var}(\underset{\sim}{l}^T\underset{\sim}{X}) = \underset{\sim}{l}^T \Sigma \underset{\sim}{l} \geq 0 \quad \forall \ \underset{\sim}{l} \neq 0$

$\Rightarrow \Sigma$ is positive semi-definite.

Theorem $P((X-\mu) \in C(\Sigma)) = 1$.

Proof

$X \in \mathbb{R}^n$, $E(\underset{\sim}{X}) = \mu \in \mathbb{R}^n$, $D(\underset{\sim}{X}) = \Sigma$

To show: $P((X-\mu) \in C(\Sigma)) = 1$.

It is equivalent to show if $l \in (C(\Sigma))^{\perp}$, then $l \perp (\underset{\sim}{X} - \underset{\sim}{\mu})$.

$\underset{\sim}{l} \in (C(\Sigma))^{\perp} \iff \underset{\sim}{l}^T \Sigma = \underset{\sim}{0}^T \iff \underset{\sim}{l}^T \Sigma \underset{\sim}{l} = \underset{\sim}{0}^T \underset{\sim}{l} = 0$

$\iff \text{Var}(\underset{\sim}{l}^T\underset{\sim}{X}) = 0$

$\iff \text{Var}(\underset{\sim}{l}^T(\underset{\sim}{X} - \underset{\sim}{\mu})) = 0$

Now, $E(\underset{\sim}{X} - \underset{\sim}{\mu}) = 0 \Rightarrow E(\underset{\sim}{l}^T(\underset{\sim}{X} - \underset{\sim}{\mu})) = 0$.

We have

$E(\underset{\sim}{l}^T(\underset{\sim}{X} - \underset{\sim}{\mu})) = 0$

$\text{Var}(\underset{\sim}{l}^T(\underset{\sim}{X} - \underset{\sim}{\mu})) = 0$. This implies,

$P(\underset{\sim}{l}^T(\underset{\sim}{X} - \underset{\sim}{\mu}) = 0) = 1 \Rightarrow P(\underset{\sim}{l} \perp (\underset{\sim}{X} - \underset{\sim}{\mu})) = 1$.

Hence, proved.

Some cases to consider

$A = I_n \quad E(\underset{\sim}{X}) = \underset{\sim}{\mu}, \; D(\underset{\sim}{X}) = \Sigma$

Case-I :

$$E(\underset{\sim}{X}^T A \underset{\sim}{X}) = E(\underset{\sim}{X}^T \underset{\sim}{X}) = E\left(\sum_{i=1}^{n} x_i^2\right)$$
$$= \sum_{i=1}^{n} (E(x_i^2))$$
$$= \sum_{i=1}^{n} (\mu_i^2 + \sigma_{ii})$$
$$= \sum_{i=1}^{n} \sigma_{ii} + \sum_{i=1}^{n} \mu_i^2$$

$$E(\underset{\sim}{X}^T A \underset{\sim}{X}) = tr(\Sigma I_n) + \underset{\sim}{\mu}^T I_n \underset{\sim}{\mu}$$

Case-II :

If $x_i \overset{iid}{\sim} N(0,1)$

$$E(\underset{\sim}{X}^T \underset{\sim}{X}) = E\left(\sum_{i=1}^{n} x_i^2\right) = n.$$

$\chi_n^2 \sim gamma(\alpha = n/2, \lambda = 1/2)$

$E[G] = \alpha/\lambda = n \; ; \; V(G) = \alpha/\lambda^2 = 2n.$

Hence, $E(X^T X) = n, \quad V(X^T X) = 2n.$

$$E(\underset{\sim}{X}^T A \underset{\sim}{X}) = E \sum_i \sum_j a_{ij} E(x_i x_j)$$
$$= \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} (\sigma_{ij} + \mu_i \mu_j)$$
$$= \sum_i \sum_j a_{ij} \sigma_{ij} + \sum_i \sum_j a_{ij} \mu_i \mu_j$$
$$= tr(\Sigma A) + \underset{\sim}{\mu}^T A \underset{\sim}{\mu}$$

$$E(\underset{\sim}{X}^T A \underset{\sim}{X}) = E[tr(\underset{\sim}{X}^T A \underset{\sim}{X})]$$
$$(\because \underset{\sim}{X}^T A \underset{\sim}{X} \text{ is a scalar quantity})$$
$$= E[tr(A X X^T)]$$
$$(\because tr(AB) = tr(BA))$$
$$= tr[E(A(XX^T))]$$
$$= tr[A(\Sigma + \mu\mu^T)] = tr(A\Sigma + A\mu\mu^T)$$

$$= \text{tr}(A\Sigma) + \text{tr}(\mu^T A \mu)$$
$$= \text{tr}(A\Sigma) + \mu^T A \mu$$
($\because \mu^T A \mu$ is a scalar quantity)

Multivariate Normal Distribution

$$f(x) = \frac{\exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\}}{(2\pi)^{n/2}|\Sigma|^{1/2}}$$

for some $\mu \in \mathbb{R}^n$ and p.s.d. Σ

Univariate Normal: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{1}{2\sigma^2}\cdot(x-\mu)^2}$

If Σ is a diagonal matrix,

$$f(x) = \frac{\exp\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_i)^2/\sigma_i^2\}}{(\sqrt{2\pi})^n(\prod_{i=1}^{n}\sigma_i)}$$

If $X \sim \mathcal{N}(\mu, \Sigma)$ then $AX \sim \mathcal{N}(A\mu, A\Sigma A^T)$

Ex: If (X,Y) follow Bivariate Normal, find $f(x,y)$, i.e., pdf of (X,Y).

Ex: Show that $Y|X=x$ follows $\mathcal{N}(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x), (1-\rho^2)\sigma_y^2)$ if (X,Y) follows Bivariate Normal distn.

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right]$$

Mahalanobis distance

Notion of distance when there is a fluctuation

$(X-\mu)^T\Sigma^{-1}(X-\mu)$

$\sum_{i=1}^{n}(x_i-\mu_i)^2/\sigma_i^2$ for diagonal Σ.

Suppose $Z = \begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right)$

$l_1 = (1,0)^T \quad l_1^T Z \sim N(\mu_x, \sigma_x^2)$

$l_2 = (0,1)^T \quad l_2^T Z \sim N(\mu_y, \sigma_y^2)$

Chi-Squared distribution

If $x_i \overset{iid}{\sim} N(0,1)$, then $\sum_{i=1}^{n} x_i^2$ follow chi-square distn. with n degree of freedom

If $x \sim N(\mu, I_n)$, then $x^T x$ follows chi-square distn. with n degrees of freedom and **non-centrality parameter** $\mu^T\mu$

$E(x^T x) = n + \mu^T\mu$.

for $n=1$,

$x \sim N(0,1)$

$x^2 \sim \chi_1^2 \quad E(x^2) = 1$

$Y \sim N(\mu, 1)$, then $E(Y^2) = \mu^2 + 1$.

Theorem If $x \sim N(\mu, I_n)$ then x^TAx has chi-sq. iff A is idempotent

Moreover, $x^TAx \sim \chi^2_{df=\text{Rank}(A), ncp=\mu^TA\mu}$

Cochran's Theorem

Let $x \sim N(\mu, I_n)$ and $x^TAx \equiv \sum_{i=1}^{k} x^TA_ix$ where A_i's are symmetric and A is an idempotent matrix. Then $x^TA_ix \sim \chi^2_{\text{Rank}(A_i), \mu^TA_i\mu}$ and they are independent

$Z \sim N(\underset{\sim}{0}, I_n)$

$X = Z + \mu, \quad \mu \neq 0$

$Z^T I_n Z \sim \chi^2_{df=n, ncp=0}$

$X^T X \sim \chi^2_{df=n, ncp=\mu^T\mu>0}$

Ncp parameter has important significance while performing tail tests.

Construction of t-test and F-test

Let $x_i \overset{iid}{\sim} N(0, \sigma^2)$. Define $\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $s^2 = \sum_{i=1}^{n}(x_i - \hat{x})^2$

$Z \sim N(0,1)$, $Y \sim \chi^2_{k,ncp}$ } independent. Then, $\frac{Z}{\sqrt{Y/k}} \sim t_{k,ncp}$

$x_i \overset{iid}{\sim} N(\mu, \sigma^2)$. Then $\underset{\sim}{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $\underset{\sim}{x} \sim N[\underset{\underset{\sim}{\mu}}{\mu \cdot \underset{\sim}{1}}, \underset{\Sigma}{\sigma^2 I_n}]$

$(\ldots \underset{\sim}{1} = (1\ 1\ 1 \ldots 1)^T)$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\underset{\sim}{1}^T\underset{\sim}{x} = l^T\underset{\sim}{x} \quad (\ldots l = 1/n \cdot \underset{\sim}{1}^T)$

$\bar{x} \sim N(l^T\underset{\sim}{\mu}, l^T\Sigma l) = N(\mu, \sigma^2/n)$

As a consequence,

$\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma} \sim N(0,1)$

$s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \underset{\sim}{x}^T(I_n - 1/n \cdot \underset{\sim}{1}\cdot\underset{\sim}{1}^T)\underset{\sim}{x}$

$P = \left(I_n - \frac{1}{n} \cdot 1 \cdot 1^T\right)$

Show that : i) $P^T = P$ ii) $P^2 = P$ iii) $\underset{\sim}{l}^T P = \underset{\sim}{0}^T$

$\frac{1}{n} \cdot 1^T (I_n - \frac{1}{n} 11^T) = \frac{1}{n}(1^T - \frac{1}{n} \cdot n \cdot 1^T) = \frac{1}{n}(1^T - 1^T)$

$S^2 \sim \chi^2_{df = \text{rank}(P), ncp = 0}$

Tutorial Class :

1. Prove that $C(AA^T) = C(A)$

We prove, i) $C(AA^T) \subseteq C(A)$

ii) $C(A) \subseteq C(AA^T)$

Let $\underset{\sim}{l} \in [C(AA^T)]^{\perp}$

$\Leftrightarrow \underset{\sim}{l}^T AA^T = \underset{\sim}{0}^T$

$\Leftrightarrow \underset{\sim}{l}^T AA^T \underset{\sim}{l} = \underset{\sim}{0}^T \underset{\sim}{l} = 0$

$\Leftrightarrow (A^T \underset{\sim}{l})^T (A^T \underset{\sim}{l}) = 0$

$\Leftrightarrow A^T \underset{\sim}{l} = \underset{\sim}{0}$

$\Leftrightarrow \underset{\sim}{l}^T A = \underset{\sim}{0}^T$

$\Leftrightarrow \underset{\sim}{l} \in [C(A)]^{\perp}$

$\therefore [C(A)]^{\perp} \subseteq [C(AA^T)]^{\perp}$ and vice versa.

$\Rightarrow C(AA^T) = C(A)$

2 a) $V = \mathbb{R}^3$, $S = \{(x_1, x_2, x_3) : 2x_1 + x_2 + x_3 = 1\}$ No

b) $V = \mathbb{R}^2$, $S = \{(x_1, x_2) : x_1, x_2 \geq 0\}$

→ S over $\mathbb{R}$? No

→ S over $\mathbb{R} \cup \{0\}$? Yes

→ S over $\mathbb{Z}_n$? Yes

Show that $\bar{x}$ and s^2 independently distribute

$x_i \overset{iid}{\sim} N(\mu, \sigma^2) \iff \underset{\sim}{x} \sim N(\underset{\sim}{\mu}, \sigma^2 I_n)$

$\underset{\sim}{1}^T = \frac{1}{n}\underset{\sim}{1}^T$

$\bar{x} = \frac{1}{n}\underset{\sim}{1}^T\underset{\sim}{x}$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$s^2 = \underset{\sim}{x}^T P \underset{\sim}{x}$

$P = (I_n - \frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}^T)$

$P^T = P$

$P^2 = P$

$\underset{\sim}{1}^T P = 0^T$

} Homework

$\underset{\sim}{l}^T\underset{\sim}{x} \sim N(\underset{\sim}{l}^T\underset{\sim}{\mu}, \underset{\sim}{l}^T\Sigma\underset{\sim}{l})$

$\Rightarrow \bar{x} \sim N(\mu, \sigma^2/n)$

As $x \overset{iid}{\sim} N(\mu, \sigma^2)$

$\underset{\sim}{y} = \left(\frac{\underset{\sim}{x}}{\sigma}\right) \sim N\left(\frac{\underset{\sim}{\mu}}{\sigma}, I_n\right)$

$s^2 = \underset{\sim}{x}^T P \underset{\sim}{x} = \sigma^2\left(\frac{\underset{\sim}{x}}{\sigma}\right)^T P \left(\frac{\underset{\sim}{x}}{\sigma}\right) = \sigma^2(\underset{\sim}{y}^T P \underset{\sim}{y})$

As $x_i \overset{iid}{\sim} N(\mu, \sigma^2)$

$y_i \overset{iid}{\sim} N(\mu/\sigma, 1)$

$\therefore \underset{\sim}{y}^T P \underset{\sim}{y} \sim \chi^2_{df=\text{rank}(P), ncp=(\mu/\sigma)^T P (\mu/\sigma)}$

$\text{Rank}(P) = (n-1)$

$\left(\frac{\mu^2}{\sigma^2}\right)\underset{\sim}{1}^T P \underset{\sim}{1} = \underset{\sim}{0}^T\underset{\sim}{1} = 0$

$I_n = \left(I_n - \frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}^T\right) + \frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}^T$

$\bar{x} \sim N(\mu, \sigma^2/n)$

$s^2 = \sigma^2 \chi^2_{df=n-1, ncp=0}$

As a result, $\bar{x}$ and S^2 are independent.

$$Z = \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma} \sim N(0,1)$$

$$S^2/\sigma^2 \sim \chi^2_{n-1,ncp=0}$$

independent

t-statistic

$$\frac{Z}{\sqrt{(S^2/\sigma^2)/(n-1)}} = \frac{\sqrt{n}(\bar{x}-\mu)/\sigma}{\sqrt{(S^2/\sigma^2)/(n-1)}} \sim t_{n-1,ncp=0}$$

$x_i \overset{iid}{\sim} N(\mu, \sigma^2) \iff \underset{\sim}{x} \sim N(\underset{\sim}{\mu}, \sigma^2 I_n)$

$S^2 = \sigma^2 \underset{\sim}{Y}^T P \underset{\sim}{Y}$

$$t = \frac{\sqrt{n}(\bar{x}-\mu)/\sigma}{\sqrt{(S^2/\sigma^2)/(n-1)}}$$

$$t^2 = \frac{(\sqrt{n}(\bar{x}-\mu)/\sigma)^2/1}{(S^2/\sigma^2)/(n-1)} = \frac{Z^2/1}{\chi^2_{n-1}/(n-1)} \sim F_{1,n-1}$$

Formulation of Hypothesis,

$H_0: \mu = 5$

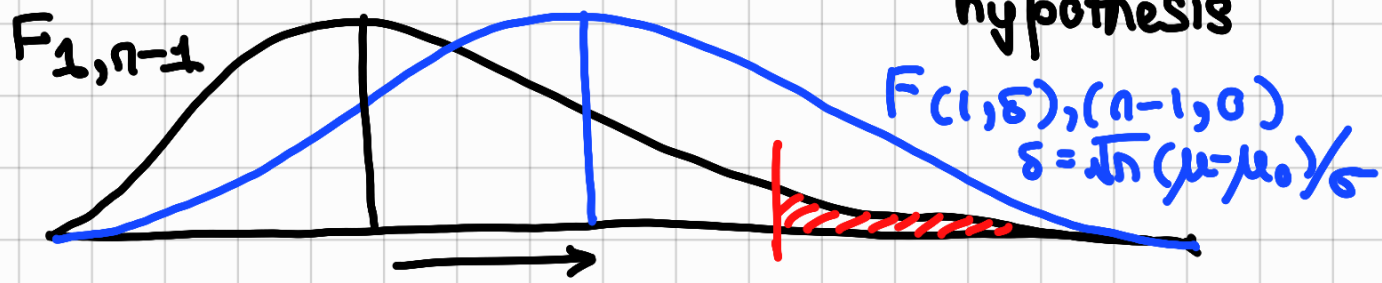
$H_1: \mu \neq 5$

σ^2 unknown

Suppose we have the above hypothesis. What type of deformation would happen to F-distn. if the mean is shifted?

A non-centrality parameter would be introduced

in the numerator, and the expectation would increase. This forms a good measure to evaluate hypothesis



We construct t-statistic from the data.
used as Upper tail test

Upper Tail test

Under H_1,

$$\bar{X} \sim N(\mu-\mu_0, \sigma^2/n)$$

$$\frac{\sqrt{n}\,\bar{X}}{\sigma} \sim N\left(\frac{\sqrt{n}}{\sigma}(\mu-\mu_0), 1\right)$$

References

Mathematical Statistics, Miller

Simple Linear Regression

$$y = \beta_0 + \beta_1 x + \epsilon$$

relation b/w response variable y and regressor variable x is linear in parameter, then it is called simple linear regression model.

Even $y = \beta_0 + \beta_1 e^x + \epsilon$ is a simple linear regression model.

But, $y = \frac{1}{\beta_0 + \beta_1 x} + \epsilon$, $y = \Phi(\beta_0 + \beta_1 x) + \epsilon$ is not.

Linear Model $y_i = \underset{\sim}{\beta}^T \underset{\sim}{x}_j + \epsilon_i$

Gauss-Markov Model $y_i = \beta_0 + \beta_1 x + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Here, $\beta_0 \in \mathbb{R}$, $\beta_1 \in \mathbb{R}$, $\sigma > 0$ are unknown model parameters.

$E(y_i) = \beta_0 + \beta_1 x_i$

$Var(y_i) = \sigma^2$

$\longrightarrow y_i \sim N(\underset{\sim}{\beta}^T \underset{\sim}{x}_j, \sigma^2)$

For simple linear regression,

$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

$\underset{\sim}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ and σ^2 unknown

We define

$\underset{\sim}{Y} = (Y_1 Y_2 \dots Y_n)^T$; $\underset{\sim}{\epsilon} = (\epsilon_1 \epsilon_2 \dots \epsilon_n)^T$

$\underset{\sim}{\beta} = (\beta_0 \ \beta_1)^T$; $\underset{\sim}{x} = (x_1 x_2 \dots x_n)^T$

$X = [1 | \underset{\sim}{x}]$

$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$

$\underset{\sim}{Y} \sim N(X\underset{\sim}{\beta}, \sigma^2 I_n)$ as $\underset{\sim}{\epsilon} \sim N(\underset{\sim}{0}, \sigma^2 I_n)$

It is necessary to define a model with the same metric throughout

Estimation of model parameters

least squared model

- Data: $\{(x_i, y_i) : i = 1, 2, \dots, n\}$
- Model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ assumed
- Metric: square distance assumed

least squared condition

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

$$S(\underset{\sim}{\beta}) = (\underset{\sim}{Y} - X\underset{\sim}{\beta})^T (\underset{\sim}{Y} - X\underset{\sim}{\beta})$$

$$= \underset{\sim}{Y}^T\underset{\sim}{Y} - \underset{\sim}{Y}^T X\underset{\sim}{\beta} - \underset{\sim}{\beta}^T X^T \underset{\sim}{Y} + \underset{\sim}{\beta}^T X^T X \underset{\sim}{\beta}$$

$$\frac{\partial S(\beta)}{\partial \underset{\sim}{\beta}} = \underset{\sim}{0}$$

$$\left(\frac{\partial x^T A x}{\partial x} = x^T (A + A^T) \right)$$

$$\left(\frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a^T \right)$$

This implies,

$$2\times 2 \quad (X^T X)\underset{\sim}{\beta} = X^T \underset{\sim}{Y}$$

Assuming $X^T X$ is invertible,

$$\underset{\sim}{\beta} = (X^T X)^{-1} X^T \underset{\sim}{Y}$$

When will $X^T X$ not be invertible?

- When $\text{rank}(X^T X) < 2$, i.e., all x_i values are same.

Homework Find β_0, β_1 from the above result and compare it with the result which we got from the usual derivative.

$$\hat{\beta}_1 = S_{xy}/S_{xx}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Hence, $\hat{y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$

$$= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_{new}$$

$$= \bar{y} + \hat{\beta}_1 (x_{new} - \bar{x})$$

passing through $(\bar{x}, \bar{y})$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x}) y_i$$

$$= \sum_{i=1}^{n} x_i (y_i - \bar{y})$$

$\hat{\beta}_1 = S_{xy}/S_{xx}$; $S_{xx} = \sum_{i=1}^{n}(x_i-\bar{x})^2$

non-random

$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{S_{xx}}$

$= \underset{\sim}{l}^T\underset{\sim}{Y}$; $l_i = (x_i-\bar{x})/S_{xx}$

$\underset{\sim}{l} = (l_1\ l_2\ \cdots\ l_n)^T$

Linear Estimator

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

$\hat{\beta}_1 = \underset{\sim}{l}^T\underset{\sim}{Y} \sim N(\underset{\sim}{l}^T X\underset{\sim}{\beta}, \sigma^2 \underset{\sim}{l}^T I_n \underset{\sim}{l})$

$E(\hat{\beta}_1) = \underset{\sim}{l}^T X\underset{\sim}{\beta}$

$= [0\ 1]^T\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} = \beta_1$

$var(\hat{\beta}_1) = (\underset{\sim}{l}^T\underset{\sim}{l})\sigma^2 = \sigma^2\frac{\sum_{i=1}^{n}(x_i-\bar{x})}{S_{xx}^2} = \sigma^2/S_{xx}$

$\hat{\beta}_1$ is an unbiased estimator of β_1.

