A Brief Summary of Statistics Course

统计学课程知识总结

Vincent*

Department of Physics, Tsinghua University

2021 年 9 月 27 日

目录

*V1ncent19@outlook.com

Chapter. I 概率论部分

Instructor: Wanlu Deng

Section 1.1 Some Important Distributions

X	$p_X(k)//f_X(x)$	E	σ^2	PGF	MGF
$B(p)$		p	pq		$q+pe^s$
$B(n,p)$	$C_n^k p^k(1-p)^{n-k}$	np	npq	$(q+ps)^n$	$(q+pe^s)^n$
$G(p)$	$(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{q}{p^2}$	$\frac{ps}{1-qs}$	$\frac{pe^s}{1-qe^s}$
$H(n,M,N)$	$\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}$	$n\frac{M}{N}$	$\frac{nM(N-n)(N-M)}{N^2(n-1)}$		
$P(\lambda)$	$\frac{\lambda^k}{k!}e^{-\lambda}$	λ	λ	$e^{\lambda(s-1)}$	$e^{\lambda(e^s-1)}$
$U(a,b)$	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$		$\frac{e^{sb}-e^{sa}}{(b-a)s}$
$N(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2		$e^{\frac{\sigma^2 s^2}{2}+\mu s}$
$\epsilon(\lambda)$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$		$\frac{\lambda}{\lambda-s}$
$\Gamma(\alpha,\lambda)$	$\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$		
$B(\alpha,\beta)$	$\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$		
χ_n^2	$\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}$	n	$2n$		
t_ν	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1+\frac{x^2}{\nu})^{-\frac{\nu+1}{2}}$	0	$\frac{\nu}{\nu-2}$		
$F_{m,n}$	$\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\frac{m^{\frac{m}{2}}n^{\frac{n}{2}}x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}}$	$\frac{n}{n-2}$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$		

Definition of PGF, MGF, CF see section 1.5.

More Properties of χ^2, t, F see section 1.8.2.

Section 1.2 Probability and Probability Model

What is **Probability**?

A 'belief' in the chance of an event occurring?

1.2.1 Sample and σ-Field

Def. sample space Ω: The set of all possible outcomes of one particular experiment.

Def. $\mathscr{F}$ a σ-field(or a σ-algebra) as a collection of some subsets of Ω **if**

- $\Omega \in \mathscr{F}$
- if $A \in \mathscr{F}$,then $A^C \in \mathscr{F}$

- if $A_n \in \mathscr{F}$, then $\bigcup\limits_{n=1}^{\infty} A_n \in \mathscr{F}$

And $(\Omega, \mathscr{F})$ is a measurable space.

1.2.2 Axioms of Probability

P is probability measure (or probability function) defined on $(\Omega, \mathscr{F})$, satisfying

- Nonnegativity

$$P(A) \geq 0 \qquad \forall A \in \Omega \tag{1.1}$$

- Normalization

$$P(\Omega) = 1 \tag{1.2}$$

- Countable Additivity

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots \quad (A_i \parallel A_j \quad \forall i \neq j) \tag{1.3}$$

Then $(\Omega, \mathscr{F}, P)$ is probability space.

Properties of Probability:

- Monotonicity

$$P(A) \leq P(B) \quad \text{for } A \subset B \tag{1.4}$$

- Finite Subadditivity (Boole Inequality)

$$P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i) \tag{1.5}$$

- Inclusion-Exclusion Formula

$$\begin{aligned} P(\Cup_{i=1}^{n} A_i) = & \sum_{1 \leq i \leq n} P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j) \\ & + \sum_{1 \leq i < j < k \leq n} P(A_i \cap A_j \cap A_k) - \cdots \\ & + (-1)^{n-1} P(A_1 \cap A_2 \cap \cdots \cap A_n) \end{aligned}$$

- Borel-Cantelli Lemma

$$\sum_{n=1}^{\infty} P(A_n) < \infty \Rightarrow P(\lim_{n \to \infty} \sup A_n) = 0$$

$$\sum_{n=1}^{\infty} P(A_n) = \infty \Rightarrow P(\lim_{n \to \infty} \sup A_n) = 1 \quad \text{if } A_i \text{ independent}$$

1.2.3 Conditional Probability

Def. **Conditional Probability** of B given A:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{1.6}$$

(Actually a change of σ-field from Ω to B)

Application of conditional probability:

- Multiplication Formula
$$P(\Cap_{i=1}^{n} A_i) = P(A_1) \prod_{i=2}^{n} P(A_i | A_1 \cap A_2 \cap \cdots \cap A_{i-1}) \tag{1.7}$$

- Total Probability Thm.
$$P(B) = \sum_{i=1}^{n} P(A_i) P(B|A_i) \tag{1.8}$$
where $\{A_i\}$ is a partition of Ω.

- Bayes's Rule
$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{n} P(A_j)P(B|A_j)} \tag{1.9}$$
where $\{A_i\}$ is a partition of Ω.

- Statistically Independence
$$P(A \cap B) = P(A)P(B), \text{ for } A \parallel B \tag{1.10}$$

Section 1.3 Properties of Random Variable and Vector

1.3.1 Random Variable

Def. Random Variable: a **function** X defined on sample space Ω, mapping from Ω to some $\mathscr{X} \in \mathbb{R}$.

Then def. Cumulative Distribution Function (CDF).

$$F_X(x) = P(X \leq x) \tag{1.11}$$

For Discrete case, consider CDF as right-continuity.

- PMF:
$$p_X(x) = F_X(x^+) - F_X(x^-) \tag{1.12}$$
PDF:
$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x} \tag{1.13}$$

- Indicator function:
$$I_{x \in A}(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases} \tag{1.14}$$

- Convolution
  - $W = X + Y$
$$f_W(w) = \int_{-\infty}^{\infty} f_X(x) f_Y(w - x) \mathrm{d}x \tag{1.15}$$

– $V = X - Y$

$$f_V(v) = \int_{-\infty}^{\infty} f_X(x) f_Y(x - v) \mathrm{d}x \tag{1.16}$$

– $Z = XY$

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|x|} f_X(x) f_Y(\frac{z}{x}) \mathrm{d}x \tag{1.17}$$

- Order Statistics

  Def $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ as order statistics of $\vec{X}$

$$g_{X_{(i)}} = n! \prod f(x_i) \qquad \text{for } x_1 < x_2 \cdots < x_n \tag{1.18}$$

  PDF of $X_{(k)}$

$$g_k(x_k) = nC_{n-1}^{k-1} [F(x_k)]^{k-1} [1 - F(x_k)]^{n-k} f(x_k) \tag{1.19}$$

- p-fractile

$$\xi_p = F^{-1}(p) = \inf\{x | F(x) \geq p\} \tag{1.20}$$

1.3.2 Random Vector

A general case of random variable.

n-dimension Random Vector $\vec{X} = (X_1, X_2, \ldots, X_n)$ defined on $(\Omega, \mathscr{F}, P)$.

CDF $F(x_1, \ldots, x_n)$ defined on $\mathbb{R}^n$:

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n) \tag{1.21}$$

Joint PDF of random vector:

$$f(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \ldots \partial x_n} \tag{1.22}$$

k-dimensional Marginal Distribution: For $1 \leq k < n$ and index set $S_k = \{i_1 \ldots, i_k\}$, distribution of $\vec{X} = (X_{i_1}, X_{i_2}, \ldots, X_{i_k})$

$$F_{S_k}(x_{i_1}, X_{i_2} \leq x_{i_2} \ldots, x_{i_k}) = P(X_{i_1} \leq x_{i_1}, \ldots, X_{i_k} \leq x_{i_k}; X_{i_{k+1}}, \ldots, X_{i_n} \leq \infty) \tag{1.23}$$

Marginal distribution:

$$g_{S_k}(x_{i_1}, \ldots, x_{i_k}) = \int_{\mathbb{R}^{n-k}} f(x_1, \ldots, x_n) \mathrm{d}x_{i_{k+1}} \ldots \mathrm{d}x_{j_n} = \frac{\partial^{n-k} F(x_1, \ldots, x_n)}{\partial x_{i_{k+1}} \ldots \partial x_{i_n}} \tag{1.24}$$

△ Function of r.v.

For $\vec{X} = (X_1, X_2, \cdots, X_n)$ with PDF $f(\vec{X})$ and define

$$\vec{Y} = (Y_1, Y_2, \cdots, Y_n) = (y_1(\vec{X}), y_2(\vec{X}), \cdots, y_n(\vec{X})) \tag{1.25}$$

with inverse mapping

$$\vec{X} = (X_1, X_2, \cdots, X_n) = (x_1(\vec{Y}), x_2(\vec{Y}), \cdots, x_n(\vec{Y})) \tag{1.26}$$

then

$$g(\vec{Y}) = f(x_1(\vec{Y}), x_2(\vec{Y}), \cdots, x_n(\vec{Y})) \left| \frac{\partial \vec{X}}{\partial \vec{Y}} \right| I_{D_Y} \tag{1.27}$$

(Intuitively: $g(\vec{Y}) \mathrm{d}\vec{Y} = \mathrm{d}P = f(\vec{X}) \mathrm{d}\vec{X}$)

Section 1.4 Properties of E , σ^2 and cov

Expectation and Variance of common distributions see sec.1.1.

1.4.1 Expection

Expectation of r.v. $g(X)$ def.:

$$E[g(X)] = \begin{cases} \int_{\Omega} g(x) f_X(x) \mathrm{d}x = \int_{\Omega} g(x) \mathrm{d}F(x) \\ \sum_{\Omega} g(X) f_X(x) \end{cases} \tag{1.28}$$

Properties of expectation $E(\cdot)$:

- Linearity of Expectation

$$E(aX + bY) = aE(X) + bE(Y) \tag{1.29}$$

- Conditional Expectation

$$E(X|A) = \frac{E(XI_A)}{P(A)} \tag{1.30}$$

  Note: if take A as Y is also a r.v. then

$$m(Y) = E(X|Y) = \int x f_{X|Y}(x) \mathrm{d}x \tag{1.31}$$

  is actually a function of Y

- Law of Total Expectation

$$E\{E[g(X)|Y]\} = E[g(X)] \tag{1.32}$$

- r.v.& Event

$$P(A|X) = E(I_A|X) \Rightarrow E[P(A|X)] = E(I_A) = P(A) \tag{1.33}$$

-

$$E[h(Y)g(X)|Y] = h(Y)E[g(X)|Y] \tag{1.34}$$

1.4.2 Variance

Variance of r.v. X:

$$var(X) = E[(X - E(X))^2] = E(X^2) - (E(X))^2 \tag{1.35}$$

(sometimes denoted as σ_X^2.)

Properties:

- Linear combination of Variance

$$var(aX + b) = a^2 var(X) \tag{1.36}$$

- Conditional Variance

$$var(X|Y) = E[X - E(X|Y)]^2|Y \tag{1.37}$$

- Law of Total Variance

$$var(X) = E[var(X|Y)] + var[E(X|Y)] \tag{1.38}$$

Standard Deviation def. as :

$$\sigma_X = \sqrt{var(X)} \tag{1.39}$$

Then can construct **Standardization** of r.v.

$$Y = \frac{X - E(X)}{\sqrt{var(X)}} \tag{1.40}$$

1.4.3 Covariance and Correlation

Covariance of r.v. X and Y:

$$cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - E(X)E(Y) \tag{1.41}$$

And (Pearson's) Correlation Coefficient

$$r = \rho_{X,Y} = corr(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}} \tag{1.42}$$

Remark: correlation $\nRightarrow$ cause and effect. Detail on causal effect topic see section **??**.

Properties:

- Bilinear of Covariance

$$cov(X + Y, Z) = cov(X, Z) + cov(Y, Z)$$
$$cov(X, Y + Z) = cov(X, Y) + cov(X, Z)$$

- Variance and Covariance

$$var(X + Y) = var(X) + var(Y) + 2cov(X, Y) \tag{1.43}$$

- Covariance Matrix

  Def $\Sigma = E[(X - \mu)(X - \mu)^T] = \{\sigma_{ij}\}$ (where X should be considered as a column vector)

$$\Sigma = \begin{pmatrix} var(X_1) & cov(X_1, X_2) & \dots & cov(X_1, X_n) \\ cov(X_2, X_1) & var(X_2) & \dots & cov(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_n, X_1) & cov(X_n, X_2) & \dots & var(X_n) \end{pmatrix} \tag{1.44}$$

Attachment: Independence:

$$X_i || X_j \Rightarrow \begin{cases} f(x_1, x_2, \cdots, x_n) = \prod f(x_i) \\ F(x_1, x_2, \cdots, x_n) = \prod F(x_i) \\ E(\prod X_i) = \prod E(X_i) \\ var(\sum X_i) = \sum var(X_i) \end{cases} \tag{1.45}$$

Section 1.5 PGF, MGF and C.F

Generating Function: Representation of P in function space. $P \Leftrightarrow$ Generating Function.

1.5.1 Probability Generating Function

PGF: used for non-negative, integer X

$$g(s) = E(s^X) = \sum_{j=0}^{\infty} s^j P(X = j), s \in [-1, 1] \tag{1.46}$$

Properties

- $P(X = k) = \dfrac{g^{(k)}(0)}{k!}$
- $E(X) = g^{(1)}(1)$
- $var(X) = g^{(2)}(1) + g^{(1)}(1) - [g^{(1)}(1)]^2$
- For $X_1, X_2, \cdots, X_n$ independent with $g_i(s) = E(s^{X_i}), Y = \sum_{i=1}^{n} X_i$, then
$$g_Y(s) = \prod_{i=1}^{n} g_i(s), s \in [-1, 1] \tag{1.47}$$
- For X_i i.i.d with $\psi(s) = E(s^{X_i}), Y$ with $G(s) = E(s^Y), W = X_1 + X_2 + \cdots + X_Y$,then
$$g_W(s) = G[\psi(s)] \tag{1.48}$$
- 2-Dimensional PGF of (X, Y)
$$g(s, t) = E(s^X t^Y) = \sum_{i=o}^{\infty} \sum_{j=0}^{\infty} P_{(X,Y)}(X = i, Y = j) s^i t^j,\ s, t \in [-1, 1] \tag{1.49}$$

1.5.2 Moment Generating Function

MGF:

$$M_X(s) = E(e^{sX}) = \begin{cases} \sum_j e^{sx} P(X = x_j) \\ \int_{-\infty}^{\infty} e^{sx} f_X(x) \mathrm{d}x \end{cases} \tag{1.50}$$

Properties

- MGF of $Y = aX + b$: $M_Y(s) = e^{sb} M(sa)$
- $E(X^k) = M^{(k)}(0)$
- $P(X = 0) = \lim_{s \to -\infty} M(s)$
- For $X_1, X_2, \cdots, X_n$ independent with $M_{X_i}(s) = E(e^{sX_i}), Y = \sum_{i=1}^{n} X_i$, then
$$M_Y(s) = \prod_{i=1}^{n} M_{X_i}(s) \tag{1.51}$$

1.5.3 Characteristic Function

C.F is actually the Fourier Transform of f.

$$\phi(t) = E(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} f_X(x)\mathrm{d}x \tag{1.52}$$

Properties

- if $E(|X|^k) < \infty$,then
$$\phi^{(k)}(t) = i^k E(X^k e^{itX}) \qquad \phi^{(k)}(0) = i^k E(X^k) \tag{1.53}$$
- For $X_1, X_2, \cdots, X_n$ independent with $\phi_{X_i}(t) = E(e^{itX_i})$, $Y = \sum\limits_{i=1}^{n} X_i$, then
$$\phi_Y(t) = \prod_{i=1}^{n} \phi_{X_i}(t) \tag{1.54}$$
- Inverse (Fourier) Transform
$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx}\phi(t)\mathrm{d}t \tag{1.55}$$

Section 1.6 Convergence and Limit Distribution

1.6.1 Convergence Mode

$$\begin{cases} \text{Convergence in Distribution} & X_n \xrightarrow{\mathscr{L}} X : \lim\limits_{n\to\infty} F_n(x) = F(x) \\ \text{Convergence in Probability} & X_n \xrightarrow{p} X : \lim\limits_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0\,, \forall \varepsilon > 0 \\ \text{Almost Sure Convergence} & X_n \xrightarrow{\text{a.s.}} X : P(\lim\limits_{n\to\infty} X_n = X) = 1 \\ L_p \text{ Convergence} & X_n \xrightarrow{L_p} X : \lim\limits_{n\to\infty} E(|X_n - X|^p) = 0 \end{cases} \tag{1.56}$$

Relations between convergence:

$$\begin{tikzcd}
\boxed{X_n \xrightarrow{\text{a.s.}} X} \arrow[dr] & & \\
& \boxed{X_n \xrightarrow{p} X} \arrow[r] & \boxed{X_n \xrightarrow{\mathscr{L}} X} \\
\boxed{X_n \xrightarrow{L_p} X} \arrow[ur] & &
\end{tikzcd}$$

Useful Thm.:

- Continuous Mapping Thm.: For continuous function $g(\cdot)$

1. $X_n \xrightarrow{\text{a.s.}} X \Rightarrow g(X_n) \xrightarrow{\text{a.s.}} g(X)$

2. $X_n \xrightarrow{p} X \Rightarrow g(X_n) \xrightarrow{p} g(X)$
3. $X_n \xrightarrow{\mathscr{L}} X \Rightarrow g(X_n) \xrightarrow{\mathscr{L}} g(X)$

- Slutsky's Thm.: For $X_n \xrightarrow{\mathscr{L}} X, Y_n \xrightarrow{p} c$
  1. $X_n + Y_n \xrightarrow{\mathscr{L}} X + c$
  2. $X_n Y_n \xrightarrow{\mathscr{L}} cX$
  3. $X_n / Y_n \xrightarrow{\mathscr{L}} X/c$
- Continuity Thm.

$$\lim_{n\to\infty} \phi_n(t) = \varphi(t) \Leftrightarrow X_n \xrightarrow{\mathscr{L}} X \tag{1.57}$$

1.6.2 Law of Large Number & Central Limit Theorem

- WLLN

$$\frac{1}{n}\sum X_i \xrightarrow{p} E(X_1) \tag{1.58}$$

- SLLN

$$\frac{1}{n}\sum X_i \xrightarrow{\text{a.s.}} C \tag{1.59}$$

- CLT

$$\frac{1}{\sigma\sqrt{n}}\sum(X_k - \mu) \xrightarrow{\mathscr{L}} N(0,1) \tag{1.60}$$

- de Moivre-Laplace Thm.

$$P(k \le S_n \le m) \approx \Phi(\frac{m+0.5-np}{\sqrt{npq}}) - \Phi(\frac{k-0.5-np}{\sqrt{npq}}) \tag{1.61}$$

- Stirling Eqa.

$$\frac{\lambda^k}{k!}e^{-\lambda} \approx \frac{1}{\sqrt{\lambda}\sqrt{2\pi}}e^{-\frac{(k-\lambda)^2}{2\lambda}} \xrightarrow[\lambda=n]{k=n} n! \approx \sqrt{2\pi n}(\frac{n}{e})^n \tag{1.62}$$

Section 1.7 Inequalities

- Cauchy-Schwarz Inequality

$$|E(XY)| \le \sqrt{E(X^2)E(Y^2)} \tag{1.63}$$

- Bonferroni Inequality

$$P(\bigcup_{i=1}^{n} A_i) \ge \sum_{1\le i\le n} P(A_i) + \sum_{1\le i<j\le n} P(A_i \cap A_j) \tag{1.64}$$

- Markov Inequality

$$P(|X| \ge \epsilon) \le \frac{E(|X|^\alpha)}{\epsilon^\alpha} \tag{1.65}$$

- Chebyshev Inequality

$$P(|X - E(X)| \ge \epsilon) \le \frac{var(X)}{\epsilon^2} \tag{1.66}$$

- Jensen Inequality: For convex function $g(x)$:

$$E[g(X)] \ge g(E(X)) \tag{1.67}$$

Section 1.8 Multivariate Normal Distribution

General Case and more discussion see section4.2.1.

Distribution of Normal $X(\mu,\sigma^2)$:

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}$$

For $X_1,X_2,\cdots,X_n$ independent and $X_k\sim N(\mu_k,\sigma_k^2)$, $k=1,\cdots,n$, $T=\sum_{k=1}^{n}c_kX_k$, (c_k const), then

$$T\sim N(\sum_{k=1}^{n}c_k\mu_k,\sum_{k=1}^{n}c_k^2\sigma_k^2) \tag{1.68}$$

Deduction in some special cases:

- Given $\mu_1=\mu_2=\cdots=\mu_n=\mu$, $\sigma_1^2=\sigma_2^2=\cdots=\sigma_n^2=\sigma^2$, i.e. X_k i.i.d., then

$$T\sim N(\mu\sum_{k=1}^{n}c_k,\sigma^2\sum_{k=1}^{n}c_k^2) \tag{1.69}$$

- Further take $c_1=c_2=\cdots=c_n=\frac{1}{n}$, i.e. $T=\sum_{k=1}^{n}X_k/n=\bar{X}$, then

$$T=\bar{X}\sim N(\mu,\frac{\sigma^2}{n}) \tag{1.70}$$

1.8.1 Linear Transform

First consider $\epsilon_1,\epsilon_2,\cdots,\epsilon_m$ i.i.d. $\sim N(0,1)$, $n\times 1$ const column vector $\vec{\mu}$, $n\times m$ const matrix $\mathbf{B}=\{b_{ij}\}$, def.$X_i=\sum_{j=1}^{m}b_{ij}\epsilon_j$, i.e.

$$\vec{X}=\begin{pmatrix}X_1\\X_2\\\vdots\\X_n\end{pmatrix}=\begin{pmatrix}b_{11}&b_{12}&\dots&b_{1m}\\b_{21}&b_{22}&\dots&b_{2m}\\\vdots&\vdots&\ddots&\vdots\\b_{n1}&b_{n2}&\dots&b_{nm}\end{pmatrix}\begin{pmatrix}\epsilon_1\\\epsilon_2\\\vdots\\\epsilon_m\end{pmatrix}+\vec{\mu} \tag{1.71}$$

We have: $\vec{X}\sim N(\vec{\mu},\Sigma)$, where Σ, as defined in eqa.1.44 is

$$\Sigma=E[(\vec{X}-\vec{\mu})(\vec{X}-\vec{\mu})^T]=\mathbf{B}\mathbf{B}^T=\begin{pmatrix}var(X_1)&cov(X_1,X_2)&\dots&cov(X_1,X_n)\\cov(X_2,X_1)&var(X_2)&\dots&cov(X_2,X_n)\\\vdots&\vdots&\ddots&\vdots\\cov(X_n,X_1)&cov(X_n,X_2)&\dots&var(X_n)\end{pmatrix}=\{\sigma_{ij}\} \tag{1.72}$$

Furthur Consider $\vec{Y}=(Y_1,\cdots,Y_n)^T$, $n\times n$ const square matrix $\mathbf{A}=\{a_{ij}\}$ and def. $\vec{Y}=\mathbf{A}\vec{X}$ i.e.

$$\begin{pmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{pmatrix}=\begin{pmatrix}a_{11}&a_{12}&\dots&a_{1n}\\a_{21}&a_{22}&\dots&a_{2n}\\\vdots&\vdots&\ddots&\vdots\\a_{n1}&a_{n2}&\dots&a_{nn}\end{pmatrix}\begin{pmatrix}X_1\\X_2\\\vdots\\X_n\end{pmatrix} \tag{1.73}$$

Then $\vec{Y} \sim N(\mathbf{A}\vec{\mu}, \mathbf{A}\Sigma\mathbf{A}^T)$

Special case: $X_1, \cdots, X_n$ i.i.d. $\sim N(\mu, \sigma^2)$, $\vec{X} = (X_1, \cdots, X_n)^T$,

$$\begin{aligned} E(Y_i) =& \mu \sum_{k=1}^{n} a_{ik} \\ var(Y_i) =& \sigma^2 \sum_{k=1}^{n} a_{ik}^2 \\ cov(Y_i, Y_j) =& \sigma^2 \sum_{k=1}^{n} a_{ik}a_{jk} \end{aligned}$$

Specially when $\mathbf{A} = \{a_{ij}\}$ orthonormal, we have $Y_1, \cdots, Y_n$ independent

$$Y_i \sim N(\mu \sum_{k=1}^{n} a_{ik}, \sigma^2) \tag{1.74}$$

1.8.2 Distributions of Function of Normal Variable: χ^2, t & F

Consider $X_1, X_2, \ldots, X_n$ i.i.d. $\sim N(0,1)$; $Y, Y_1, Y_2, \ldots, Y_m$ i.i.d. $\sim N(0,1)$

- χ^2 Distribution: Def. χ^2 distribution with degree of freedom n:

$$\xi = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2 \tag{1.75}$$

  PDF of χ_n^2:

$$g_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2} e^{-x/2} I_{x>0} \tag{1.76}$$

  Properties

  - E and var of $\xi \sim \chi_n^2$

$$E(\xi) = n \qquad var(\xi) = 2n \tag{1.77}$$

  - For independent $\xi_i \sim \chi_{n_i}^2$, $i = 1, 2, \ldots, k$:

$$\xi_0 = \sum_{i=1}^{k} \xi_i \sim \chi^2_{n_1+\ldots+n_k} \tag{1.78}$$

  - Denoted as $\Gamma(\alpha, \lambda)$:

$$\xi = \sum_{i=1}^{n} X_i \sim \Gamma(\frac{n}{2}, \frac{1}{2}) = \chi_n^2 \tag{1.79}$$

- t Distribution: Def. t distribution with degree of freedom n:

$$T = \frac{Y}{\sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n}}} = \frac{Y}{\sqrt{\frac{\xi}{n}}} \sim t_n \tag{1.80}$$

  (Usually take ν instead of n)

  PDF of t_ν:

$$t_\nu(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{1.81}$$

Denote: Upper α-fractile of t_ν, satisfies $P(T \geq c) = \alpha$:

$$c = t_{\nu,\alpha} \tag{1.82}$$

(Similar for χ^2_n and $F_{m,n}$ etc.)