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \frac{1}{n}1^T\underset{\sim}{Y} - \frac{\bar{x}}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})y_i$

$= \frac{1}{n}1^T\underset{\sim}{Y} - \bar{x}\,\underset{\sim}{l}^T\underset{\sim}{Y}$

$= (\frac{1}{n}1^T - \bar{x}\,\underset{\sim}{l}^T)\underset{\sim}{Y}$

$= \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\bar{x}(x_i-\bar{x})}{S_{xx}}\right)y_i$

SLR

$\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx})$; $\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}))$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$
$$\hat{y} = X\hat{\beta} = [X(X^T X)^{-1} X^T] Y = P_X Y$$

$P_X^T = P_X$ and $P_X^2 = P_X$ (Homework)

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$
$$\Rightarrow y_i \overset{\text{independent}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$$

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

Estimated Error

$$e_i = y_i - \hat{y}_i \quad \forall\ i = 1, 2, \ldots, n$$
$$e = (e_1, e_2, \ldots, e_n)^T$$
$$e = (y - \hat{y}) = (I - P_X) Y$$

$$e \sim N(0, \sigma^2 (I_n - P_X))$$ no more iid

Using,
$Z \sim N(\mu, \Sigma) \Rightarrow AZ \sim N(A\mu, A\Sigma A^T)$

Squared estimate error

$$\sum_{i=1}^{n} e_i^2 = e^T e$$
$$= [(I_n - P_X) Y]^T (I_n - P_X) Y$$
$$= Y^T (I_n - P_X)(I_n - P_X) Y$$
$$= Y^T (I_n - P_X) Y$$

$$\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \sim \chi^2_{df = \text{rank}(I_n - P_X),\ ncp = \frac{1}{\sigma^2}[(X\beta)^T (I_n - P_X) X\beta]}$$

$e^Te \sim \chi^2_{(n-2)}$ $\ldots (I_n = P_X + (I_n - P_X) \therefore n = 2 + (n-2))$

$ncp = \frac{1}{\sigma^2} \beta^T X^T (I_n - P_X) X \beta$

$= \frac{1}{\sigma^2} \beta^T [X^TX - X^TP_XX] \beta$

$P_X = X^T(X^TX)^{-1}X$
orthogonal projection matrix for the column space of X.
$\therefore P_XX = X$

Hence, $ncp = \frac{1}{\sigma^2} \beta^T 0 \beta = 0. \Rightarrow \frac{e^Te}{\sigma^2} \sim \chi^2_{(n-2),0}$

$E\left(\frac{e^Te}{\sigma^2}\right) = n-2 \Rightarrow E\left(\frac{e^Te}{n-2}\right) = \sigma^2$

$C(X) = \text{span}\{x_1, x_2, \ldots, x_n\}$
XX^T: projection matrix of $C(X)$.
$P_X = X(X^TX)^{-1}X^T$: orthogonal projection matrix
Prediction $\hat{y} = P_X y$
Estimated error $e = (I_n - P_X)y$

$\hat{y}^Te = y^TP_X(I_n - P_X)y$
$= y^T(P_X - P_X)y$
$= y^T0y = 0$

$\hat{y}$ and e are independently distributed.
$Cov(\hat{y}, e) = P_X(\sigma^2 I_n)(I - P_X)^T = 0_{n \times n}$

$\hat{y}$ and $e^Te = y^T(I_n - P_X)y$ are also independently distributed.

Homework: Verify that

$$\underset{\sim}{Y}^T(I_n - P_x)(I_n - P_x)\underset{\sim}{Y} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

One major advantage of this expression (RHS) is that we can readily estimate the error despite fitting the model.

Exercise: Find the MLE of $\hat{\beta}_0, \hat{\beta}_1, \sigma^2$.

Likelihood fn:

$$L(\underset{\sim}{\beta}) = \prod_{i=1}^{n} \exp\left\{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2\right\} \Big/ \sigma\sqrt{2\pi}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

$\hat{\beta}_{1\,LS} = \hat{\beta}_{1\,MLE}$; $\hat{\beta}_{0\,LS} = \hat{\beta}_{0\,MLE}$

Homework

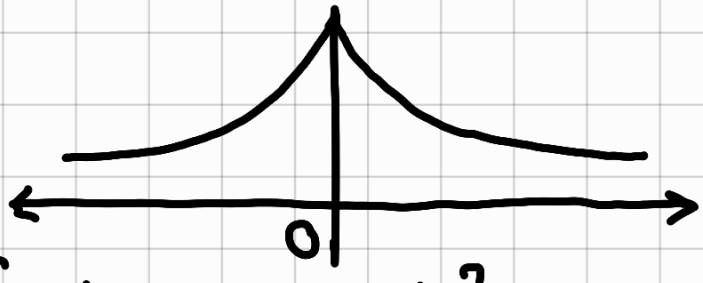
$\hat{\sigma}^2_{LS} = \frac{\underset{\sim}{e}^T\underset{\sim}{e}}{n-2}$ unbiased

but

$\hat{\sigma}^2_{MLE} = \frac{\underset{\sim}{e}^T\underset{\sim}{e}}{n}$ biased

Suppose,

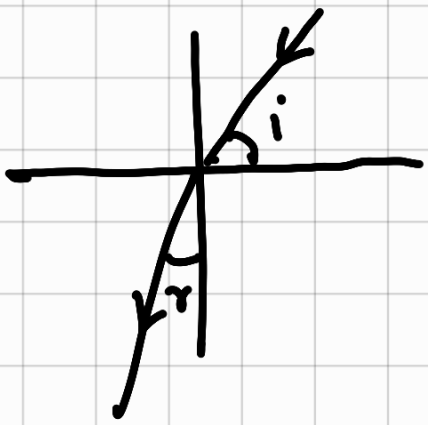
$\epsilon_i \overset{iid}{\sim} f(b) = \frac{e^{-|t|/\lambda}}{2\lambda}$



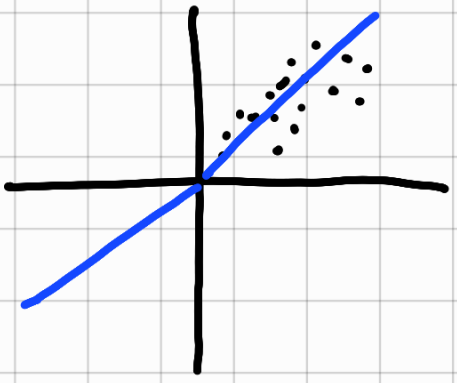
$y_i \sim g(y) = \frac{1}{2\lambda} \cdot \exp\left\{\frac{-|y_i - \beta_0 - \beta_1 x_i|}{\lambda}\right\}$

Obtain the MLE of $\hat{\beta}_0, \hat{\beta}_1$ and λ.

Testing for Simple Linear Regression



$$\frac{\sin i}{\sin r} = e.$$



$H_0: \beta_0 = 0 (= b_0)$
$H_1: \beta_0 \neq 0 (= b_0)$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

Step-1

$$\rightarrow \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim N(0,1)$$

Step-2

$$\hat{\sigma}^2 = \frac{e^T e}{n-2} = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right)$$

estimation of unknown parameter

Step-3

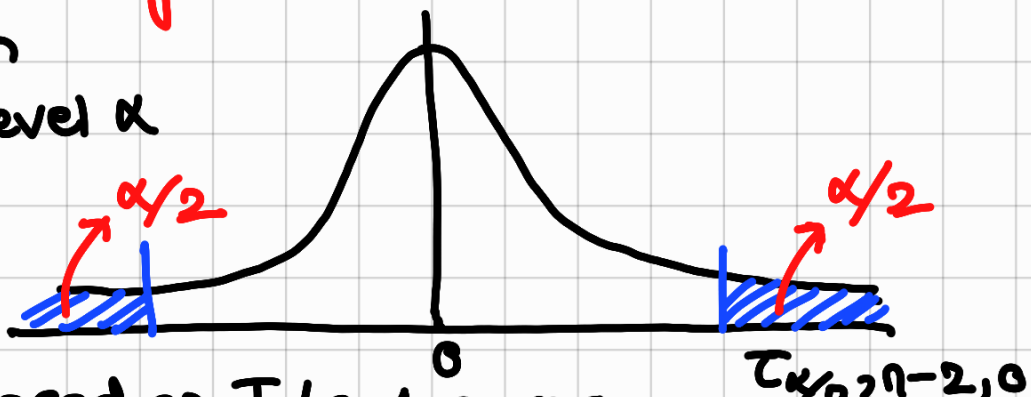
$$T = \frac{\hat{\beta}_0 - b_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2,0}$$ under H_0.

Step-4

Calculate the value $T_{observed}$

Step-5 Rejection condition

based on H_1 and level α



α is based on Type-1 error

P_{H_0}(rejecting H_0) $\leq \alpha$

... (probability of rejecting H_0 when H_0 is true)

$\alpha = 0.01, 0.05$ or 0.1

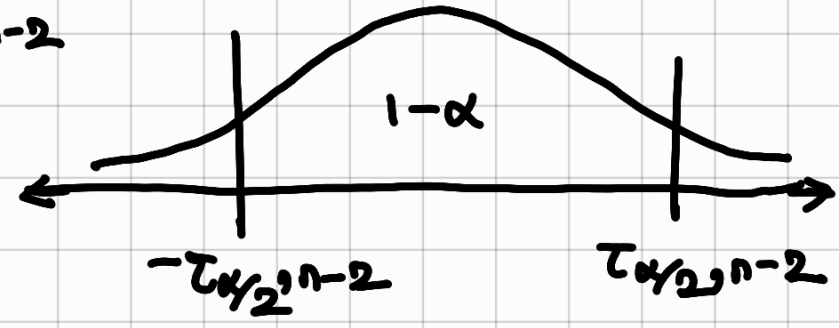
Step 6 Test Rule

We reject H_0 in favour of H_1 at level α.

if $|T_{observed}| > \tau_{\alpha/2,n-2}$

95% confidence interval of β_0

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}$$



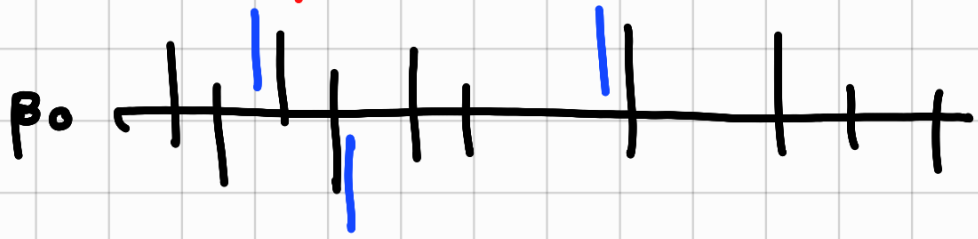
$$P\left(-\tau_{\alpha/2,n-2} < T < \tau_{\alpha/2,n-2}\right) = 1-\alpha$$

$$P\left(-\tau_{\alpha/2,n-2} < \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2\left(1/n + \bar{x}^2/S_{xx}\right)}} < \tau_{\alpha/2,n-2}\right) = 1-\alpha$$

$$L(\underline{x},\underline{y}) = \hat{\beta}_0 - T_{\alpha/2,n-2}\,\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}$$

$$U(\underline{x},\underline{y}) = \hat{\beta}_0 + T_{\alpha/2,n-2}\,\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}$$

Interpretation of confidence interval



Homework Find the 95% CI of σ^2.

$D = \{(x_i, y_i) : i = 1,2,\ldots,n\}$

$\hat{\beta}_0, \hat{\beta}_1$ obtained from data

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$x_0 \notin \{x_1, x_2, \ldots, x_n\}$

95% CI of $E(y_0)$.

for the same model

$$E(y_0) = \beta_0 + \beta_1 x_0$$

for parameter, we use term confidence interval (.... CI of $E(y_0)$)

for variable, we use term prediction interval (... PI of y_0)

$$y_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2(\ \ ))$$

$$\begin{aligned} \operatorname{var}(\hat{y}_0) &= \operatorname{var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\ &= \operatorname{var}(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0) \\ &= \operatorname{var}(\bar{y} + \hat{\beta}_1 (x_0 - \bar{x})) \\ &= \operatorname{var}(\bar{y}) + (x_0 - \bar{x})^2 \operatorname{var}(\hat{\beta}_1) + \operatorname{Cov}(\bar{y}, \hat{\beta}_1) \end{aligned}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\underset{\sim}{1}^T\underset{\sim}{y}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n}\frac{(x_i-\bar{x})y_i}{S_{xx}} = \underset{\sim}{l}^T\underset{\sim}{y}$$

$$\text{cov}(\bar{y}_0, \hat{\beta}_1) = \frac{1}{n}\underset{\sim}{1}^T(\sigma^2 I_n)\underset{\sim}{l} = \frac{\sigma^2}{n}\underset{\sim}{1}^T\underset{\sim}{l} = 0$$

$$l_i = (x_i-\bar{x})/S_{xx}$$

$$y_0 \sim N(\beta_0+\beta_1 x_0, \sigma^2)$$

$$\hat{y}_0 = \hat{\beta}_0+\hat{\beta}_1 x_0 \sim N\left(\beta_0+\beta_1 x_0, \sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)\right)$$

$$\frac{\hat{y}_0-(\beta_0+\beta_1 x_0)}{\hat{\sigma}\sqrt{\left(\frac{1}{n}+\left(\frac{x_0-\bar{x}}{S_{xx}}\right)^2\right)}} \sim t_{n-2}$$

95% of CI of $E(y_0)$ is

$$\hat{y} \mp t_{\alpha/2,n-2}\,\hat{\sigma}\sqrt{\frac{1}{n}+\left(\frac{\bar{x}-x_0}{S_{xx}}\right)^2}$$

95% prediction interval of y_0

independent $\begin{cases} y_0 \sim N(\beta_0+\beta_1 x_0, \sigma^2) \\ \hat{y}_0 \sim N\left(\beta_0+\beta_1 x_0, \sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)\right) \end{cases}$

$$y_0-\hat{y}_0 \sim N\left(0, \left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)\right)$$

Another case example for testing

$(X,Y) \sim BVN(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$

$H_0: \rho = 0$ v/s $H_1: \rho \neq 0$

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

$$r = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2 \sum (y_i-\bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$H_0: \rho = \rho_0 \neq 0$ v/s $\rho \neq \rho_0$

define $Z = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) = \tanh^{-1}(r)$

$$\mu_Z = \frac{1}{2}\log_e\left(\frac{1+\rho_0}{1-\rho_0}\right) = \tanh^{-1}(\rho_0)$$

$$\hat{\sigma}_Z^2 = (n-3)^{-1}$$

test statistic $W = \frac{Z-\mu_Z}{\sigma_Z} \sim N(0,1)$ as $n \to \infty$

Multiple linear regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_K x_{ik} + \epsilon_i$$

$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

unknown parameters : $\underset{\sim}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}$ and σ^2

Polynomial regression

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_K x_i^K + \epsilon_i$$

$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Unknown : $(\underset{\sim}{\beta}, \sigma^2)$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \beta_4 x_{i1}^2 + \beta_5 x_{i2}^2 + \epsilon$$

Unknown : $(\underset{\sim}{\beta}, \sigma^2)$

$$\underset{\sim}{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} \quad \underset{\sim}{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad \underset{\sim}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

$$X = [\underset{\sim}{1}\ c_1\ c_2 \cdots c_k]$$
$$= [1\ x_{i1}\ x_{i2} \cdots x_{ik}] \rightarrow \text{ith-row}$$

$$\underset{n\times 1}{\underset{\sim}{Y}} = \underset{n\times(k+1)}{X}\ \underset{(k+1)\times 1}{\underset{\sim}{\beta}} + \underset{\sim}{\epsilon} \rightarrow n\times 1$$

Least squared condition

$$S(\beta) = (Y - X\underset{\sim}{\beta})^T(Y - X\underset{\sim}{\beta}) \rightarrow \frac{\partial S(\underset{\sim}{\beta})}{\partial \underset{\sim}{\beta}} = \underset{\sim}{0}$$

$$\rightarrow \hat{\beta} = (X^TX)^{-1}X^TY$$

If $|X^TX| \neq 0$,

$$\hat{y}_{new} = \underset{\sim}{x}_{new}^T \cdot \underset{\sim}{\hat{\beta}}$$

$$\hat{y} = X\underset{\sim}{\hat{\beta}} = [X(X^TX)^{-1}X^T]Y = P_XY$$

$P_X \leftarrow$ orthogonal projection matrix of $C(X)$.

Estimated error vector

$$\underset{\sim}{e} = \underset{\sim}{Y} - \underset{\sim}{\hat{Y}} = (I - P_X)Y$$

$\hat{\beta} = (X^TX)^{-1}X^TY$ where $\underset{\sim}{Y} \sim \mathcal{N}(X\underset{\sim}{\beta}, \sigma^2 I_n)$

$\hat{\beta} \sim \mathcal{N}(\underset{\sim}{\beta}, \sigma^2(X^TX)^{-1})$ ... $((X^TX)^{-1}(X^TX)\underset{\sim}{\beta} = \underset{\sim}{\beta})$

$\Rightarrow \hat{\beta}$ is an unbiased estimator.

Estimation of σ^2

$$\sigma^2\left(\frac{\underset{\sim}{e}^T\underset{\sim}{e}}{\sigma^2}\right) = \left[(I_n - P_X)\frac{\underset{\sim}{Y}}{\sigma}\right]^T\left[(I_n - P_X)\frac{\underset{\sim}{Y}}{\sigma}\right]$$

$$\left[\ldots \frac{\underset{\sim}{Y}}{\sigma} \sim N\left(\frac{X\underset{\sim}{\beta}}{\sigma}, I_n\right)\right]$$

$$= \sigma^2\left[\frac{\underset{\sim}{Y}^T}{\sigma}\cdot(I_n - P_X)\frac{\underset{\sim}{Y}}{\sigma}\right]$$

$$\left[\ldots (I_n - P_X)^T = I_n - P_X,\ (I_n - P_X)^2 = I_n - P_X\right]$$

$$\sigma^2\left(\frac{e^Te}{\sigma^2}\right) \sim \sigma^2\chi^2$$

$df = \text{rank}(I - P_X) = n - 1$

$$ncp = \left(\frac{X\underset{\sim}{\beta}}{\sigma}\right)^T(I_n - P_X)\left(\frac{X\underset{\sim}{\beta}}{\sigma}\right) = 0$$

$$E[\underset{\sim}{e}^T\underset{\sim}{e}] = \sigma^2(n-k+1)$$

$$\Rightarrow E\left[\frac{\underset{\sim}{e}^T\underset{\sim}{e}}{n-k-1}\right] = \sigma^2$$

Unbiased estimator of σ^2 is $\hat{\sigma}^2 = \frac{\underset{\sim}{e}^T\underset{\sim}{e}}{n-k-1}$.

ncp: $\left(\frac{X\underset{\sim}{\beta}}{\sigma}\right)^T(I_n - P_X)\left(\frac{X\underset{\sim}{\beta}}{\sigma}\right)$

$$= \frac{1}{\sigma^2}\underset{\sim}{\beta}^T\left[X^T(I_n - P_X)X\right]\beta$$

$$= \frac{1}{\sigma^2}\underset{\sim}{\beta}^T\left[X^TX - X^TP_XX\right]\underset{\sim}{\beta} = \frac{1}{\sigma^2}\underset{\sim}{\beta}^T\left[X^TX - X^TX\right]\underset{\sim}{\beta}$$

Homework MLE of $\sigma^2 = \underset{\sim}{\varepsilon}^T \underset{\sim}{\varepsilon}/n$

Hence, it is not an unbiased estimator.

Testing the significance of feature(s)

1. $H_0: \beta_i = 0 \ (= b_i)$
   $H_1: \beta_i \neq 0 \ (\neq b_i)$

   Testing of linear combination of β_i's, $i = 1, \ldots$

2. $H_0: \beta_1 + 2\beta_2 = b_0$
   $H_1: \beta_1 + 2\beta_2 \neq b_0$

   $\iff$ $H_0: \underset{\sim}{l}^T \underset{\sim}{\beta} = b_0$
   $H_1: \underset{\sim}{l}^T \underset{\sim}{\beta} \neq b_0$

3. $H_0: \begin{cases} \underset{\sim}{l_1}^T \underset{\sim}{\beta} = b_1 \\ \underset{\sim}{l_2}^T \underset{\sim}{\beta} = b_2 \end{cases}$
   H_1: H_0 is not true.

   How to combine both the hypothesis ?

   Develop a methodology to simultaneously test both of them together

4. $H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$
   H_1: H_0 is not true

   ⫼

   More general case of above version

   $H_0: \begin{cases} \beta_1 = 0 \\ \beta_1 - \beta_2 = 0 \\ \beta_1 - \beta_3 = 0 \\ \vdots \\ \beta_1 - \beta_k = 0 \end{cases}$
   H_1: H_0 is not true

   $\longrightarrow$ This testing signifies if it is at all necessary to build a model this way or not.

   $\longrightarrow$ Analysis of variance.

   ANOVA regression model

Assumption : abs $|(X^T X)| >> 0$

If $|X^TX| = 0$, then any $\underset{\sim}{l}^T\underset{\sim}{\beta}$ may not be estimable i.e there will be no linear unbiased estimator of $\underset{\sim}{l}^T\underset{\sim}{\beta}$.

i.e., $\nexists\ \underset{\sim}{b}^T\underset{\sim}{y}$ such that $E[\underset{\sim}{b}^T\underset{\sim}{y}] = \underset{\sim}{l}^T\underset{\sim}{\beta}$.