- F Distribution: Def. F distribution with degree of freedom m and n:

$$F = \frac{\sum_{i=1}^m Y_i}{\sum_{i=1}^n X_i} \sim F_{m,n} \tag{1.83}$$

PDF of $F_{m,n}$:

$$f_{m,n}(x) = \frac{\Gamma(\frac{m+n}{2})m^{\frac{m}{2}}n^{\frac{n}{2}}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}x^{\frac{m}{2}-1}(mx+n)^{-\frac{m+n}{2}}I_{x>0} \tag{1.84}$$

Properties

  - If $Z \sim F_{m,n}$, then $\dfrac{1}{Z} \sim F_{n,m}$.
  - If $T \sim t_n$, then $T^2 \sim F_{1,n}$
  - $F_{m,n,1-\alpha} = \dfrac{1}{F_{n,m,\alpha}}$

□ Some useful Lemma (uesd in statistic inference, see section 2.3.3):

- For $X_1, X_2, \ldots, X_n$ independent with $X_i \sim N(\mu_i, \sigma_i^2)$, then

$$\sum_{i=1}^n \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 \sim \chi^2_n \tag{1.85}$$

- For $X_1, X_2, \ldots, X_n$ i.i.d.$\sim N(\mu, \sigma^2)$, then

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1} \tag{1.86}$$

For $X_1, X_2, \ldots, X_m$ i.i.d.$\sim N(\mu_1, \sigma^2)$, $Y_1, Y_2, \ldots, Y_n$ i.i.d.$\sim N(\mu_2, \sigma^2)$, denote sample pooled variance $S^2_\omega = \dfrac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}$, then

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_\omega} \cdot \sqrt{\frac{mn}{m+n}} \sim t_{m+n-2} \tag{1.87}$$

- For $X_1, X_2, \ldots, X_m$ i.i.d.$\sim N(\mu, \sigma^2)$, $Y_1, Y_2, \ldots, Y_n$ i.i.d.$\sim N(\mu_2, \sigma^2)$, then

$$T = \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F_{m-1,n-1} \tag{1.88}$$

- For $X_1, X_2, \ldots, X_n$ i.i.d. $\sim \epsilon(\lambda)$, then

$$2\lambda n\bar{X} = 2\lambda\sum_{i=1}^n X_i \sim \chi^2_{2n} \tag{1.89}$$

Remark: for $X_i \sim \epsilon(\lambda) = \Gamma(1, \lambda) \Rightarrow 2\lambda\sum_{i=1}^n X_i \sim \Gamma(n, 1/2) = \chi^2_{2n}$.

Chapter. II 统计推断部分

Instructor: Jiangdian Wang

Statistical Inference: use sample to estimate population.

Two main tasks of Statistical Inference:

- Parameter Estimation
  - Point Estimation: section 2.2
  - Interval Estimation: section 2.3
- Hypothesis Testing: section 2.4

Section 2.1 Statistical Model and Statistics

Random sample comes from population X. In parametric model case, we have population distribution family:

$$\mathscr{F} = \{f(x;\vec{\theta})|\vec{\theta} \in \Theta\} \tag{2.1}$$

where parameter $\vec{\theta}$ reflect some quantities of population (e.g. mean, variance, etc.), each $\vec{\theta}$ corresponds to a distribution of population X.

Sample space: Def. as $\mathscr{X} = \{\{x_1, x_2, \dots, x_n\}, \forall x_i\}$, then $\{X_i\} \in \mathscr{X}$ is random sample from population $X \sim f(x;\vec{\theta})$.

2.1.1 Statistics

Statistic(s): function of random sample $\vec{T}(X_1, X_2, \dots, X_n)$, **but not a function of parameter**.

Some useful statistics, e.g.

- Sample mean (Consider X_i i.i.d.)

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.2}$$

- Sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{2.3}$$

- Sample moments
  - Origin moment

$$a_{n,k} = \frac{1}{n}\sum_{i=1}^{k} X_i^k \qquad k = 1, 2, 3, \dots \tag{2.4}$$

  - Center moment

$$m_{n,k} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k \qquad k = 2, 3, 4, \dots \tag{2.5}$$

- Order statistics

$$(X_{(1)}, X_{(2)}, \dots, X_{(n)}), \text{ for } X_{(1)} \le X_{(2)} \le \dots \le X_{(n)} \tag{2.6}$$

- Sample p-fractile

$$m_p = X_{(m)}, \quad m = [(n+1)p] \tag{2.7}$$

- Sample coefficient of variation

$$\hat{\nu} = \frac{S}{\bar{X}} \tag{2.8}$$

- Skewness and Kurtosis

$$\hat{g}_1 = \frac{m_{n,3}}{m_{n,2}^{3/2}} \qquad \hat{g}_2 = \frac{m_{n,4}}{m_{n,2}^2} - 3 \tag{2.9}$$

☐ Properties

Statistic T is a function of random sample $\{X_i\}$, thus has distribution (say $g_T(t)$) called **Sampling Distribution**.

For X_i i.i.d. from $X \sim f(x)$ with population mean μ and variance σ^2

- Calculation of sample variance S^2

$$(n-1)S^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \tag{2.10}$$

- E and var of $\bar{X}$ and S^2

$$E(\bar{X}) = \mu \qquad var(\bar{X}) = \frac{\sigma^2}{n} \qquad E(S^2) = \sigma^2 \tag{2.11}$$

Further if X_i i.i.d. from $X \sim N(\mu, \sigma^2)$ where μ and σ^2 unknown.

- Independence of $\bar{X}$ and S^2

$$\bar{X} \text{ and } S^2 \text{ are independent} \tag{2.12}$$

  - Distribution of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \tag{2.13}$$

  - Distribution of $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{2.14}$$

2.1.2 Exponential Family

Def. $\mathscr{F} = \{f(x;\vec{\theta}|\vec{\theta} \in \Theta)\}$ is **Exponential Family** if $f(x;\vec{\theta})$ has the form as

$$f(x;\vec{\theta}) = C(\vec{\theta})h(x)\exp\left[\sum_{i=1}^{k} Q_i(\vec{\theta})T_i(x)\right] \qquad \vec{\theta} \in \Theta \tag{2.15}$$

Canonical Form: Take $Q_i(\vec{\theta}) = \varphi_i$, then $\vec{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_k)$ =$(Q_1(\vec{\theta}), Q_2(\vec{\theta}), \ldots, Q_k(\vec{\theta}))$ is a transform from Θ to Θ^*, s.t. $\mathscr{F}$ has canonical form, i.e.

$$f(x;\vec{\varphi}) = C^*(\vec{\varphi})h(x)\exp\left[\sum_{i=1}^{k} \varphi_i T_i(x)\right] \qquad \vec{\varphi} \in \Theta^* \tag{2.16}$$

Θ^* is canonical parameter space.

☐ Why we need exponential family? Have some nice properties.

2.1.3 Sufficient and Complete Statistics

Note: For simplification, the following parts denote $\vec{\theta}, \vec{T}, \ldots$ as $\theta, T, \ldots$ etc.

► A **Sufficient Statistic** $T(\vec{X})$ for θ contains all the information of sample when infer θ, i.e.

$$f(\vec{X}; T(\vec{X})) = f(\vec{X}; T(\vec{X}), \theta) \tag{2.17}$$

Properties

- **Factorization Thm.** $T(\vec{X})$ is sufficient **if and only if** $f_{\vec{X}}(\vec{x}; \theta) = f(\vec{x}; \theta)$ can be written as

$$f(\vec{x}; \theta) = g[t(\vec{x}); \theta] h(\vec{x}) \tag{2.18}$$

- If $T(\vec{X})$ sufficient, then $T'(\vec{X}) = g[T(\vec{X})]$ also.(require g single-valued and invertible)
- If $T(\vec{X})$ sufficient, then (T, T_1) also.
- Minimal sufficient statistic $T_\theta(\vec{X})$ satisfies

$$\forall \text{ sufficient statistic } S,\ \exists\, q_S(\cdot),\ \text{s.t.} T_\theta = q_S(S) \tag{2.19}$$

  A minimal sufficient statistic not always exists.

  Sufficient & Complete ⇒ Minimal sufficient.

- Usually dimension of $\vec{T_\theta}$ and θ equals.

Sufficient statistic is **not** unique.

► A **Complete Statistic** $T(\vec{X})$ for θ satisfies

$$\forall \theta \in \Theta\, ;\ \forall \varphi \text{ satisfies } E[\varphi(T(\vec{X}))] = 0, \text{ we have } P[\varphi(T) = 0; \theta] = 1 \tag{2.20}$$

Explanation: $T \sim g_T(t)$. Rewrite as

$$\int \varphi(t) g_T(t)\, \mathrm{d}t = 0\ \forall\, \theta \Rightarrow \varphi(T) = 0 \text{ a.s.} \tag{2.21}$$

i.e. span$\{g_T(t); \forall\theta\}$ is a complete space. Or to say that $\nexists$ none-zero $\varphi(t)$ so that $E(\varphi(T)) = 0$ (unbiased estimation)

$$\varphi(T) \neq 0\ \forall\theta \Rightarrow E[\varphi(T(\vec{X}))] \neq 0 \tag{2.22}$$

So make sure the uniqueness of unbiased estimation of $\hat{\theta}$ using T.

Properties

- If $T(\vec{X})$ complete, then $T'(\vec{X}) = g[T(\vec{X})]$ also.(require g measurable)
- A complete statistic not always exists.

- ► An **Ancillary Statistic** $S(\vec{X})$ is a statistic whose distribution does not depend on θ

  Basu Thm.: $\vec{X} = (X_1, X_2, \ldots, X_n)$ is sample from $\mathscr{F} = \{f(x;\theta), \theta \in \Theta\}$. $T(\vec{X})$ is a complete and minimal sufficient statistic, $S(\vec{X})$ is ancillary statistic, then $S(\vec{X}) \parallel T(\vec{X})$.

- ► Exponential family: For $\vec{X} = (X_1, X_2, \ldots, X_n)$ from exponential family with canonical form, i.e.

$$f(\vec{x};\theta) = C(\theta)h(\vec{x})\exp\left[\sum_{i=1}^{k}\theta_i T_i(\vec{x})\right], \quad \theta \in \Theta \tag{2.23}$$

  Then if $\Theta \in \mathbb{R}^k$ interior point exists, then $T(\vec{X}) = (T_1(\vec{X}), T_2(\vec{X}), \ldots, T_k(\vec{X}))$ is sufficient & complete statistic.

Section 2.2 Point Estimation

For parametric distribution family $\mathscr{F} = \{f(x,\theta), \theta \in \Theta\}$, random sample $\vec{X} = (X_1, X_2, \ldots, X_n)$ from $\mathscr{F}$. $g(\theta)$ is a function defined on Θ.

Mission: use sample $\{X_i\}$ to estimate $g(\theta)$, called **Parameter Estimation**.

$$\text{Parameter Estimation}\begin{cases}\text{Point Estimation} & \surd \\ \text{Interval Estimation}\end{cases} \tag{2.24}$$

Point estimation: when estimating θ or $g(\theta)$, denote the estimator (defined on sample space $\mathscr{X}$) as

$$\hat{\theta}(\vec{X}) \qquad \hat{g}(\vec{X}) \tag{2.25}$$

Estimator is a statistic, with sampling distribution.

2.2.1 Optimal Criterion

Some nice properties of estimators (that we expect)

- Unbiasedness

$$E(\hat{\theta}) = \theta \quad \text{or} \quad E(\hat{g}(\vec{X})) = g(\theta) \tag{2.26}$$

  Otherwise, say $\hat{\theta}$ or $\hat{g}$ biased. Def. **Bias**: $E(\hat{\theta}) - \theta$

  Asymptotically unbiasedness

$$\lim_{n\to\infty} E(\hat{g}_n(\vec{X})) = g(\theta) \tag{2.27}$$

- Efficiency: say $\hat{g}_1(\vec{X})$ is more efficient than $\hat{g}_2(\vec{X})$, if

$$var(\hat{g}_1) \le var(\hat{g}_2) \quad \forall\theta \in \Theta \tag{2.28}$$

- Mean Squared Error (MSE): Bias-Variance Trade-Off

$$\text{MSE} = E[(\hat{\theta} - \theta)^2] = var(\hat{\theta}) + [Bias(\hat{\theta})]^2 \tag{2.29}$$

  For unbiased estimator, i.e. $Bias(\hat{\theta}) = 0$, we have

$$\text{MSE} = E[(\hat{\theta} - \theta)^2] = var(\hat{\theta}) \tag{2.30}$$

- (Weak) Consistency

$$\lim_{n\to\infty} P(|\hat{g}_n(\vec{X}) - g(\theta)| \geq \varepsilon) = 0 \quad \forall \varepsilon > 0 \tag{2.31}$$

- Asymptotic Normality

2.2.2 Method of Moments

Review: Population moments & Sample moments

$$\alpha_k = E(X^k) \qquad\qquad \mu_k = E[(X - E(X))^k]$$
$$a_{n,k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k \qquad\qquad m_{n,k} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$$

Property: $a_{n,k}$ is the unbiased estimator of α_k.(while $m_{n,k}$ unually biased for μ_k)

For sample $\vec{X} = (X_1, X_2, \dots, X_n)$ from $\mathscr{F} = \{f(x;\theta, \theta \in \Theta)\}$, unknown parameter (or its function) $g(\theta)$ can be written as

$$g(\theta) = G(\alpha_1, \alpha_2, \dots, \alpha_k; \mu_2, \mu_3, \dots, \mu_l) \tag{2.32}$$

Then its **Moment Estimate** $\hat{g}(\vec{X})$ is

$$\hat{g}(\vec{X}) = G(a_{n,1}, a_{n,2}, \dots, a_{n,k}; m_{n,2}, m_{n,3}, \dots, m_{n,l}) \tag{2.33}$$

Example: coefficient of variance & skewness

$$\hat{\nu} = \frac{S}{\bar{X}} \qquad \hat{\beta}_1 = \frac{m_{n,3}}{m_{n,2^{3/2}}} = \sqrt{n}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{[\sum_{i=1}^{n}(X_i - \bar{X})^2]^{\frac{3}{2}}} \tag{2.34}$$

☐ Note:

- G may not have explicit expression.
- Moment estimate may not be unique.
- If $G = \sum_{i=1}^{k} c_i\alpha_i$ (linear combination of α, without μ), then $\hat{g}(\vec{X}) = \sum_{i=1}^{k} c_i a_{n,i}$ unbiased.

  Usually $\hat{g}(\vec{X})$ is asymptotically unbiased.
- For small sample, not so accurate.
- May not contain all the information about θ, i.e. may not be sufficient statistic.
- Do not require a statistic model.

2.2.3 Maximum Likelihood Estimation

For sample $\vec{X} = (X_1, X_2, \ldots, X_n)$ with distribution $f(\vec{x};\theta)$ from $\mathscr{F} = \{f(x;\theta), \theta \in \Theta\}$, def. **Likelihood Function** $L(\theta;\vec{x})$, defined on Θ (as a function of θ)

$$L(\theta;\vec{x}) = f(\vec{x};\theta) \qquad \theta \in \Theta,\, \vec{x} \in \mathscr{X} \tag{2.35}$$

Also def. log-likelihood function $l(\theta;\vec{x}) = \ln L(\theta;\vec{x})$.

If estimator $\hat{\theta} = \hat{\theta}(\vec{X})$ satisfies

$$L(\hat{\theta};\vec{x}) = \sup_{\theta\in\Theta} L(\theta;\vec{x}), \quad \vec{x} \in \mathscr{X} \tag{2.36}$$

Or equivalently take $l(\theta;\vec{x})$ instead of $L(\theta;\vec{x})$.

Then $\hat{\theta} = \hat{\theta}(\vec{X})$ is a **Maximum Likelihood Estimate**(MLE) of $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$

How to identify MLE?

- Differentiation: Fermat Lemma

$$\left.\frac{\partial L}{\partial \theta_i}\right|_{\theta=\hat{\theta}} = 0 \qquad \left.\frac{\partial^2 L}{\partial \theta_i \partial \theta_j}\right|_{\theta=\hat{\theta}} \text{negative definite} \qquad \forall i, j = 1, 2, \ldots, k \tag{2.37}$$

- Graphing method.
- Numerically compute maximum.

☐ Some properties of MLE

- (Depend on the case, not always) unbiased.
- Invariance of MLE: If $\hat{\theta}$ is MLE of θ, invertible function $g(\theta)$, then $g(\hat{\theta})$ is MLE of $g(\theta)$.
- MLE and Sufficiency: $T = T(X_1, X_2, \ldots, X_n)$ is a sufficient statistic of θ, if MLE of θ exists, say $\hat{\theta}$, then $\hat{\theta}$ is a function of T, i.e.

$$\hat{\theta} = \hat{\theta}(\vec{X}) = \hat{\theta}^*(T(\vec{X})) \tag{2.38}$$

- Asymptotic Normality:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma_\theta^2), \quad \sigma_\theta^2 = \frac{1}{E_\theta[\frac{\partial}{\partial\theta} \ln f(\vec{X};\theta)]^2} \tag{2.39}$$

i.e.

$$\hat{\theta}_n \xrightarrow{d} N(\theta, \frac{\sigma_\theta^2}{n}) \tag{2.40}$$

☐ Comparison: MoM and MLE

- MoM do not require statistic model; MLE need to know PDF.
- MoM is more robust than MLE.

MLE in Exponential Family:

For sample $\vec{X} = (X_1, X_2, \ldots, X_n)$ from canonical exponential family $\mathscr{F} = \{f(x;\theta), \theta \in \Theta\}$

$$f(x;\theta) = C(\theta)h(x)\exp\left[\sum_{i=1}^{k}\theta_i T_i(x)\right] \quad \theta = (\theta_1, \ldots, \theta_k) \in \Theta \tag{2.41}$$

Likelihood function $L(\theta, \vec{x}) = \prod_{j=i}^{n} f(x_j;\theta)$ and log-likelihood function $l(\theta, \vec{x})$

$$L(\theta, \vec{x}) = C^n(\theta)\prod_{j=1}^{n} h(x_j)\exp\left[\sum_{i=1}^{k}\theta_i\sum_{j=1}^{n}T_i(x_j)\right]$$
$$l(\theta, \vec{x}) = n\ln C(\theta) + \sum_{j=1}^{n}\ln h(x_j) + \sum_{i=1}^{k}\theta_i\sum_{j=1}^{n}T_i(x_j)$$

Solution of MLE: (Require $\hat{\theta} \in \Theta$)

$$\frac{n}{C(\theta)}\frac{\partial C(\theta)}{\partial\theta_i}\bigg|_{\theta=\hat{\theta}} = -\sum_{j=1}^{n}T_i(x_j), \quad i = 1, 2, \ldots, k \tag{2.42}$$

2.2.4 Uniformly Minimum Variance Unbiased Estimator

MSE: For $\hat{g}(\vec{X})$ is estimate of $g(\theta)$,then MSE

$$\text{MSE}(\hat{g}(\vec{X})) = E[(\hat{g}(\vec{X}) - g(\theta))^2] = var(\hat{g}) + [Bias(\hat{g})]^2 \tag{2.43}$$

Note: Unbiased estimator (i.e. $Bias(\hat{g}) = 0$) not unique; not always exist.

Now only consider unbiased estimators of $g(\theta)$ exists, say $\hat{g}(\vec{X})$, then

$$\text{MSE}(\hat{g}(\vec{X})) = var(\hat{g}(\vec{X})) \tag{2.44}$$

If $\forall$ unbiased estimate $\hat{g}\prime(\vec{X})$, $\hat{g}$ satisfies

$$var[\hat{g}(\vec{X})] \le var[\hat{g}\prime(\vec{X})] \tag{2.45}$$

☐ Then $\hat{g}(\vec{X})$ is **Uniformly Minimum Variance Unbiased Estimator(UMVUE)** of $g(\theta)$

How to determine UMVUE? (Not an easy task)

- Zero Unbiased Estimate Method
- Sufficient and Complete Statistic Method
- Cramer-Rao Inequality

1. **Zero Unbiased Estimate Method**

   Let $\hat{g}(\vec{X})$ be an unbiased estimate with $var(\hat{g}) < \infty$. If $\forall$ $E(\hat{l}(\vec{X})) = 0$, $\hat{g}$ holds that

$$cov(\hat{g}, \hat{l}) = E(\hat{g}\cdot\hat{l}) = 0, \quad \forall\theta \in \Theta \tag{2.46}$$

   Then $\hat{g}$ is a UMVUE of $g(\theta)$ (sufficient & necessary).

2. **Sufficient and Complete Statistic Method**

   For $T(\vec{X})$ sufficient statistic, $\hat{g}(\vec{X})$ unbiased estimate of $g(\theta)$, then

$$h(T) = E(\hat{g}(\vec{X})|T) \tag{2.47}$$

   is an unbiased estimate of $g(\theta)$ and $var(h(T)) \leq var(\hat{g})$.

   Remark:

   - A method to improve estimator.
   - A UMVUE has to be a function of sufficient statistic.

   Lehmann-Scheffé Thm.: For $\vec{X} = (X_1, X_2, \ldots, X_n)$ from population $X \sim \mathscr{F} = \{f(x, \theta, \theta \vec{\in} \Theta)\}$. $T(\vec{X})$ sufficient and complete, and $\hat{g}(T(\vec{X}))$ be an unbiased estimator, then $\hat{g}(T(\vec{X}))$ is the unique UMVUE.

   Can be used to construct UMVUE: given $T(\vec{X})$ sufficient and complete and some unbiased estimator $\hat{g}\prime(\theta)$ then

$$\hat{g}(T) = E(\hat{g}\prime|T) \tag{2.48}$$

   is the unique UMVUE.

3. **Cramer-Rao Inequality**

   Core idea: determine a lower bound of $var(\hat{g})$.

   Consider $\theta = \theta$ (One dimension parameter); For $\{X_i\}$ i.i.d. $f(x, \theta)$: def.

   - **Score function**: Reflects the steepness/slope of likelihood function f.

$$S(\vec{x};\theta) = \frac{\partial \ln f(\vec{x};\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(x_i;\theta)}{\partial \theta} \tag{2.49}$$

$$E[S(\vec{X};\theta)] = 0 \tag{2.50}$$

   - **Fisher Information**: Variance of $S(\vec{x};\theta)$, reflects the accuracy to conduct estimation, i.e. reflects information of statistic model that sample brings.

$$I(\theta) = E\left[\left(\frac{\partial \ln f(\vec{x};\theta)}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \ln f(\vec{x};\theta)}{\partial \theta^2}\right] \tag{2.51}$$

   Consider $\mathscr{F}$ satisfies some regularity conditions(in most cases, regularity conditions do hold), then the lower bound of $var(\hat{g})$ satisfies **Cramer-Rao Inequality**:

$$var(\hat{g}(\vec{X})) \geq \frac{[g'(\theta)]^2}{nI(\theta)} \tag{2.52}$$

   Special case: $g(\theta) = \theta$ then

$$var(\hat{\theta}) \geq \frac{1}{nI(\theta)} \tag{2.53}$$

   note:

   - C-R Inequality determine a lower bound, not the infimum(i.e. UMVUE$\nRightarrow$ $var(\hat{g}(\vec{X})) = \dfrac{[g'(\theta)]^2}{nI(\theta)}$).