We have

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

$$\underset{\sim}{\hat{\beta}} \sim N(\underset{\sim}{\beta}, \sigma^2(X^TX)^{-1})$$

$$\hat{\sigma}^2 = \frac{\underset{\sim}{e}^T\underset{\sim}{e}}{n-k-1}$$

$$(X^TX)^{-1} = C_{(k+1)\times(k+1)} = ((c_{ij}))$$

$i,j = 0,1,2,\ldots,k$

Q1 $H_0: \beta_i = b_i$

$H_1: \beta_i \neq b_i$

$\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii})$

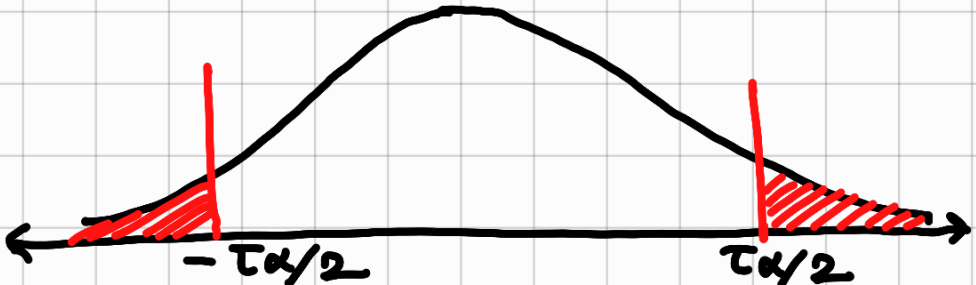
$$\rightarrow \frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{c_{ii}}} \sim N(0,1)$$

Under H_0,

$$\frac{\hat{\beta}_i - b_i}{\sigma\sqrt{c_{ii}}} \sim N(0,1)$$

$$\Rightarrow T = \frac{\hat{\beta}_i - b_i}{\hat{\sigma}\sqrt{c_{ii}}} \sim t_{n-k-1}$$

Reject H_0 if $|T_{obs}| > T_{\alpha/2, n-k-1}$



Q2 $H_0: a\beta_i + b\beta_j = d$

$H_1: a\beta_i + b\beta_j \neq d$

$$a\hat{\beta}_i + b\hat{\beta}_j \sim \mathcal{N}(a\beta_i + b\beta_j, (a^2c_{ii} + b^2c_{jj} + 2abc_{ij})\sigma^2)$$

Under H_0:

$$\frac{(a\hat{\beta}_i + b\hat{\beta}_j - d)}{\sigma\sqrt{a^2c_{ii} + b^2c_{jj} + 2abc_{ij}}} \sim \mathcal{N}(0,1)$$

$$\rightarrow T = \frac{a\hat{\beta}_i + b\hat{\beta}_j - d}{\hat{\sigma}\sqrt{a^2c_{ii} + b^2c_{jj} + 2abc_{ij}}} \overset{H_0}{\sim} t_{n-k-1}$$

We reject H_0 in favor of H_1 at level α if

$|T_{obs}| > T_{\alpha/2, n-k-1}$

Q3. $H_0: \begin{cases} 2\beta_1 + 3\beta_2 = 1 \\ 4\beta_3 - 5\beta_6 = 7 \end{cases}$ $\begin{bmatrix} 0 & 2 & 3 & 0 & \dots & 0 \\ 0 & 0 & 0 & 4 & 0 & -5 \end{bmatrix} \underset{\sim}{\beta} = \begin{bmatrix} 1 \\ 7 \end{bmatrix}$

H_1: H_0 is not true

$A\underset{\sim}{\beta} = \underset{\sim}{b}$

$$\hat{\beta} \sim \mathcal{N}(\underset{\sim}{\beta}, C\sigma^2)$$

$$A\hat{\beta} \sim \mathcal{N}(A\underset{\sim}{\beta}, ACA^T\sigma^2)$$

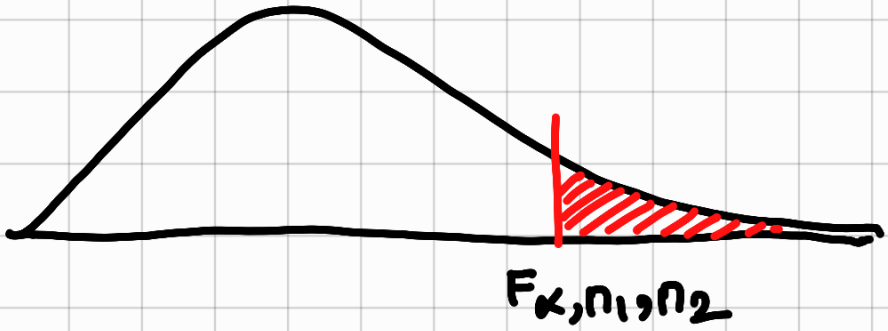
(...we can no longer make it into t-statistic since this is a vector)

$$(A\hat{\beta} - b)^T \left[\frac{(ACA^T)^{-1}}{\sigma^2}\right](A\hat{\beta} - b) \overset{H_0}{\sim} \chi^2_{rank(ACA^T)}$$

Note: $|\Sigma| > 0$, $\underset{\sim}{Z} \sim \mathcal{N}(\underset{\sim}{\mu}, \Sigma)$

$(\underset{\sim}{Z} - \underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{Z} - \underset{\sim}{\mu})$ follows $\chi^2_{rank(\Sigma)}$

$$(A\hat{\beta}-b)^T \frac{(AC A^T)^{-1}}{\hat{\sigma}^2} (A\hat{\beta}-b) \sim F_{\text{rank}(ACA^T),\, n-k-1}$$



Reject H_0 if $F_{obs} > F_{\alpha,n_1,n_2}$

Q4. ANOVA

H_0: $\beta_1 = \beta_2 = \dots = \beta_k = 0$

H_1: H_0 is not true

$\longrightarrow$ "do we even need to fit a model this way or not"

$$SSTotal = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \underset{\sim}{Y}^T (I_n - \tfrac{1}{n} 1 1^T) \underset{\sim}{Y}$$

$$SSError = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \underset{\sim}{Y}^T (I_n - P_X) \underset{\sim}{Y}$$

$$SSRegression = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \underset{\sim}{Y}^T (P_X - \tfrac{1}{n} 1 1^T) \underset{\sim}{Y}$$

$$I_n - \tfrac{1}{n} 1 1^T = (I_n - P_X) + (P_X - \tfrac{1}{n} 1 1^T)$$

$$SSTotal = SSError + SSRegression$$

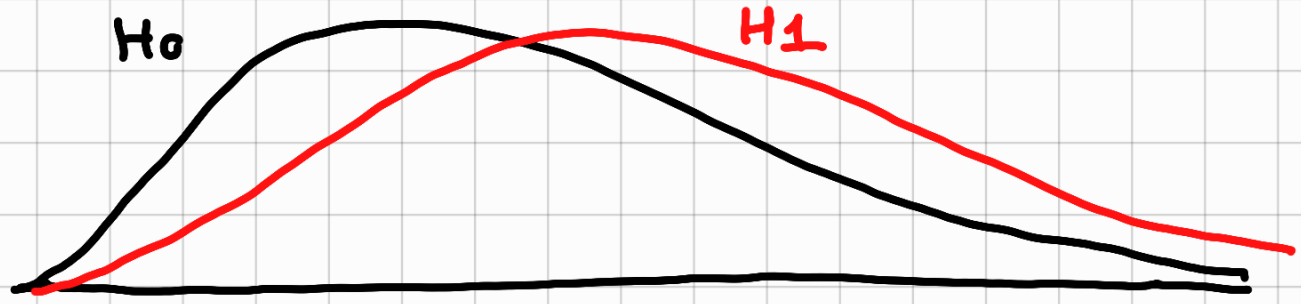
$$SSError/\sigma^2 \sim \chi^2_{n-k-1,\, ncp=0}$$

$$SSReg./\sigma^2 \sim \chi^2_{k,\, ncp=\lambda} \quad \dots \text{Cochran's thm.} \longrightarrow$$

$$SSTotal/\sigma^2 \sim \chi^2_{n-1,ncp=\lambda}$$

$$F = \frac{(SSReg/\sigma^2)/k}{(SSError/\sigma^2)/n-k-1} \sim F_{(k,\lambda),(n-k-1,0)}$$

If H_0 is true, $\lambda = 0$.
otherwise, $\lambda > 0$.



Exercise. Try ANOVA in the case of simple linear regression and the line passing through origin.

Rough Outline for the Proof of (*)

$$\text{NCP of } (SSReg./\sigma^2) = \left(\frac{X\underset{\sim}{\beta}}{\sigma}\right)^T \left(P_X - \frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}^T\right)\left(\frac{X\underset{\sim}{\beta}}{\sigma}\right)$$

$$= \frac{1}{\sigma^2}\underset{\sim}{\beta}^T\left[X^T\left(P_X - \frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}^T\right)X\right]\underset{\sim}{\beta}$$

$$= \frac{1}{\sigma^2}\underset{\sim}{\beta}^T\left[X^TX - \frac{1}{n}X^T\underset{\sim}{1}\underset{\sim}{1}^TX\right]\underset{\sim}{\beta}$$

$$\underset{\sim}{\beta} = \begin{pmatrix}\beta_0 \\ \beta_R\end{pmatrix} \qquad X = \left[\underset{\sim}{1} : X_R\right]$$

Hence,

$$\text{NCP} = \frac{1}{\sigma^2}\underset{\sim}{\beta}^T\left[\begin{pmatrix} n & \underset{\sim}{1}^TX_R \\ X_R^T\underset{\sim}{1} & X_R^TX_R\end{pmatrix} - \begin{pmatrix} n & \underset{\sim}{1}^TX_R \\ X_R^T\underset{\sim}{1} & \frac{1}{n}X_R^T\underset{\sim}{1}\underset{\sim}{1}^TX_R\end{pmatrix}\right]\underset{\sim}{\beta}$$

$$= \frac{1}{\sigma^2}\begin{pmatrix}\beta_0\\ \beta_R\end{pmatrix}^T\begin{pmatrix}0 & 0^T\\ 0 & A\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_R\end{pmatrix} = \frac{1}{\sigma^2}\beta_R^T A \beta_R$$

not a function of β_0

$$A = X_R^T X_R - \frac{1}{n} X_R^T \mathbf{1}\mathbf{1}^T X_R,$$

Variance - covariance matrix of $X_1\ X_2\ \dots\ X_k$

What if (X^TX) is not invertible? The rest of the part deals with this question.

Some definitions and theorems :

Linear Unbiased Estimator

$l^Ty = \sum_i l_i y_i$ is said to be LUE of $p^T\beta$ if $E[l^Ty] = p^T\beta\ \forall\ \beta \in \mathbb{R}^{k+1}$.

Linear Zero function

$l^Ty = \sum_i l_i y_i$ is said to be LZF if $E(l^Ty) = 0$

- $p^T\beta$ is estimable if it has a LUE
- Best LUE (BLUE) is LUE $p^T\beta$ with min. variance.

Theorem :

A linear function is BLUE of its expectation iff it is uncorrelated with all LZF.

$p^T\beta$ is estimable iff $\exists\ l^Ty$ such that $E(l^Ty) = p^T\beta\ \forall\ \beta \in \mathbb{R}^{k+1}$

$$\Leftrightarrow\ l^TX\beta = p^T\beta\ \forall\ \beta \in \mathbb{R}^{k+1}$$

$$\boxed{l^TX = p^T} \qquad \boxed{p \in C(X^T)}$$

$$p = X^Tl$$

LZF $E(l^T y) = 0 \quad \forall \beta \in \mathbb{R}^{k+1}$

$\Rightarrow l^T X \beta = 0 \quad \forall \beta \in \mathbb{R}^{k+1}$

$\Rightarrow \boxed{l^T X = 0^T} \iff l \in (C(X))^\perp$

$\Rightarrow l \in C(I_n - P_X)$

$l^T X = m^T (I_n - P_X) X \quad (l = (I_n - P_X) m)$

$E[l^T Y] = m^T (I_n - P_X) X \beta$

$= 0^T \beta$

Theorem :
A linear function is BLUE of its expectation iff it is uncorrelated with all LZF.

given $E(l^T Y) = b^T \beta \Rightarrow l^T X \beta = b^T \beta$

$l^T Y = l^T I_n Y = l^T (P_X + I_n - P_X) Y$

$= (l^T P_X Y) + l^T (I_n - P_X) Y$ (LZF)

$E(l^T P_X Y) = l^T P_X X \beta = l^T X \beta = b^T \beta$

$\text{cov}(l^T P_X Y, l^T (I_n - P_X) Y)$

$= \sigma^2 (l^T P_X (I_n - P_X) l)$

$= \sigma^2 (l^T 0 l)$

$= 0.$

BLUE $(l^T P_X Y)$

See the prof's slides for recommended books.

Coefficient of determination

a) R^2 = Variation in Y explained / Total variation in Y

$$= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \Big/ \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$= \frac{SS\,Model}{SS\,Total} = 1 - \frac{SS\,Error}{SS\,Total}$$

$$= 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

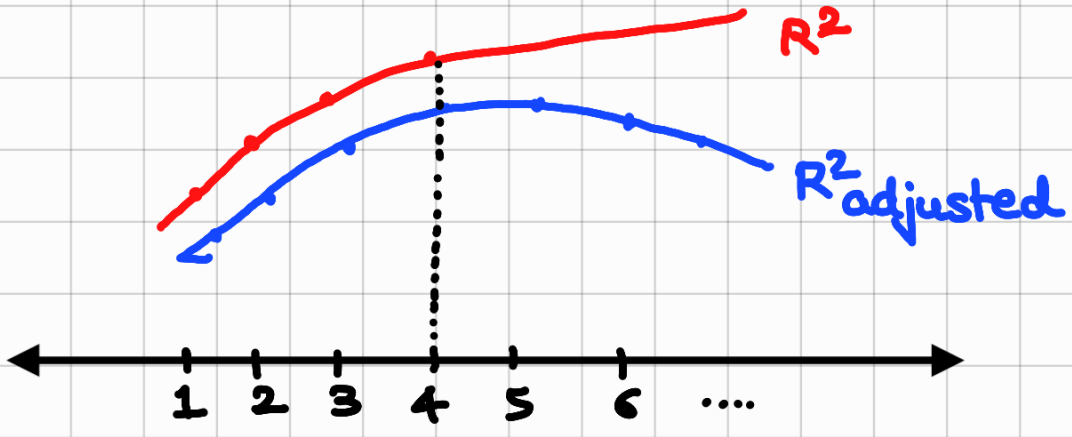
→ $R^2 \in (0,1)$

→ increasing fn. of no. of variables

b) Adjusted R^2:

$$R^2_{adjusted} = 1 - \frac{\sum_{i=1}^{n} e^2/df}{\sum_{i=1}^{n}(y_i - \bar{y})^2/df}$$

$$= 1 - \frac{\sum_{i=1}^{n} e_i^2/(n-k-1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/n-1} < R^2$$



→ find the maxima using local search heuristics.

{ even the courses of both the facads go hand-in-hand :) }

Polynomial Regression

→ analogous to multiple linear regression

$$Y = X\beta + \epsilon$$

$$X = (x_1, x_2, \ldots, x_n)^T \text{ with } x_j = (1, x_i, x_i^2, \ldots, x_i^k)^T$$

$$\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$$

k+2 unknown model parameters : β and $\sigma^2 > 0$.

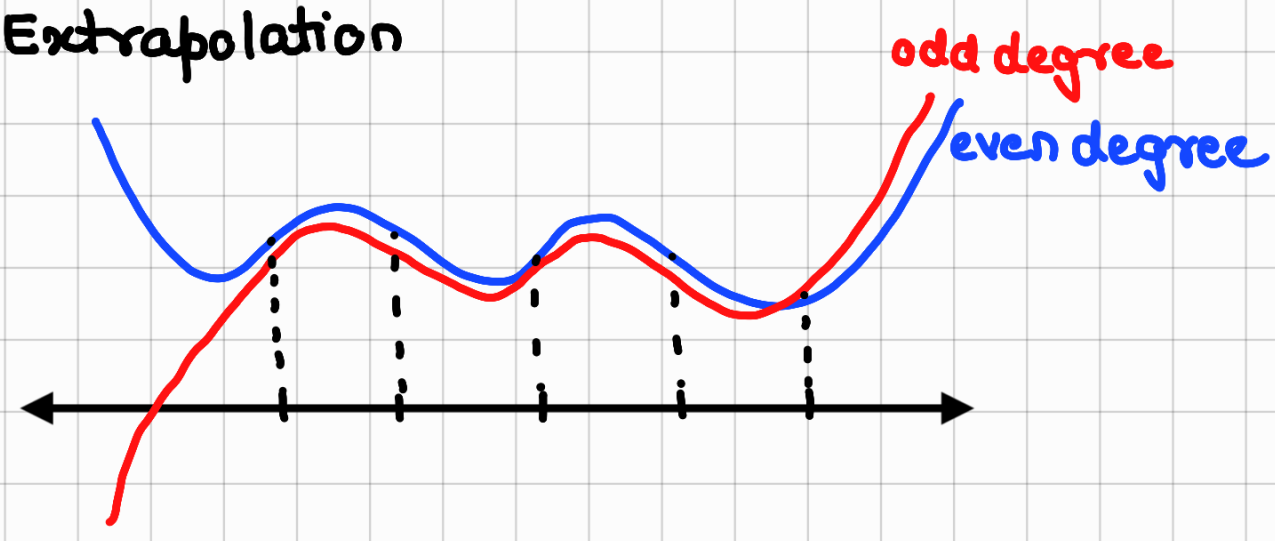
$$Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$$

Problems of polynomial regression

→ Finding order of the model

- forward selection
  - → continue fitting the order and computed adjusted R^2 values.
  - → find $\text{argmax}(R^2_{adj})$
- backward elimination

→ Extrapolation



→ Ill-conditioning

X^TX may be computationally singular

→ Hierarchy

not necessary that all degrees till k (1,2,...,k) should be present to fit the model.

regression model won't do so.

Orthogonal Polynomial

$u^T v = \sum_{i=1}^{n} u_i v_i = 0$ orthogonal vectors

$\langle u(t), v(t) \rangle = 0$ orthogonal functions

e.g. $\int_0^1 u(t) v(t) dt = 0$

given data $\{(x_i, y_i), i = 1, 2, \ldots, n\}$

$P_0(x_i) = 1 \quad \forall i = 1, \ldots, n$

$$\sum_{i=1}^{n} P_j(x_i) P_k(x_i) = \underset{\sim}{P}_j^T \underset{\sim}{P}_k = 0 \text{ for all } j \neq k.$$

$$\underset{\sim}{P}_j = \begin{pmatrix} P_j(x_1) \\ P_j(x_2) \\ \vdots \\ P_j(x_n) \end{pmatrix} \qquad \underset{\sim}{P}_k = \begin{pmatrix} P_k(x_1) \\ P_k(x_2) \\ \vdots \\ P_k(x_n) \end{pmatrix}$$

$P_j(x) = \sum_{m=0}^{j} \theta_m x^m$ is the jth degree orthogonal polynomial

To fit a k-degree orthpoly to the data

$$y_i = \sum_{j=0}^{k} \alpha_j (P_j(x_i)) + \epsilon_i$$

$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Unknown $\{(\alpha_0, \alpha_1, \ldots, \alpha_k), \sigma^2\}$

$$X_0 = \begin{pmatrix} P_0(x_1) & P_1(x_1) & \cdots & P_k(x_1) \\ \vdots & \vdots & & \vdots \\ P_0(x_n) & P_1(x_n) & \cdots & P_k(x_n) \end{pmatrix}$$

Hence,

$\underset{\sim}{y} = X_0 \underset{\sim}{\alpha} + \underset{\sim}{\epsilon} \quad \underset{\sim}{\epsilon} \sim N(0, \sigma^2 I_n) \Rightarrow \underset{\sim}{Y} \sim N(X_0 \underset{\sim}{\alpha}, \sigma^2 I_n)$

$$\hat{\underset{\sim}{\alpha}} = (X_0^T X_0)^{-1} X_0^T \underset{\sim}{Y}$$

$$(X_0^T X_0) = \begin{bmatrix} \sum_{i=1}^{n} P_0(x_i)^2 & & 0 \\ & \sum_{i=1}^{n} P_1(x_i)^2 & \\ 0 & & \ddots \\ & & \sum_{i=1}^{n} P_k(x_i)^2 \end{bmatrix}$$

$$\hat{\alpha_j} = \sum_{i=1}^{n} P_j(x_i) y_i \Big/ \sum_{i=1}^{n} P_j^2(x_i)$$

It is a linear estimator of $\hat{\alpha_j} = \underset{\sim}{P_j}^T \underset{\sim}{Y} \Big/ \underset{\sim}{P_j}^T \underset{\sim}{P_j} = \underset{\sim}{l}^T \underset{\sim}{Y}$

$$l_i = P_j(x_i) / (\underset{\sim}{P_j}^T \underset{\sim}{P_j})$$

$$\hat{\alpha_j} \sim \mathcal{N}\left(\frac{\underset{\sim}{P_j}^T [X_0 \alpha]}{\underset{\sim}{P_j}^T \underset{\sim}{P_j}}, \sigma^2 \left(\frac{\underset{\sim}{P_j}^T I_n \underset{\sim}{P_j}}{(\underset{\sim}{P_j}^T \underset{\sim}{P_j})^2} \right) \right)$$

$$\equiv \mathcal{N}\left(\alpha_j, \sigma^2 \Big/ \underset{\sim}{P_j}^T \underset{\sim}{P_j} \right)$$

SS Error

$$SSE = \underset{\sim}{Y}(I_n - P_{X_0}) \underset{\sim}{Y} \quad ; \quad P_{X_0} = X_0 (X_0^T X_0)^{-1} X_0^T$$

$$= \underset{\sim}{Y}^T \underset{\sim}{Y} - \underset{\sim}{Y}^T P_{X_0} \underset{\sim}{Y}$$

$$= \sum_{i=1}^{n} y_i^2 - \underset{\sim}{Y}^T X_0 \hat{\underset{\sim}{\alpha}}$$

$$\cdots \left(P_{X_0} Y = X_0 (X_0^T X_0)^{-1} X_0^T Y = X_0 \hat{\underset{\sim}{\alpha}} \right)$$

$$= \sum_{i=1}^{n} y_i^2 - \sum_{j=0}^{k} \hat{\alpha_j} \left(\sum_{i=1}^{n} P_j(x_i) y_i \right)$$

$$= \sum_{i=1}^{n} y_i^2 - \left(\hat{\alpha}_0 \sum_{i=1}^{n} P_0(x_i) y_i \right) - \sum_{j=1}^{k} \hat{\alpha}_j \left(\sum_{i=1}^{n} P_j(x_i) y_i \right)$$

$$\left(\because \sum_{i=1}^{n} P_0(x_i) y_i = n\bar{y} \right)$$

$$SSE = \sum y_i^2 - n\bar{y}^2 - \sum_{j=1}^{k} \hat{\alpha}_j \left(\sum_{i=1}^{n} P_j(x_i) y_i \right)$$

$$= \sum_{i=1} (y_i - \bar{y})^2 - \sum_{j=1}^{k} \hat{\alpha}_j \left(\sum_{i=1}^{n} P_j(x_i) y_i \right)$$

$$SSE = SS_{Total} - SS_{Model}$$

Multicollinearity

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon} \; ; \; \underset{\sim}{\beta} \in \mathbb{R}^{k+1}$$

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

If $|X^T X| = 0$ or $|X^T X| \approx 0$ ← ill-conditioned

$$\hat{\underset{\sim}{\beta}} = \frac{Adj(X^T X)}{|X^T X|} X^T \underset{\sim}{Y}$$

$\Rightarrow$ $Var(\hat{\beta}_j)$ may be unbounded.