- Take '=': Only some cases in Exponential family.
- **Efficiency**: How good the estimator is.

$$e_{\hat{g}(\vec{X})}(\theta) = \frac{[g'(\theta)]^2/(nI(\theta))}{var(\hat{g}(\vec{X}))} \tag{2.54}$$

4. **Multi-Dimensional Cramer-Rao Inequality** ReDef. Fisher Information:

$$\mathbf{I}(\theta) = \{I_{ij}(\theta)\} = \{E\left[\left(\frac{\partial \ln f(\vec{x};\theta)}{\partial \theta_i}\right)\left(\frac{\partial \ln f(\vec{x};\theta)}{\partial \theta_j}\right)\right]\} \tag{2.55}$$

Then covariance matrix $\Sigma(\theta)$ satisfies **Cramer-Rao Inequality**

$$\Sigma(\theta) \geq (n\mathbf{I}(\theta))^{-1} \tag{2.56}$$

Note: '$\geq$' holds for all diagonal elements, i.e.

$$var(\hat{\theta}_i) \geq \frac{I^*_{ii}(\theta)}{n}, \quad \forall\, i = 1, 2, \ldots, k \tag{2.57}$$

2.2.5 MoM and MLE in Linear Regression

Note: More detailed knowledge see sec.3 Linear Regression Analysis.

☐ Linear Regression Model(1-dimension case):

$$y_i = \beta_0 + \beta_1 x_0 + \epsilon_i \tag{2.58}$$

where β_0, β_1 are regression coefficient, and ϵ_i are unknown random **error**.

Basic Assumptions (Guass-Markov Assumption):

$$\begin{aligned} \text{Zero-Mean: } & \epsilon_i \text{ are i.i.d.} \\ \text{Homogeneity of Variance: } & E(\epsilon_i|x_i) = 0 \\ \text{Independent: } & var(\epsilon_i) = \sigma^2 \end{aligned}$$

Mission: use data $\{(x_i, y_i)\}$ to estimate β_0, β_1(i.e. regression line), and error ϵ_i.

1. OLS (Ordinary Least Squares): Take β_0, β_1 so that MSE min, i.e. SSE min

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.59}$$

(Express in Matrix Notation (eqa.2.74), so that it can be generalized to multidimensional case) SSE can be expressed as the **Eucliean Distance** between $\{y_i\}$ and $\{\hat{\beta}_0 + \hat{\beta}_1 x_i\}$, i.e.

$$\arg\min D(y, X\hat{\beta}) \tag{2.60}$$

i.e. $\hat{\beta}$ is the Projection of y onto hyperplane X, then

$$(X\hat{\beta})^T(y - X\hat{\beta}) = 0 \Rightarrow \hat{\beta} = (X^TX)^{-1}X^Ty \tag{2.61}$$

Solution for 2-D case:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \bar{y} - \hat{\beta}_1 \bar{x} \\ \dfrac{\sum\limits_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \end{bmatrix} \tag{2.62}$$

So get regression line:$y = \hat{\beta}_0 + \hat{\beta}_1 x$

Def. Residuals

$$e_i = \hat{\epsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \tag{2.63}$$

Residuals can be used to estimate ϵ_i: $E[(\epsilon_i)^2] = \sigma^2$

$$\hat{\sigma^2} = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \tag{2.64}$$

2. MoM: Consider r.v. $\epsilon \sim f(\varepsilon; x, y, \beta_0, \beta_1)$, sample $\{\epsilon_i | \epsilon_i = y_i - \beta_0 - \beta_1 x_i\}$, then obviously

$$\bar{\epsilon} = \bar{y} - \beta_0 - \beta_1 \bar{x} \tag{2.65}$$

Take moment estimate of ϵ, we have

$$E(\epsilon_i) = 0 \qquad E(\epsilon_i x_i) = 0 \text{ (note that} E(\epsilon | x) = 0) \tag{2.66}$$

$$\text{i.e.} \begin{cases} \dfrac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \dfrac{1}{n} \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases} \tag{2.67}$$

Solution:

$$\begin{cases} \hat{\beta}_0 & = \bar{y} - \beta_1 \bar{x} \\ \hat{\beta}_1 & = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{cases} \tag{2.68}$$

(Same as OLS)

Moment estimate of σ^2

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \tag{2.69}$$

3. MLE: Assume $\epsilon_i \sim N(0, \sigma^2)$, then $y_i | x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Get likelihood function:

$$L(\beta_0, \beta_1, \sigma^2; x_1, \ldots, x_n, y_1, \ldots, y_n) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)}{2\sigma^2} \right] \tag{2.70}$$

Log-likelihood:

$$l(\beta_0, \beta_1, \sigma^2; x_1, \ldots, x_n, y_1, \ldots, y_n) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.71}$$

MLE, use Fermat Lemma:

$$\begin{cases} \dfrac{\partial l}{\partial \beta_0} = 0 & \Rightarrow -\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \dfrac{\partial l}{\partial \beta_1} = 0 & \Rightarrow -\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \dfrac{\partial l}{\partial \sigma^2} = 0 & \Rightarrow -\dfrac{n}{2}\dfrac{1}{\sigma^2} + \dfrac{1}{2(\sigma^2)^2}\displaystyle\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases} \tag{2.72}$$

Solution:

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \beta_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \end{aligned}$$

□ Linear Regression Model(Multi-dimension case):

Detailed derivation see section 3.3

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \tag{2.73}$$

Denote: $\vec{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$, $\vec{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$, then for each i: $y_i = x_i^T \beta + \epsilon_i$

Further denote: Matrix form:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} = X\vec{\beta} + \vec{\epsilon} \tag{2.74}$$

Basic Assumptions: Gauss-Markov Assumptions

- OLS unbiased

$$E(\epsilon_i|x_i) = 0 \qquad E(y_i|x_i) = x_i^T\beta \tag{2.75}$$

- Homogeneity of ϵ_i

$$var(\epsilon_i) = \sigma^2 \tag{2.76}$$

- Independent of ϵ

- (For MLE) ϵ_i i.i.d. $\sim N(0, \sigma^2)$

Residuals:

$$e_i = \hat{\epsilon}_i = y_i - \hat{y}_i = y_i - x_i^T\beta \tag{2.77}$$

Def. Error Sum of Squares (SSE)

$$\text{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 \tag{2.78}$$

Estimator exists and unique:($\hat{\sigma}^2$ is after bias correction)

$$\begin{aligned}\hat{\beta} &= (X^TX)^{-1}X^Ty \\ \hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 \\ \hat{\sigma}^2 &= \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2\end{aligned} \tag{2.79}$$

2.2.6 Kernel Density Estimation

Given random sample $\{X_i\}$. Def. Empirical CDF.

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i) \tag{2.80}$$

Problem: Overfitting when getting $\hat{f}$. Solution: Using **Kernel Estimate**, replace $I_{(-\infty,x]}(\cdot)$ with Kernel function $K(\cdot)$, then

$$\hat{f}_n(x) = \frac{F_n(x+h_n) - F - n(x-h_n)}{2h_n} = \frac{1}{nh_n}\sum_{i=1}^{n} K(\frac{x-X_i}{h_n}) \tag{2.81}$$

where h_n is **bandwidth**. Take proper kernel function K to get estimate of f.

Can be considered as a convolution of sample $\{X_i\}$ and kernel function K.

Useful Kernel Functions:

- $K(x) = \frac{1}{2}I_{[-\frac{1}{2},\frac{1}{2}]}$
- $K(x) = (1-|x|)I_{[-1,1]}$
- $K(x) = \frac{1}{2\pi}e^{-\frac{x^2}{2}}$
- $K(x) = \frac{1}{\pi(1+x^2)}$
- $K(x) = \frac{1}{2\pi}\text{sinc}^2(\frac{x}{2})$

Section 2.3 Interval Estimation

$$\text{Parameter Estimation}\begin{cases}\text{Point Estimation} \\ \text{Interval Estimation} \quad \surd\end{cases} \tag{2.82}$$

Interval Estimation: to estimate $g(\theta)$, give **two** estimators $\hat{g}_1(\vec{X})$, $\hat{g}_2(\vec{X})$ defined on $\mathscr{X}$ as the two ends of interval (i.e. give an interval $[\hat{g}_1(\vec{X}),\ \hat{g}_2(\vec{X})]$), then random interval $[\hat{g}_1(\vec{X}),\ \hat{g}_2(\vec{X})]$ is an **Interval Estimation** of $g(\theta)$.

2.3.1 Confidence Interval

How to judge an interval estimation?

- Reliability

$$P(g(\theta) \in [\hat{g}_1, \hat{g}_2]) \tag{2.83}$$

- Precision

$$E(\hat{g}_2 - \hat{g}_1) \tag{2.84}$$

Trade off: (in most cases)

Given a level of reliability, find an interval with the highest precision above the level

□ For a given $0 < \alpha < 1$, if

$$P(\hat{g}_1 \le g(\theta) \le \hat{g}_2) \ge 1 - \alpha \tag{2.85}$$

then $[\hat{g}_1, \hat{g}_2]$ is a **Confidence Interval** for $g(\theta)$, with **Confidence Level** $1 - \alpha$.

Confidence Coefficient:

$$\inf_{\forall\theta\in\Theta} P(\theta \in \text{CI}) \tag{2.86}$$

Other cases:

- **Confidence Limit**:Upper/Lower Confidence Limit

$$P(g \le \hat{g}_U) \ge 1 - \alpha$$
$$P(\hat{g}_L \le \theta) \ge 1 - \alpha$$

- **Confidence Region**: For high dimensional parameters $\vec{g} = (g_1, g_2, \ldots, g_k)$

$$P(\vec{g} \in S(\vec{X})) \ge 1 - \alpha \quad \forall\theta \in \Theta \tag{2.87}$$

Mission: Determine $\hat{g}_1, \hat{g}_2$.

2.3.2 Pivot Variable Method

Idea: Based on point estimation, construct a new variable and thus find the interval estimation.

Def. **Pivot Variable** T, satisfies:

- Expression of T contains θ (thus T is not a statistic).
- Distribution of T independent of θ.

In different cases, construct different pivot variable, usually base on sufficient statistics and transform.

Knowing a proper pivot variable $T = T(\hat{\varphi}, g(\theta)) \sim f$, ($f$ is some distribution independent of θ), $\hat{\varphi}$ is a sufficient statistic), then we can take T satisfies:

$$P(f_{1-\frac{\alpha}{2}} \le T \le f_{\frac{\alpha}{2}}) = 1 - \alpha \tag{2.88}$$

Construct the inverse mapping of $T = T(\hat{\varphi}, g(\theta)) \rightleftarrows g(\theta) = T^{-1}(T, \hat{\varphi})$, we get

$$P[T^{-1}(f_{1-\frac{\alpha}{2}}, \hat{\varphi}) \le \hat{g} \le T^{-1}(f_{\frac{\alpha}{2}}, \hat{\varphi})] = 1 - \alpha \tag{2.89}$$

Thus get a confidence interval for θ with confidence coefficient $1 - \alpha$.

2.3.3 Confidence Interval for Common Distributions

Some important properties of χ^2, t and F see section 1.8.2.

1. Single normal population: $\vec{X} = \{X_1, X_2, \ldots, X_n\} \in \mathscr{X}$ i.i.d from Normal Distribution population $N(\mu, \sigma^2)$. Denote sample mean and sample variance:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad S_\mu = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2, \text{(for } \mu \text{ known)} \tag{2.90}$$

Estimating μ & σ^2: construction of pivot variable under different circumstances:

Estimation	Pivot Variable	Confidence Interval
σ^2 known, estimate μ	$T = \dfrac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1)$	$\left[\bar{X} - \dfrac{\sigma}{\sqrt{n}} N_{\frac{\alpha}{2}}, \bar{X} + \dfrac{\sigma}{\sqrt{n}} N_{\frac{\alpha}{2}}\right]$
σ^2 unknown, estimate μ	$T = \dfrac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}$	$\left[\bar{X} - \dfrac{S}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}}, \bar{X} + \dfrac{S}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}}\right]$
μ known, estimate σ^2	$T = \dfrac{nS_\mu^2}{\sigma^2} \sim \chi_n^2$	$\left[\dfrac{nS_\mu^2}{\chi^2_{n,\frac{\alpha}{2}}}, \dfrac{nS_\mu^2}{\chi^2_{n,1-\frac{\alpha}{2}}}\right]$
μ unknown, estimate σ^2	$T = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$	$\left[\dfrac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}, \dfrac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right]$

2. Double normal population: $\vec{X} = \{X_1, X_2, \ldots, X_m\}$ i.i.d. from $N(\mu_1, \sigma_1^2)$; $\vec{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ i.i.d. from $N(\mu_2, \sigma_2^2)$

   Denote sample mean, sample variance and pooled sample variance:

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{n} X_i \qquad S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2 \qquad S_{\mu_1}^2 = \frac{1}{m}\sum_{i=1}^{m}(X_i - \mu_1)^2, (\mu_1 \text{ known})$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad S_{\mu_2}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_2)^2, (\mu_2 \text{ known})$$

$$S_\omega^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

   Estimating $\mu_1 - \mu_2$:

   When $\sigma_1^2 \neq \sigma_2^2$ unknown, estimate $\mu_1 - \mu_2$: Behrens-Fisher Problem, remain unsolved, but can deal with simplified cases.

Estimation	Pivot Variable	Confidence Interval
σ_1^2 & σ_2^2 known, estimate $\mu_1-\mu_2$	$T=\dfrac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}}}\sim N(0,1)$	$\left[\bar{X}-\bar{Y}-N_{\frac{\alpha}{2}}\sqrt{\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}},\ \bar{X}-\bar{Y}+N_{\frac{\alpha}{2}}\sqrt{\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}}\right]$
$\sigma_1^2=\sigma_2^2$ unknown, estimate $\mu_1-\mu_2$	$T=\dfrac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_\omega\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}}\sim t_{m+n-2}$	$\left[\bar{X}-\bar{Y}-S_\omega t_{m+n-2,\frac{\alpha}{2}}\sqrt{\dfrac{1}{m}+\dfrac{1}{n}},\ \bar{X}-\bar{Y}+S_\omega t_{m+n-2,\frac{\alpha}{2}}\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}\right]$
Welch's t-Interval (when m, n large enough)	$T=\dfrac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\dfrac{S_X^2}{m}+\dfrac{S_Y^2}{n}}}\xrightarrow{\mathscr{L}}N(0,1)$	$\left[\bar{X}-\bar{Y}-N_{\frac{\alpha}{2}}\sqrt{\dfrac{S_1^2}{m}+\dfrac{S_2^2}{n}},\ \bar{X}-\bar{Y}+N_{\frac{\alpha}{2}}\sqrt{\dfrac{S_1^2}{m}+\dfrac{S_2^2}{n}}\right]$

Estimating $\dfrac{\sigma_1^2}{\sigma_2^2}$:

Estimation	Pivot Variable	Confidence Interval
μ_1,μ_2 known, estimate $\dfrac{\sigma_1^2}{\sigma_2^2}$	$T=\dfrac{S_{\mu_2}^2}{S_{\mu_1}^2}\dfrac{\sigma_1^2}{\sigma_2^2}\sim F_{n,m}$	$\left[\dfrac{S_{\mu_1}^2}{S_{\mu_2}^2}\dfrac{1}{F_{m,n,\frac{\alpha}{2}}},\dfrac{S_{\mu_1}^2}{S_{\mu_2}^2}\dfrac{1}{F_{m,n,1-\frac{\alpha}{2}}}\right]$ or $\left[\dfrac{S_{\mu_1}^2}{S_{\mu_2}^2}F_{m,n,\frac{\alpha}{2}},\dfrac{S_{\mu_1}^2}{S_{\mu_2}^2}F_{n,m,\frac{\alpha}{2}}\right]$
μ_1,μ_2 unknown, estimate $\dfrac{\sigma_1^2}{\sigma_2^2}$	$T=\dfrac{S_Y^2}{S_X^2}\dfrac{\sigma_1^2}{\sigma_2^2}\sim F_{n-1,m-1}$	$\left[\dfrac{S_X^2}{S_Y^2}\dfrac{1}{F_{m-1,n-1,\frac{\alpha}{2}}},\dfrac{S_X^2}{S_Y^2}\dfrac{1}{F_{m-1,n-1,1-\frac{\alpha}{2}}}\right]$ or $\left[\dfrac{S_X^2}{S_Y^2}\dfrac{1}{F_{m-1,n-1,\frac{\alpha}{2}}},\dfrac{S_X^2}{S_Y^2}F_{n-1,m-1,\frac{\alpha}{2}}\right]$

3. Non-normal population:

Estimation	Pivot Variable	Confidence Interval
Uniform Distribution: $\vec{X}$ i.i.d. from $U(0,\theta)$	$T=\dfrac{X_{(n)}}{\theta}\sim U(0,1)$	$\left[X_{(n)},\dfrac{X_{(n)}}{\sqrt[n]{\alpha}}\right]$
Exponential Distribution: $\vec{X}$ i.i.d. from $\epsilon(\lambda)$	$T=2n\lambda\bar{X}\sim\chi_{2n}^2$	$\left[\dfrac{\chi_{2n,1-\frac{\alpha}{2}}^2}{2n\bar{X}},\dfrac{\chi_{2n,\frac{\alpha}{2}}^2}{2n\bar{X}}\right]$
Bernoulli Distribution: $\vec{X}$ i.i.d. from $B(1,\theta)$	$T=\dfrac{\sqrt{n}(\bar{X}-\theta)}{\sqrt{\bar{X}(1-\bar{X})}}\xrightarrow{\mathscr{L}}N(0,1)$	$\left[\bar{X}-N_{\frac{\alpha}{2}}\sqrt{\dfrac{\bar{X}(1-\bar{X})}{n}},\bar{X}+N_{\frac{\alpha}{2}}\sqrt{\dfrac{\bar{X}(1-\bar{X})}{n}}\right]$
Poisson Distribution: $\vec{X}$ i.i.d. from $P(\lambda)$	$T=\dfrac{\sqrt{n}(\bar{X}-\lambda)}{\sqrt{\bar{X}}}\xrightarrow{\mathscr{L}}N(0,1)$	$\left[\bar{X}-N_{\frac{\alpha}{2}}\sqrt{\dfrac{\bar{X}}{n}},\bar{X}+N_{\frac{\alpha}{2}}\sqrt{\dfrac{\bar{X}}{n}}\right]$

4. General Case: Use asymptotic normality of MLE to construct CLT for large sample. MLE of θ satisfies:

$$\sqrt{n}(\hat{\theta}^* - \theta) \xrightarrow{\mathscr{L}} N(0, \frac{1}{I(\theta)}) \tag{2.91}$$

where $\hat{\theta}^*$ is MLE of θ. Replace $\dfrac{1}{I(\theta)}$ by $\sigma^2(\hat{\theta}^*)$, then

$$T = \frac{\sqrt{n}(\hat{\theta}^* - \theta)}{\sigma(\hat{\theta}^*)} \xrightarrow{\mathscr{L}} N(0,1) \tag{2.92}$$

confidence interval:

$$\left[\hat{\theta}^* - \frac{N_{\frac{\alpha}{2}}}{\sqrt{n}}\sigma(\hat{\theta}^*), \hat{\theta}^* + \frac{N_{\frac{\alpha}{2}}}{\sqrt{n}}\sigma(\hat{\theta}^*)\right] \tag{2.93}$$

2.3.4 Fisher Fiducial Argument*

Idea: When sample is known, we can get '**Fiducial Probability**' of θ, thus can find an interval estimation based on fiducial distribution.(Similar to the idea of MLE)

Remark: Fiducial probability (denoted as $\tilde{P}(\theta)$) is 'probability of parameter', in the case that sample is known. **Fiducial probability is different from Probability**.

Thus get

$$\tilde{P}(\hat{g}_1 \leq g(\theta) \leq \hat{g}_2) = 1 - \alpha \tag{2.94}$$

Section 2.4 Hypothesis Testing

Hypothesis is a statement about the characteristic of population, e.g. distribution form, parameters, etc.

Mission: Use sample to test the hypothesis, i.e. judge whether population has some characteristic.

2.4.1 Basic Concepts

Parametric hypothesis testing.

For random sample $\vec{X} = (X_1, X_2, \ldots, X_n) \in \mathscr{X}$ i.i.d. from $\mathscr{F} = \{f(x;\theta); \theta \in \Theta\}$

- Null Hypothesis H_0 & Alternative Hypothesis H_1: Wonder whether a statement is true. Def. **Null Hypothesis**: $H_0 : \theta \in \Theta_0 \subset \Theta$, **a statement that we try to reject based on sample**; $H_1 : \theta \in \Theta_1 = \Theta - \Theta_0$ is **Alternative Hypothesis**.

  ☐ Note: **Cannot** exchange H_0 and H_1, because when the evidence is ambiguity, we have to accept H_0, regardless of what H_0 is. So it is **very important** to pick the proper H_0.

  Thus Hypothesis Testing:

$$H_0 : \theta \in \Theta_0 \longleftrightarrow H_1 : \theta \in \Theta_1 \tag{2.95}$$

- Rejection Region R & Acceptance Region R^C: Judge whether to reject H_0 from sample, Def. **Rejection Region**:

$$R \subset \mathscr{X}: \text{reject } H_0 \text{ if } \vec{X} \in R \tag{2.96}$$

  Acceptance Region: accept H_0 if $\vec{X} \in R^C$

- Test Function: Describe how to make a decision.
  - Continuous Case:

$$\varphi(\vec{X}) = \begin{cases} 1, & \vec{X} \in R \\ 0, & \vec{X} \in R^C \end{cases} \tag{2.97}$$

    i.e. $R = \{\vec{X} : \varphi(\vec{X}) = 1\}$. Where R to be determined.
  - Discrete Case: Randomized Test Function

$$\varphi(\vec{X}) = \begin{cases} 1, & \vec{X} \in R - \partial R \\ r, & \vec{X} \in \partial R \\ 0, & \vec{X} \in R^C \end{cases} \tag{2.98}$$

    Where R and r to be determined.

- Type I Error & Type II Error: Sample is random, possible to make a wrong judge.
  - Type I Error (弃真): H_0 is true but sample falls in R, thus H_0 is rejected.

$$P(\text{type I error}) = P(\vec{X} \in R|H_0) = \alpha(\theta) \tag{2.99}$$

  - Type II Error (取伪): H_0 is wrong but sample falls in R^C, thus H_0 is accepted.

$$P(\text{type II error}) = P(\vec{X} \notin R|H_1) = \beta(\theta) \tag{2.100}$$

	Judgement		
		Accept H_0	Reject H_0
Real Case	H_0	$\surd$	Type I Error
	H_1	Type II Error	$\surd$

表 1: 'Confusion Matrix'

Impossible to make probability of Type I & II Error small simultaneously, how to pick a proper test $\varphi(\vec{x})$?

☐ **Neyman-Pearson Principle**: First control $\alpha \le \alpha_0$, then take min β.

How to determine α_0? Depend on specific problem.[1]

- p-value: probability to get larger bias than observed $\vec{x}_0$ under H_0.

  e.g. For reject region $R = \{\vec{X}|T(\vec{X}) \ge C\}$, p-value:

$$p(\vec{x}) = P[T(\vec{X}) \ge t(\vec{x}_0)|H_0] \tag{2.101}$$

  Remark: Under H_0, the probability to get a **worse** result than $\vec{x}_0$.

  Rule: Reject H_0 if $p(\vec{x}_0) \le \alpha_0$

[1] In most cases, take $\alpha_0 = 0.05$.