$\hat{y}_0 = (1 . \underset{\sim}{x_0}^T)\hat{\beta}$ will also have large variance.

Why should this problem arise at all?

→ one regressor is a linear combination of the rest.

→ improper collection of the data may lead to multicollinearity.

→ Model building depending on the magnitude

of the regressor.

Consider the expected sq. error for $\underset{\sim}{\beta}$.

$$E\{(\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta})^T(\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta})\}$$
$$= E\sum_{j=0}^{k}(\hat{\beta_j}-\beta_j)^2$$
$$= \sum_{j=0}^{k} E(\hat{\beta_j}-\beta_j)^2 = \sum_{j=0}^{k} \mathrm{Var}(\hat{\beta_j})$$
$$= \mathrm{tr}\{\sigma^2(X^TX)^{-1}\}$$

$$\left[\because \hat{\underset{\sim}{\beta}} \sim N(\underset{\sim}{\beta}, \sigma^2(X^TX)^{-1})\right]$$

Note.

1) X^TX is a symmetric matrix.
2) $(X^TX)^{-1}$ is also a symmetric matrix, if exists.
3) X^TX is positive semidefinite matrix.
4) X^TX has non-negative evs.
   $\lambda_0 > \lambda_1 > \lambda_2 > \ldots \geq 0$.
5) $$\lambda_{max} = \max_{\|z\|\neq 0} \frac{\underset{\sim}{z}^T(X^TX)\underset{\sim}{z}}{\underset{\sim}{z}^T\underset{\sim}{z}}$$
   $$\lambda_{min} = \min_{\|z\|\neq 0} \frac{\underset{\sim}{z}^T(X^TX)\underset{\sim}{z}}{\underset{\sim}{z}^T\underset{\sim}{z}}$$
6) $\mathrm{tr}(X^TX) = \sum \lambda_i$
7) $\mathrm{tr}((X^TX)^{-1}) = \sum 1/\lambda_i$
8) $|X^TX| = \prod_i \lambda_i$
9) $|(X^TX)^{-1}| = \prod_i 1/\lambda_i$ $\uparrow \infty$ as some $\lambda_i \downarrow 0$

Hence,

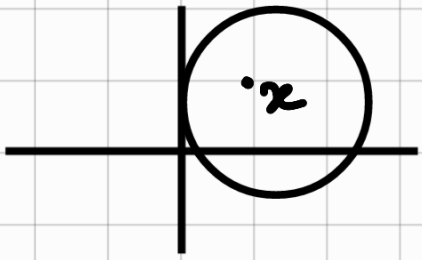
$$E\{(\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta})^T(\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta})\} = \mathrm{tr}\{\sigma^2(X^TX)^{-1}\}$$
$$= \sigma^2 \sum_i (1/\lambda_i)$$

$$\hat{\underset{\sim}{\beta}} \sim N(\underset{\sim}{\beta}, \sigma^2(X^TX)^{-1})$$

Possible Solutions

→ Principal Component Regression

→ restrict variance in euclidean norm/ℓ^2-norm



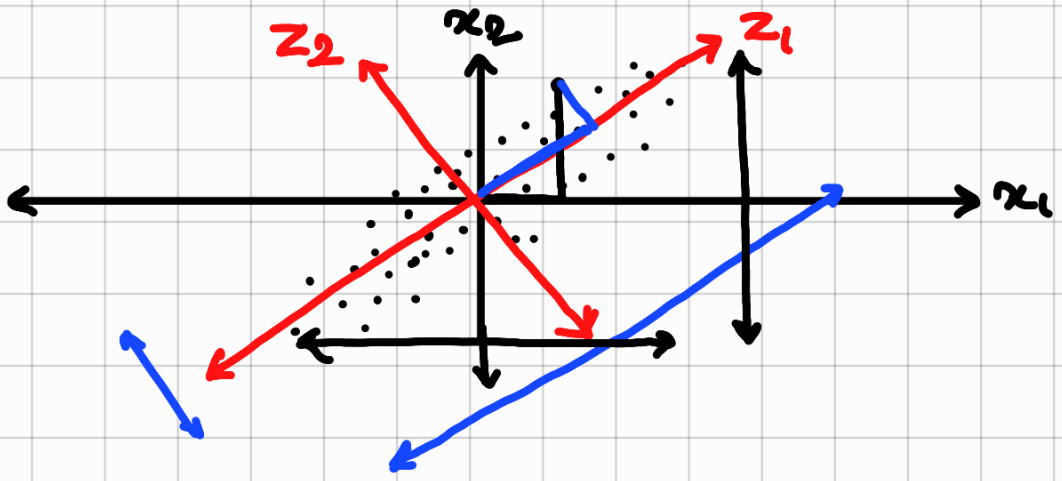
$L_1 \rightarrow$ LASSO

$L_2 \rightarrow$ RIDGE

→ Principal Component Regression

SVD — Math

PCA — Statistics



→ one method of dimension reduction

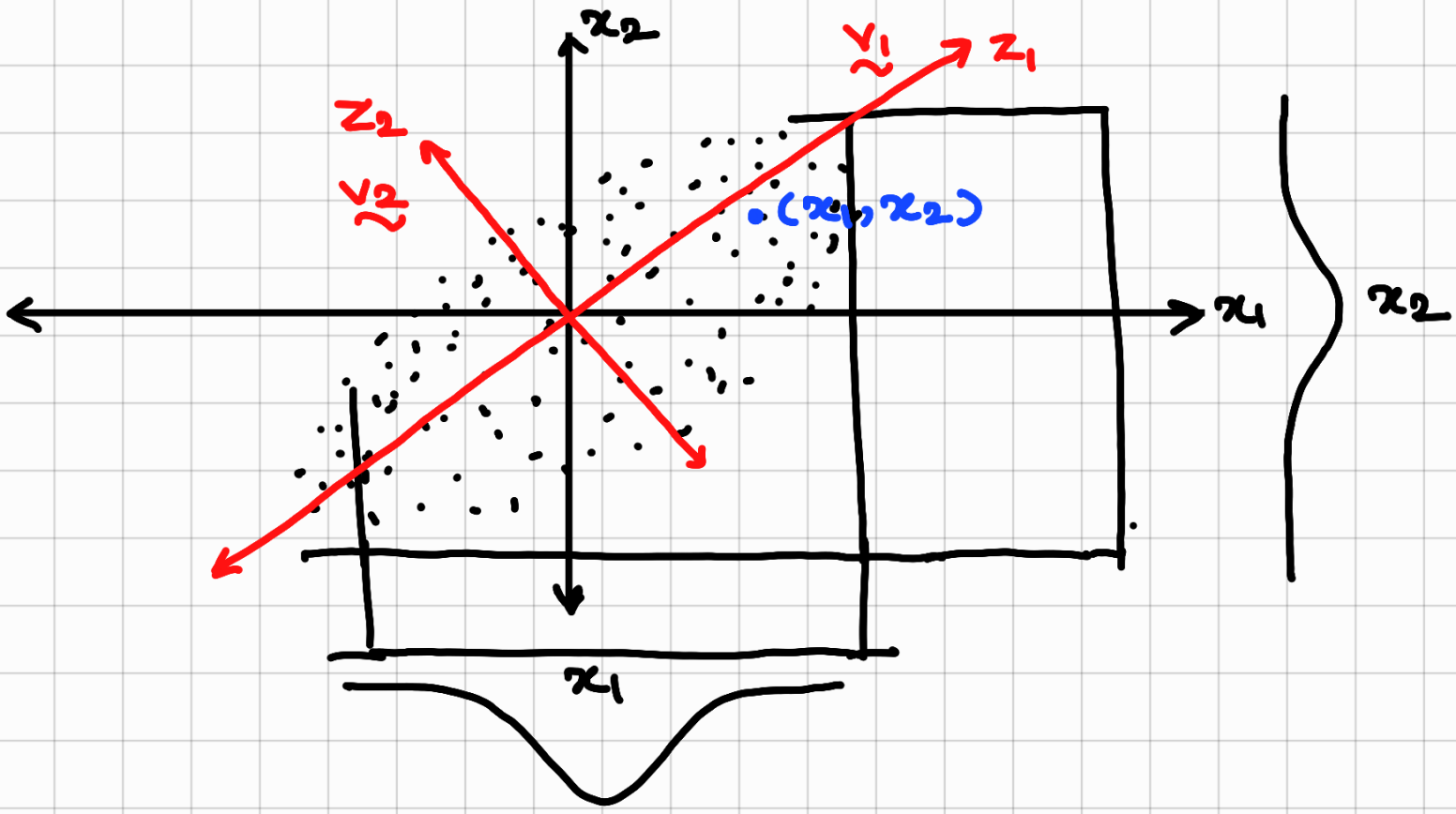
Tentatively, Midsem syllabus is upto and including Polynomial Regression.

Principal Component Regression

$|X^TX| \approx 0$

X^TX p.s.d

$\lambda_0 > \lambda_1 > \lambda_2 \dots \lambda_k \geq 0$



$$e_1 = \begin{pmatrix}1\\0\end{pmatrix}, e_2 = \begin{pmatrix}0\\1\end{pmatrix} \qquad \begin{pmatrix}x_1\\x_2\end{pmatrix} = x_1 \cdot \begin{pmatrix}1\\0\end{pmatrix} + x_2 \cdot \begin{pmatrix}0\\1\end{pmatrix}$$

$$\langle \underset{\sim}{x}, \underset{\sim}{e}_1 \rangle \cdot \begin{pmatrix}1\\0\end{pmatrix} + \langle \underset{\sim}{x}_2, \underset{\sim}{e}_2 \rangle \cdot \begin{pmatrix}0\\1\end{pmatrix} = \underset{\sim}{x}$$

$$\begin{pmatrix}x_1\\x_2\end{pmatrix} = \langle x, \underset{\sim}{v}_1 \rangle \underset{\sim}{v}_1 + \langle x, \underset{\sim}{v}_2 \rangle \underset{\sim}{v}_2$$
$$= z_1 \underset{\sim}{v}_1 + z_2 \underset{\sim}{v}_2$$

$$\lambda_0 > \lambda_1 > \lambda_2 \ldots\ldots > \lambda_k \geq 0$$
$$\begin{bmatrix} \underset{\sim}{v}_0 & \underset{\sim}{v}_1 & \underset{\sim}{v}_2 & \ldots\ldots & \underset{\sim}{v}_k \end{bmatrix} = P$$

$X^T X \longrightarrow (\lambda_i, \underset{\sim}{v}_j)_{i=0,1,2,\ldots,k}$

$X^T X = P D P^T$

D: diagonal matrix

P: orthogonal matrix

$P^T P = P P^T = I$.

Example of orthogonal matrix: $\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}$

orth. matrix for reflection →

$$Y = X\beta + \epsilon$$

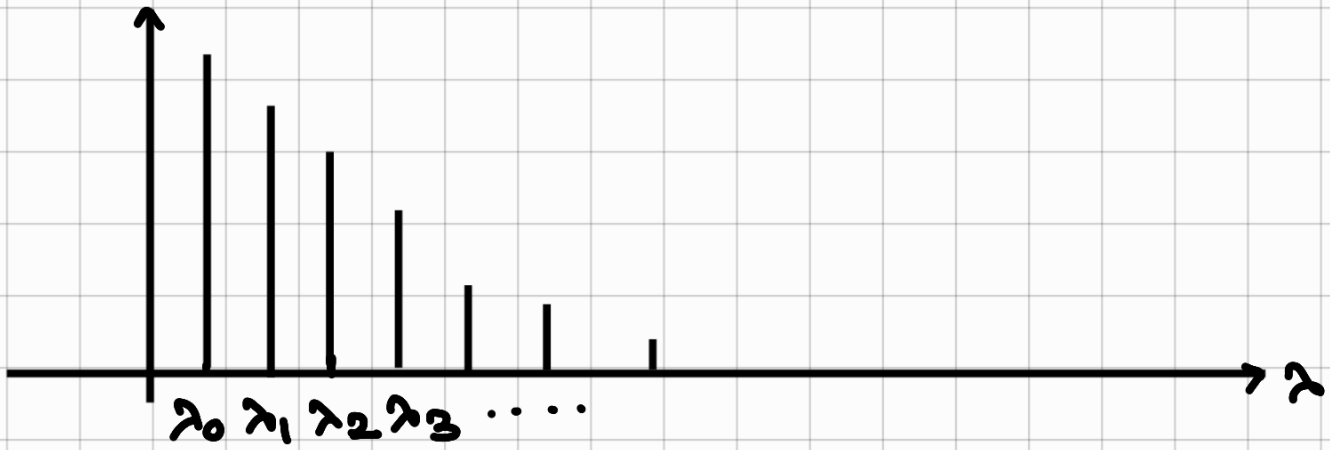
$$Y = XPP^T\beta + \epsilon \Longrightarrow Y = Z\alpha + \epsilon$$

$$\Rightarrow \hat{\alpha} = (Z^TZ)^{-1}Z^TY$$

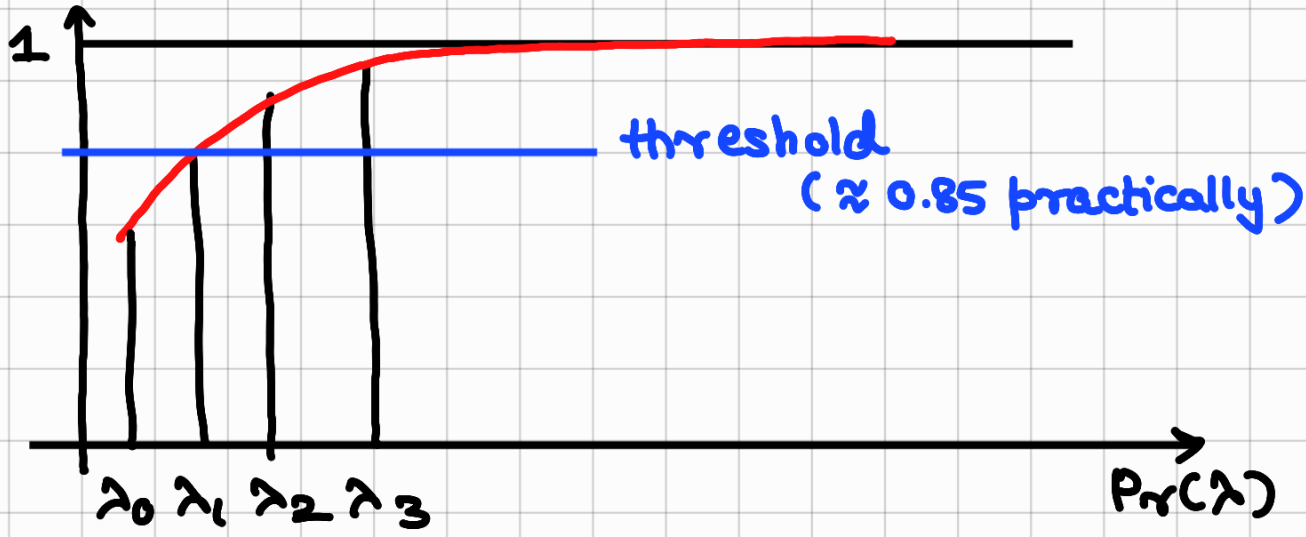
As a process,

$$\alpha_{(r)} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_r \end{pmatrix}$$

(reducing the dimensions)



$$P_r(\lambda) = \sum_{i=0}^{t} \lambda_i \Big/ \sum_{i=0}^{k} \lambda_i$$



$Z_{(r)}$ is the generated matrix from X and

$[V_0\ V_1\ \dots\ V_r]$

$$Z_{(r)} = \underset{n\times(k+1)}{X}\ [V_0\ V_1\ \dots\ V_r]$$

$$\hat{\underset{\sim}{\alpha}}_{(r)} = (Z_{(r)}^T Z_{(r)})^{-1} Z_{(r)}^T Y$$

Transform $\hat{\alpha}_{(r)}$ to β,

$\underset{\sim}{\beta} = I\underset{\sim}{\beta}$

$\Rightarrow \underset{\sim}{\beta} = PP^T\underset{\sim}{\beta}$

$\Rightarrow \underset{\sim}{\beta} = P.\underset{\sim}{\alpha}$

$$\Rightarrow \underset{(k+1)}{\hat{\underset{\sim}{\beta}}_{PC}} = P.\begin{pmatrix} \hat{\alpha}_{(r)} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Tentatively, Midsem syllabus is upto and including Polynomial Regression.

$|X^TX| \approx 0$

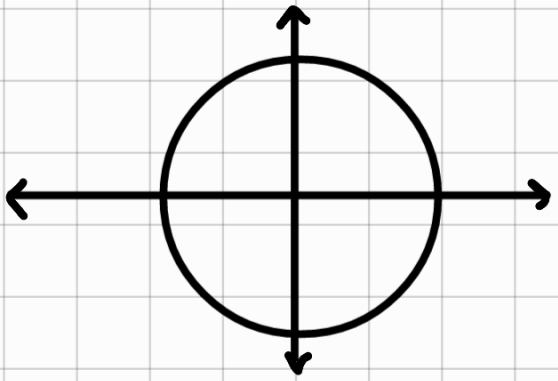
$\longrightarrow$ PCR

$\longrightarrow$ Shrinkage estimation or regularization

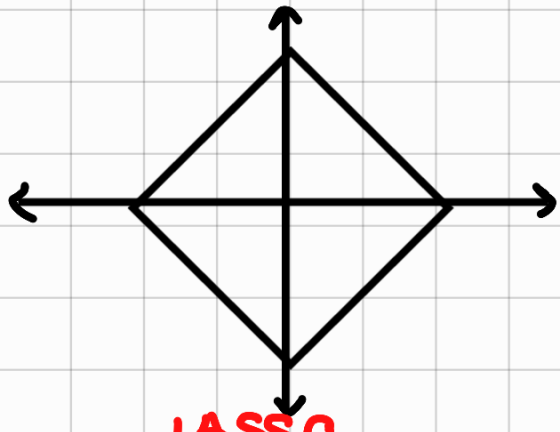
Shrinkage estimation / regularization

As $|X^TX| \approx 0$, the variance of $\hat{\underset{\sim}{\beta}}_{LS}$ is large. Hence, we want some bound on the norm of $\hat{\beta}$.

L_2 : $\|\hat{\underset{\sim}{\beta}}\|_2^2 \leq c_2$ $\Longleftrightarrow$ $\sum_{i=0}^{k} |\beta_i|^2 \leq c_2$ or L_1 : $\|\hat{\underset{\sim}{\beta}}\|_1 \leq c_1$ $\Longleftrightarrow$ $\sum_{i=0}^{k} |\beta_i| < c_1$



RIDGE



LASSO

Minimization of LS condition

$S(\beta) = (y - X\beta)^T(y - X\beta)$ w.r.t. $\sum_{i=0}^{k} |\beta_i|^2 < c_2$

Using Lagrangian multiplier

$$\hat{\beta}_R = \text{argmin}\ (S(\beta) + \lambda(\beta^T\beta - c_2))$$

$$\hat{\beta}_R = (X^TX + \lambda I_{k+1})^{-1} X^Ty$$

Hoerl, Kennard, Baldon (1975)

$$\lambda = (k+1)\hat{\sigma}_{LS} / \hat{\beta}_{LS}^T \hat{\beta}_{LS}$$

These $\hat{\sigma}_{LS}$, $\hat{\beta}_{LS}$ are generated from original model with/without using PCR depending on $(X^TX)^{-1}$.

Note :

1) $\lambda \to \infty \Rightarrow \hat{\beta}_R = 0$
2) $\lambda \to 0 \Rightarrow \hat{\beta}_R = \hat{\beta}_{LS}$

Is $\hat{\beta}_R$ an unbiased estimator of β?

$$\hat{\beta}_R = (X^TX + \lambda I_{k+1})^{-1}(X^TX)(X^TX)^{-1}X^Ty$$
$$= \underbrace{[(X^TX + \lambda I_{k+1})^{-1}(X^TX)]}_{\text{not an identity matrix}} \hat{\beta}_{LS}$$

$\hat{\beta}_R$ is not an unbiased estimator.

Notations :

$W = (X^TX + \lambda I)^{-1}$

$S = (X^TX)$

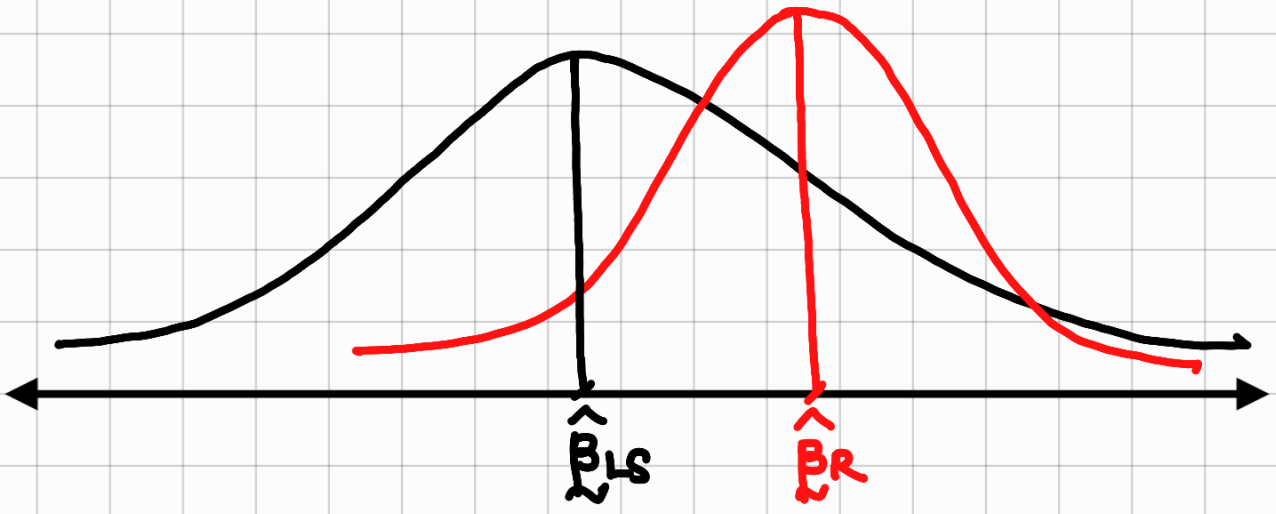
$\therefore \hat{\beta}_R = WS\hat{\beta}_{LS}$

Does RIDGE estimate have less variance?

$$D(\hat{\beta}_{LS}) - D(\hat{\beta}_R) = \sigma^2(S^{-1} - WSS^{-1}SW)$$
$$= \sigma^2(S^{-1} - WSW)$$
$$= \sigma^2 W(W^{-1}S^{-1}W^{-1} - S)W$$
$$= \sigma^2 W((S+\lambda I)S^{-1}(S+\lambda I) - S)W$$
$$= \sigma^2 W(2\lambda I + \lambda^2 S^{-1})W$$

psd psd psd

Hence, $D(\hat{\beta}_{LS}) - D(\hat{\beta}_R)$ is a psd matrix.
This implies RIDGE estimate has less variance.

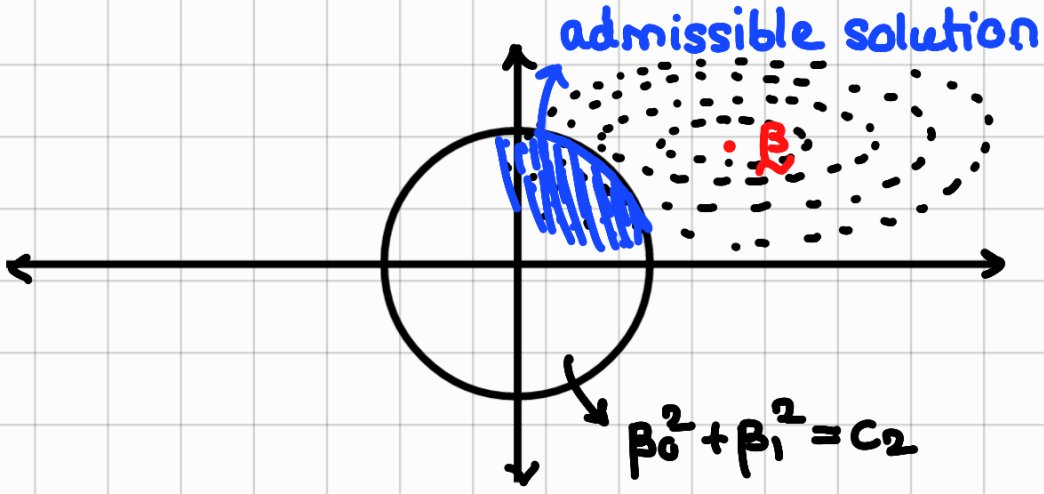


Bias $E(\hat{\beta}_R) - \hat{\beta} = (WS - I)\beta$

$$MSE(\hat{\beta}_R) = tr[D(\hat{\beta}_R)] + \beta^T W^2 \beta \lambda^2$$
$$= \sigma^2 \sum_{i=0}^{k} \lambda_i/(\lambda_i+\lambda)^2 + \lambda^2 \sum_{i=0}^{k} \beta_i^2/(\lambda_i+\lambda)^2$$
$$= \sigma^2 \sum_{i=0}^{k} \lambda_i/(\lambda_i+\lambda)^2 + \sum_{i=0}^{k} \beta_i^2/(1+\lambda_i/\lambda)^2$$

λ_i's are eigenvalues of X^TX.

Solution space for RIDGE Geometrical Intuition



Similar steps for LASSO (differentiation is a bit complex)

Variable selection

$$Y = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$$
$$= [X_p | X_r] \begin{pmatrix} \underset{\sim}{\beta_p} \\ \hline \underset{\sim}{\beta_r} \end{pmatrix} + \underset{\sim}{\epsilon}$$

$(p+r=k+1)$

Data we received : (Y, X_p)
We need r more features to fit data into true model

True Model $Y = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$
Reduced Model $Y = X_p\underset{\sim}{\beta_p} + \underset{\sim}{\epsilon}$ reduced according to available data

Q1. Is $\hat{\underset{\sim}{\beta_p}}$ an unbiased estimator?
Q2. Does $\hat{\underset{\sim}{\beta_p}}$ have less variation compared to the true one?
Q3. Is σ_p^2 (estimate value of σ^2 in reduced model) an unbiased estimator of σ^2?

Note: $\Sigma = \left(\begin{array}{c|c} \Sigma_{pp} & \Sigma_{rp} \\ \hline \Sigma_{pr} & \Sigma_{rr} \end{array}\right)$

reduced model, Σ_{pp} is to be used.

In reduced model,

$$\hat{\underset{\sim}{\beta}}_p = (X_p^T X_p)^{-1} X_p^T \underset{\sim}{Y}$$

$$\begin{aligned} E(\hat{\underset{\sim}{\beta}}_p) &= (X_p^T X_p)^{-1} X_p^T E(\underset{\sim}{Y}) \\ &= (X_p^T X_p)^{-1} X_p^T [X_p \underset{\sim}{\beta}_p + X_r \underset{\sim}{\beta}_r] \\ &= \underset{\sim}{\beta}_p + (X_p^T X_p)^{-1}(X_p^T X_r)\underset{\sim}{\beta}_r \end{aligned}$$

Hence, $\hat{\underset{\sim}{\beta}}_p$ is an unbiased estimator if $X_p^T X_r = 0$
This would happen in case of PCA and orthogonal polynomials.

A1. Not true in general

$$\sigma_p^2 = \frac{\underset{\sim}{Y}^T (I_n - X_p (X_p^T X_p)^{-1} X_p^T)\underset{\sim}{Y}}{n-p}$$

$$\begin{aligned} E(\sigma_p^2) &= [\sigma^2(n-p) + \sigma^2(ncp)]/(n-p) \\ &= \sigma^2 + \sigma^2(ncp)/(n-p) \end{aligned}$$

$$ncp = \frac{1}{\sigma^2}(X\beta)^T (I_n - X_p(X_p^T X_p)^{-1} X_p^T) X\underset{\sim}{\beta}$$

$$= 0+0+0+\frac{1}{\sigma^2}(\underset{\sim}{\beta}_r^T X_r^T (I_n - P_{X_p}) X_r \underset{\sim}{\beta}_r)$$

{ expand $X\underset{\sim}{\beta} = X_p\underset{\sim}{\beta}_p + X_r\underset{\sim}{\beta}_r$
all X_p terms get cancelled }

$$ncp = \frac{1}{\sigma^2}(\underset{\sim}{\beta}_r^T X_r^T (I_n - P_{X_p}) X_r \underset{\sim}{\beta}_r)$$

A3. No (it is unbiased only when we capture all features)

Note: $\Sigma = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right)$. Then,

$\Sigma^{-1} = (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}$, and

$(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$

given in prof's slides

Project (Credits : Anubhab Mandal)

20/30 marks

Group size → 1, 5 or 6

Report → 5/6 pages ①

+ code ② (analysis of dataset using things taught in class)

Justification of maths + dataset description

+ PDF of ③ executed code (Python or R)

30 marks → Video recording ④ of project (10 mins)

Upload all the stuff in a drive folder

All group members must submit project folder by the same name so that checked together

Submission → 15th April 2024

Generalized Linear Model

In linear model : $y_i = \underset{\sim}{\beta}^T \underset{\sim}{x}_i + \epsilon_i$, $i = 1, 2, \ldots, n$

$E(\epsilon_i) = 0 \quad V(\epsilon_i) = \sigma^2$

$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

$E(y_i | \underset{\sim}{x}_i) = \underset{\sim}{\beta}^T \underset{\sim}{x}_i = I(\underset{\sim}{\beta}^T \underset{\sim}{x}_i)$ ($I(x) = x$.)

For categorical variables as output, we need

$E(y_i | \underset{\sim}{x}_i) = g(\underset{\sim}{\beta}^T \underset{\sim}{x}_i)$

$\underset{\sim}{x}_i \in \mathbb{R}^{k+1} \xrightarrow{\text{linear combination}} \underset{\sim}{x}_i^T \underset{\sim}{\beta} \in \mathbb{R} \xrightarrow{g} (0,1) \longrightarrow \{0,1\}$ or $\{A, B, C, D\}$

GLM — classification problem

$$g(x) = \begin{cases} \dfrac{e^z}{1+e^z} & z \in \mathbb{R} \quad \text{logit-model} \\ \text{any cdf of a cont. random variable} \\ \Phi(z) & z \in \mathbb{R} \quad \text{probit-model} \end{cases}$$

Logistic Regression

$$P(y_i = 1) = 1 - P(y_i = 0) = \pi_i = E(y_i | \underset{\sim}{x}_i) = \frac{e^{\underset{\sim}{x}_i^T \underset{\sim}{\beta}}}{1 + e^{\underset{\sim}{x}_i^T \underset{\sim}{\beta}}}$$

$$\iff \log_e\left(\frac{\pi_i}{1-\pi_i}\right) = \underset{\sim}{x}_i^T \underset{\sim}{\beta} \in \mathbb{R}$$

$\pi_i/(1-\pi_i)$ is known as 'odd' of π_i

Two treatment problem; often defined as $\log_e\left(\frac{p_A/1-p_A}{p_B/1-p_B}\right)$

→ log of odds ratio.

≈ asymptotically follow normal distribution.

$$\psi = \log_e\left(\frac{p_A/1-p_A}{p_B/1-p_B}\right) \; ; \quad \hat{\psi} \sim N(0, \sigma_y^2) \text{ for large } n.$$

Let $y \overset{iid}{\sim}$ Bernoulli(p)

$f(y) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$ joint pdf

$$= p^{\sum_i y_i}(1-p)^{n-\sum_i y_i} = \left(\frac{p}{1-p}\right)^{\sum_i y_i}(1-p)^n$$

$$\log_e f(y) = \left(\sum_{i=1}^{n} y_i\right)\left(\log_e\left(p/1-p\right)\right) + n\log(1-p) + 0$$

$$= D_1(y)\,D_2(p) + D_3(p) + D_4(y)$$

For any distribution (almost any), we are able to decompose $\log_e f(y)$ in this form.

$D_1(y)$ → sufficient statistic (all info. of data)

$D_2(p)$ → natural parameter structure

$D_3(p)$ → efficient parameter estimation from data

$D_4(y)$ → carries certain info. about ancillary part depending on family of distribution. does not depend upon parameter

Due to this natural structure, logistic model is preferred over other models.

$$\left(\begin{array}{l} \theta = \log p/1-p \\ p = e^{\theta}/1+e^{\theta} \end{array}\right)$$

Now,

$$E(y_i|x_i) = \pi_i = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}$$

Likelihood fn. of (β)

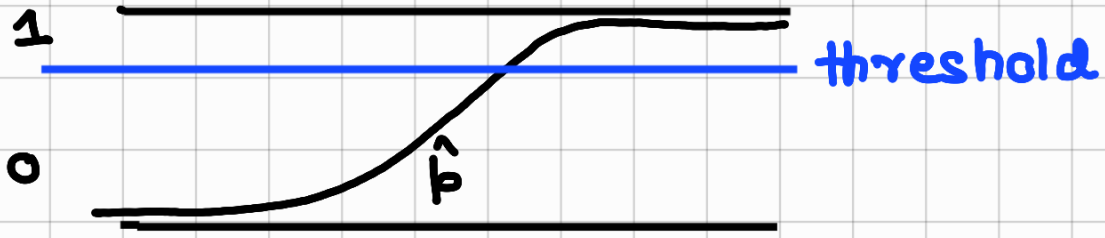
$$L(\beta|data) = \pi_{i=1}^{n} \left\{ \pi_i^{y_i} (1-\pi_i)^{1-y_i} \right\}$$

$$\log(L(\beta|y)) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) \right] + \sum_{i=1}^{n} \log(1-\pi_i)$$

$$L(\beta) = \sum_{i=1}^{n} [y_i (x_i^T \beta)] - \sum_{i=1}^{n} \log_e (1 + e^{x_i^T \beta})$$

$$\frac{\partial L(\beta)}{\partial \beta} = 0$$ we get $\hat{\beta}$ as an iterative solution

$$\hat{p} = \frac{e^{\hat{\beta} x}}{1 + e^{\hat{\beta} x}}$$



Error Analysis of MLR

$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ assumption

$y_i = x_i^T \beta + \epsilon_i$ data $\qquad H \equiv P_x$

$e_i = y_i - \hat{y}_i$ estimated error

$$e = (y - \hat{y}) \sim N(0, \sigma^2 (I_n - P_x))$$

$$\hat{\sigma}^2 = \frac{e^T e}{n-k-1}$$ unbiased estimate of σ^2

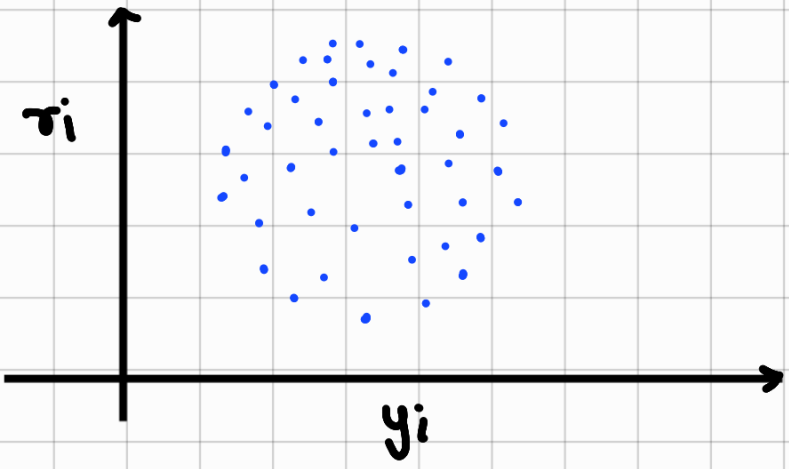
$$cov(e_i, e_j) = \begin{cases} \sigma^2 (1 - h_{ii}) & i = j \\ -\sigma^2 h_{ij} & i \neq j \end{cases}$$

$$MSError = \frac{SSError}{n-k-1}$$

Standardized Residual $d_i = e_i / \sqrt{MSError}$

Studentized Residual $r_i = e_i / \sqrt{MSError(1-h_{ii})}$

If all the analysis is correct, the pattern resembles:



Homework: Show that

$$r_i = \frac{e_i}{\left(MSError\left[1-\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right]\right)^{1/2}} \text{ for SLR.}$$

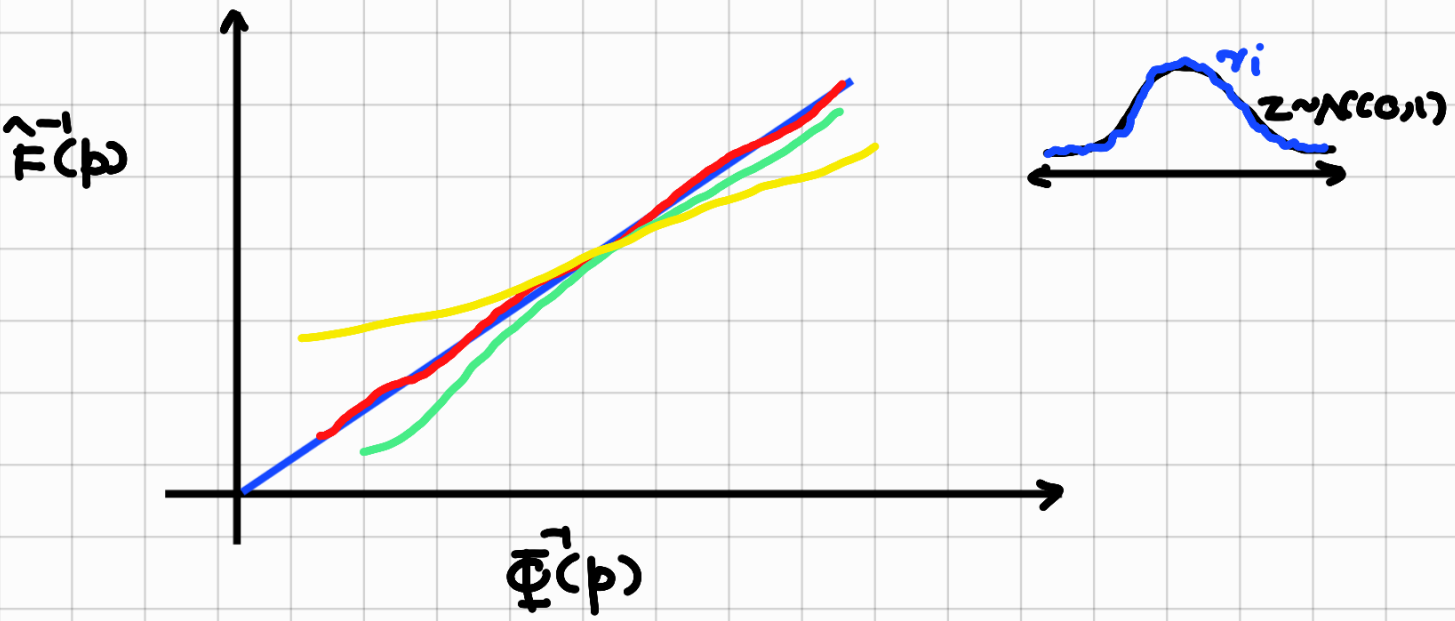
If $n \uparrow$ to ∞ and $k < \infty$, $\hat{\sigma}^2 = MSError$ will converge to σ^2 with probability 1.

$\Rightarrow r_i \overset{a}{\sim} N(0,1)$

To validate this,

- ⟶ Q-Q plot
- ⟶ χ^2 test
- ⟶ KS test

Q-Q Plot



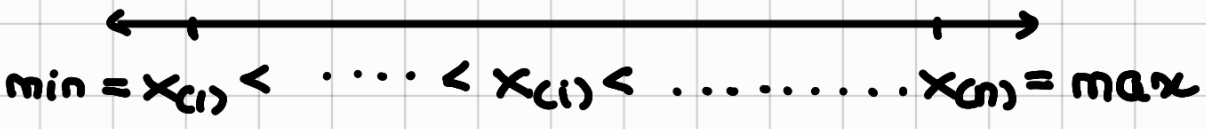
Tail part can most likely fluctuate. The middle part more or less follows the plot

$\Phi^{-1}(p)$ cdf of $N(0,1)$

$\hat{F}$ is the empirical cdf of plot

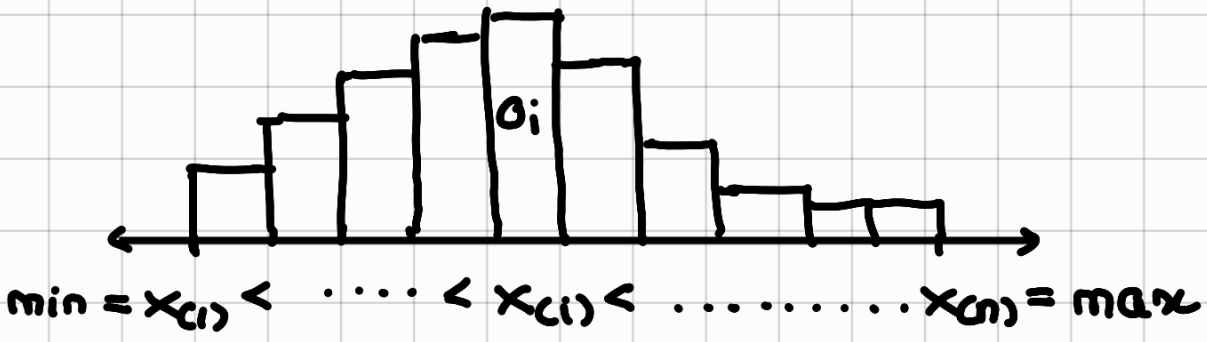
χ^2-test (goodness of fit)

pdf based test



→ arrange the data in an increasing order

→ divide the range $(x_{(1)}, x_{(n)})$ into k-parts, such that each subinterval has more than 5 data



→ Let there are O_i many observations in the ith interval

→ E_i = expected no. of observations within ith interval

$= n \times p_i$

= (total obs) × prob. of ith interval

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$$

df = $k-1$ when parameters are known

df gets decremented with increasing no. of unknown parameters.