- Power Function: (when H_0 is given), probability to reject H_0 by sampling.

$$\pi(\theta)=\begin{cases}P(\text{type I error}), & \theta\in\Theta_0\\ 1-P(\text{type II error}), & \theta\in\Theta_1\end{cases}=\begin{cases}\alpha(\theta), & \theta\in\Theta_0\\ 1-\beta(\theta), & \theta\in\Theta_1\end{cases} \tag{2.102}$$

Express as test function:

$$\pi(\theta)=E[\varphi(\vec{X})|\theta] \tag{2.103}$$

A nice test: $\pi(\theta)$ small under H_0, large under H_1 (and grows very fast at the boundary of H_0 and H_1).

☐ **General Steps of Hypothesis Testing:**

1. Propose H_0 & H_1.
2. Determine R (usually in the form of a statistic, e.g. $R=\{\vec{X}:T(\vec{X})\geq c\}$).
3. Select a proper α (to determine c).
4. Sampling, get sample (as well as $t(\vec{x})$), then
   - compare with R and determine whether to reject/accept H_0, or
   - calculate p-value and determine whether to reject/acceptH_0

2.4.2 Hypothesis Testing of Common Distributions

For some common distribution populations, determine rejection region R under certain H_0 with confidence coefficient α.

Definition of necessary statistics see section 2.3.3.

1. Single normal population:

Condition	H_0	H_1	Testing Statistic T	Rejection Region R
σ^2 known, test μ	$\mu=\mu_0$	$\mu\neq\mu_0$	$T=\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}\sim N(0,1)$	$\|T\|>N_{\frac{\alpha}{2}}$
	$\mu\leq\mu_0$	$\mu>\mu_0$		$T>N_\alpha$
	$\mu\geq\mu_0$	$\mu<\mu_0$		$T<-N_\alpha$
σ^2 unknown, test μ	$\mu=\mu_0$	$\mu\neq\mu_0$	$T=\frac{\sqrt{n}(\bar{X}-\mu_0)}{S}\sim t_{n-1}$	$\|T\|>t_{n-1,\frac{\alpha}{2}}$
	$\mu\leq\mu_0$	$\mu>\mu_0$		$T>t_{n-1,\alpha}$
	$\mu\geq\mu_0$	$\mu<\mu_0$		$T<-t_{n-1,\alpha}$
μ known, test σ^2	$\sigma^2=\sigma_0^2$	$\sigma^2\neq\sigma_0^2$	$T=\frac{nS_\mu^2}{\sigma_0^2}\sim\chi_n^2$	$T<\chi^2_{n,1-\frac{\alpha}{2}}\cup T>\chi^2_{n,\frac{\alpha}{2}}$
	$\sigma^2\leq\sigma_0^2$	$\sigma^2>\sigma_0^2$		$T>\chi^2_{n,\alpha}$
	$\sigma^2\geq\sigma_0^2$	$\sigma^2<\sigma_0^2$		$T<\chi^2_{n,1-\alpha}$
μ unknown, test σ^2	$\sigma^2=\sigma_0^2$	$\sigma^2\neq\sigma_0^2$	$T=\frac{(n-1)S^2}{\sigma_0^2}\sim\chi^2_{n-1}$	$T<\chi^2_{n-1,1-\frac{\alpha}{2}}\cup T>\chi^2_{n-1,\frac{\alpha}{2}}$
	$\sigma^2\leq\sigma_0^2$	$\sigma^2>\sigma_0^2$		$T>\chi^2_{n-1,\alpha}$
	$\sigma^2\geq\sigma_0^2$	$\sigma^2<\sigma_0^2$		$T<\chi^2_{n-1,1-\alpha}$

2. Double normal population:

Condition	H_0	H_1	Testing Statistic T	Rejection Region R
σ_1^2, σ_2^2 known, test $\mu_1 - \mu_2$	$\mu_1 - \mu_2 = \mu_0$ $\mu_1 - \mu_2 \leq \mu_0$ $\mu_1 - \mu_2 \geq \mu_0$	$\mu_1 - \mu_2 \neq \mu_0$ $\mu_1 - \mu_2 > \mu_0$ $\mu_1 - \mu_2 < \mu_0$	$T = \dfrac{\bar{X} - \bar{Y} - \mu_0}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}} \sim N(0,1)$	$\|T\| > N_{\frac{\alpha}{2}}$ $T > N_\alpha$ $T < -N_\alpha$
σ_1^2, σ_2^2 unknown, test $\mu_1 - \mu_2$	$\mu_1 - \mu_2 = \mu_0$ $\mu_1 - \mu_2 \leq \mu_0$ $\mu_1 - \mu_2 \geq \mu_0$	$\mu_1 - \mu_2 \neq \mu_0$ $\mu_1 - \mu_2 > \mu_0$ $\mu_1 - \mu_2 < \mu_0$	$T = \dfrac{\bar{X} - \bar{Y} - \mu_0}{S_\omega}\sqrt{\dfrac{mn}{m+n}} \sim t_{m+n-2}$	$\|T\| > t_{m+n-2,\frac{\alpha}{2}}$ $T > t_{m+n-2,\alpha}$ $T < -t_{m+n-2,\alpha}$
μ_1, μ_2 known, test $\dfrac{\sigma_1^2}{\sigma_2^2}$	$\sigma_1^2 = \sigma_2^2$ $\sigma_1^2 \geq \sigma_2^2$ $\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$ $\sigma_1^2 < \sigma_2^2$ $\sigma_1^2 > \sigma_2^2$	$T = \dfrac{S_{\mu_2}^2}{S_{\mu_1}^2} \sim F_{n,m}$	$T < F_{n,m,1-\frac{\alpha}{2}}$ $\cup T > F_{n,m,\frac{\alpha}{2}}$ $T > F_{n,m,\alpha}$ $T < F_{n,m,1-\alpha}$
μ_1, μ_2 unknown, test $\dfrac{\sigma_1^2}{\sigma_2^2}$	$\sigma_1^2 = \sigma_2^2$ $\sigma_1^2 \geq \sigma_2^2$ $\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$ $\sigma_1^2 < \sigma_2^2$ $\sigma_1^2 > \sigma_2^2$	$T = \dfrac{S_2^2}{S_2^2} \sim F_{n-1,m-1}$	$T < F_{n-1,m-1,1-\frac{\alpha}{2}}$ $\cup T > F_{n-1,m-1,\frac{\alpha}{2}}$ $T > F_{n-1,m-1,\alpha}$ $T < F_{n-1,m-1,1-\alpha}$

3. None normal population:

Condition	H_0	H_1	Testing Statistic T	Rejection Region R
$\vec{X}$ from $B(1,p)$, test p	$p = p_0$	$p \neq p_0$	$T = \dfrac{\sqrt{n}(\bar{X} - p_0)}{\sqrt{p_0(1-p_0)}} \xrightarrow{\mathscr{L}} N(0,1)$	$\|T\| > N_{\frac{\alpha}{2}}$
$\vec{X}$ from $P(\lambda)$, test λ	$\lambda = \lambda_0$	$\lambda \neq \lambda_0$	$T = \dfrac{\sqrt{n}(\bar{X} - \lambda_0)}{\sqrt{\lambda_0}} \xrightarrow{\mathscr{L}} N(0,1)$	$\|T\| > N_{\frac{\alpha}{2}}$

4. More than two normal population: Analysis of Variance.

2.4.3 Likelihood Ratio Test

Idea: To test $H_0 : \theta \in \Theta_0 \longleftrightarrow H_1 : \theta \in \Theta_1$ known $\vec{x}$, examine the likelihood function $L(\theta; \vec{x})$ and **compare** $L_{\theta\in\Theta_0}$ and $L_{\theta\in\Theta}$ to see the likelihood that H_0 is true.

Def. **Likelihood Ratio** (LR):

$$\Lambda(\vec{x}) = \frac{\sup\limits_{\theta\in\Theta_0} L(\theta; \vec{x})}{\sup\limits_{\theta\in\Theta} L(\theta; \vec{x})} \tag{2.104}$$

Reject H_0 if $\Lambda(\vec{x}) < \Lambda_0$. Or equivalently: Reject H_0 if $-2\ln\Lambda(\vec{x}) > C(= -2\ln\Lambda_0)$.

where Λ_0 (or equivalently $C = -2\ln\Lambda_0$) satisfies:

$$E_{\Theta_0}[\varphi(\vec{X})] \leq \alpha, \quad \forall\theta \in \Theta_0 \tag{2.105}$$

LR and sufficient statistic: $\Lambda(\vec{x})$ can be expressed as $\Lambda(\vec{x}) = \Lambda^*(T(\vec{x}))$, where $T(\vec{X})$ is sufficient statistic.

☐ LRT for one-sample t-test: For $X_1, X_2, \ldots, X_n$ i.i.d. $\sim N(\mu, \sigma^2)$, test

$$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu \neq \mu_0 \quad \text{when } \sigma^2 \text{ unknown}$$

Can prove:

$$\Lambda^{2/n} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}{\sum\limits_{i=1}^{n}(x_i - \mu_0)^2}$$

Denote $T = \dfrac{\sqrt{n}(\bar{x} - \mu_0)}{S}$, then LRT is

$$\Lambda = \left(1 + \frac{T^2}{n-1}\right)^{-n/2}$$

The Multivariate case see section 4.2.4, where T^2 itself is the Hotelling's T^2 statistic.

☐ Limiting Distribution of LRT: Wilks' Thm.

If $\dim \Theta = k > \dim \operatorname{span}\{\Theta_0\} = s$[2], then under $H_0 : \theta \in \Theta_0$:

$$\Lambda_{\theta \in \Theta_0}(\vec{x}) = -2 \ln \lambda(\vec{x}) \xrightarrow{\mathscr{L}} \chi^2_{k-s} \tag{2.106}$$

2.4.4 Uniformly Most Powerful Test

Idea: Neyman-Pearson Principle: control α, find $\min \beta$. i.e. control α, find $\max \pi(\theta)$

Def. **Uniformly Most Powerful Test** (UMP) φ_{UMP} with level of significance α satisfies

$$\pi_{\text{UMP}}(\theta) \geq \pi(\theta),\ \forall \theta \in \Theta_1 \tag{2.107}$$

Neyman-Pearson Lemma: For $\vec{X} = (X_1, X_2, \ldots, X_n)$ i.i.d. from $f(\vec{x}; \theta)$.

Test hypothesis $H_0 : \theta = \theta_0 \longleftrightarrow H_1 : \theta = \theta_1$. Def. test function φ as:

$$\varphi(\vec{x}) = \begin{cases} 1, & \dfrac{f(\vec{x};\theta_1)}{f(\vec{x};\theta_0)} > C \\ r, & \dfrac{f(\vec{x};\theta_1)}{f(\vec{x};\theta_0)} = C \\ 0, & \dfrac{f(\vec{x};\theta_1)}{f(\vec{x};\theta_0)} < C \end{cases} \tag{2.108}$$

Then there exists C and r such that

- $E[\varphi(\vec{x})|\theta_0] = P(\dfrac{f(\vec{x};\theta_1)}{f(\vec{x};\theta_0)} > C) + rP(\dfrac{f(\vec{x};\theta_1)}{f(\vec{x};\theta_0)} = C) = \alpha$
- This φ is UMP of level of significance α

Actually kind of 1-dimensional case of LRT.

Note: UMT exist for **simple** H_0, H_1, otherwise may not exist.

UMP and sufficient statistics: Test function $\varphi(\vec{X}$ given by eqa(2.108) is function of sufficient statistics $T(\vec{X})$, i.e. $\varphi(\vec{X}) = \varphi^*(T(\vec{X}))$.

[2]Here 'dimension' refers to 'degree of freedom'.

UMP and Exponential Family: For sample $\vec{X} = (X_1, X_2, \dots, X_n)$ from exponential family:

$$f(\vec{x};\theta) = C(\theta)h(\vec{x})\exp\{Q(\theta)T(\vec{x})\} \tag{2.109}$$

Test single hypothesis $H_0 : \theta = \theta_0 \longleftrightarrow H_1 : \theta = \theta_1$, (where $\theta_0 < \theta_1$). If

- θ_0 is inner point of Θ
- $Q(\theta)$ monotone increase with θ

Then UMP exists, in the form of:

$$\varphi(\vec{x}) = \begin{cases} 1, & T(\vec{x}) > C \\ r, & T(\vec{x}) = C \\ 0, & T(\vec{x}) < C \end{cases} \tag{2.110}$$

where C and r satisfies $E[\varphi(\vec{x})|\theta_0] = \alpha$.

Note: or take $Q(\theta)$ mono decreased, then in eqa(2.110), take opposite inequality operators.

☐ **General Steps of UMP**:

1. Find a point $\theta_0 \in \Theta_0$ and a point $\theta_1 \in \Theta_1$. (Note: **one** point)
2. Construct test function in the form of eqa.(2.108), use $E[\varphi(\vec{x})|\theta_0] = \alpha$ to determine C and r.
3. Get R and $\varphi(\vec{x})$.
4. If φ does **not** depend on θ_1, then H_1 can be generalized to $H_1 : \theta \in \Theta_1$.
5. If φ satisfies $E_{\theta\in\Theta_0}(\varphi) \leq \alpha$, then H_0 an be generalized to $H_0 : \theta \in \Theta_0$.

2.4.5 Duality of Hypothesis Testing and Interval Estimation

- Thm.: $\forall\theta_0 \in \Theta$ there exists hypothesis testing $H_0 : \theta = \theta_0 \longleftrightarrow H_1 : \theta \neq \theta_0$ of level α with rejection region R_{θ_0}. Then

  $$C(\vec{X}) = \{\theta : \vec{X} \in R_\theta^C\} \tag{2.111}$$

  is a $1 - \alpha$ confidence region for θ

- Thm.: $C(\vec{X})$ is a $1 - \alpha$ confidence region for θ. Then $\forall\theta_0 \in C(\vec{X})$, the rejection region of hypothesis testing $H_0 : \theta = \theta_0 \longleftrightarrow H_1 : \theta \neq \theta_0$ of level α satisfies

  $$R_{\theta_0}^C = \{\vec{X} : \theta_0 \in C(\vec{X})\} \tag{2.112}$$

☐ Idea:

$$H_0 : \theta = \theta_0 \longleftrightarrow H_1 : \theta \neq \theta_0$$

$$\updownarrow \tag{2.113}$$

$$P(R^C(\vec{X})|H_0) = P(R^C(\vec{X})|\theta_0) = 1 - \alpha$$

$$\updownarrow \tag{2.114}$$

$$\text{Confidence Interval: } \theta_0 \in R^C(\vec{X})$$

Similar for Confidence Limit and One-Sided Testing.

2.4.6 Introduction to Non-Parametric Hypothesis Testing

Motivation: Usually distribution form unknown, cannot use parametric hypothesis testing.

Useful Method:

- Sign Test: Used for paired comparison $\vec{X} = (X_1, X_2, \ldots, X_n), \vec{Y} = (Y_1, Y_2, \ldots, Y_n)$.

  Take $Z_i = Y_i - X_i$ i.i.d., denote $E(Z) = \mu$. Test $H_0 : \mu = 0 \longleftrightarrow H_1 : \mu \neq 0$.

  Denote $n_+ = \#(\text{positive } Z_i)$ and $n_- = \#(\text{negative } Z_i)$, $n_0 = n_+ + n_-$. Then $n_+ \sim B(n_0, \theta)$, test $H_0 : \theta = \frac{1}{2} \longleftrightarrow H_1 : \theta \neq \frac{1}{2}$

  Then use Binomial Testing or large sample CLT Normal Testing.

  Remark:

  - Also can test $H_0 : \theta \leq \frac{1}{2} \longleftrightarrow H_1 : \theta > \frac{1}{2}$
  - Drawback: ignores magnitudes.

- Wilcoxon Signed Rank Sum Test: Improvement of Sign Test. Base on order statistics.

  Order Statistics of Z_i: $Z_{(1)} < Z_{(2)} < \ldots < Z_{(n)}$, where each $Z_{(j)}$ corresponds to some Z_i, denote as $Z_i = Z_{(R_i)}$, then R_i is the rank of Z_i.[3]

  Def. $\vec{R} = (R_1, R_2, \ldots, R_n)$ is **Rank Statistics** of $(Z_1, Z_2, \ldots, Z_n)$

  Def. **Sum of Wilcoxon Signed Rank**:

$$W^+ = \sum_{i=1}^{n_0} R_i I_{Z_i > 0} \tag{2.115}$$

  Distribution of W^+ is complex. E and var of W^+ under H_0:

$$E(W^+) = \frac{n_0(n_0+1)}{4} \qquad var(W^+) = \frac{n_0(n_0+1)(2n_0+1)}{24} \tag{2.116}$$

  Usually consider large sample CLT, construct normal approximation:

$$T = \frac{W^+ - E(W^+)}{\sqrt{var(W^+)}} \xrightarrow{\mathscr{L}} N(0,1) \tag{2.117}$$

  Rejection Region: $R = \{|T| > N_{\frac{\alpha}{2}}\}$

- Wilcoxon Two-Sample Rank Sum Test: Used for two independent sample comparison.

  Assume $\vec{X} = (X_1, \ldots, X_m)$ i.i.d. $\sim f(x)$; $\vec{Y} = (Y_1, \ldots, Y_n)$ i.i.d. $\sim f(x - \theta)$, test $H_0 : \theta = 0 \longleftrightarrow H_1 : \theta \neq 0$.

  Rank X_i and Y_i as:

$$Z_1 \leq Z_2 \leq \ldots \leq Z_{m+n} \tag{2.118}$$

[3] If some $X_i, X_j, \ldots$ equal, then take same rank $R = \text{mean}\{R_i, R_j, \ldots\}$.

in which denote rank of Y_i as R_i, and def. **Wilcoxon two-sample rank sum**:

$$W = \sum_{i=1}^{n} R_i \tag{2.119}$$

E and var of W under H_0:

$$E(W) = \frac{n(m+n+1)}{2} \qquad var(W) = \frac{mn(n+m+1)}{12} \tag{2.120}$$

Use large sample approximation, construct CLT:

$$T = \frac{W - E(W)}{\sqrt{var(W)}} \xrightarrow{\mathscr{L}} N(0,1) \tag{2.121}$$

- Goodness-of-Fit Test: For $\vec{X} = (X_1, X_2, \dots, X_n)$ i.i.d. from some certain population X. Test $H_0 : X \sim F(x)$. where F is theoretical distribution, can be either parametric or non-parametric.

  Idea: Define some *quantity* $D = D(X_1, \dots, X_n; F)$ to measure the difference between F and sample. And def. *Goodness-of-fit* when observed value of D (say d_0) is given:

  $$p(d_0) = P(D \geq d_0 | H_0) \tag{2.122}$$

  Goodness-of-Fit Test: Reject H_0 if $p(d_0) < \alpha$.

  Pearson χ^2 Test: Usually used for discrete case.

  Test $H_0 : P(X_i = a_i) = p_i$, $i = 1, 2, \dots, r$. Denote $\#(X_j = a_i) = \nu_i$, take D as:

  $$K_n = K_n(X_1, \dots, X_n; F) = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} \tag{2.123}$$

  Pearson Thm.: For K_n defined as eqa(2.123), then under H_0:

  $$K_n \xrightarrow{\mathscr{L}} \chi^2_{r-1-s} \tag{2.124}$$

  Here s is number of unknown parameter, $r - 1 - s$ is the degree of freedom.

  Note:

  - a_i must **not** depend on sample.
  - For continuous case, construct division:

    $$\mathbb{R} \to (-\infty, a_1, a_2, \dots, a_{r-1}, \infty = a_r) \tag{2.125}$$

    and test $H_0 : P(X \in I_j) = p_j$

    Criterion: Pick proper interval so that np_i and ν_i both ≥ 5.

- Contingency Table Independence & Homogeneity Test

– Independence Test:

Test a two-parameter sample and to see whether these two parameters(features) are independent. Denote $Z = (X, Y)$ are some 'level' of sample, n_{ij} is number of sample with level (i, j)

Contingency Table:

X \ Y	1	$\dots$	j	$\dots$	s	$\sum$
1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1s}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{is}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	n_{r1}	$\dots$	n_{rj}	$\dots$	n_{rs}	$n_{r\cdot}$
$\sum$	$n_{\cdot 1}$	$\dots$	$n_{\cdot j}$	$\dots$	$n_{\cdot s}$	n

Test $H_0 : X \,\&\, Y$ are independent. i.e. $H_0 : P(X = i, Y = j) = P(X = i)P(Y = j) = p_{i\cdot}p_{\cdot j}$.

Construct χ^2 test statistic:

$$K_n = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{[n_{ij} - n(\frac{n_{i\cdot}}{n})(\frac{n_{\cdot j}}{n})]^2}{n(\frac{n_{i\cdot}}{n})(\frac{n_{\cdot j}}{n})} = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - 1\right) \tag{2.126}$$

Then under H_0, $K_n \xrightarrow{\mathscr{L}} \chi^2_{rs-1-(r+s-2)} = \chi^2_{(r-1)(s-1)}$

Reject H_0 if $p(k_0) = P(K_n \geq k_0) < \alpha$

– Homogeneity Test:

Test R groups of sample with category rank, to see whether these groups has similar rank distribution.

Group \ Category	Category 1	$\dots$	Category j	$\dots$	Category C	$\sum$
Group 1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1C}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Group i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{iC}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Group R	n_{R1}	$\dots$	n_{Rj}	$\dots$	n_{RC}	$n_{R\cdot}$
$\sum$	$n_{\cdot 1}$	$\dots$	$n_{\cdot j}$	$\dots$	$n_{\cdot C}$	n

Denote $P(\text{Category } j|\text{Group } i) = p_{ij}$. Test $H_0 : p_{ij} = p_j, \forall 1 \leq i \leq R$.

Construct χ^2 test statistic:

$$D = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{[n_{ij} - n(\frac{n_{i\cdot}}{n})(\frac{n_{\cdot j}}{n})]^2}{n(\frac{n_{i\cdot}}{n})(\frac{n_{\cdot j}}{n})} = n\left(\sum_{i=1}^{R}\sum_{j=1}^{C} \frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - 1\right) \tag{2.127}$$

Then under H_0, $D \xrightarrow{\mathscr{L}} \chi^2_{R(C-1)-(C-1)} = \chi^2_{(R-1)(C-1)}$

- Test of Normality: normality is a good & useful assumption.

  For $\vec{Y} = (Y_1, Y_2, \ldots, Y_n)$,

  Test H_0 : exists μ & σ^2 such that Y_i i.i.d. $\sim N(\mu, \sigma^2)$.

  – Kolmogorov-Smirnov Test: Assume $\vec{X}$ form population CDF $F(x)$, test $H_0 : F(x) = F_0(x)$(where can take $F_0 = \Phi$ or some other known CDF).

    use $F_n(x)$ (as defined in eqa(2.80)) as approx. to $F(x)$, test

    $$D_n = \sum_{-\infty<x<+\infty} |F_n(x) - F_0(x)| \tag{2.128}$$

    Reject H_0 if $D_n > c$

    or use goodness-of-fit: denote observed value of D_n as d_n. Reject H_0 if

    $$p(d_n) = P(D_n > d_n|H_0) < \alpha \tag{2.129}$$

  – Shapiro-Wilk Test:

    Test H_0 : exists μ & σ^2 such that X_i i.i.d. $\sim N(\mu, \sigma^2)$.

    Denote $Y_{(i)} = \dfrac{X_{(i)} - \mu}{\sigma}$, $m_i = E(Y_{(i)})$

    Under H_0, $(X_{(i)}, m_i)$ falls close to straight line. Test Statistic: Correlation

    $$R^2 = \frac{\left(\sum_{i=1}^n (X_{(i)} - \bar{X})(m_i - \bar{m})\right)^2}{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (m_i - \bar{m})^2} = corr(X_{(i)}, m_i) \tag{2.130}$$

    Reject H_0 if $R^2 < c$

    Shapiro-Wilk correction:

    $$W = \frac{\left(\sum_{i=1}^{[n/2]} a_i (X_{(n+1-i)} - X_{(i)})\right)^2}{\sum_{i=1}^n (X_{(i)} - \bar{X})^2} \tag{2.131}$$

☐ Summary: Useful Non-Parameter Hypothesis Testing.

- Non-Parameter Hypothesis Testing
  - One Population Sample
    - χ^2 Test
    - Binomial Test
    - One-Sample K-S Test
    - Wilcoxon Sign Test
    - Runs Test
  - Two Population Sample
    - Independent Sample
      - Mann-Whitney Test
      - K-S Test
      - Wald-Wolfowitz Test
      - Moses Test of Extreme Reactions
    - Relative Sample
      - Sign Test
      - McNemar Test
      - Wilcoxon Rank Sum Test
      - Marginal Homogeneity Test
  - Multi-Population Sample
    - Independent Sample
      - Median Test
      - K-W One-Way ANOVA Test
      - Jonckheere-Terpstra Test
    - Relative Sample
      - Friedman Rank Sum Test
      - Kendall's Coefficient of Concordance Test
      - Cochran Q Test

Chapter. III 线性回归分析部分

Instructor: Zaiying Zhou

☐ Steps in Regression Analysis

1. Statement of the problem;
2. Selection of potentially relevant **variables**;
3. Data collection;
4. Exploratory Data Analysis (**EDA**)
5. **Model** specification;
6. Choice of fitting method;
7. Model fitting;
8. Model validation and criticism;
9. Using the chosen model(s) for the solution of the posed problem;
10. **Explain** the result.

R. Code for EDA

```
libaray('GGally')
head(df)
ggpairs(df)
str(df)
summary(df)
```

☐ Used Packages in R.

```
library('ggplot2')
libaray('GGally')
library('car')
library('moments')
library('lmtest')
library('nortest')
library('MASS')
library('tseries')

source('package.r')
```

Section 3.1 Regression Model

In regression model, we will observe pairs of variables, called 'cases'(样本点). A sample is $(X_1;Y_1),\dots,(X_n;Y_n)$, where X_i can be multivariate $X_i=\vec{X}_i=(X_{i1},X_{i2},\dots,X_{ip})$.

If X is continuous `numeric` variable, use Regression Model(s), else if X is discrete `factor` variable, use Factor Model(s).

▷ `R.` **Code**

Example data import:

```
df <- read.table('dataset/testdata.txt',header=FALSE,
    sep=',',col.names = c('y','x1','x2'))
```

3.1.1 Linear Regression Model

Regression Model focuses on how Y changes with continuous variables $X\in\mathbb{R}$. As a basic situation, we use **Linear Regression**, i.e. $Y\sim X$ in linear relation.

☐ Sample Geometry Notation

For most general case, in sample matrix notation:

$$Y=X\beta+\varepsilon \leftrightharpoons Y_j=X\beta_j+\varepsilon_j,\ \forall j=1,2,\dots,q \tag{3.1}$$

in Einstein Summation Convention:

$$Y_{ij}=X_{ij'}\beta_{j'j}+\varepsilon_{ij} \tag{3.2}$$

Why we need ε as 'random error term'?

- It represents the intrinsic random property of the model.
- Based on ε, we can take r.v. into our statistic model.

where

$$\underset{n\times q}{Y}=\begin{bmatrix} y_{11} & y_{12} & \dots & y_{1q}\\ y_{21} & y_{22} & \dots & y_{2q}\\ \vdots & \vdots & \ddots & \vdots\\ y_{n1} & y_{n2} & \dots & y_{nq}\end{bmatrix}=\begin{bmatrix}y_1,y_2,\dots,y_q\end{bmatrix} \qquad y_j=\begin{bmatrix}y_{1j}\\ y_{2j}\\ \vdots\\ y_{nj}\end{bmatrix} \tag{3.3a}$$

$$\underset{n\times(p+1)}{X}=\begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p}\\ 1 & x_{21} & x_{22} & \dots & x_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & x_{n2} & \dots & x_{np}\end{bmatrix}=\begin{bmatrix}x'_1\\ x'_2\\ \vdots\\ x'_n\end{bmatrix} \qquad x_i=\begin{bmatrix}1\\ x_{i1}\\ \vdots\\ x_{ip}\end{bmatrix} \tag{3.3b}$$

$$\underset{(p+1)\times q}{\beta}=\begin{bmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0q}\\ \beta_{11} & \beta_{12} & \dots & \beta_{1q}\\ \beta_{21} & \beta_{22} & \dots & \beta_{2q}\\ \vdots & \vdots & \ddots & \vdots\\ \beta_{p1} & \beta_{p2} & \dots & \beta_{pq}\end{bmatrix}=\begin{bmatrix}\beta_1,\beta_2,\dots,\beta_q\end{bmatrix} \qquad \beta_j=\begin{bmatrix}\beta_{j0}\\ \beta_{j1}\\ \vdots\\ \beta_{jp}\end{bmatrix} \tag{3.3c}$$

$$\underset{n\times q}{\varepsilon} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \dots & \varepsilon_{1q} \\ \varepsilon_{21} & \varepsilon_{22} & \dots & \varepsilon_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \dots & \varepsilon_{nq} \end{bmatrix} = \begin{bmatrix} \varepsilon_1, \varepsilon_2, \dots, \varepsilon_q \end{bmatrix} \qquad \varepsilon_j = \begin{bmatrix} \varepsilon_{1j} \\ \varepsilon_{2j} \\ \vdots \\ \varepsilon_{nj} \end{bmatrix} \tag{3.3d}$$

with Guass-Markov Assumption:

$$\begin{aligned} \text{Zero-Mean: } & E(\epsilon_i|X_i) = 0 \\ \text{Homogeneity of Variance: } & var(\epsilon_i) = \sigma^2 \\ \text{Independent: } & \epsilon_i \text{ i.i.d. } \sim \varepsilon \end{aligned} \tag{3.4}$$

and Normality Error Assumption:

$$\text{Normality: } \varepsilon_i \text{ i.i.d. } \sim N(0, \sigma^2) \tag{3.5}$$

Under matrix notation, model and assumptions eqa(3.4)(3.5) can be expressed in condensed notation:

$$Y_j = X\beta_j + \varepsilon_j \sim N_n(X\beta_j, \sigma_j^2 I_n), \quad j = 1, 2, \dots, q \tag{3.6}$$

△ **Note:** In this section we only focus on $q = 1$, i.e.

$$\underset{n\times 1}{Y} = \underset{n\times(p+1)}{X}\ \underset{(p+1)\times 1}{\beta} + \underset{n\times 1}{\varepsilon} \tag{3.7}$$

3.1.2 Factor Analysis Model

Regression Model focuses on continuous variables $X \in \mathbb{R}$ while factor model focus on discrete variable. More specifically, the 'value' of X is just a label, not necessarily a 'numeric value'.

Here only introduce one-way factor analysis,(single factor analysis) i.e. Y with only one factor with r levels: fac $= 1, 2, \dots, r$. Re-denote Y_{ij} = the observation outcome of the j^{th} item labelled the i^{th} level.

Model:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij},\ i = 1, \dots, r,\ j = 1, \dots, n_i$$

where μ is the average effect of all r factor levels, τ_i is the level effect of the i^{th} factor level, and ε i.i.d. $\sim N(0, \sigma^2)$ is noise error.

In matrix notation:

$$Y = \begin{bmatrix} y_{11} & \dots & y_{1n_1} & y_{21} & \dots & y_{2n_2} & \dots \end{bmatrix}^T \tag{3.8a}$$

$$X = \begin{bmatrix} 1 & 1 & 0 & \dots \\ 1 & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots \\ 1 & 0 & 1 & \dots \\ 1 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \ddots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \tag{3.8b}$$

$$\tau = \begin{bmatrix} \mu & \tau_1 & \tau_2 & \dots \end{bmatrix}^T \tag{3.8c}$$

$$\varepsilon = \begin{bmatrix} \varepsilon_{11} & \dots & \varepsilon_{1n_1} & \varepsilon_{21} & \dots & \varepsilon_{2n_2} & \dots \end{bmatrix}^T \tag{3.8d}$$

For more factor model e.g. two-way factor analysis with k denoting item and i, j denoting factor:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

cannot be simply expressed in matrix notation $\longrightarrow$ use index notation.

Section 3.2 Monovariate Linear Regression Model

First focus on the simpliest monovariate case $\vec{X}_i = X_i$. Monovariate Linear Model[4] with Gauss-Markov assumption & Normal Error assumption:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ \varepsilon_i \text{ i.i.d.} \sim N(0, \sigma^2) \tag{3.9}$$

What does Linear Regression do? Under Linear Model, try to estimate

- β_0 (intercept) ;
- β_1 (slope) ;
- σ^2 (variance of error).

(Thus Linear Regression is also a Statistics Inference process: deduce properties of model from data)

3.2.1 The Ordinary Least Square Estimation

Aim: use (x_i, y_i) to estimate $\beta_0, \beta_1, \sigma^2$. The idea is to define a 'loss function' to reflect the 'distance' from sample point to estimation point.

Estimate Principle: [5]

- Ordinary Least Squares:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min \sum_{i=1}^{n} (y - \beta_0 - \beta_1 x_i)^2 \tag{3.10}$$

- MLE or MoM Estimation.

And get $\hat{\beta}_1, \hat{\beta}_0$ as well as $\hat{\sigma^2}$(see eqa(3.15):[6]

[4] Here in linear regression, we consider X_i only as real number, **without** randomness. So here Y_i can be considered as an r.v. with X_i as parameter, i.e. $Y_i|_{X_i=x_i}$

[5] Detailed Definition and Derivation see section 2.2.5 or section 3.3.

[6] A memory trick: use $\dfrac{Y}{\sqrt{s_Y}} = r_{XY}\dfrac{X}{\sqrt{s_X}}$ to get formular of $Y \sim X$:

$$\hat{\beta}_1 = r_{XY}\frac{\sqrt{s_Y}}{\sqrt{s_X}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \tag{3.11}$$

$$
\begin{aligned}
\hat{\beta}_1 =& \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
\hat{\beta}_0 =& \bar{y}-\hat{\beta}_1\bar{x} \\
\hat{\sigma^2} =& \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2
\end{aligned}
\tag{3.12}
$$

Def. **Residual**: distance from sample point to estimate point, to reflect how the sample points fit the model.

$$
e_i = y_i - \hat{y}_i = \text{observed value of } \varepsilon_i \tag{3.13}
$$

Note: under least square estimation, we have[7]

$$
\sum e_i = 0 \qquad \sum x_i e_i = 0 \tag{3.14}
$$

Then use e_i to estimate σ^2 (because it is ε_0 that are i.i.d., not Y_i), where $(n-p-1)$ is Degree of Freedom (df or dof)[8]

$$
\begin{aligned}
\hat{\sigma_n^2} &= \frac{1}{n}\sum e_i^2 \quad \text{(use MLE or MoM)} \\
\hat{\sigma^2} &= \frac{1}{n-p-1}\sum e_i^2 = \frac{1}{n-2}\sum e_i^2 \quad \text{(use OLS, unbiased)}
\end{aligned}
\tag{3.15}
$$

Degree of Freedom of a Quadric Form:

- Intuitively: the number of independent variable;
- Rigorously: for quadric SS $= x'Ax$:

$$
dof_{SS} = \text{rank}(A) \tag{3.16}
$$

▷ R. **Code**

```
lmfit <- lm(formula,df)
summary(lmfit,cor=TRUE)
ggcoef(lmfit)
```

`lmfit` includes parameters `lmfit$coefficient` and `lmfit$residuals`

Example `lm()` output:

```
    Call:
    lm(formula = y ~ x, data = df)

    Residuals:
         Min         1Q    Median         3Q        Max
```

[7]Intuitively, they each means '$E(\varepsilon)=0$' and '$X \parallel \varepsilon$'.

[8]Generally, MLE and OLSE are different.

Comment from R.A.Fisher: $\sum e_i^2$ should be divided by 'number of e_i^2 that contribute to variance'. Here $(n-p-1)$ corresponds to 'degree of freedom' $=(n-2)$, $p=1$ corresponds to 'one' variable (see sec.2.2.5, eqa(2.79)), and correponds to the two equations of e_i, eqa(3.14)

```
    -16.1368  -6.1968  -0.5969   6.7607  23.4731

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 156.3466     5.5123   28.36   <2e-16 ***
x            -1.1900     0.0902  -13.19   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.173 on 58 degrees of freedom
Multiple R-squared:  0.7501,    Adjusted R-squared:  0.7458
F-statistic: 174.1 on 1 and 58 DF,  p-value: < 2.2e-16
```

3.2.2 Statistical Inference to β_0, β_1, σ^2, e_i

☐ Sampling Distribution of $\hat{\beta}_1, \hat{\beta}_0$

Consider $\hat{\beta}_1, \hat{\beta}_0$ as statistics of sample, then we can examine the sampling distribution of $\hat{\beta}_1, \hat{\beta}_0$. Their randomness comes from

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{3.17}$$

(The following part treats $\hat{\beta}_1, \hat{\beta}_0$ as r.v., and note that X_i are **not** r.v.. And for convenience and conciseness, denote $S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2$)

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} \frac{X_i - \bar{X}}{S_{XX}} \varepsilon_i$$
$$\hat{\beta}_0 = \beta_0 + \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{S_{XX}} \right) \varepsilon_i$$

Denote corresponding variance as $\sigma^2_{\hat{\beta}_1}$ and $\sigma^2_{\hat{\beta}_0}$, using eqa(1.69) to get:

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{S_{XX}} \qquad \sigma^2_{\hat{\beta}_0} = \sigma^2 (\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}) \tag{3.18}$$

And under normal error assumption, distribution of $\hat{\beta}_1, \hat{\beta}_0$ are

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1}) = N(\beta_1, \frac{\sigma^2}{S_{XX}})$$
$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2_{\hat{\beta}_0}) = N(\beta_0, \sigma^2 (\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}))$$

Based on sampling distribution of $\hat{\beta}_1, \hat{\beta}_0$, we can conduct statistical inference, including CI and HT.[9]

Note: In linear regression model, we usually focus more on β_1. And note that when 0 is **not** within the fitting range,β_0 is not so important.[10]

[9] Detail see sec.2.4, estimating/testing $\hat{\beta}_1, \hat{\beta}_0$ usually corresponds to 'estimate μ, with σ^2 unknown'.

[10] Two reason:

☐ Sampling Distribution of e_i Consider e_i as r.v. satisfies

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \tag{3.19}$$

and get the expression of $\hat{e}_i$

$$\hat{e}_i = \varepsilon_i - \sum_{k=1}^{n}\left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{XX}}\right)\varepsilon_k \tag{3.20}$$

$$E(e_i) = 0 \qquad \sigma^2_{e_i} = \sigma^2\left(1 - \frac{1}{n} - \frac{(X_i - \bar{X})^2}{S_{XX}}\right) \tag{3.21}$$

Under normal assumption:

$$\varepsilon_i \sim N(0, \sigma^2\left(1 - \frac{1}{n} - \frac{(X_i - \bar{X})^2}{S_{XX}}\right)) \tag{3.22}$$

Further we can get $\hat{\sigma}^2 = E(\frac{1}{n-2}\sum_{i=1}^{n} e_i^2)$ where $e_i^2 \sim \sigma^2\left(1 - \frac{1}{n} - \frac{(X_i - \bar{X})^2}{S_{XX}}\right)\chi^2$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sigma^2\sum_{i=1}^{n}(1 - \frac{1}{n} - \frac{(X_i - \bar{X})^2}{S_{XX}}) = \sigma^2 \tag{3.23}$$

More definition of refined residuals see sec.3.4.3 in page 3.4.3.

☐ Why we choose OLS to get regression coefficients?

Gauss-Markov Thm.: the OLS estimator has the lowest sampling variance within the class of linear unbiased estimators, i.e. OLS is the Best Linear Unbiased Estimator(BLUE).[11]

3.2.3 Prediction to Y_h

For a new X_h at which we wish to **predict** the corresponding Y_h (based on other known point (X_i, Y_i)), denote the estimator as $\hat{\mu}_h$:

$$\hat{\mu}_h = \hat{\beta}_1 X_h + \hat{\beta}_0 = \beta_1 X_h + \beta_0 + \sum_{i=1}^{n}\left(\frac{1}{n} + \frac{(X_i - \bar{X})(X_h - \bar{X})}{S_{XX}}\right)\varepsilon_i \tag{3.24}$$

Thus we can get[12]

$$E(\hat{\mu}_h) = \beta_1 X_h + \beta_0 \qquad \sigma^2_{\hat{\mu}_h} = \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}}\right)\sigma^2 \tag{3.25}$$

Under Normal assumption:

$$\hat{\mu}_h \sim N(\beta_1 X_h + \beta_0, \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}}\right)\sigma^2) \tag{3.26}$$

Base on distribution we can give CI and HT.

Note: We can either consider

- **Y_h itself as an r.v.** : Confidence Interval of Y_h;

  And we can just use $\sigma^2_{\hat{\mu}_h}$ to construct CI;

  ▷ **R. Code**

- The etimation error of Y from $\hat{\beta}_1$ increases with $X_h - \bar{X}$;
- $\beta_1 == 0$ is important: decides whether linear model can be used.

[11] This Thm. does **not** require normal error assumption.

[12] So $\sigma^2(\hat{\mu}_h)$ increases with $X_h - \bar{X}$. Intuitively it make sense, because $(\bar{X}, \bar{Y})$ must falls on regression line.

```r
predict(lmfit,newdata = 40),
    interval="confidence",level=0.95)
```

- **predicted Y_h from other sample points**: Prediction Interval of Y_h

  Need to have the randomness of $\hat{\beta}_0, \hat{\beta}_1$ considered(if they are unknown).

  Def. Prediction Error: Y_h itself is an Y of the linear model, i.e. $Y_i = \beta_0 + \beta_1 X_h + \varepsilon_h$, we can and define **Prediction Error**:

$$d_h = Y_h - \hat{\mu}_h \tag{3.27}$$

$$E(d_h) = 0 \qquad \sigma^2_{d_h} = var(Y_h - \hat{\mu}_h) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})}{S_{XX}}\right] > \sigma^2_{\hat{\mu}_h} \tag{3.28}$$

▷ R. **Code**

```r
predict(lmfit,newdata = 40),
    interval="prediction",level=0.95)
```

☐ Simultaneous Confidence Band(SCB)

Confidence Band is **not** the CI at each point, but really a **band** for the **entire** regression line.[13]

Aim: Find lower and upper function $L(x)$ and $U(x)$ such that

$$P[L(x) < (\beta_0 + \beta_1 x) < U(x), \forall x \in I_x] = 1 - \alpha \tag{3.29}$$

and get **Confidence Band**:

$$\{(x, y) | L(x) < y < U(x) | \forall x \in I_x\} \tag{3.30}$$

Where $(L(x), U(x))$ can be derived as

$$(L(x), U(x)) = \hat{\mu}_x \pm s_{\hat{\mu}_x} W_{2,n-2,1-\alpha} \tag{3.31}$$

Where W correponds to W distribution: $W_{m,n} = \sqrt{2F_{m,n}}$

Small sample case: Bonferroni correction.

▷ R. **Code**

```r
library(ggplot2)
ggplot(df,aes(x,y))+geom_point()+geom_smooth(method='lm',formula=y~x)
```

[13]Why they are different? We require the confidence band have a **simultaneous** converage probability. For the same band $(L(x), U(x))$, $P(\text{the whole line}) < P(\text{each point})$, so Confidence Band is wider than $\bigcup$CIs to hold the same $1 - \alpha$.

Also, we will see that for linear model, the boundary of SCB forms hyperbola, which make sense considering its asymptotic line.

3.2.4 Analysis of Variance: Monovariate

ANalysis **O**f **VA**riance (ANOVA): One-sample t test $\rightsquigarrow$ Two sample t test $\rightsquigarrow$ More-sample: ANOVA

☐ **Key Idea Of ANOVA**: Test whether the mean of some groups are the same, i.e. $\mu_1 = \mu_2 = \ldots = \mu_r$

In linear regression model, modified as testing $\beta_1 == 0$. Conduction: Take Partition of Total Sum of Square To Examine **Variation**. Because Y_i are not i.i.d. (different mean value $X\beta$), so has different parts of variation from Regression Model/Error Term.

Measure of Variation: Sum of Square (SS) & Mean Sum of Square (MS).

MS: Divide each SS by corresponding dof. Definition of dof see eqa(3.16).

$$\text{MS} = \frac{\text{SS}}{dof} \tag{3.32}$$

- SST: Total Sum of Squares

$$\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad dof_{\text{SST}} = n - 1 \tag{3.33}$$

- SSRegression: Variation due to Regression Model (which is explained by regression line);[14]

$$\text{SSR} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \qquad dof_{\text{SSR}} = 1 \tag{3.34}$$

- SSError: Variation attribtes to ε (which is reflected by residuals).

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i) \qquad dof_{\text{SSE}} = n - 2 \tag{3.35}$$

△ **IMPORTANT:** In some books

- SSRegression → SSExplained or SSModel;
- SSError → SSResidual.

And Cause **Confusion**! In this summary we take the former.

Idea: take partition of SST. i.e.

$$Y_i - \bar{Y} = (Y_i - \hat{Y}) + (\hat{Y} - \bar{Y}) = e_i \tag{3.36}$$

And we can prove that

$$\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \text{SSR} + \text{SSE} \tag{3.37}$$

That is: we **partition** SST into two parts, so that we can examine them seperately.

☐ ANOVA Table

Source	dof	SS	MS	F-Statistic
SSRegression	1	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	SSR/dof_R	MSR/MSE
SSError	$n-2$	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	SSE/dof_E	
SSTotal	$n-1$	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$	SST/dof_T	

▷ `R.` **Code**

[14] $\text{SSR} = \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$, so $dof_R = 1$

```
anova(lmfit)
```

Properties:

$$E(\text{MSE}) = \sigma^2 \qquad E(\text{MSR}) = \sigma^2 + \beta_1^2 S_{XX} \tag{3.38}$$

Section 3.3 Multivariate Linear Regression Model

As a more general case of $\vec{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$, Multivariate Linear Model is expressed as in eqa(3.7):

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

3.3.1 The Ordinary Least Estimation

To conduct OLS

$$\hat{\beta} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{p+1}} (Y - X\beta)^T (Y - X\beta) \tag{3.39}$$

Here we introduce two approaches:

- Analytical: Take matrix differciation (See section 4.1.2 eqa(4.27))

$$\begin{aligned} 0 &= \frac{\partial (Y - X\beta)^T (Y - X\beta)}{\partial \beta} = \frac{\partial}{\partial \beta} (Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta) \\ &= -X^T Y - X^T Y + (X^T X + X X^T)\beta = -2X^T (Y - X\beta) \end{aligned}$$

  Thus we get OLS:

$$\hat{\beta} = (X'X)^{-1} X'Y \tag{3.40}$$

- Geometric/Algebraical: Use hyper-projection.

$$\hat{\beta} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{p+1}} d(Y, X\beta) \tag{3.41}$$

  i.e. $\hat{\beta}$ is the (hyper-)projection of Y onto X (within Euclidean Space), naturally we have

$$(X\beta)^T (Y - X\beta) = 0 \Rightarrow \hat{\beta} = (X'X)^{-1} X'Y \tag{3.42}$$

☐ Matrix Notation of OLS Estimator:

$$\hat{\beta} = (X'X)^{-1} X'Y \tag{3.43}$$

3.3.2 Statistical Inference to β, σ^2, e

Properties & Extrapolation

- Sampling Distribution of $\hat{\beta}$: (Here consider normal case $Y \sim N(X\beta, \sigma^2 I_n)$, and use eqa(4.36))

$$\hat{\beta} = (X'X)^{-1}X'Y \sim N_n(\beta, \sigma^2(X'X)^{-1}) \tag{3.44}$$

Comment: $cov(\beta_i, \beta_j)$ are generally not 0, $\Rightarrow \beta_i, \beta_j$ dependent.

- Predicted Response & Hat Matrix H:

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y \equiv HY = P_X Y \tag{3.45}$$

where **Hat Matrix**/Influence matrix/Projection matrix $H = P_X = X(X'X)^{-1}X'$, with properties

  – Symmertric: $H^T = H$;

  – Idempotence: $H^2 = H$

  – Rank: $\text{rk}(H) = tr(H) = \text{rk}(X)$

  – H and self-influene factor h_{ii}: Note the linearity of $\hat{Y}$ on Y

$$\hat{Y} = HY \Rightarrow H = \frac{\partial \hat{Y}}{\partial Y} \tag{3.46}$$

  The diagonal elements of H is

$$h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i} = X_i(X'X)^{-1}X_i' \tag{3.47}$$

  Comment on h_{ii}: $var(e_i) = \sigma^2(1 - h_{ii})$, for $h_{ii} \to 1$, i.e. the regression line always pass y_i, thus it's 'influential'.