If we use $\hat{E}_i = n \times \hat{p}_i$, $\hat{p}_i$ is the probability after estimating parameters from the data, then it will follow $\chi^2_{(k-1)-(\text{no. of parameters estimated})}$

KS Test

Assume that the target distribution specification is completely known.

$$Z_i \sim F(\text{completely known})$$

Z_i are continuous r.v.

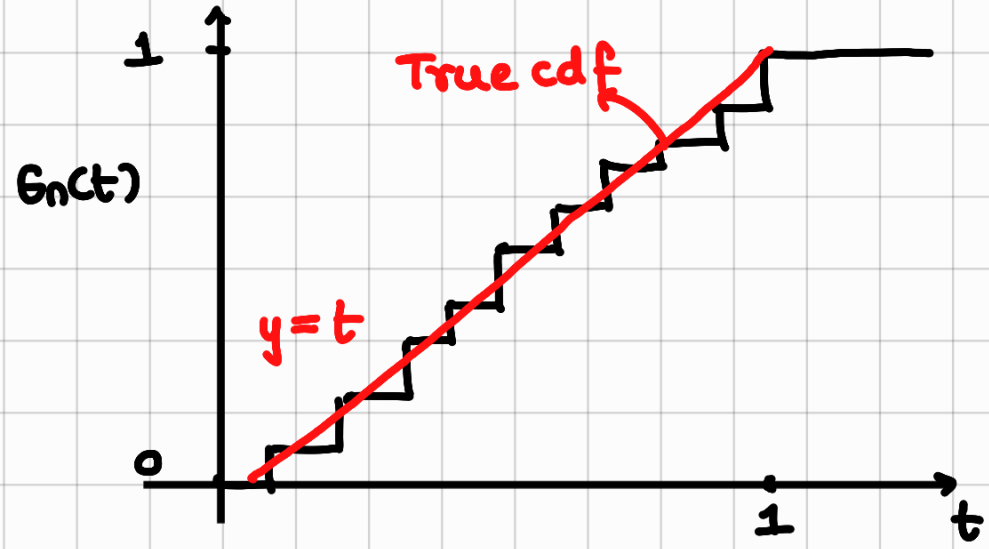
→ test only applicable to continuous r.v.

$$W_i = F(Z_i) \overset{iid}{\sim} U(0,1)$$

Empirical cdf of W is given as

$$G_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{w_i \leq t\}}$$

Plot of $G_n(t)$ v/s t looks like:



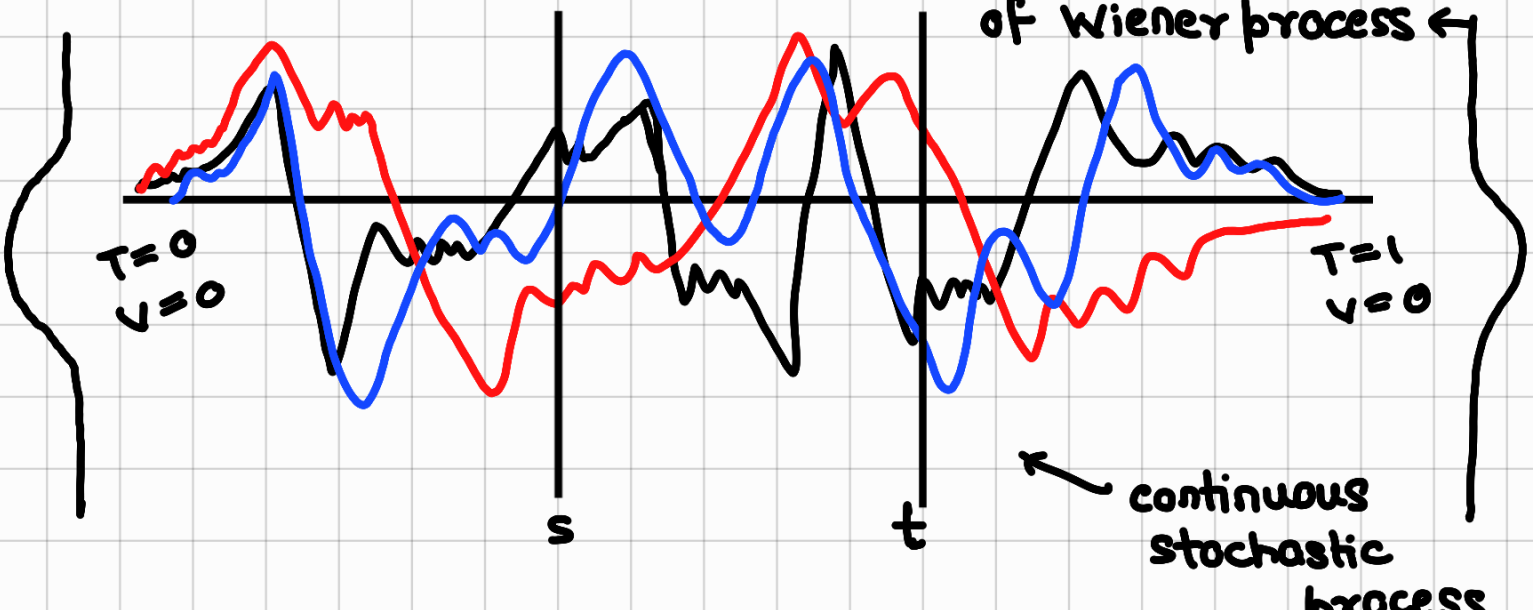
Test statistic: $\sup_{0 \le t \le 1} \sqrt{n}\,|G_n(t)-t| = D$

D is said to follow Kolmogrove distribution as $n \uparrow \infty$.

If $Z_i \sim F$ then $|G_n(t)-t| \to 0$ a.s. (almost surely)
{Measure Theory PTSD}

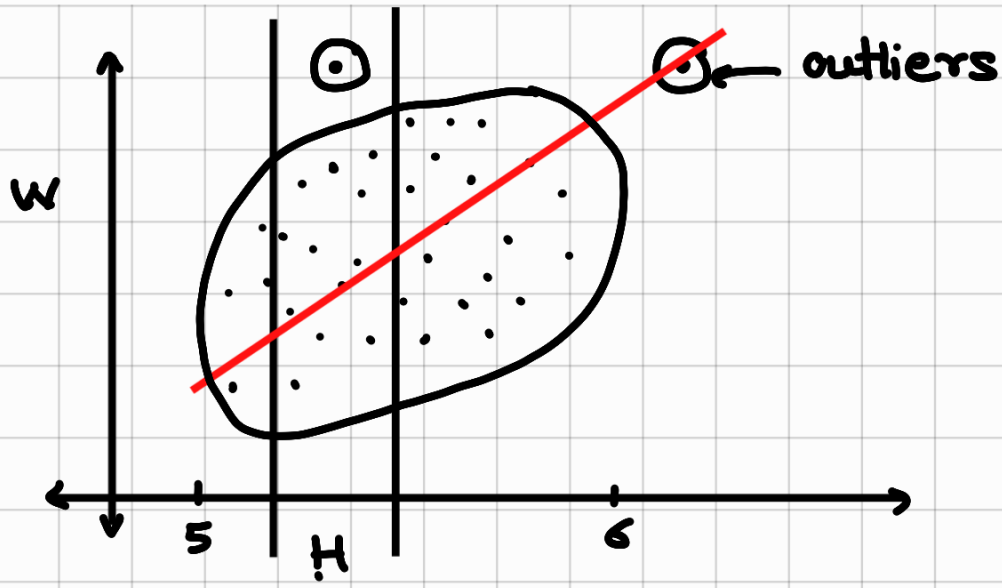
Kolmogorov Distribution:

Brownian Bridge:



$$\begin{pmatrix} B_0(s) \\ B_0(t) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} s(1-s) & s(1-t) \\ s(1-t) & t(1-t) \end{pmatrix} \right)$$

$K = \sup_{0 \le t \le 1} |B_0(t)|$ ← r.v. of Kolmogorov distn.



$e_i = y_i - \hat{y}_i$

$= y_i - \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}$

we used all $(y_i, \underset{\sim}{x}_i)$ to get $\hat{\underset{\sim}{\beta}}$.

$n-1$ { ~~$(y_1, \underset{\sim}{x}_1)$~~, $(y_2, \underset{\sim}{x}_2)$, ⋮, $(y_n, \underset{\sim}{x}_n)$ }

get $\hat{\underset{\sim}{\beta}}_{(1)}$

$\hat{y}_{(1)} = \underset{\sim}{x}_1^T \hat{\underset{\sim}{\beta}}_{(1)}$

$e_{(1)} = (y_1 - \hat{y}_{(1)}) = (y_1 - \underset{\sim}{x}_1^T \hat{\underset{\sim}{\beta}}_{(1)})$

$n-1$ { $(y_1, \underset{\sim}{x}_1)$, ~~$(y_2, \underset{\sim}{x}_2)$~~, ⋮, $(y_n, \underset{\sim}{x}_n)$ }

get $\hat{\underset{\sim}{\beta}}_{(2)}$

$\hat{y}_{(2)} = \underset{\sim}{x}_2^T \hat{\underset{\sim}{\beta}}_{(2)}$

$e_{(2)} = (y_2 - \hat{y}_{(2)}) = (y_2 - \underset{\sim}{x}_2^T \hat{\underset{\sim}{\beta}}_{(2)})$

Jack knife method

for i=1 to n

1) remove the observation
2) do the analysis for (n-1) data
3) Compute the estimated error for ith data

$$e_{(i)} = y_i - \hat{y}_{(i)} = y_i - \hat{\beta}_{(i)}^T x_i$$

Prediction Residual sum of square (PRESS)

$$\sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} (y_{(i)} - y_i)^2$$

It can be shown that $\frac{e_{(i)}}{\sqrt{Var(e_{(i)})}} = \frac{e_i}{\sqrt{V(e_i)}}$

Hence, fitting the whole n observations serves the purpose.

$$e_{(i)} = \frac{e_i}{1-h_{ii}} \sim N\left(0, \frac{\sigma^2}{1-h_{ii}}\right)$$

$$\frac{e_{(i)}}{\sqrt{Var\, e_{(i)}}} = \frac{e_i/(1-h_{ii})}{\sqrt{\sigma^2/(1-h_{ii})}} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}} = r_i$$

$\hat{\sigma}^2$ = MSError may not be a good estimator when the ith observation has been removed

We can use

$$\hat{\sigma}^2 = S_{(i)}^2 = \frac{(n-k-1)MSError - (e_i^2/1-h_{ii})}{n-k-2}$$

$$= \frac{\sum_{i=1}^{n} e_i^2 - e_i^2/1-h_{ii}}{n-k-2}$$

$$T = \frac{e_{(i)}}{S_{(i)}^2/(1-h_{ii})} \sim t_{n-k-2} \text{ under } H_0$$

H_0: ith observation is NOT an outlier

H_1: ith observation is an outlier

$$\underset{(k+1)\times(n-1)\ \ (n-1)\times(k+1)}{[X_{(i)}^T X_{(i)}]^{-1}} = [\underset{(k+1)\times n\ \ n\times(k+1)}{X^T X} - \underset{(k+1)\times 1\ \ 1\times(k+1)}{x_{(j)} x_{(i)}^T}]^{-1}$$

$$(A + \underset{\sim}{u}\underset{\sim}{v}^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + \underset{\sim}{v}^T A^{-1}\underset{\sim}{u}}$$

For related proofs, refer to Appendix C.7 and C.8 from the reference book { The prof is not accountable for the proof :) }

Regression syllabus is completed.

Time-Series

Relation b/w exponential and geometric distribution:

$X \sim \exp(\lambda) \longrightarrow$ continuous

$F(x) = 1 - e^{-\lambda x}$

Consider $Y = [X]$

$$\begin{aligned} P(Y = r) &= P([X] = r) \\ &= P(r \le X < r+1) \\ &= P(X < r+1) - P(X < r) \\ &= (1 - e^{-\lambda(r+1)}) - (1 - e^{-\lambda r}) \\ &= (e^{-\lambda})^r (1 - e^{-\lambda}) \\ &\sim q^r p \longleftarrow \text{geometric}(p = 1 - e^{-\lambda}) \end{aligned}$$

$\longrightarrow$ discrete

This implies that the adjectives of time (discrete, continuous) depend upon the context of approximation of time, and not the data.

Hence, discrete/continuous are adjectives of time and not the data.

Example of time-series

eg 1 White Noise

A time series $\{W_t\}$ is said to follow white noise if

- $E(W_t) = 0$
- $V(W_t) = \sigma_\omega^2$
- and they are UNCORRELATED

$W_t \sim WN(0, \sigma_\omega^2)$ (WN stands for white noise)

- $x_i \overset{iid}{\sim} N(0, \sigma^2)$
- $x_i \sim \begin{cases} N(0,1) & \text{when } i \text{ is even} \\ \exp(1) - 1 & \text{when } i \text{ is odd} \end{cases}$

Homework: Construct uncorrelated but dependent WN for both the cases.

(i) WN need not be normally distributed

(ii) WN need not be iid

(iii) iid sequence with zero mean and finite variance are always WN.

(iv) WN is weakly stationary.

(v) WN with normal distn. are strongly stationary.

eg.2 Binary process

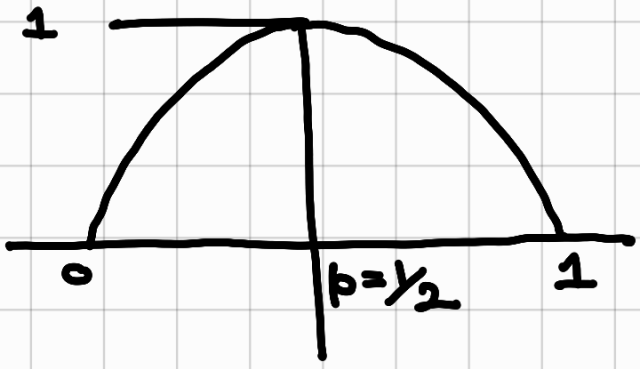
$$X_t = \begin{cases} +1 & \text{with prob. } p = 1/2 \\ -1 & \text{with prob. } 1-p = 1/2 \end{cases}$$

$E(X_t) = 2p - 1$

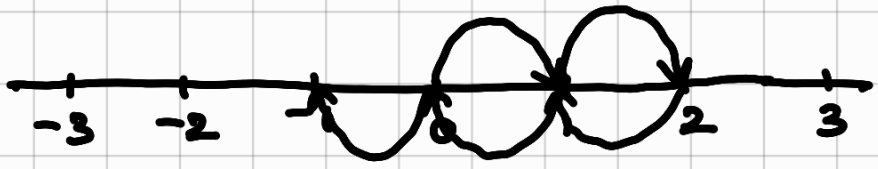
$Var(X_t) = 4p(1-p)$

$Z_t \overset{iid}{\sim} \text{Bernoulli}(p)$

$X_t = 2Z_t - 1$



e.g 3 Random walk on $\mathbb{Z}$



$x_0 = 0$

$x_t = x_0 + \sum_{t=1}^{T} W_t$

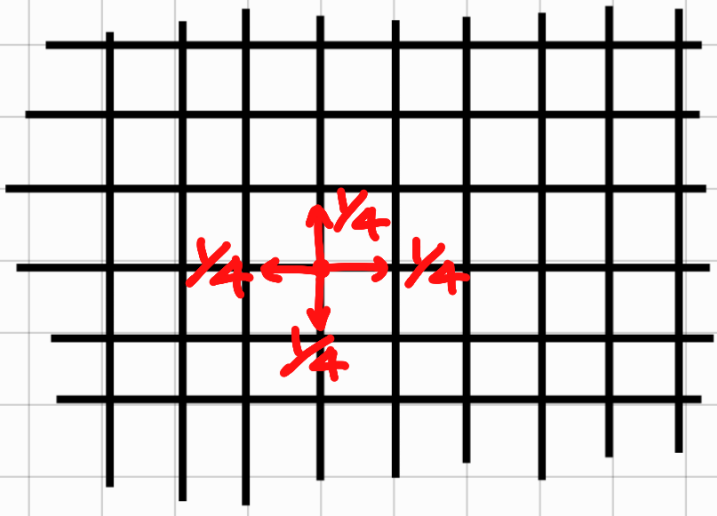
where W_t is a binary process with prob.(p).

$\binom{2n}{n} p^n (1-p)^n$: probability that it returns to 0 in $2n$ steps

Apply stirlings approximation and check for $n \to \infty$.

(i) If ($p = 1/2$) then sequence will eventually return to zero

(ii) Random walk on $\mathbb{Z}^2$



In $\mathbb{Z}^2$ with equal probability $1/4$ the sequence will eventually return to $(0,0)$.

Reference for proofs of above results;

Markov Chains by Norris

e.g.4. Random walk width drift

$W_t \overset{iid}{\sim}$ $E(W_t) = \delta$

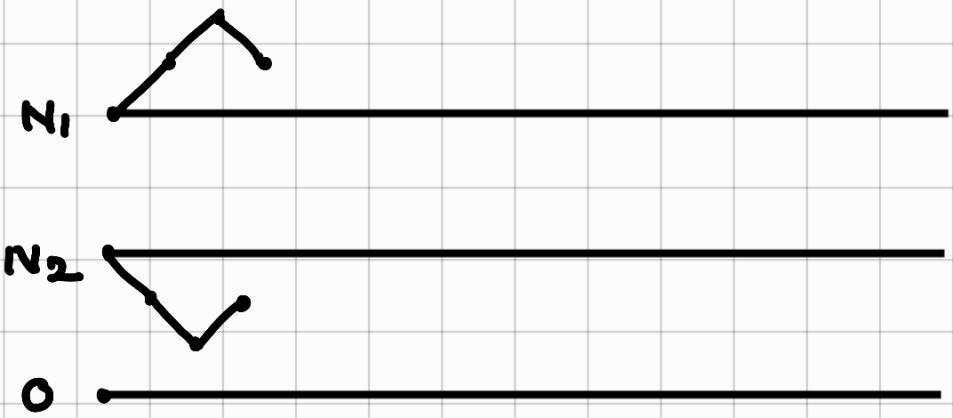
$V(W_t) = \sigma_w^2$

$x_t = x_0 + \sum_{i=1}^{t} W_i$

$E(x_t) = \delta t$

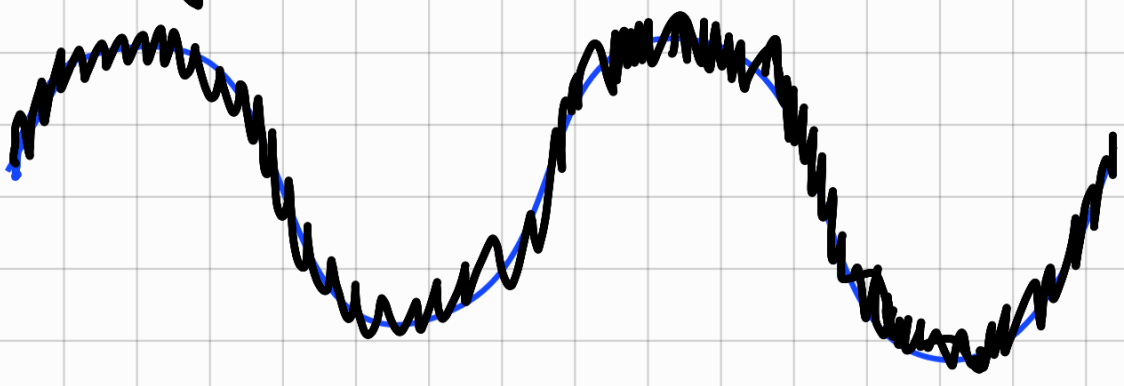
$V(x_t) = \sigma^2 t$

Gambler's ruin problem



First player which reaches 0 coins loses.

e.g.5 Signal with noise



$x_t = A\sin(2\pi f t + \phi) + W_t$

$W_t \sim WN(0, \sigma_w^2)$

e.g.6 Moving average prices (order one)

MA(1) process

Let $W_t \sim WN(0, \sigma^2)$

$x_t = 1W_t + \theta W_{t-1}$

e.g. 7 Autoregressive process (order one)
AR(1)

$x_t = \phi x_{t-1} + w_t \quad |\phi| < 1, \ \phi \neq 0$

→ analogous / similar to the regression problem:

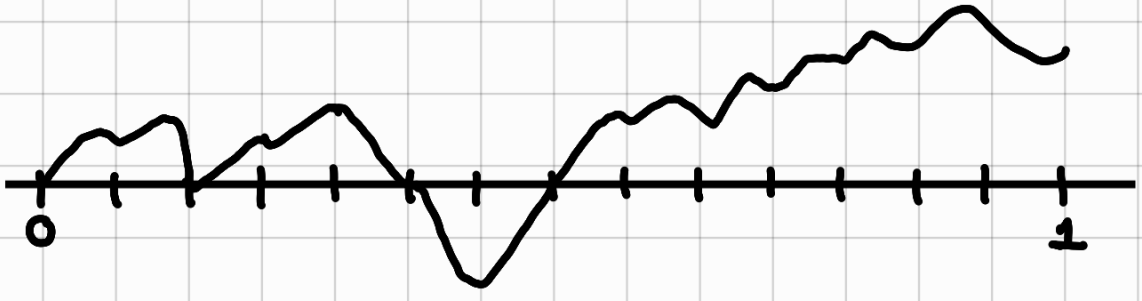
$y_i = \beta^T x_i + \epsilon_i$

Wiener process (Brownian motion)
For definition, refer to the slides.
Generating data from Brownian motion (#)

$T_1 = \sup_{0<t<1} W_t$

$T_2 = \int_0^1 \omega_t^2 \, dt$

→ statistics used to measure qualities of brownian motion



$W_t \sim$ Brownian motion

$y = x^2$ on $(0,1)$

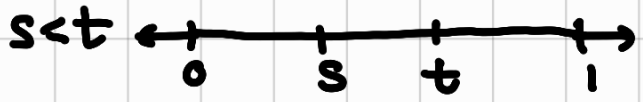
Remark: Let $\{W_i\}$ be a sequence of iid rvs with $E(W_i) = 0$ and $Var(W_i) = 1$ then $Y_n(t) = \frac{1}{\sqrt{n}} \sum_{k=1}^{[nt]} W_k$ converges to Wiener process X_t on $[0,1]$ for large n.

$$Y_n(t) = \frac{1}{\sqrt{n}} \sum_{k=1}^{[nt]} W_k$$

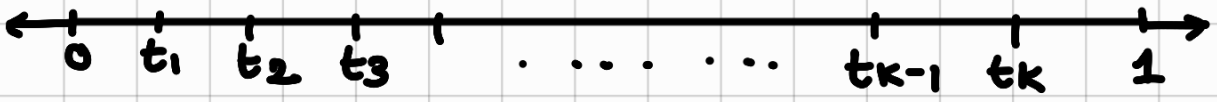
$E(Y_n(t)) = 0$

$$\lim_{n\uparrow\infty} (\mathrm{Var}(Y_n(t)) = \sqrt{\frac{[nt]}{n}} \left(\frac{1}{\sqrt{[nt]}} \sum_{k=1}^{[nt]} W_k \right) \xrightarrow{d} \sqrt{t}\, N(0,1) \equiv N(0,t)$$

Brownian Bridge:



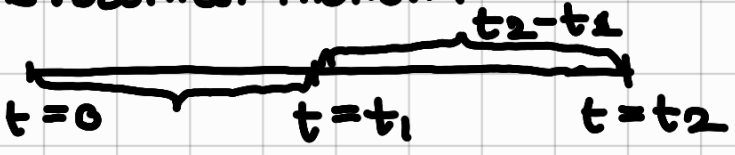
$$\begin{pmatrix} Y_n(s) \\ Y_n(t) \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} s & \min(s,t) \\ \min(s,t) & t \end{pmatrix} \right)$$



$$\begin{pmatrix} Y_n(t_1) \\ Y_n(t_2) \\ \vdots \\ Y_n(t_k) \end{pmatrix} \sim N\left(\underset{\sim}{0}, \begin{pmatrix} t_1 & \cdots & \min(t_1,t_j) & \cdots \\ \vdots & & \vdots & \\ \cdots & \min(t_k,t_j) & \cdots & t_k \end{pmatrix} \right)$$

Description of Brownian motion:

$t = 0, X_0 = 0$



$t = t_1, X_{t_1} = X_0 + Z_1$ where $Z_1 \sim N(0, t_1)$

$t = t_2, X_{t_2} = X_{t_1} + Z_2$ where $Z_2 \sim N(0, t_2 - t_1)$

$\mathrm{Var}(X_{t_2}) = 0 + t_1 + t_2 - t_1 = t_2$

If we fix the initial and final time and we simulate a Brownian motion in b/w these, it forms a Brownian bridge.

Brownian Bridge ($B_0(t)$):

Let $B(t)$ be a brownian motion

$B(t_1) = a$ and $B(t_2) = b$, $t_1 < t_2$

$$\text{mean} = a + \frac{t-t_1}{t_2-t_1}(b-a) \equiv \text{function of } t$$

$$\text{Var} = \frac{(t_2-t)(t-t_1)}{(t_2-t_1)}$$

$$\text{cov} = \frac{(t_2-t)(s-t_1)}{(t_2-t_1)} \quad \text{between } B_0(t) \text{ and } B_0(s)$$

$$B_0(t) = B(t) - \frac{t}{T} B(T)$$

Standard Brownian bridge on $[0,1]$

$(t=0, B_0(t)=0)$ and $(t=1, B_0(t)=0)$ $\begin{cases} B_0(t) = B(t) - tB(1) \\ 0 \le t \le 1 \end{cases}$

$$E(B_0(t)) = 0 + \frac{t-0}{1}(0-0) = 0$$

$$\text{Var}(B_0(t)) = \frac{(1-t)(t-0)}{1} = t(1-t)$$

$$\text{Cov}(B_0(s), B_0(t)) = \frac{(1-t)(s)}{1} = s(1-t) \quad \underline{s \le t}$$

Brownian Bridge ($B_0(t)$)

→ KS Test for goodness of fit

→ Change point detection in temporal data

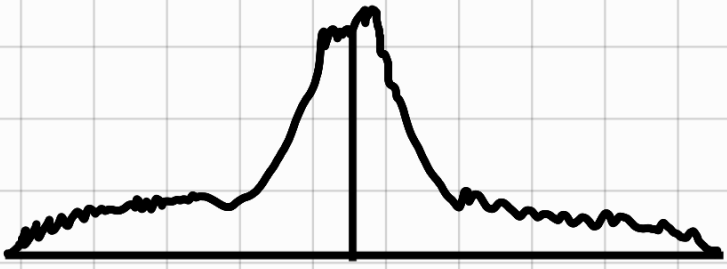
$$S_t := \frac{1}{\sqrt{n}} \sum_{i=1}^{[tn]} \frac{(x_i - \bar{x})}{\hat{\sigma}} \; ; 0 \le t \le 1 \quad \underline{\text{CUSUM statistic}}$$

↳ estimated process variance

$S_t \Rightarrow B_0(t)$

$|S(t)|$

$T = \sup_{0 \le t < 1} |S(t)|$



⟶ these statistics are such that software cannot directly estimate.

⟶ we need to simulate the data to find the cut-off values.

$$T = \sup_{0 \le t < 1} |S(t)| \longrightarrow \sup_{0 \le t \le 1} |B_0(t)|$$

Hypothesis Testing:

$H_0: E(X_1) = E(X_2) = \dots = E(X_n)$

$H_1: E(X_1) = E(X_2) = \dots = E(X_k) \neq E(X_{k+1}) \dots = E(X_n)$

Formulation of variance and covariance:

$X_1, X_2, \dots, X_n \overset{iid}{\sim} F$ continuous distribution

$Y = F(X) \sim U(0,1)$

$E\left(\sum_{i=1}^{n} I_{(Y_i < t)}\right) = nt, \quad 0 \le t \le 1$

$V\left(\sum_{i=1}^{n} I_{(Y_i < t)}\right) = nt(1-t)$

$\text{Var}\left(\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} I_{(Y_i < t)} - \sqrt{n}t\right)\right) = t(1-t)$

$H_n(t) = \sqrt{n}\left(\frac{1}{n}\sum I_{(Y_i < t)} - t\right)$

$\text{Cov}(H_n(s), H_n(t)) \quad \underline{s < t}$

$= \text{cov}\left(\frac{1}{\sqrt{n}}\sum_{i}^{n} I_{(Y_i < t)}, \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{(Y_i < s)}\right)$

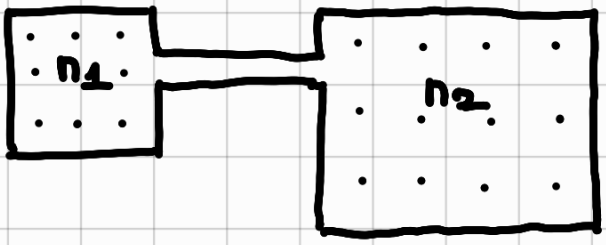
$= \frac{1}{n}\sum_{i}\sum_{j} \text{cov}\left[\left(I_{(Y_i < t)}\right), I_{(Y_j < s)}\right]$

$$\text{Cov} = \min(s,t) - st$$
$$= s(1-t)$$
$$\text{Cov} = \frac{1}{n}.n\{\min(s,t)-st\} = s(1-t)$$

Stationary Timeseries

→ strong stationarity
→ weak stationarity

e.g.1 $X_t \sim \text{bin}(n_1+n_2, 1/2)$ as $t \to \infty$

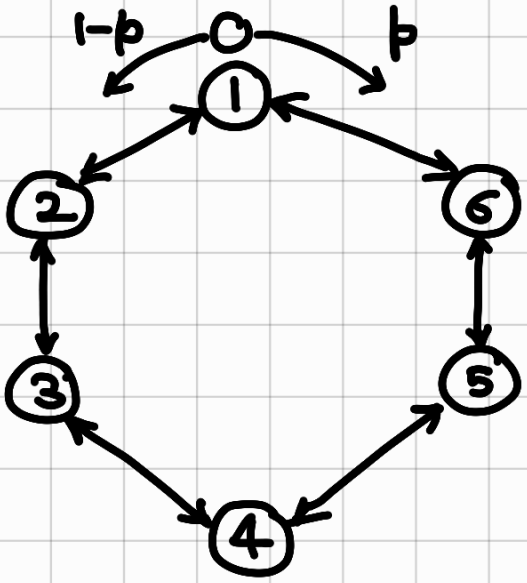


Box 1 Box 2

$X_0 = n_1$

X_t = # balls on box 1 at time t

e.g.2



Limiting distribution of $\{X_t\}$ will follow uniform on $\{1,2,3,4,5,6\}$ $P(X_t=1) = 1/6$.

$X_t :=$ location of the ball at time t.

Transition probability matrix :

	1	2	3	4	5	6
1	0	q	0	0	0	p
2	p	0	q	0	0	0
3	0	p	0	q	0	0
4	0	0	p	0	q	0
5	0	0	0	p	0	q
6	q	0	0	0	p	0

Strongly stationary timeseries

Let $\{X_t\}$ be a time series with a joint distribution fn. of $(X_1, X_2, \ldots, X_n)$ as

$P(X_1 \leq a_1, X_2 \leq a_2, \ldots, X_n \leq a_n) = F_n(a_1, a_2, \ldots, a_n)$

Now, $\forall n \in \mathbb{N}$, $\forall k \in \mathbb{Z}$, $\forall h \in \mathbb{Z}$ and $\forall a_i \in \mathbb{R}$,

If $P(X_{k+1} \leq a_1, X_{k+2} \leq a_2, \ldots, X_{k+n} \leq a_n)$

$= P(X_{k+h+1} \leq a_1, X_{k+h+2} \leq a_2, \ldots, X_{k+h+n} \leq a_n)$

$= F(a_1, a_2, \ldots, a_n)$,

then $\{X_t\}$ is strongly stationary.

Weakly stationary timeseries

A time series $\{X_t\}$ is said to be weakly stationary if

i) $\mu_t = E(X_t)$ is free from t.

ii) $\text{Cov}(X_t, X_{t+h})$ is free from 't' but can be a function of 'h'.

→ this is similar to saying that this is upto second moment condition.

If a TS is strongly stationary with finite 2nd order moment, then it is weakly stationary.

Now, suppose $\{X_t\}$ is atleast weakly stationary.

$$\gamma_x(h) = \text{Cov}(X_t, X_{t+h})$$
$$= E\{(X_t - \mu)(X_{t+h} - \mu)\}$$
$$\dots\ t+h = s \text{ (say)}$$
$$= E\{(X_{s-h} - \mu)(X_s - \mu)\}$$
$$= \gamma_x(-h)$$

'h' is the lag.

Autocorrelation fn.

Auto-correlation coefficient or an 'atleast' weakly stationary timeseries is defined as

$$\rho_x(h) = \gamma_x(h) / \gamma_x(0).$$

Significance/Formulation :

correlation coeff. $\rho_x(h) = \dfrac{\text{cov}(X_{t+h}, X_t)}{\sqrt{\text{var}(X_{t+h})\,\text{var}(X_t)}}$

$$= \gamma_x(h) \Big/ \sqrt{\gamma_x(0)\,\gamma_x(0)}$$
$$= \gamma_x(h) / \gamma_x(0).$$

e.g.1 $\{X_t\} \overset{\text{WN or iid}}{\sim} E(X_t) = 0,\ \text{Var}(X_t) = \sigma^2$

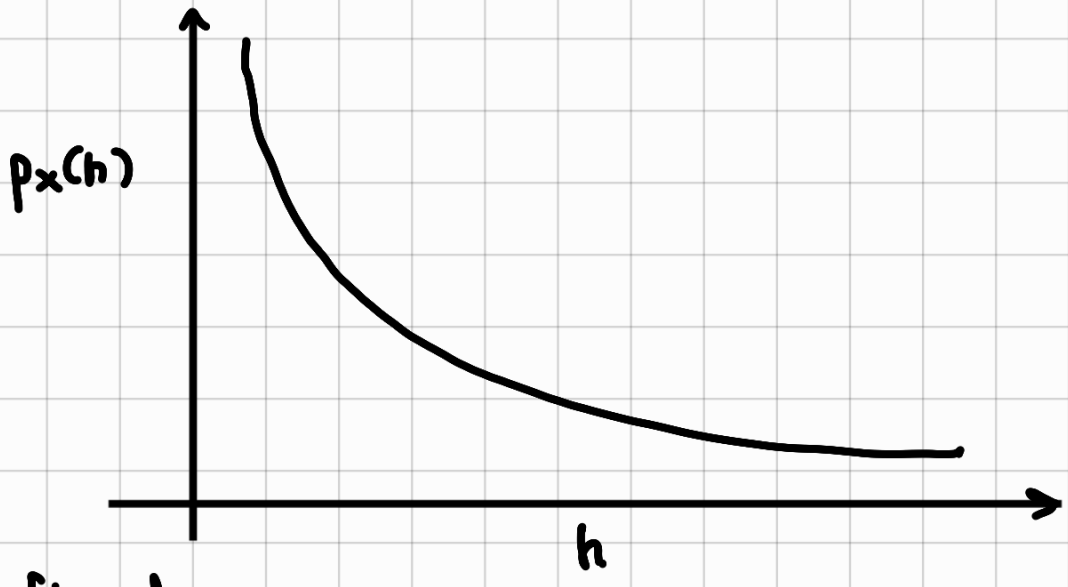
$$\gamma_x(h) = \begin{cases} \sigma^2 & h = 0 \\ 0 & h \neq 0 \end{cases}$$

e.g.2 $S_t = \sum_{i=1}^{t} w_i \quad w_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$

Random walk $\gamma_x(h) = \begin{cases} t\sigma^2 & h = 0 \\ t\sigma^2 & h > 0 \end{cases}$

$$\rho_x(h) = \frac{\operatorname{cov}(S_t, S_{t+h})}{\sqrt{V(S_t)V(S_{t+h})}} = \frac{t\sigma^2}{\sigma^2\sqrt{t(t+h)}}$$

$$\rho_x(h) = \sqrt{t/_{t+h}}$$



t is fixed.

e.g.3 MA(1) Moving averages order one

$Z_t = W_t + \theta W_{t-1}$ where $W_t \overset{iid}{\sim} N(0, \sigma^2)$

$E(Z_t) = 0$

$V(Z_t) = (1+\theta^2)\sigma^2$

$$\gamma_Z(h) = \begin{cases} (1+\theta^2)\sigma^2 & h = 0 \\ \theta\sigma^2 & h = \pm 1 \\ 0 & |h| > 1 \end{cases}$$

e.g.4 Autoregressive process of order one AR(1)

$Z_t = \varphi Z_{t-1} + W_t$ $\begin{cases} W_t \overset{iid}{\sim} N(0, \sigma^2) \\ |\varphi| < 1 \\ \varphi = 0 \\ Z_t \text{ is weakly stationary} \end{cases}$

$E(Z_t) = \varphi E(Z_{t-1}) + E(W_t)$

$$E(Z_t) = \varphi E(Z_{t-1}) + 0$$
$$\Rightarrow E(Z_t) = 0.$$

$$V(Z_t) = E(Z_t^2) = E(\varphi^2 Z_{t-1}^2 + W_t^2 + 2\varphi Z_{t-1} W_t)$$
$$= \varphi^2 E(Z_{t-1}^2) + E(W_t^2) + 0$$
$$= \varphi^2 E(Z_{t-1}^2) + \sigma^2$$
$$E(Z_t^2) = \sigma^2/(1-\varphi^2)$$
$$V(Z_t) = \sigma^2/(1-\varphi^2)$$

$$E(Z_t) = 0$$
$$V(Z_t) = \sigma^2/(1-\varphi^2) > \sigma^2$$

→ this is similar to BVN case.

$(X, Y) \sim BVN$

$V(Y|X) = \sigma_y^2(1-\rho^2)$

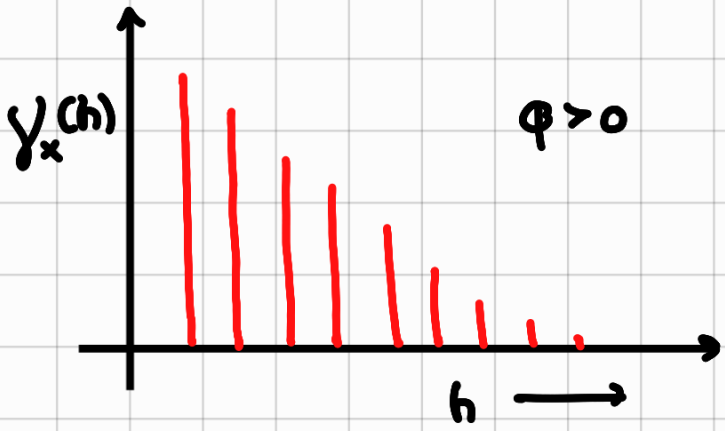
$V(Y) = \sigma_y^2$

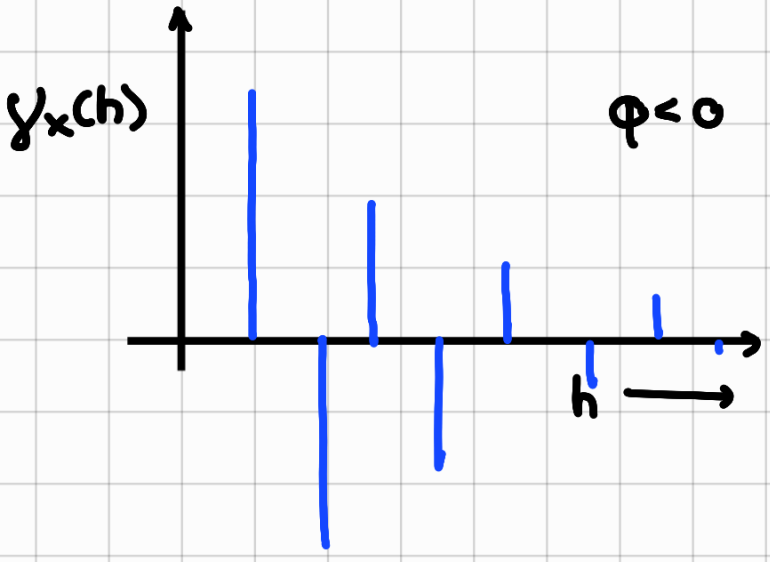
$$\gamma_Z(h) = \begin{cases} (\sigma^2/(1-\varphi^2)) & h = 0 \\ (\sigma^2/(1-\varphi^2))\varphi^h & h \neq 0 \end{cases}$$

$$\gamma_Z(h) = \text{cov}(Z_{t+h}, Z_t)$$
$$= \text{cov}(\varphi Z_{t+h-1} + W_{t+h}, Z_t)$$
$$= \varphi \,\text{cov}(Z_{t+h-1}, Z_t)$$
$$= \varphi^2 \,\text{cov}(Z_{t+h-2}, Z_t)$$
$$\cdots$$
$$= \varphi^h \,\text{cov}(Z_t, Z_t)$$
$$= \varphi^h \left(\frac{\sigma^2}{1-\varphi^2}\right) = \gamma_Z(-h)$$

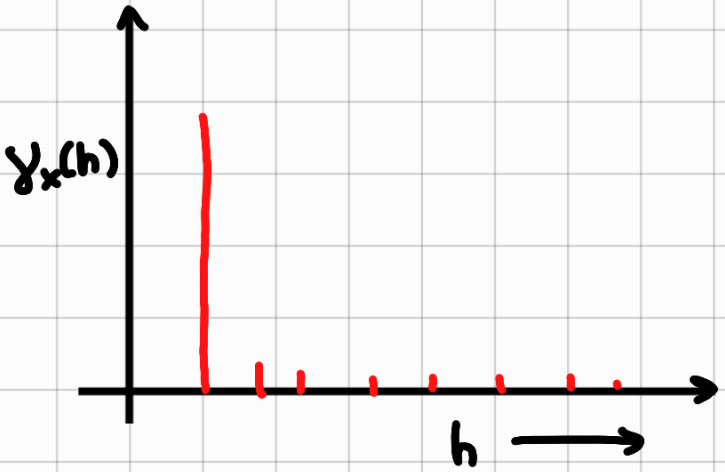
AR(1)

$$\gamma_X(h) = \gamma_X(-h) = \varphi^{|h|}(\sigma^2/(1-\varphi^2))$$





MA(1)



Auto-correlation fn.