- Residual:

$$e = Y - \hat{Y} = (I - H)Y \sim N_n\left(0, \sigma^2(I - H)\right) \tag{3.48}$$

Covariance Matrix of Residual:

$$cov(e) = \sigma^2(I - H) = \sigma^2 \begin{bmatrix} 1 - h_{11} & -h_{12} & \dots & -h_{1n} \\ -h_{21} & 1 - h_{22} & \dots & -h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -h_{n1} & -h_{n2} & \dots & 1 - h_{nn} \end{bmatrix} \tag{3.49}$$

- Estimator and Distribution of σ^2:

First use eqa(4.37) to get [15]

$$E(\text{SSE}) = E(e'e) = E(Y'(I - H)Y) = (X\beta)'(I - H)X\beta + tr((I - H)\sigma^2 I_n) = \sigma^2(n - p - 1) \tag{3.51}$$

dof of Residual e (use definition eqa(3.16)):

$$dof_e = dof_{(I-H)Y} = \text{rank}(I - H) = n - p - 1 \tag{3.52}$$

[15] Also we need the property of idmpotnet matrix

$$\lambda_i = 0 \text{ or } 1 \Rightarrow tr(H) = \text{rank}(H) = \sum_{i=1}^{n} \lambda_i = \#(\lambda = 1) \tag{3.50}$$

Thus the unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \text{MSE} = \frac{e'e}{n-p-1} = \frac{Y'(I-H)Y}{n-p-1} \tag{3.53}$$

Distribution (under normal assumption):

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1} \tag{3.54}$$

- Gauss-Markov Thm.: OLS Estimator of β is the BLUE Estimator.

More hypothesis testing to β see section 4.2.4.

3.3.3 Prediction to Y_h

For a new $\vec{X}_h$ at which we wish to **predict** the corresponding Y_h (based on other known point (X_i, Y_i)), denote the estimator as $\hat{\mu}_h$:

$$\hat{\mu}_h = X_h'\hat{\beta} = X_h'(X'X)^{-1}X'Y$$

thus we get

$$E(\hat{\mu}_h) = X_h'\beta \qquad \sigma^2_{\hat{\mu}_h} = \sigma^2(1 + X_h'(X'X)^{-1}X_h)$$

under normal assumption:

$$\hat{\mu}_h \sim N(X'\beta, \sigma^2(1 + X_h'(X'X)^{-1}X_h))$$

3.3.4 Analysis of Variance: Multivariate

Sampling Notation see eqa(3.3), still consider $(p+1)$ -dim $(\mathbf{1}_n, X_i)$ v.s. 1-dim Y, and $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$

- SST:

$$\text{SST} = (Y - \bar{Y}\mathbf{1}_n)'(Y - \bar{Y}\mathbf{1}_n) \qquad dof_{\text{SST}} = n-1 \tag{3.55}$$

- SSR:

$$\text{SSR} = (\hat{Y} - \bar{Y}\mathbf{1}_n)'(\hat{Y} - \bar{Y}\mathbf{1}_n) \qquad dof_{\text{SSR}} = p \tag{3.56}$$

Denoted in hat matrix H and $\mathcal{J}$ in eqa(4.9)

$$\text{SSR} = Y'(H - \frac{1}{n}\mathcal{J})Y \tag{3.57}$$

- SSE:

$$\text{SSE} = (Y - \hat{Y})'(Y - \hat{Y}) \qquad dof_{\text{SSE}} = n-p-1 \tag{3.58}$$

Denoted in residual e and hat matrix H:

$$\text{SSE} = e'e = Y'(I-H)Y \tag{3.59}$$

☐ ANOVA Table

Source	dof	SS	MS	F-Statistic
SSRegression	p	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	SSR/dof_R	MSR/MSE
SSError	$n-p-1$	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	SSE/dof_E	
SSTotal	$n-1$	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$	SST/dof_T	

▷ R. **Code**

```
anova(lmfit)
```

Section 3.4 Diagnostics

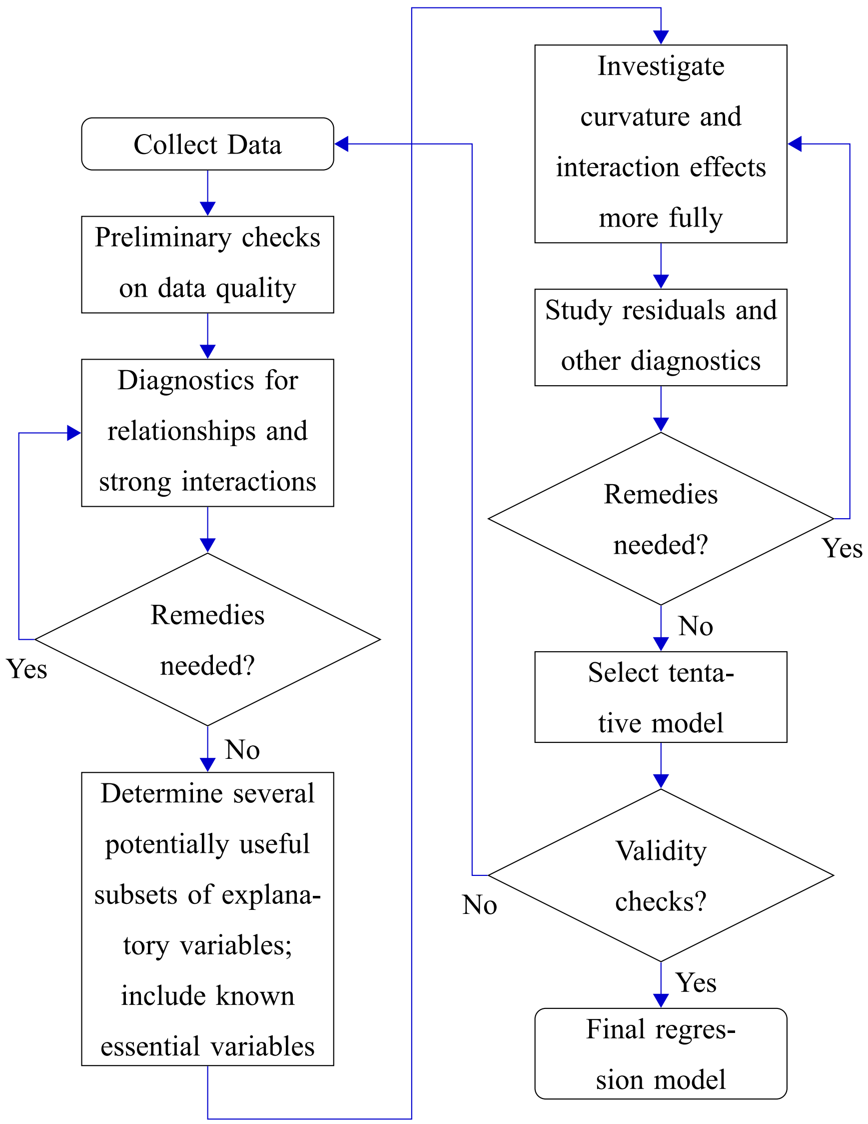


图 1: Diagnostics and Remedies for Regression Model

To apply OLS, we need the basic Gauss-Markov Assumption eqa(3.4); or we further need better properties of the model, e.g. take Normal Assumption.

Assumptions:

$$\begin{aligned}\text{Zero-Mean: } & E(\epsilon_i|X_i) = 0\\ \text{Homogeneity of Variance: } & var(\epsilon_i) = \sigma^2\\ \text{Independent: } & \epsilon_i \text{ i.i.d. } \sim \varepsilon\\ \text{Normal: } & \varepsilon \sim N(0, \sigma^2)\end{aligned} \tag{3.60}$$

Or sum up as

$$Y \sim N_n(X\beta, \sigma^2 I_n) \tag{3.61}$$

Thus we need to conduct Diagnostics and Remedies to

- examine whether these assumptions are satisfies;
- perform correction to regression method.

Preliminary Diagnostics:

▷ R. **Code**

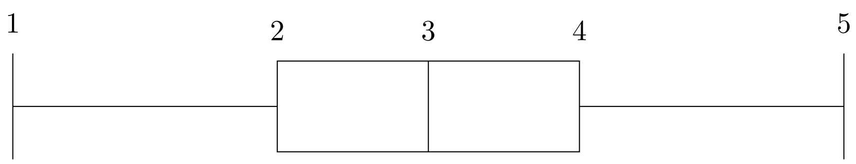
```
lmfit <- lm(y~x,lmfit)
par(mfrow = c(2, 2))
plot(lmfit)
```

3.4.1 Useful Diagnostics Plots

- BoxPlot: to examine the similarity of distribution.

  Notation:

  1. min point above 25% quantile-1.5IQR;
  2. 25% quantile;
  3. median;
  4. 75% quantile;
  5. max point below 75% quantile+1.5IQR.



- Histogram Plots: Frequency distribution (can deal with many-peak)
- Quantile-Quantile Plots: Examine the similarity between distribution.

  For two CDF $q = F(x)$ and $q = G(x)$(where q for quantile), with $x = F^{-1}(q)$, $x = G^{-1}(q)$. And Plot $F^{-1}(q)$-$G^{-1}(q)$.

  Usually test normality, take $G = \Phi$

- Partial Regression Plot: Test non-linearity/heterogeneous-variance.

  For each X_i variable:

  - Use other $X_{(\wedge i)}$ to predict Y, get residual $e_Y|X_{(\wedge i)}$;
  - Use other $X_{(\wedge i)}$ to predict X_i, get residual $e_{X_i}|X_{(\wedge i)}$

  Plot $(e_Y|X_{(\wedge i)})$-$(e_{X_i}|X_{(\wedge i)})$ as Added Variable Plot (AV Plot). Used for testing non-linearity/heterogeneous-variance.

▷ R. **Code**

```
boxplot(df$x)

hist(df$x)

hist(df$x,freq=FALSE)
lines(density(df$x))

stem(df$x)

qqnorm(df$x)
qqline(df$x,col='red')

library(car)
avPlots(lmfit)
```

3.4.2 Diagnostics to X Distribution

Considering the dependence of Y_i on X_i, to get a more reliable $\hat{\beta}_1$, we cannot just focus on the (marginal) distribution of Y_i, we would also need a better 'distribution' of X_i

- Plots: BoxPlot/QQPlot
- 4 statistics(parameters);[16]
  - Mean: Location;

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.62}$$

  - Standard Deviation: Variability;

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{3.63}$$

[16] See sec.2.1.1

- Skewness: Lack of Symmertry;

$$\hat{g}_1 = \frac{m_{n,3}}{m_{n,2}^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})\right)^{3/2}} \tag{3.64}$$

Adjusted Skewness (Least MSE):

$$\frac{\sqrt{n(n-1)}}{n-2}\hat{g}_1 \tag{3.65}$$

  * $\hat{g}_1 > 0$: Right skewness, longer right tail;
  * $\hat{g}_1 < 0$: Left skewness, longer left tail.

  Fisher-Pearson coefficient of skewness: $\dfrac{3(\text{mean} - \text{median})}{\sigma}$.

- Kurtosis: Heavy/Light Tailed.

$$\hat{g}_2 = \frac{m_{n,4}}{m_{n,2}^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2} - 3 \tag{3.66}$$

$\hat{g}_2 = 0 \Rightarrow$ similar to normal.

  * $\hat{g}_2 > 0$: Leptokurtic, heavy tail, slender;
  * $\hat{g}_2 < 0$: Platykurtic, light tail, broad.

  Note: In expression of $\hat{g}_1$ and $\hat{g}_2$, we already divide the variance. So Skewness and Kurtosis only reflect the difference from normal, but **not** related to variance.

  Best tool to determine Kurtosis: QQ-Plot.

▷ `R.` **Code**

```
summary(df$x)
```

Other moments use package `moments`

- Bias: Inspect the design methodology
  - Selection Bias: Not completely random sampling;
  - Information Bias: Difference between ’designed’ and ’get’, e.g. no response;
  - Confounding: Exist another important variable, while the model actually focuses on a less important variable, or even reverse the causality.

3.4.3 Diagnostics to Residual

☐ Residual Reflects the properties of ε

- **Linearity** : use Residual Plot/AV Plot to Reflect the linearity and variance assumption.

▷ `R.` **Code**

```r
lmfit <- lm(y~x,df)
scatter(df$x,lmfit$residuals)
abline(h=0)

library(car)
avPlots(lmfit)
```

- The Assumption of **Equal Variances** :
  - AV Plot , e.g. test the R^2 of $(e_Y|X_{(\wedge i)})$-$(e_{X_i}|X_{(\wedge i)})$ relation.
  - Bartlett's test:

    Idea: divide the sample into groups g, and get each MSE

$$\mathrm{MSE}_g = \frac{1}{n_g}\sum_{i=1}^{n_g}(Y_{gi} - \hat{Y}_g)^2 \tag{3.67}$$

    and take statistic

$$S = -\frac{(N-g)\ln\left[\sum_g \frac{n_g}{N-n_g}\mathrm{MSE}_g\right] - \sum_g (n_g-1)\ln\frac{n_g}{N-n_g}\mathrm{MSE}_g}{1+\frac{1}{3(G-1)}\sum_g\left(\frac{1}{n_g-1}-\frac{1}{N-G}\right)} \sim \chi^2 \tag{3.68}$$

    to conduct test.

    Note: **sensitive** to normal assumption, not robust. Used when normal assumption is satisfied.
  - Levene's test: Divide the sample into G groups. Denote **mean** of residual within each group as $\tilde{e}_g$, and in each group compute

$$d_{ig} = |e_{ig} - \tilde{e}_g| \Rightarrow \bar{d}_g = \frac{1}{n_g}\sum_{j=1}^{n_g} d_{ig} \tag{3.69}$$

    Then conduct ANOVA to d_{ig}.

    If $G = 2$: 2-sample t-test,

$$T = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \xrightarrow{\mathscr{L}} t_{n-2} \qquad s^2 = \frac{\sum(d_{i1}-\bar{d}_1)^2 + \sum(d_{i2}-\bar{d}_2)^2}{n-2} \tag{3.70}$$

  - Brown-Forsythe's Test (Modified Levene's test): For skewed sample, take the **mean** as **median**, more robust.
  - ⋆ Breusch-Pagan Test:

    Assume variance of ε_i dependent on X_i as m^{th} polynomial:

$$\sigma_i^2 = \alpha_0 + \sum_{k=1}^{m}\alpha_k X_i^k \tag{3.71}$$

    and test

$$H_0 : \alpha_k = 0\,\forall k = 1,2,\ldots,m \longleftrightarrow H_1 \tag{3.72}$$

Method: First conduct OLS to get regression line $\hat{l}_1$ and residuals e_i and SSE, and conduct regression of e_i^2 over X_i to get another regression line $\hat{l}_2$ and corresponding SSR*.

Then statistic

$$S = \frac{\text{SSR}^*/2}{(\text{SSE}/n)^2} \xrightarrow{\mathscr{L}} \chi^2_m \tag{3.73}$$

▷ `R.` **Code**

Example for $G = 2$:

```
group <-factor(rep(c(1,2),length.out=length(df$x),
    each=(ceiling(length(df$x)/2))))

bartlett.test(lmfit$residuals~group,group)

library(car)
leveneTest(lmfit$residuals~group,group,center=mean)
leveneTest(lmfit$residuals~group,group,center=median)

library(lmtest)
bptest(lmfit)
```

- The Assumption of **Normality** :

  In most case we use S-W Test($n < 2000$) and K-S Test($n > 2000$):

  – QQ-plot of ordered residuals.

  ⋆ Shapiro-Wilk Test (Most Powerful)[17]: To test $H_0 : \exists\sigma^2,\ s.t.\, \varepsilon \sim N_n(0, \sigma^2 I_n)$, denote

  $$m_i = E(\frac{\varepsilon_{(i)}}{\sigma})$$

  then under H_0, $\varepsilon_{(i)} \sim m_i \to$ linear, thus test correlation

  $$R^2 = \frac{\left(\sum_{i=1}^n (e_{(i)} - \bar{e})(m_i - \bar{m})\right)^2}{\sum_{i=1}^n (e_i - \bar{e})^2 \sum_{i=1}^n (m_i - \bar{m})^2} = corr(e_{(i)}, m_i) \tag{3.74}$$

  – Kolmogorov-Smirnov Test:

  $$D_n = \sum_e |F_n(e) - \Phi(e)| \tag{3.75}$$

  – Cramér-von Mises Test:

  $$T = n \int_{-\infty}^{\infty} (F_n(e) - \Phi(e))^2 \, \mathrm{d}\Phi(e) \tag{3.76}$$

  – Anderson-Darling Test:

  $$A^2 - n \int_{-\infty}^{\infty} (F_n(e) - \Phi(e))^2 \frac{1}{\Phi(e)(1 - \Phi(e))} \, \mathrm{d}\Phi(e) \tag{3.77}$$

[17] Detail of S-W Test and K-S Test see Test of Normality in sec.2.4.6

- Jarque-Bera Test , using skewness $\hat{g}_1$ and kurtosis $\hat{g}_2$ of $\vec{e}$

$$\text{JB} = \frac{n}{6}(\hat{g}_1^2 + \frac{1}{4}\hat{g}_2^2) \xrightarrow{\mathscr{L}} \chi_2^2 \tag{3.78}$$

▷ R. **Code**

```
qqnorm(lmfit$residuals)
qqline(lmfit$residuals)

qqp <- qqnorm(lmfit$residuals)
cor(qqp$x,qqp$y)

shapiro.test(lmfit$residuals)

ks.test(jitter(lmfit$residuals),pnorm,mean(lmfit$residuals),
    sd(lmfit$residuals))

library(nortest)
cvm.test(lmfit$residuals)

ad.test(lmfit$residuals)

library(tseries)
jarque.bera.test(lmfit$residuals)
```

- The Assumption of **Independence** :
  - Durbin-Watson Test:

$$d = \frac{\sum_{j=2}^{n}(e_j - e_{j-1})^2}{\sum_{j=1}^{n} e_j^2} \tag{3.79}$$

  $d \in (1.5, 2.5)$ is fine.

  - Ljung-Box Test:

$$Q = n(n+2)\sum_{k=1}^{n} \frac{\hat{\rho}_k^2}{n-k} \tag{3.80}$$

▷ R. **Code**

```
dwtest(lmfit)
```

3.4.4 Diagnostics to Influentials

An intuitive explanation to extreme values:

- Outliers: Extreme case for Y;

- High Leverage: Extreme case for X;
- Influentials: Cases that influence the regression line.

☐ Influentials = Outliers ∩ High Leverage

In section 3.3, we got the $\hat{\beta}$ as $\hat{\beta} = (X'X)^{-1}X'Y = HY$ and got $\hat{Y}$ as

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = \hat{H}Y \tag{3.81}$$

where hat matrix $H \equiv X(X'X)^{-1}X' = \dfrac{\partial \hat{Y}}{\partial Y}$

Also we got statistical inference to β, σ^2, e

$$\hat{\beta} = (X'X)^{-1}X'Y \sim N(\beta, \sigma^2(X'X)^{-1}) \tag{3.82}$$

$$e = Y - \hat{Y} = (I - H)Y \sim N(0, \sigma^2(I - H)) \tag{3.83}$$

$$\hat{\sigma}^2 = \text{MSE} = \frac{e'e}{n-p-1} = \frac{Y'(I-H)Y}{n-p-1} \tag{3.84}$$

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1} \tag{3.85}$$

The diagonal elements of $\hat{H}$ is self-sensitivity h_{ii}

$$h_{ii} = X_i'(X'X)^{-1}X_i$$

☐ Some refined residuals to help conduct Diagnostics:

- Standardized Residual:

$$e_{\text{sd}i} = \frac{e_i}{\sigma_{e_i}} = \frac{e_i}{\sigma\sqrt{1-h_{ii}}} \tag{3.86}$$

- (Internal) Studentized Residual: replace σ with $s = \hat{\sigma}$

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} = \frac{e_i}{\sqrt{\text{MSE}}\sqrt{1-h_{ii}}} \tag{3.87}$$

- Deleted Residual:[18]

---

[18] *Proof:*

Lemma: $(A+B)^{-1} = A^{-1} - \dfrac{1}{1+tr(BA^{-1})}A^{-1}BA^{-1}$, where $\text{rk}(B) = 1$.

$$\hat{\beta}_{(\wedge i)} = (X'_{(\wedge i)}X_{(\wedge i)})^{-1}X'_{(\wedge i)}Y_{(\wedge i)}$$

Using the above lemma: (here for aesthetic purpose, treat X_i as row vector)

$$\begin{aligned}
(X'_{(\wedge i)}X_{(\wedge i)})^{-1} =& (X'X - X_i'X_i)^{-1} \\
=& (X'X)^{-1} + \frac{1}{1 - tr[X_i'X_i(X'X)^{-1}]}(X'X)^{-1}X_i'X_i(X'X)^{-1} \\
=& (X'X)^{-1} + \frac{1}{1-h_{ii}}(X'X)^{-1}X_i'X_i(X'X)^{-1} \\
X_{(\wedge i)}Y_{(\wedge i)} =& X'Y - X_i'Y_i
\end{aligned}$$

then calaulate $\hat{\beta}_{(\wedge i)}$:

$$\hat{\beta}_{(\wedge i)} = (X'_{(\wedge i)}X_{(\wedge i)})^{-1}X'_{(\wedge i)}Y_{(\wedge i)}$$

$$d_i = Y_i - \hat{Y}_{i(\wedge i)} = \frac{e_i}{1-h_{ii}} \tag{3.89}$$

where $\hat{Y}_{i(\wedge i)}$ is predicted Y value at X_i obtained from the regression of dataset with the i case (X_i, Y_i) removed:

$$\hat{\beta}_{(\wedge i)} = (X'_{(\wedge i)}X_{(\wedge i)})^{-1}X'_{(\wedge i)}Y_{(\wedge i)} \qquad \hat{Y}_{i(\wedge i)} = X'_i\hat{\beta}_{(\wedge i)}$$

- (External) Studentized Residual: To avoid self-influence, take deleted residual in eqa(3.87)

$$t_i = \frac{d_i}{\hat{\sigma}_{(\wedge i)}\sqrt{1-h_{ii}}} = \frac{d_i}{\sqrt{\text{MSE}_{(\wedge i)}}\sqrt{1-h_{ii}}} \sim t_{n-p-2} \tag{3.90}$$

  Relation between MSE and $\text{MSE}_{(\wedge i)}$:

$$(n-p-1)\text{MSE} = (n-p-2)\text{MSE}_{(\wedge i)} + \frac{e_i^2}{1-h_{ii}}$$

- Diagnostics to **Outlier**: use external studentized residual for t-test with Bonferroni adjustment. Declare the i^{th} case an outlier if:

$$|t_i| > t_{\alpha/2n, n-p-2}$$

- Diagnostics to **Leverage**: use hat matrix H/self-sensitivity h_{ii}.

$$\sum_{i=1}^{n} h_{ii} = tr(H) = p+1 \Rightarrow \bar{h} = \frac{p+1}{n}$$

  Declare the i^{th} case a leverage if:

$$h_{ii} > \kappa\bar{h} = \kappa\frac{p+1}{n}$$

  where usually take $\kappa = 2$ or 3.

$$\begin{aligned}
&= \left[(X'X)^{-1} + \frac{(X'X)^{-1}X'_iX_i(X'X)^{-1}}{1-h_{ii}}\right](X'Y - X'_iY_i) \\
&= \hat{\beta} + \frac{(X'X)^{-1}X'_iX_i(X'X)^{-1}X'Y}{1-h_{ii}} - (X'X)^{-1}X'_iY_i - \frac{(X'X)^{-1}X'_iX_i(X'X)^{-1}X'_iY_i}{1-h_{ii}} \\
&= \hat{\beta} + \frac{(X'X)^{-1}X'_i\hat{Y}_i}{1-h_{ii}} - \frac{(X'X)^{-1}X'_iY_i(1-h_{ii})}{1-h_{ii}} - \frac{(X'X)^{-1}X'_iY_i}{1-h_{ii}}h_{ii} \\
&= \hat{\beta} + \frac{(X'X)^{-1}X'_i}{1-h_{ii}}(\hat{Y}_i - Y_i) \\
&\Rightarrow \hat{\beta} - \hat{\beta}_{(\wedge i)} = (X'X)^{-1}X'_i\frac{e_i}{1-h_{ii}}
\end{aligned} \tag{3.88}$$

Then

$$\begin{aligned}
Y_i - \hat{Y}_{i(\wedge i)} &= Y_i - \hat{Y}_i + \hat{Y}_i - \hat{Y}_{i(\wedge i)} \\
&= e_i + X_i(\hat{\beta} - \hat{\beta}_{(\wedge i)}) \\
&= e_i + X_i(X'X)^{-1}X'_i\frac{e_i}{1-h_{ii}} \\
&= \frac{e_i}{1-h_{ii}}
\end{aligned}$$

- Diagnostics to **Influential**: Studentized DIFference caused to FITted values (DIFFITS)

  DIFFIT:

$$\text{DIFFIT}_i = \hat{Y}_i - \hat{Y}_i(\wedge i) = e_i \frac{h_{ii}}{1 - h_{ii}}$$

  DIFFITS:

$$\text{DIFFITS}_i = \frac{\text{DIFFIT}_i}{s(\hat{Y}_i)} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

  Declare the i^{th} case an influential if:

$$\begin{cases} \text{DIFFITS}_i > 1 & \text{small/medium data} \\ \text{DIFFITS}_i > 2\sqrt{\dfrac{p+1}{n}} & \text{large data} \end{cases}$$

- Diagnostics to **Influential**: Cook's Distance, by quantifying the 'influence' to $\hat{\beta}$.