Notations of operators

1. Backshift operator (B)

$$B^h x_t \equiv x_{t-h}$$

2. Difference operator (∇)

$$\nabla x_t \equiv x_t - x_{t-1} \equiv (I-B)x_t$$

$$\boxed{\nabla = I - B}$$

$$\nabla^h x_t = (I-B)^h x_t = \sum_{k=0}^{h} \binom{h}{k} (-1)^{h-k} (B^{h-k} x_t)$$

3. Seasonal Operator

$$\nabla_s = (I - B^s)$$

$$\nabla_s x_t = x_t - x_{t-s}$$

Consider $f(x) = a_0 + a_1 x + a_2 x^2 = 1 + 2x + 3x^2$

$x = 1, f(1) = 6$

$x = 2, f(2) = 17$

$x = 3, f(3) = 34$

$x = 4, f(4) = 57$

	∇	∇^2	∇^3
6			
17	11		
34	17	6	0
57	23	6	0

Linear process

A time series $\{X_t\}$ is a linear process if it can be represented as :

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j} \quad \text{where } t \in \mathbb{Z},$$

$$= \mu + \sum_{j=-\infty}^{\infty} \psi_j B^j Z_t$$

$Z_t \sim WN(0, \sigma^2)$,

$\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$

$$= \mu + \left(\sum_{j=-\infty}^{\infty} \psi_j B^j \right) Z_t$$

$$= \mu + \Psi(B) Z_t$$

$$\Psi(B) = \sum_{-\infty}^{\infty} \psi_j B^j$$

$$X_t = \{ \Psi(B) Z_t \} + \mu$$

Note. $E(X)$ exists if $E|X| < \infty$, i.e.,

$\sum_i |x_i| f(x_i) < \infty$ (or, $\int |x| f(x) dx < \infty$)

Note. Absolutely summable series is always summable

Ex 1. $X_t \sim WN(0, \sigma^2)$

$$\psi_j = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{o.w.} \end{cases}$$

Ex 2. $X_t \sim MA(1)$

$$\psi_j = \begin{cases} 1 & \text{if } j = 0 \\ \theta & \text{if } j = 1 \\ 0 & \text{o.w.} \end{cases}$$

Ex 3. $X_t \sim AR(1)$

$$X_t = \varphi X_{t-1} + Z_t$$

$$\psi_j = \begin{cases} \varphi^j & \text{if } j \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

MA(q) process $q \in \mathbb{N}$

MA(q) process is a linear process where

$$\begin{cases} \psi_0 = 1 \\ \psi_j = \begin{cases} \theta_j & 1 \leq j \leq q \\ 0 & \text{o.w.} \end{cases} \end{cases}$$

$$\begin{aligned} X_t &= Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j} \\ &= Z_t + \left(\sum_{j=1}^{N} \theta_j B^j\right) Z_t \\ &= \left(I + \sum_{j=1}^{q} \theta_j B^j\right) Z_t \equiv \Theta_q(B) Z_t. \end{aligned}$$

$Z_t \sim WN(0, \sigma^2)$

$$\text{cov}(X_t, X_s) = \begin{cases} 0 & \text{lag } (|t-s| > q) \\ \text{can be non-zero} & \text{o.w.} \end{cases}$$

If $Z_t \overset{iid}{\sim}$ then, (X_t, X_j) are independent if $|t-s| > q$.

Autoregressive Process of order (p) $\equiv$ AR(p) $p \in \mathbb{N}$

A process is said to be an AR(p) process if it can be represented as

$$X_t = \sum_{j=1}^{p} \varphi_j X_{t-j} + Z_t$$

where $Z_t \sim WN(0, \sigma^2)$.

$\Rightarrow (x_t) - (\sum_{j=1}^{p} \varphi_j B^j) x_t = Z_t.$

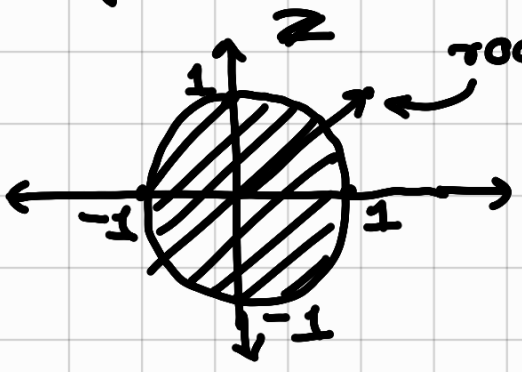
$\Rightarrow (I - \sum_{j=1}^{p} \varphi_j B^j) x_t = Z_t.$

Let $\Phi_p(u) = (1 - \sum_{j=1}^{p} \varphi_j u^j)$, then

$\Phi_p(B) x_t = Z_t$

$\Rightarrow x_t = \left(\frac{1}{\Phi_p(B)}\right) Z_t. \longrightarrow (*)$

$\frac{1}{\Phi_p(B)} = \frac{1}{1 - \sum_{j=1}^{p} \varphi_j B^j}$. This should satisfy unit root condition for $(*)$ to be feasible, that is, for eqn. $1 - gu \equiv 0$, root $u = 1/g > 1$.
(for our case, $gu \equiv \sum_{j=1}^{p} \varphi_j B^j$)
Also, $|g| < 1$, $g \neq 0$.



AR(1) process can be thought as MA(∞) process.
$AR(1) = \lim_{q \to \infty} MA(q)$.

We can represent AR(1) process as MA(∞) process

$x_t = \varphi x_{t-1} + Z_t \quad |\varphi| < 1 \quad \varphi \neq 0$

$\rightarrow (I - \varphi B) x_t = Z_t$

$\rightarrow x_t = \left(\frac{1}{1 - \varphi B}\right) Z_t$

$\Rightarrow x_t = \left(\sum_{j=0}^{\infty} (\varphi B)^j\right) Z_t$

$\Rightarrow x_t = \sum_{j=0}^{\infty} \varphi^j Z_{t-j}$

$$x_t = \left(\sum_{j=0}^{k} \varphi^j z_{t-j}\right) + \varphi^{k+1} x_{t-(k+1)}$$

$$E\left(x_t - \sum_{j=0}^{k} \varphi^j z_{t-j}\right)^2 = E\left\{\varphi^{k+1} x_{t-(k+1)}\right\}^2$$

$$\lim_{k\to\infty} E\left(x_t - \sum_{j=0}^{k} \varphi^j z_{t-j}\right) = \lim_{k\to\infty} E\left\{\varphi^{k+1} x_{t-(k+1)}\right\}^2$$

$$= \lim_{k\to\infty} \left(\varphi^{2k+2}\right) E\left(x_{t-(k+1)}^2\right)$$

$\varphi^{2k+2} \to 0$; $E\left(x_{t-(k+1)}^2\right)$ bounded

$$\left(\lim_{n\to\infty} E|x_n - y|^2 \to 0 \implies x_n \xrightarrow{d} y.\right)$$

Hence, $x_t \overset{d}{=} \sum_{j=0}^{\infty} \varphi^j z_{t-j}$

$$\lim_{n\uparrow\infty} \left(x_t - \sum_{j=0}^{k} \varphi^j z_{t-j}\right)^2 = \lim_{k\uparrow\infty} \left\{\varphi^{2k} E\left(x_{t-k}^2\right)\right\}$$

x_t : AR(1) ; $\sum_{j=0}^{k} \varphi^j z_{t-j}$: MA(k) ; $\varphi^{2k} \to 0$; $E\left(x_{t-k}^2\right)$ bounded

$Y_k = \sum_{j=0}^{k} \varphi^j z_{t-j}$ such that $E|x_t - Y_k(t)|^2 \to 0$

$\implies Y_k(t) \xrightarrow{d} x_t$ as $k \uparrow \infty$.

Note. $x_1, x_2, \ldots, x_n$ iid $E(x) = \mu$ $V(x) = \sigma^2$

$T_k = \sqrt{k}(\bar{x} - \mu)/\sigma$

$Z \sim N(0,1)$

Then $E(T_k - Z)^2 \to 0 \implies T_k \xrightarrow{d} Z$

Covariance of a Linear Process

Let $\{x_t\}$ be a weakly stationary process with $E(x_t) = 0$ and $\gamma_x(h) = \text{cov}(x_t, x_{t+h})$ exists.

If $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, then defining $Y_t = \sum_{j=-\infty}^{\infty} \psi_j x_{t-j}$, we have

$\to E(Y_t) = 0$

$\to \gamma_Y(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_x(h+k-j)$

Proof.

$$E|Y_t| = E\left|\sum_{j=-\infty}^{\infty}\psi_j X_{t-j}\right| \le \sum_{j=-\infty}^{\infty}|\psi_j| E|X_{t-j}| \le m_1 \sum_{j=-\infty}^{\infty}|\psi_j| < \infty$$

where $E|X_t| < m_1$.

Homework. Let Z be a random variable. Prove that

$$E(|Z|^r) \le \infty \implies E(|Z|^s) < \infty, \quad r \ge s$$

Hence, $E|Y_t|$ exists.

$$E(Y_t) = E\left(\sum_{j=-\infty}^{\infty}\psi_j X_{t-j}\right) = \sum_{j=-\infty}^{\infty}\psi_j E(X_{t-j}) = 0.$$

$$E|Y_{t+h}Y_t| = E\left|\sum_j \psi_j X_{t+h-j} \sum_k \psi_k X_{t-k}\right|$$

$$= E\left|\sum_j\sum_k \psi_j\psi_k X_{t+h-j}X_{t-k}\right|$$
$$\le \sum_j\sum_k |\psi_j||\psi_k| E(|X_{t+h-j}||X_{t-k}|)$$
$$\le \sum_j\sum_k |\psi_j||\psi_k| m_2$$

where $\max_{t,h\in\mathbb{Z}} E|X_{t+h-j}X_{t-k}| < m_2$.

So,

$$E|Y_{t+h}Y_t| \le \sum_j\sum_k |\psi_j||\psi_k| m_2 = m_2\left(\sum_j|\psi_j|\sum_k|\psi_k|\right) < \infty.$$

$$\gamma_Y(h) = \operatorname{cov}(Y_{t+h}Y_t)$$
$$= \operatorname{cov}\left(\sum_j \psi_j X_{t+h-j} \sum_k \psi_k X_{t-k}\right)$$
$$= E\left(\sum_j\sum_k \psi_j\psi_k X_{t+h-j}X_{t-k}\right)$$
$$= \sum_j\sum_k \psi_j\psi_k E(X_{t+h-j}X_{t-k})$$
$$= \sum_j\sum_k \psi_j\psi_k \operatorname{cov}(X_{t+h-j}, X_{t-k})$$
$$= \sum_j\sum_j \psi_j\psi_k\gamma_X(h-j)$$

Homework. $X_t \sim WN(0,\sigma^2)$ and Y_t is a linear process.

Then, $\gamma_Y(h) = \sigma^2\sum_{j=-\infty}^{\infty}\psi_j\psi_{j+h}$.

Homework. $Y_t \sim AR(1)$. $\gamma_Y(h) = \phi^h \sigma^2 / (1-\phi^2)$.

ARMA process

ARMA (p,q)

A time series $\{x_t\}$ is said to follow ARMA (p,q) process if we have the representation

$$(x_t - \Sigma_{j=1}^{p} \phi_j x_{t-j}) = z_t + \Sigma_{j=1}^{q} \theta_j z_{t-j}$$
$$\forall\ t \in \mathbb{Z},\ z \sim WN.$$

$$\rightarrow (I - \Sigma_{j=1}^{p} \phi_j B^j) x_t = (I + \Sigma_{j=1}^{q} \theta_j B^j) z_t$$
$$\rightarrow \Phi_p(B) x_t = \mathcal{H}_q(B) z_t \text{ where}$$

$$\Phi_p(x) = 1 - \sum_{j=1}^{p} \phi_j x^j \text{ and } \mathcal{H}_q(x) = 1 + \sum_{j=1}^{q} \theta_j x^j$$

$\Phi_p(x)$ and $\mathcal{H}_q(x)$ should not have any common root.

For example, ARMA(1,1)

$$x_t - \phi x_{t-1} = z_t + \theta z_{t-1}$$
$$\Rightarrow x_t = \left(\frac{I + \theta B}{I - \phi B}\right) z_t$$

$$\begin{cases} |\phi| < 1,\ \phi \neq 0 \\ \theta + \phi \neq 0 \end{cases}$$

$$x_t = \{(1+\theta B)(\Sigma_{j=0}^{\infty} (\phi B)^j)\} z_t = \Psi(B) z_t$$
$$= (\psi_0 + \Sigma_{j=1}^{\infty} \psi_j B^j) z_t$$

$$\psi_0 = 1$$
$$\psi_j = \begin{cases} (\theta+\phi)\phi^{j-1} \end{cases} \forall\ j \geq 1$$

Linear process representation

$$x_t = z_t + (\theta+\phi) \Sigma_{j=1}^{\infty} \phi^{j-1} z_{t-j}$$
$$\gamma_x(0) = \sigma^2 \Sigma_{j=0}^{\infty} \psi_j^2 = \sigma^2 [1 + (\theta+\phi)^2 \Sigma_{j=1}^{\infty} \phi^{j-2}]$$
$$= \sigma^2 \left[1 + \frac{(\theta+\phi)^2}{1-\phi^2}\right]$$

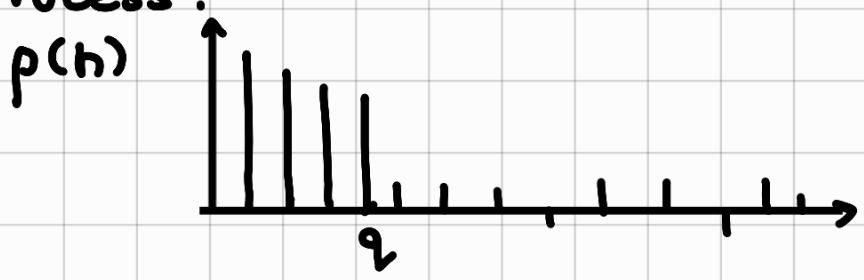
$$\gamma_x(1) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+1} = \sigma^2 \left[(\theta+\phi)^2 + \frac{(\theta+\phi)^2 \phi}{1-\phi^2} \right]$$

$\gamma_x(h) = \phi^{h-1} \gamma_x(1)$. Homework.

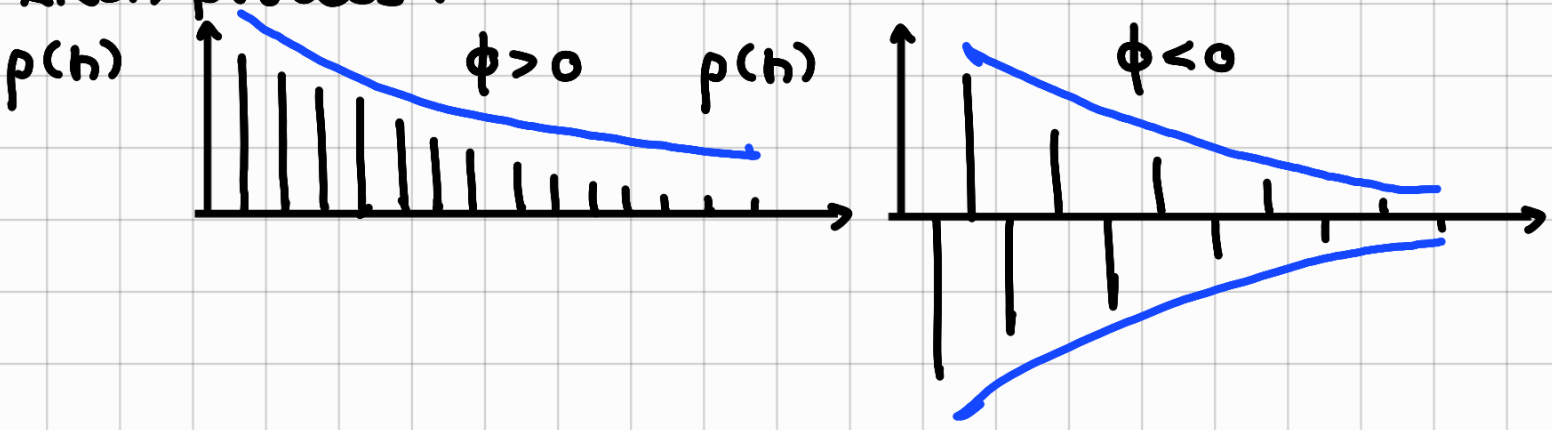
Autocorrelation coefficient of weakly stationary process.

$$\rho(h) = \frac{\gamma_x(h)}{\gamma_x(0)}$$

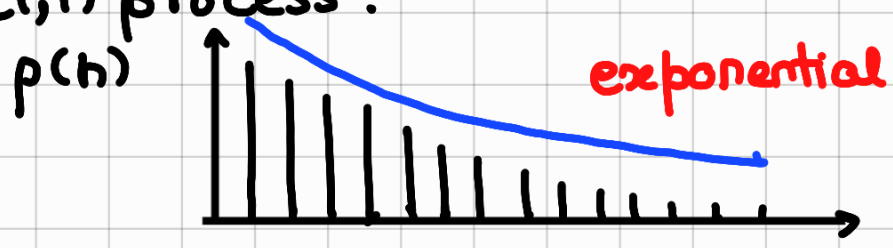
MA(q) process:



AR(1) process:



ARMA(1,1) process:



ACF can be used to distinguish b/w MA and (AR/ARMA)
But it cannot distinguish b/w AR & ARMA.
To do so, we need partial auto correlation coefficient. (PACF)

Partial correlation Coefficient

given, $\begin{pmatrix} Y \\ Z \\ \underset{\sim}{X} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_y \\ \mu_z \\ \underset{\sim}{\mu_x} \end{pmatrix}, \left(\begin{array}{c|c} \Sigma_{yz} & \Sigma_{yz\underset{\sim}{x}}^T \\ \hline \Sigma_{yz\underset{\sim}{x}} & \Sigma_{\underset{\sim}{x}} \end{array}\right) \right)$ $\quad y \in \mathbb{R}^1, \; z \in \mathbb{R}^1, \; \underset{\sim}{x} \in \mathbb{R}^p$

$E(Y) = \mu_y \quad E(Z) = \mu_z \quad E(\underset{\sim}{X}) = \underset{\sim}{\mu_x}$

$D\begin{pmatrix} Y \\ Z \end{pmatrix} = \Sigma_{yz} = \begin{pmatrix} \sigma_{yy} & \sigma_{yz} \\ \sigma_{yz} & \sigma_{zz} \end{pmatrix}_{2\times 2} \quad D(X) = \Sigma_{\underset{\sim}{x}} = ((\sigma_{x_i x_j}))_{p\times p}$

$$\sigma_{y\underset{\sim}{x}} = cov(Y, \underset{\sim}{X}) = \begin{pmatrix} \sigma_{yx_1} \\ \sigma_{yx_2} \\ \vdots \\ \sigma_{yx_p} \end{pmatrix}$$

$$\sigma_{z\underset{\sim}{x}} = cov(Z, \underset{\sim}{X}) = \begin{pmatrix} \sigma_{zx_1} \\ \sigma_{zx_2} \\ \vdots \\ \sigma_{zx_p} \end{pmatrix}$$

$\Sigma_{yz\underset{\sim}{x}} = (\underset{\sim}{\sigma}_{y\underset{\sim}{x}}, \underset{\sim}{\sigma}_{z\underset{\sim}{x}})_{p\times 2} \quad \Sigma_{yz\underset{\sim}{x}}^T = \begin{pmatrix} \underset{\sim}{\sigma}_{y\underset{\sim}{x}}^T \\ \underset{\sim}{\sigma}_{z\underset{\sim}{x}}^T \end{pmatrix}_{2\times p}$

Partial correlation coefficient is the correlation b/w Y and Z after removing the effect of $\underset{\sim}{X}$ and is denoted as $\rho_{yz.\underset{\sim}{x}}$

$$\rho_{yz.\underset{\sim}{x}} = \frac{cov(Y - E(Y|x), Z - E(Z|x))}{\sqrt{Var(Y - E(Y|x)) \, Var(Z - E(Z|x))}}$$

$$\Rightarrow \rho_{yz.\underset{\sim}{x}} = cov(e_{y|x}, e_{z|x}) / \sqrt{Var(e_{y|x}) Var(e_{z|x})}$$

$$= \sigma_{yz.x} / \sqrt{\sigma_{yy.x}\,\sigma_{zz.x}}$$

When we use it for time series we call it PACF.

conditional distribution of $\left(\begin{smallmatrix} Y \\ Z \end{smallmatrix}\right) \Big| X = \underset{\sim}{x}$.

This also follow normal distribution,

$$N\left(\underset{2\times 1}{\begin{pmatrix} \mu_y \\ \mu_z \end{pmatrix}} + \underset{2\times p}{\Sigma_{yzx}^{T}} \underset{p\times p}{\Sigma_x^{-1}} \underset{p\times 1}{(\underset{\sim}{x} - \underset{\sim}{\mu_x})},\ \underset{2\times 2}{\Sigma_{yz}} - \underset{2\times p}{\Sigma_{yzx}^{T}} \underset{p\times p}{\Sigma_x^{-1}} \underset{p\times 2}{\Sigma_{yzx}}\right)$$

Case: For Bivariate Normal,

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim BVN\left(\begin{pmatrix} \mu_u \\ \mu_v \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix}\right)$$

$$V|U=x \sim N\left(\mu_v + \rho\frac{\sigma_v}{\sigma_u}(x-\mu_u),\ (1-\rho^2)\sigma_u^2\right)$$

$$E(V|U=u) = \mu_v + \rho\frac{\sigma_u\sigma_v}{\sigma_u^2}(x-\mu_u)$$

$$= \mu_v + \frac{cov(u,v)}{V(u)}(x-\mu_u)$$

$$= \mu_v + cov(u,v)(V(u))^{-1}(x-\mu_u)$$

$$V(V|U=u) = \sigma_v^2 - cov(u,v)(V(u))^{-1}cov(u,v)$$
$$= \sigma_v^2 - \rho\sigma_u\sigma_v(\sigma_u^2)^{-1}\rho\sigma_u\sigma_v$$
$$= \sigma_v^2 - \rho^2\sigma_v^2$$
$$= (1-\rho^2)\sigma_v^2$$

Mean and the variance-covariance structure will remain same, if we do best linear prediction of $\left(\begin{smallmatrix} Y \\ Z \end{smallmatrix}\right)$ based on

$\underset{\sim}{x}$, even when $\begin{pmatrix} y \\ \underset{\sim}{x} \end{pmatrix}$ are not jointly normally distributed but 2nd order moments exists.