  Using eqa(3.44)(3.54) we could construct the following Cook's Distance[19]

$$D_i = \frac{||X(\hat{\beta} - \hat{\beta}_{(\wedge 1)})||^2}{(p+1)\hat{\sigma}^2} = \frac{e_i^2}{(p+1)\hat{\sigma}^2} \frac{h_{ii}}{(1 - h_{ii})^2} \qquad \frac{1 - h_{ii}}{h_{ii}} D_i \sim F_{p+1, n-p-1}$$

  Comment:

$$D_i = \frac{e_i^2}{(p+1)\hat{\sigma}^2} \left[\frac{h_{ii}}{(1 - h_{ii})^2} \right] = \frac{1}{p+1} \frac{h_{ii}}{1 - h_{ii}} \times r_i^2 \tag{3.91}$$

  where $\frac{1}{p+1}\frac{h_{ii}}{1-h_{ii}}$ correponds to hige leverage, and r_i^2 correponds to outliers, multiply to get influentials.
  Declare the i^{th} case an influential if

$$D_i > \frac{4}{n}$$

  Or conduct F-test using the distribution of D_i, with $\alpha \sim 20\%$.

- Diagnostics to **Influential**: Studentized DiFference in BETA estimates (DFBETAS). Use eqa(3.44), define

$$var(\hat{\beta}_k) = \sigma^2 (X'X)^{-1}_{kk} := \sigma^2 c_{kk}$$

  And studentize difference in $\hat{\beta}$ with i^{th} case removed: $\hat{\beta}_k - \hat{\beta}_{k(\wedge i)}$

$$\text{DFBETAS}_{k(\wedge i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(\wedge i)}}{\sqrt{\text{MSE}_{(\wedge i)} c_{kk}}}, \; k = 1, 2, \ldots, p$$

  Declare the i^{th} case an influential if

$$\begin{cases} \text{DFBETAS}_i > 1 & \text{small/medium data} \\ \text{DFBETAS}_i > \dfrac{2}{\sqrt{n}} & \text{large data} \end{cases}$$

▷ `R.` **Code**

[19] Proof uses eqa(3.88).

```
rstudent(lmfit)
library(car)
outlierTest(lmfit)

hatvalues(lmfit)

cooks.distance(lmfit)
plot(lmfit,which=4)

dfbetas(lmfit)
```

Leverage and Mahalanobis Distance:

Mahalanobis Distance between X and Y as defined in eqa(4.19)

$$d_M(\vec{x}) = \sqrt{(\vec{x}-\vec{\mu})^T S^{-1}(\vec{x}-\vec{\mu})} \tag{3.92}$$

And we can proof d_M of a case item $X_{i\cdot} = (1, X_{i1}, X_{i2}, \dots, X_{ip})$ is

$$d_M^2(X_{i\cdot}) = (n-1)(h_{ii} - \frac{1}{n}) \tag{3.93}$$

3.4.5 Extra Sum Of Square

Def. Extra SS: the part of SSE explained by a new X_2 when adding to model $Y \sim X_1$:

$$\text{SSR}(X_2|X_1) = \text{SSE}(X_1) - \text{SSE}(X_1, X_2) = \text{SSR}(X_1, X_2) - \text{SSR}(X_1) \tag{3.94}$$

where $\text{SS}(\cdot)$ represents the SS when the model contains variable $\cdot$.[20]

(The following part use model (Y, X_1, X_2) as example.)

We could use extra SS to examine the proper regression model: examine the `F value` and `Pr(>F)` in the output.

▷ **`R.` Code**

```
lm(y~x1+x2+x1:x2) %>% anova
```

Note: Three types of SS

Term	Type I SS[21]	Type II SS	Type III SS
X_1	$\text{SSR}(X_1)$	$\text{SSR}(X_1\|X_2)$	$\text{SSR}(X_1\|X_2, X_1X_2)$
X_2	$\text{SSR}(X_2\|X_1)$	$\text{SSR}(X_2\|X_1)$	$\text{SSR}(X_2\|X_1, X_1X_2)$
X_1X_2	$\text{SSR}(X_1X_2\|X_1, X_2)$	Assume no interaction term	$\text{SSR}(X_1X_2\|X_1, X_2)$
Language.Function	`R.anova`	`python.`	`SPSS,SAS,R.lm`

[20] $\text{SSE}(1) = \text{SST}$, where 1 correponds to intercept.

To get Type II and III anova, use `Anova(lmfit,type='III')` in `'car'` package.

Hierarchical Principle: the interaction term X_1X_2 should always come in **after** marginal term X_1 and X_2.

▷ `R.` **Code**

```
libaray('car')
Anova(lmfit,type='II')
Anova(lmfit,type='III')
```

3.4.6 Hypotheses Testing to Slope

Main focus: whether the linear relation exist:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \longleftrightarrow H_1 : \exists \beta_i \neq 0,\ i = 1, 2, \ldots, p \tag{3.95}$$

As for general case $H_0 : \underset{q\times(p+1)}{C} \beta - \underset{(p+1)\times 1}{t} = 0$, use General Linear Test.

- ANOVA F-Test:

  We can examine
  $$F = \frac{\text{MSR}}{\text{MSE}} \sim F_{p,n-p-1} \tag{3.96}$$

- General Linear Test (GLT)

  First we introduce the examine models:

  – Full model: Include all variable/parameters to be examined, with p variables.
  $$Y = X\beta + \varepsilon \tag{3.97}$$
  And define SSE_F with $dof_\text{F} = n - p - 1$ under Full Model.

  – Reduced model: Apply the Null Hypothesis to Full Model, with $\tilde{p}$ variables
  $$Y_i = \tilde{X}\tilde{\beta} + \varepsilon \tag{3.98}$$
  And define SSE_R with $dof_\text{R} = n - \tilde{p} - 1$ under Reduced Model.

  Then conduct test to the difference between Full model and Reduced model through SSE_F and SSE_R.

  – One dimensional case: $H_0 : \beta_1 = 0$

  Examine
  $$F = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{F})/(dof_\text{R} - dof_\text{F})}{\text{SSE}_\text{F}/dof_F} \sim F_{1,n-2} \tag{3.99}$$

  ▷ `R.` **Code**

```
fullmodel <- lmfit
nullmodel <- lm(y ~ 1,df)
anova(nullmodel,fullmodel)
```

- General case: Test $H_0: \underset{q\times(p+1)}{C}\beta - \underset{(p+1)\times 1}{t} = 0$, construct F statistics as

$$F = \frac{(C\hat{\beta}-t)'\left[C(X'X)^{-1}C'\right]^{-1}(C\hat{\beta}-t)}{q\hat{\sigma}^2} \sim F_{q,n-q}$$

- Pearson Correlation Coefficient r and Coefficient of Multiple Determination R^2:

Pearson's r:

$$r = c\hat{o}v(Y,\hat{Y}) = \frac{\sum_{i=1}^{n}(Y_i-\bar{Y})(\hat{Y}_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}\sqrt{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{Y}-\bar{Y})^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{3.100}$$

CMD R^2:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{3.101}$$

Adjusted R^2:

$$R_{\text{a}}^2 = 1 - \frac{\text{MSE}}{\text{MST}} = 1 - \frac{n-1}{n-p-1}\frac{\text{SSE}}{\text{SST}} \tag{3.102}$$

- Relation between r and R^2: Under Simple Linear Model, we have

$$R^2 = r^2 \tag{3.103}$$

- Relation between R^2 and F-Statistic:

$$F = \frac{R^2}{1-R^2}\frac{n-p-1}{n-1} \tag{3.104}$$

Hypothesis testing for r:

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

▷ R. **Code**

```
cor.text(df$x,df$y)
```

- Coefficient of Partial Determination $R^2_{Yk|\wedge X_k}$ and Coefficient of Multiple Determination R^2: CMD reflects the interpretability of the model, to examine the interpretability of each variable, use coef. partial determination

$$R^2_{YX_k|X_1,\dots,X_{k-1},X_{k+1},\dots,X_p} = R^2_{YX_k|\wedge X_k} = \frac{\text{SSR}(X_k|X_1,\dots,X_{k-1},X_{k+1},\dots,X_p)}{\text{SSE}(X_1,\dots,X_{k-1},X_{k+1},\dots,X_p)} = \frac{\text{SSR}(X_k|\wedge X_k)}{\text{SSE}(\wedge X_k)} \tag{3.105}$$

Note: Coef. Partial determination can also be used for X_i, X_j: $R^2_{X_iX_j|\wedge X_i,X_j}$

Sometimes we use $\eta_k^2 = R^2_{YX_k|\wedge k} = R^2_{Yk.\wedge k}$

- Coefficient of Partial Correlation: Measures the strength of linear relation, $\pm$ sign depend on posi./nega. correction.

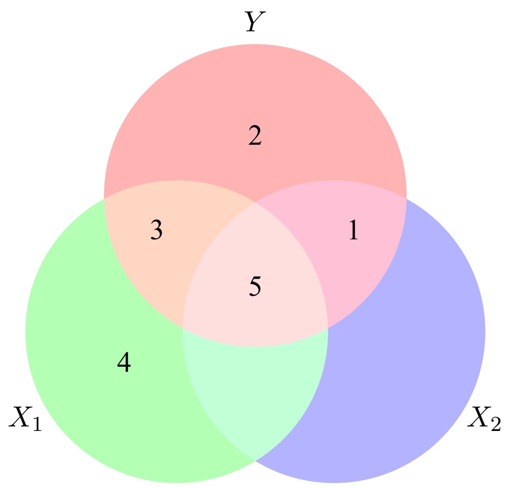
$$\eta_k = \pm\sqrt{\eta_k^2} \tag{3.106}$$

▷ R. **Code**

```r
library('heplots')
etasq(lmfit)
```

3.4.7 Diagnostics to Multi-Collinearity

- Venn Diagram for Multi-Linear Regression: Used to show the interpretability of variables.



Explanation of each region:

  - 1/3: Variation in Y uniquely attributes to X_2/X_1;
  - 2: Variation in Y that cannot be explained by regression to X_1, X_2, corresponds to ε;
  - 5: Cross term of X_1, X_2, **cannot** verify the orientation, corresponds to **Multi-Collinearity**.

In the presence of multi-collinearity, i.e. X is column singular ($\dfrac{S_5}{S_1 \text{ or } S_3}$ large), the regression parameter

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{3.107}$$

Issue of multi-collinearity:

- Statistically:'better' prediction, worse interpretability;
- Numerically: Calculation of $(X'X)^{-1}$ becomes unstable/ill-posed/NAN.

Use Variance Inflation Factor (VIF) to detect multi-collinearity.

First construct R_k^2, $k = 1, 2, \ldots p$: Regress X_k against other $p-1$ variable X_is and get corresponding R_k^2, and

$$\mathrm{VIF}_k = (1 - R_k^2)^{-1} \tag{3.108}$$

$$\overline{\mathrm{VIF}} = \frac{1}{p}\sum_{k=1}^{p} \mathrm{VIF}_k \tag{3.109}$$

If $||\mathrm{VIF}_i||_\infty > 10$ or $\overline{\mathrm{VIF}} > 1$, then we identify an excessive multi-collinearity.

▷ `R.` **Code**

```r
library('car')
vif(lmfit)
```

3.4.8 Diagnostics to Model Variable Selection

In Multi-variate regression, proper explanatory variables form a subset of all available variables.

Aim: Avoid over-fitting, get a simple explanatory model.

Comment: If we consider the model with all $p_{\max}$ variables as full, unbiased model, then model selection is a kind of **Bias-Variance Trade-Off**.

☐ Model Validation: k-Fold Cross Validation(CV):

1. Separate the dataset size n into k parts;
2. pick the i^{th} part as test set y_i, and the other $k-1$ part as train set $y_{\wedge i}$(to conduct regression, etc); then conduct prediction of model $y_{\wedge i}$ to part y_i ane get MSE_i;
3. Take average of MSE_i as the measure of validity.

☐ Evaluation Criteria

Useful model validation approach: To check a model with $p-1$ variable (this part p for 1+# variable)

- Traditional way: Test R^2, R_a^2, p-value, etc.
- Mallow's C_p: For a model with p variable:

$$\hat{Y}^p = X_p(X_p'X_p)^{-1}X_p'Y = H_pY$$

Denote:

$$E(\hat{Y}^p) = H_pE(Y) \equiv H_p\mu \qquad var(\hat{Y}^p) = H_p\sigma^2 I_n H_p' = \sigma^2 H_p$$

Recall the MSE expansion of bias-variance trade-off in eqa(2.29)[22]

---

[22]Derivtion:

– Bias part: (Here use eqa(4.37) in 3rd line; use eqa(3.51) in 4th line.)

$$\begin{aligned}\sum_{i=1}^{n}[e(\hat{Y}_i^p-\mu_i)]^2 =&\mu'(H_p-I)'(H_p-I)\mu\\ =&\mu(I-H_p)\mu'\\ =&E(Y'(I-H_p)Y)-tr[(I-H_p)\sigma^2]\\ =&E(\text{SSE}(p))-(n-p)\sigma^2\end{aligned}$$

– Variance part:

$$\sum_{i=1}^{n}var(\hat{Y}_i^p) =tr(var(\hat{Y}^p)) = \sigma^2 tr(H_p) = p\sigma^2$$

Then

$$\frac{\sum_{i=1}^{n}E[(\hat{Y}_i^p-\mu_i)^2]}{\sigma^2} = \frac{E(\text{SSE}(p))}{\sigma^2} - (n-2p)$$

$$\sum_{i=1}^{n} E[(\hat{Y}_i^p - \mu_i)^2] = \sum_{i=1}^{n} E(\hat{Y}_i^p) - \mu_i]^2 + \sum_{i=1}^{n} var(\hat{Y}_i^p)$$
$$\Rightarrow = E(\text{SSE}(p)) - (n-2p)\sigma^2$$

Sum Squared Prediction Error (SSPE):

$$\Gamma_0 \equiv \frac{\sum_{i=1}^{n} E[(\hat{Y}_i^p - \mu_i)^2]}{\sigma^2} = \frac{E(\text{SSE}(p))}{\sigma^2} - (n-2p) \tag{3.110}$$

And construct Mallow's $C-p$: Estimation of Γ_p

$$C_p = \hat{\Gamma}_p = \frac{E(SSE(p))}{\hat{\sigma}^2} - (n-2p) \tag{3.111}$$

where SSE(p) $= Y'(I-H_p)Y$.

When the model is unbiased, then $E(SSE(p)) \to n-p$, use C_p-p plot to pick proper p:

- $C_p \approx p$: Model unbiased, then choose model with smaller C_p;
- $C_p \gg p$: Significant biased, miss some important predictors;
- $C_p \ll p$: Overfitting.

- Akaike Information Criterion (AIC): Euivalent to Mallow's C_p for gaussion regression model.

$$\text{AIC}(p) = -2\log(\hat{L}) + 2p$$

where $\hat{L}$ is the maximum likelihood, for linear regression case

$$\text{AIC}(p) = n\log\left(\frac{\text{SSE}(p)}{n}\right) + 2p$$

Select the model that minimizes AIC(p).

- Bayesian Information Criterion (BIC)/Schwarz's Bayesian Criterion (SBC):

$$\text{BIC}(p) = -2\log(\hat{L}) + p\log n$$

where $\hat{L}$ is the maximum likelihood, for linear regression case

$$\text{BIC}(p) = n\log\left(\frac{\text{SSE}(p)}{n}\right) + p\log n$$

Select the model that minimizes BIC(p).

- PRESS Creterion (Predictive Residual Error Sum of Squares): A kind of within-model cross validation

$$\text{PRESS}(p) = \sum_{i=1}^{n}(Y_i - \hat{Y}_{i(\wedge i)})^2$$

where

$$\begin{aligned}\hat{Y}_{i(\wedge i)} =&(1, X_{i1}, \ldots, X_{ip})\hat{\beta}_{(\wedge i)}\\ \hat{\beta}_{(\wedge i)} =&(X'_{(\wedge i)}X_{(\wedge i)})^{-1}X'_{(\wedge i)}Y_{(\wedge i)}\end{aligned}$$

where $\hat{\beta}_{(\wedge i)}$ as in eqa(3.89), is the estimated β with (X_i, Y_i) removed from X.[23]

Select the model that minimizes PRESS(p).

▷ R. **Code**

```
library('leaps')
predictor <- df[,c('...','...',...)]
response <- df[,...]
leapSet <- leaps(x=predictor, y=response, nbest = ...)
% method=c('Cp','adjr2','r2')
leapSet$which[which.min(leapSet$Cp),]
```

`nbest` for `NUMBER_OF_BEST_MODELS`

Section 3.5 Remedies

3.5.1 Variable Transformation

The goal of Transformation:

- Stablize Variance;
- Improve Normality;
- Simplify the Model.

☐ Transformation Methods:

- Variance Stabilizing Transformations: For $E(Y_X) = \mu_X$, $var(Y_X) = h(\mu_X)$, take transformation $f(Y)$ such that $var(f(Y)) = \text{const}$, satisfies

$$f(\mu) = \int \frac{c\,\mathrm{d}\mu}{\sqrt{h(\mu)}} \tag{3.112}$$

Examples:

$$\begin{aligned}h(\mu) =&\mu^2 \Rightarrow f(\mu) = \ln\mu\\ h(\mu) =&\mu^{2\nu} \Rightarrow f(\mu) = \mu^{1-\nu}\end{aligned}$$

- Box-Cox Transformation: Take

$$Y^* = \frac{Y^\lambda - 1}{\lambda} \tag{3.113}$$

[23] A useful thm.: Deleted Residual

$$d_i := Y_i - \hat{Y}_{i(\wedge i)} = \frac{e_i}{1 - h_{ii}}$$

Examples:

$$\lambda = 1 \Rightarrow Y^* \sim Y$$
$$\lambda = 0.5 \Rightarrow Y^* \sim \sqrt{Y}$$
$$\lambda = 0 \Rightarrow Y^* \sim \ln Y$$
$$\lambda = -1 \Rightarrow Y^* \sim 1/Y$$

And conduct regression to model

$$Y^* = \beta_0 + \beta_1 X + \varepsilon_i \tag{3.114}$$

Likelihood Function

$$L(\beta, \sigma^2; \lambda) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i^* - \beta_0 - \beta_1 X_i)^2 \right) J(\frac{\partial Y^*}{\partial Y}) \tag{3.115}$$

where the Jacobi Matrix denoted in Geometric Mean $\mathrm{GM}(Y) = \prod_{i=1}^n Y_i^{1/n}$

$$J(\frac{\partial Y^*}{\partial Y}) = \prod_{i=1}^{n} Y_i^{\lambda-1} = \mathrm{GM}(Y)^{n(\lambda-1)} \tag{3.116}$$

MLE Estiamtor:

$$\hat{\beta}^* = (X'X)^{-1}X'Y^*$$
$$\hat{\sigma}_n^2 = \frac{1}{n}\mathrm{SSE}^*$$
$$\mathrm{SSE}^* = \sum_{i=1}^{n} (Y_i^* - \hat{Y}_i^*)^2$$

And when β, σ^2 take MLE estimator, $L(\beta, \sigma^2; \lambda)$ can be regarded a function of λ:

$$\ln L(\beta, \sigma^2; \lambda) = l(\lambda) = -\frac{n}{2} \ln \frac{\hat{\sigma}_n^2}{\mathrm{GM}(Y)^{2(\lambda-1)}} + \text{const} \tag{3.117}$$

For simplification, denote $Z = Y * / J^{1/n}$ and get

$$l(\lambda) = -n \ln \sigma_{n_Z}^2 + \text{const} \tag{3.118}$$

where

$$Z_i^* = \begin{cases} \dfrac{Y_i^\lambda - 1}{\lambda} \dfrac{1}{\prod\limits_{k=1}^{n} Y_k^{\frac{\lambda-1}{n}}}, & \lambda \neq 0 \\ \ln Y_i \prod\limits_{k=1}^{n} Y_k^{\frac{1}{n}}, & \lambda = 0 \end{cases} \tag{3.119}$$

Plot $l(\lambda)$-λ to determine a proper λ and transform $Y^* = \dfrac{Y^\lambda - 1}{\lambda}$:

– Selected λ should be closed to $\lambda_{\arg\max l}$, at least within CI[24]

$$\{\lambda | l(\lambda) \geq l(\lambda_{\arg\max l}) - \frac{1}{2}\chi^2_{1,1-\alpha}\} \tag{3.120}$$

[24]Here CI can be derived using Wilk's Thm.

– Should pick a λ which is **Interpretable**. e.g. If $\lambda = 1$ is within range $[0.94, 1.08]$, then take $\lambda = 1$ (does not transform).

▷ R. **Code**

```
library(MASS)
bctrans <- boxcox(y~x,df,lambda = seq(-1.5, 1.5, length = 15))
bctrans$x[which.max(bctrans$y)]
```

Note: we can transform on X or Y or simultaneously to get better regression model.

3.5.2 Weighted Least Squares Regression

To deal with heterogeneous variance, use Weighted Least Squares (WLS) instead of OLS: Minimizing

$$\sum_{i=1}^{n} e_i^2 \longrightarrow \sum_{i=1}^{n} w_i e_i^2$$

And e.g. take weight for each case as

$$w_i = \frac{1}{\sigma_i^2}$$

▷ R. **Code**

```
Wlmfit <- lm(y~x,weights=WEIGHT_VECTOR,data=df)
```

3.5.3 Remedies for Model Variable Selection

Several Algorithm to search for best variable set:

- Exhaustive Search and Test: Used for $p \leq\sim 30$
- Greedy Search: Get a locally optimal solution.
  - Forward Selection: Start with $p = 0$, add one variable each times and conduct $t/F/p$-value test until a presupposed certain limit.
  - Backward Elimination: Start with $p_{\max}$, eliminate one variable each times and conduct $t/F/p$-value test until a presupposed certain limit.
  - Stepwise Regression: Alternate forward selection & backward elimination until no add/elimination.
- Penalized Optimization:

  Recall: OLS regression model: Minimize SSE[25]

$$\hat{\beta} = \arg\min ||Y - X\beta||_2^2 \tag{3.121}$$

  Idea: Add a penalty term in SSE, such that SSE increases with # of variable.

[25] Here expressed in ℓ_p norm, definition see sec.4.1.2, Norm

– LASSO (Least Absolute Shrinkage and Selection Operator)

Penalty term: $\lambda||\beta||_1$), where λ is a proper penalty parameter.

$$\hat{\beta} = \arg\min(||Y - X\beta||_2^2 + \lambda||\beta||_1) \tag{3.122}$$

or equivlantly expressed as

$$\hat{\beta} = \arg\min ||Y - X\beta||_2^2, \text{ with } ||\beta||_1 \leq s \tag{3.123}$$

where s is a parameter correponding to λ.

– Ridge Regression/Tikhonov Regularization:

Penalty term: $\lambda||\beta||_2$), where λ is penalty parameter.

$$\hat{\beta} = \arg\min(||Y - X\beta||_2^2 + \lambda||\beta||_2) \tag{3.124}$$

or equivlantly expressed as

$$\hat{\beta} = \arg\min ||Y - X\beta||_2^2, \text{ with } ||\beta||_2^2 \leq s \tag{3.125}$$

▷ R. **Code**

```
library('MASS')
Rfit <- lm.ridge(y~x,lambda=seq(0,0.1,0.001),data=df)
summary(Rfit)
whichLambda <- which.min(Rfit$GCV)
coef(fits)[whichLambda,]

library('lars')
Lfit <- lars(x,y,type='lasso')
summary(Lfit)
whichCp <- which.min(Lfit$Cp)
Lfit$Cp[whichCp]
Lfit$beta[whichCp,]
```

- Non-parametric Regression Model: Add smooth/penalty function.

  Example: `loess` (Locally Regression), `lowess` (Locally Weighted ScatterPlot Smoother), Regression Tree.

- Other Regression Model: 、

  – Standardized Regression Model For regression model $Y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i$, $i = 1, 2, \ldots, n$, conduct Standardization (with an extra const $1/\sqrt{n-1}$) to Y and X.

$$Y_i^* = \frac{1}{\sqrt{n-1}} \frac{Y_i - \bar{Y}}{s_Y} \qquad X_{ij}^* = \frac{1}{\sqrt{n-1}} \frac{X_{ij} - \bar{X}_i}{s_{X_i}} \qquad \varepsilon_i^* = \frac{1}{\sqrt{n-1}} \frac{\varepsilon_i - \bar{\varepsilon}}{s_Y} \tag{3.126}$$

And the regreeesion model for standardized data:

$$Y_i^* = 0 + \sum_{j=1}^{n} X_{ij}^* \beta_j^* + \varepsilon_i^* \tag{3.127}$$

with

$$\beta_j^* = \frac{\beta_j s_{X_j}}{s_Y} \tag{3.128}$$

Note: set the const as $\sqrt{n-1}$ so that

$$r_{X^*X^*} = X^{*T} X^* \qquad r_{Y^*X^*} = X^{*T} Y^* \tag{3.129}$$

▷ R. **Code**

```
scaledf <- data.frame(scale(df))
scalelmfit <- lm(~,scaledf)
summary(scalelmfit)
```

– Polynomial Regression Model

▷ R. **Code**

```
polfit <- lm(y~x+I(x^2),df)
polfit <- lm(y~polym(x1,x2,degree=),df)
```

– Interaction Model Example:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Re-write as

$$\begin{aligned} Y =& \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \varepsilon \\ Y =& \beta_0 + \beta_1 X_2 + (\beta_2 + \beta_3 X_1) X_2 + \varepsilon \end{aligned}$$

test the regression coefficient dependence on another variable.

Section 3.6 Factor Analysis

Chapter. IV 多元统计分析部分

Instructor: Dong Li & Tianying Wang

Section 4.1 Multivariate Data

In this section, we consider a **Multivariate Statistic Model**. Sample comes from p dimension multivariate population $f(x_1, x_2, \ldots, x_p)$.

Notation : In this section, we still denote random variable in upper case and observed value in lower case, specially express random vector in bold font. **But** in this section we usually omit the vector symbol $\vec{\cdot}$. e.g. random vector with n **variable** is denoted as $\mathbf{X} = (X_{\cdot 1}, X_{\cdot 2}, \ldots, X_{\cdot p})$; sample of size n from the multivariate population is a $n \times p$ matrix $\{x_{ij}\}$, each sample item (a row in sample matrix) is denoted as x_i' or x_i^T.[26]

4.1.1 Matrix Representation

- Random Variable Representation
- Sample Representation
- Statistics Representation
- Sample Statistics Properties

☐ Random Variable Representation:

- Random Matrix: Definition and basic properties of r.v. see section 1.3. Now extend the definition to matrix $X = \{X_{ij}\}$.

$$X = \{X_{ij}\} = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{n2} & \ldots & X_{np} \end{bmatrix} \tag{4.1}$$

And we can further define $E(X) = \{E(X_{ij})\}$. For any const matrix A, B we have

$$E(AXB) = AE(X)B \tag{4.2}$$

- Random Vector: For a $p \times 1$ random vector $\vec{X} = (X_1, X_2, \ldots, X_p)^T$, denote (Marginal) expectation and variance, and covariance, correlation coefficient between X_i, X_j as follows:

$$\begin{aligned} \mu_i &= E(X_i) \\ \sigma_{ii} &= \sigma_i^2 = E(X_i - \mu_i)^2 \\ \sigma_{ij} &= E[(X_i - \mu_i)(X_j - \mu_j)] \\ \rho_{ij} &= \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} \end{aligned}$$

[26] Here sample item (or sample case) $x_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]^T$ is a column vector.

and we have covariance matrix (as defined in section 1.4.3, eqa.1.44)

$$\Sigma = E[(X-\mu)(X-\mu)^T] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix} \tag{4.3}$$

and Standard Deviation Matrix

$$V^{1/2} = diag\{\sqrt{\sigma_{ii}}\} \tag{4.4}$$

Based on $\vec{X} = (X_1, X_2, \dots, X_p)$, consider the linear combination:$Y = c'X = c_1X_1 + c_2X_2 + \dots c_pX_p$

$$E(y) = c'\mu \qquad var(Y) = c'\Sigma c$$

and $Z_i = \sum_{j=1}^{p} c_{ij}X_j$ (i.e. $Z = CX$):

$$\mu_Z = E(Z) = C\mu_X \qquad \Sigma_Z = C\Sigma_X C^T \tag{4.5}$$

and Correlation Matrix

$$\rho = \begin{bmatrix} \rho_{11} & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{p2} & \dots & \rho_{pp} \end{bmatrix} = V^{-1/2}\Sigma V^{-1/2} \tag{4.6}$$

☐ Sample Representation:

Sample of n items from population characterized by p variables

Variable / Item	Variable 1	Variable 2	$\dots$	Variable j	$\dots$	Variable p
Item 1	x_{11}	x_{12}	$\dots$	x_{1j}	$\dots$	x_{1p}
Item 1	x_{21}	x_{22}	$\dots$	x_{2j}	$\dots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
Item j	x_{i1}	x_{i2}	$\dots$	x_{ij}	$\dots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
Item n	x_{n1}	x_{n2}	$\dots$	x_{nj}	$\dots$	x_{np}

Or represented in condense notation:

$$X = \{x_{ij}\} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & \dots & y_p \end{bmatrix} \tag{4.7}$$

☐ Statistics Representation

- Unit 1 vector:

$$\mathbf{1}_k = (\underbrace{1, 1, \ldots, 1}_{k \text{ 1 in total}})^T \tag{4.8}$$

Unit 1 matrix:

$$\mathcal{J}_n = \{1\}_{n\times n} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 1 \end{bmatrix}_{n\times n} \tag{4.9}$$

- Sample mean:

$$\bar{x}_i = \frac{x_{1i} + x_{2i} + \ldots + x_{ni}}{n} = \frac{y_i'\mathbf{1}_n}{n} \tag{4.10}$$

- Deviation of measurement of the i^{th} variable:

$$d_i = y_i - \bar{x}_i\mathbf{1}_n = \begin{bmatrix} x_{1i} - \bar{x}_i \\ x_{2i} - \bar{x}_i \\ \vdots \\ x_{ni} - \bar{x}_i \end{bmatrix} \tag{4.11}$$

- Covariance Matrix:

  - Variance of y_i:

$$s_i^2 = s_{ii} = \frac{1}{n}d_i'd_i = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2, \quad i = 1, 2, \ldots p \tag{4.12}$$

  - Covariance between y_i and y_j:

$$s_{ij} = \frac{1}{n}d_i'd_j = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j), \quad i, j = 1, 2, \ldots p \tag{4.13}$$

  - Correlation Coefficient:

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}} = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \bar{x}_j)^2}}, \quad i, j = 1, 2, \ldots p \tag{4.14}$$

In condense notation, define Covariance Matrix from sample of size n:

$$S_n = \begin{bmatrix} s_{11} & s_{12} & \ldots & s_{1p} \\ s_{21} & s_{22} & \ldots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{p2} & \ldots & s_{pp} \end{bmatrix} \tag{4.15}$$

and sample Correlation Coefficient Matrix:

$$R_n = \begin{bmatrix} r_{11} & r_{12} & \ldots & r_{1p} \\ r_{21} & r_{22} & \ldots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{p2} & \ldots & r_{pp} \end{bmatrix} \tag{4.16}$$

- Generalized sample variance: $|S| = \lambda_1\lambda_2 \dots \lambda_p$, where λ_i are eigenvalues.
- 'Statistical Distance' between vectors: to measure the difference between two vectors $x = (x_1, x_2, \dots, x_p)$ and $y = (y_1, y_2, \dots, y_p)$.
  - Euclidean Distance:
  $$d_E(x, y) = \sqrt{(x-y)^T(x-y)} \tag{4.17}$$
  - **Mahalanobis Distance**: Scale invariant distance, and include information about relativity:
  $$d_M(x, y) = \sqrt{(x-y)'S^{-1}(x-y)} \tag{4.18}$$
  Note: P, Q are from the same distribution with covariance matrix S_p. When $S = I$, return to Euclidean distance.

    Remark: Mahalanobis distance is actually the normalized Euclidean distance in principal component space. So we can actually define the Mahalanobis distance for one sample case $\vec{x} = (x_1, x_2, \dots, x_p)$ from distribution of $(\vec{\mu}, \Sigma)$
  $$d_M(\vec{x}) = \sqrt{(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})} \tag{4.19}$$
  Note: the hyper-sruface $d_M(\vec{x})$ forms a ellipsoid.

☐ Sample Statistics Properties

Consider take an n cases sample from r.v. population $\vec{X} = (X_1, X_2, \dots, X_p)$, population mean μ and covariance matrix Σ.

- $E(\bar{\bar{X}}) = \mu$;
- $cov(\bar{X}) = \frac{1}{n}\Sigma$;
- $E(S_n) = \frac{n-1}{n}\Sigma$

4.1.2 Review: Some Matrix Notation & Lemma

- Orthonormality: For square matrix P satisfies:
$$x_i^T x_j = \delta_{ij} \tag{4.20}$$
  where x_i, x_j are columns of P.
- Eigenvalue and Eigenvector: For square matrix A, its eigenvalues λ_i and corresponding eigenvectors e_i satisfies:
$$Ae_i = \lambda_i e_i,\ \forall i = 1, 2, \dots p \tag{4.21}$$
  Denote $P = [e_1, e_2, \dots, e_p]$, which is an orthonormal matrix. And denote $\Lambda = diag\{\lambda_1, \lambda_2, \dots, \lambda_p\}$.
$$A = \sum_{i=1}^{p} \lambda_i e_i e_i^T = P\Lambda P^T = P\Lambda P^{-1} \tag{4.22}$$
  is called the Spectral Decomposition of A

- Square root matrix: Def. as

$$A^{1/2} = \sum_{i=1}^{p} \sqrt{\lambda_i} e_i e_i^T = P\Lambda^{1/2}P^T \tag{4.23}$$

Properties:

  - $A^{1/2}A^{1/2}A$;
  - $A^{-1/2} = (A^{1/2})^{-1} = PL^{-1/2}P^T$;
  - $tr(A) = \sum_{i=1}^{n} \lambda_n$;
  - $|A| = \prod_{i=1}^{n} \lambda_n$.

- (Symmetric) Positive Definite Matrix: Say A a Positive Definite Matrix if

$$x^T Ax > 0, \forall x \in \mathbb{R}^p \tag{4.24}$$

where $x^T Ax$ is called a Quadric Form.

Properties:

  - Use the Spectral Decomposition of A, we can write the Quadric Form as

$$x^T Ax = x^T P\Lambda P^T x = y^T \Lambda y = \sum_{i=1}^{p} \lambda_i y_i^2 = \sum_{i=1}^{p} (\sqrt{\lambda_i} y_i)^2 \tag{4.25}$$

  - Eigenvalues $\lambda_i > 0$, $\forall i = 1, 2, \ldots, p$
  - A can be written as product of symmetric matrix: $A = Q^T Q$ (Q is symmetric);

- Trace of Matrix: For $p \times p$ square matrix A

$$tr(A) = \sum_{i=1}^{p} a_{ii} \tag{4.26}$$

Properties:

  - $tr(AB) = tr(BA)$;
  - $x'Ax = tr(x'Ax) = tr(Axx')$

- Matrix Partition: partition matrix $\underset{p\times p}{A}$ as

$$A = \begin{bmatrix} \underset{q_1\times q_1}{A_{11}} & \underset{q_1\times q_2}{A_{12}} \\ \underset{q_2\times q_1}{A_{21}} & \underset{q_2\times q_2}{A_{22}} \end{bmatrix}$$

where $p = q_1 + q_2$

Property:

$$|A| = |A_{22}||A_{11} - A_{12}A_{22}^{-1}A_{21}| = |A_{11}||A_{22} - A_{21}A_{11}^{-1}A_{12}|$$

| A |=| A22 || A11 - A12A-122A21 |=| A11 || A22 - A21A-111A12 |

- Matrix Differentiation

  Calculus Notations: We want to take derivative of $y = (y_1, y_2, \dots, y_q)^T$ over $x = (x_1, x_2, \dots, x_p)^T$

  We use 'Denominator-layout', which is

$$\frac{\partial y}{\partial x} = \frac{\partial y^T}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \dots & \frac{\partial y_q}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \dots & \frac{\partial y_2}{\partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_p} & \frac{\partial y_2}{\partial x_p} & \dots & \frac{\partial y_q}{\partial x_p} \end{bmatrix} \tag{4.27}$$

  Properties (under denominator-layout):[27]

  - $\frac{\partial}{\partial x} Ax = A^T$;
  - $\frac{\partial}{\partial x} x^T A = A$;
  - $\frac{\partial}{\partial x} x^T x = 2x$;
  - $\frac{\partial}{\partial x} x^T Ax = Ax + A^T x$;
  - $\frac{\partial}{\partial x} \log(x^T Ax) = \frac{2Ax}{x^T Ax}$;
  - $\frac{\partial |A|}{\partial A} = |A|A^{-1}$;
  - $\frac{\partial tr(AB)}{\partial A} = B^T$;
  - $\frac{\partial tr(A^{-1}B)}{\partial A} = -A^{-1}B^T A^{-1}$

- Kronecker Product: For matrix $\underset{m\times n}{A} = \{a_{ij}\}$, $\underset{p\times q}{B} = \{b_{ij}\}$. Their Kronecker product

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m}B & a_{m2}B & \dots & a_{mn}B \end{bmatrix} \tag{4.28}$$

- Norm: Denote $x = [x1, x2, \dots, x_n]^T \in \mathbb{R}^n$, then the ℓ_p-norm of x is

$$||x||_p \equiv \left(\sum_{i=1}^{n} ||x_i|^p \right)^{1/p}$$

  Useful norm:

[27] More matrix diffrenciation equation see book [5] P49.

- ℓ_0-norm: # of none 0 elements in x;
- ℓ_1-norm: $||x||_1 = \sum_{i=1}^n |x_i|$;
- ℓ_2-norm/Euclidean norm: $||x||_2 = \sqrt{\sum_{i=1}^n x_i^2}$.

4.1.3 Useful Inequalities

- Cauchy-Schwartz Inequality:

  Let b, d are any $p \times 1$ vectors.

$$(b'd)^2 \leq (b'b)(d'd) \tag{4.29}$$

- Extended Cauchy-Schwartz Inequality:

  Let B be a positive definite matrix.

$$(b'd)^2 \leq (b'Bb)(d'B^{-1}d) \tag{4.30}$$

- Maximazation Lemma:

  d be a given vector, for any non-zero vector x,

$$\frac{(x'd)^2}{x'Bx} \leq d'B^{-1}d \tag{4.31}$$

  Take Maximum when $x = cB^{-1}d$.

Section 4.2 Statistical Inference to Multivariate Population

Statistics model: a n cases sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, where each $\mathbf{X}_i$ i.i.d. from a multivariate population (usually consider a multi-normal). i.e.

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{n2} & \ldots & X_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{bmatrix} \tag{4.32}$$

4.2.1 Multivariate Normal Distribution

Univariate Noraml Distribution: $N(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\mu)^2}{2\sigma^2} \tag{4.33}$$

Multivariate Normal Distribution: $X \sim N_p(\vec{\mu}, \Sigma)$[28]

$$f_{\mathbf{X}}(\vec{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{(\vec{x}-\vec{\mu})'\Sigma^{-1}(\vec{x}-\vec{\mu})}{2}\right) \tag{4.34}$$

Note: Here in the exp, the $(\vec{x}-\vec{\mu})'\Sigma^{-1}(\vec{x}-\vec{\mu})$ is the Mahalanobis Distance d_M defined in eqa.4.19

[28]Detailed derivation see section 1.8

Remark: A n-dimension multivariate normal has $\frac{p(p+1)}{2}$ free parameters. Thus for a very high dimension, contains too many free parameters to be determined!

Properties: Consider $X \sim N_p(\mu, \Sigma)$

- Linear Transform:
  - For a $p \times 1$ vector a:

$$X \sim N_p(\mu, \Sigma) \Leftrightarrow a'X \sim N(a'\mu, a'\Sigma a), \forall a \in \mathbb{R}^p \tag{4.35}$$

    (Proof: use characteristic function.)

  - For a $q \times p$ const matrix A:

$$AX + a \sim N_q(A\mu + a, A\Sigma A') \tag{4.36}$$

  - For a $p \times p$ square matrix A:

$$E(X'AX) = \mu' A\mu + tr(A\Sigma) \tag{4.37}$$

- Conditional Distribution: Take partition of $\underset{p\times 1}{X} \sim N(\underset{p\times 1}{\mu}, \underset{p\times p}{\Sigma})$ into $\underset{q_1\times 1}{X_1}$ and $\underset{q_2\times 1}{X_2}$, where $q_1 + q_2 = p$. Write in matrix form:

$$\underset{p\times 1}{X} = \begin{bmatrix} \underset{q_1\times 1}{X_1} \\ \underset{q_2\times 2}{X_2} \end{bmatrix} \qquad \underset{p\times 1}{\mu} = \begin{bmatrix} \underset{q_1\times 1}{\mu_1} \\ \underset{q_2\times 2}{\mu_2} \end{bmatrix} \qquad \underset{p\times p}{\Sigma} = \begin{bmatrix} \underset{q_1\times q_1}{\Sigma_{11}} & \underset{q_1\times q_2}{\Sigma_{12}} \\ \underset{q_2\times q_1}{\Sigma_{21}} & \underset{q_2\times q_2}{\Sigma_{22}} \end{bmatrix} \tag{4.38}$$

  i.e.

$$\underset{p\times 1}{X} = \begin{bmatrix} \underset{q_1\times 1}{X_1} \\ \underset{q_2\times 2}{X_2} \end{bmatrix} \sim N_{q_1+q_2}\left(\begin{bmatrix} \underset{q_1\times 1}{\mu_1} \\ \underset{q_2\times 2}{\mu_2} \end{bmatrix}, \begin{bmatrix} \underset{q_1\times q_1}{\Sigma_{11}} & \underset{q_1\times q_2}{\Sigma_{12}} \\ \underset{q_2\times q_1}{\Sigma_{21}} & \underset{q_2\times q_2}{\Sigma_{22}} \end{bmatrix} \right) \tag{4.39}$$

  Independence: $X_1 \parallel X_2 \Leftrightarrow \Sigma_{21} = \Sigma_{12}^T = 0$

  And the conditional dictribution $X_1|X_2 = x_2$ is given by [29]

$$X_1|_{X_2=x_2} \sim N_p(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \tag{4.40}$$

- Multivariate Normal & χ^2

  Let $X \sim N_p(\mu, \Sigma)$, then

$$(X-\mu)^T\Sigma^{-1}(X-\mu) \sim \chi_p^2 \tag{4.41}$$

- Linear Combination: Let $X_1, X_2 \ldots, X_n$ with $X_i \sim N_p(\mu_i, \Sigma)$ (different μ_i, same Σ). And denote $V_1 = \sum_{i=1}^n c_i X_i$, then

$$V_1 \sim N_p(\sum_{i=1}^n c_i\mu_i, \sum_{i=1}^n c_i^2 \Sigma) \tag{4.42}$$

[29] In eqa(4.36), take

$$\underset{p\times p}{A} = \begin{bmatrix} \underset{q\times q}{I} & \underset{q\times(p-q)}{-\Sigma_{12}\Sigma_{22}^{-1}} \\ \underset{(p-q)\times q}{0} & \underset{(p-q)\times(p-q)}{I} \end{bmatrix}$$

4.2.2 MLE of Multivariate Normal

Under the notation in eqa(4.32), i.e. each sample case $\mathbf{X}_i$ i.i.d. $\sim N_p(\mu, \Sigma)$, we can get the joint PDF of $\mathbf{X}$:

$$f_{\mathbf{X_1},\ldots,\mathbf{X_n};\mu,\Sigma}(x_1,\ldots,x_n) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp\left(-\sum_{i=1}^{n}\frac{(x_i-\mu)'\Sigma^{-1}(x_i-\mu)}{2}\right) \tag{4.43}$$

and at the same time get likelihood function[30]:

$$L(\mu,\Sigma;x_1,\ldots,x_n) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp\left[-\frac{1}{2}tr\left(\Sigma^{-1}\left(\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})' + n(\bar{x}-\mu)(\bar{x}-\mu)'\right)\right)\right] \tag{4.45}$$

And we can get the MLE of μ and Σ as follows[31]:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$
$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})' = \frac{n-1}{n}S$$

And we can furthur construct MLE of function of μ, Σ (use invariance property of MLE), for example

$$|\hat{\Sigma}| = |\hat{\Sigma}|$$

Note: $(\hat{\mu}, \hat{\Sigma})$ is sufficient statistic of multi-normal population.

4.2.3 Sampling distribution of $\bar{X}$ and S

$\hat{\mu} = \bar{X}$ and $\hat{\Sigma} = \dfrac{n-1}{n}S$ are statistics, with sampling distribution.

☐ Sampling distribution of $\bar{X}$

Similar to monovariate case:

$$\bar{X} \sim N_p(\mu, \frac{1}{n}\Sigma)$$

☐ Sampling distribution of S^2

- Monovariate case: Consider $(X_1, X_2, \ldots, X_n)$ i.i.d. $\sim N(\mu, \sigma^2)$

  Then
  $$\frac{(n-1)S}{\sigma^2} \sim \chi^2_{n-1}$$

- Multivariate case: Consider $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ i.i.d. $\sim N_p(\mu, \Sigma)$

  Then
  $$(n-1)S \sim W_p(n-1, \Sigma)$$

---

[30] Here we need to use the property of trace

$$x'Ax = tr(x'Ax) = tr(Ax'x) \tag{4.44}$$

[31] Detailed proof see '*Applied Multivariate Statistical Analysis*' P130

Where $W_p(n-1,\Sigma)$ is Wishart Distribution, details as follows:

For r.v. $Z_1, Z_2, \ldots, Z_m$ i.i.d. $\sim N_p(0,\Sigma)$, def p dimensional **Wishart Distribution** with dof m as $W_p(m,\Sigma)$.[32]

$$W_p = \sum_{i=1}^{n} Z_i Z_i' \tag{4.46}$$

PDF of $W_p(m,\Sigma)$:

$$f_W(w;p,m,\Sigma) = \frac{|w|^{\frac{m-p-1}{2}} \exp\left(-\frac{1}{2}tr(\Sigma^{-1}w)\right)}{2^{\frac{mp}{2}}|\Sigma|^{-1/2}\pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma(\frac{m-i+1}{2})} \tag{4.47}$$

C.F.

$$\phi(T) = |I_p - 2i\Sigma T|^{-\frac{m}{2}} \tag{4.48}$$

Properties:

– For independent $A_1 \sim W_p(m_1,\Sigma)$ and $A_2 \sim W_p(m_2,\Sigma)$, then

$$A_1 + A_2 \sim W_p(m_1+m_2,\Sigma)$$

– For $A \sim W_p(m,\Sigma)$, then

$$CAC' \sim W_p(m, C\Sigma C')$$

– Wishart distribution is the matrix generization of χ^2_n. When $p=1$, $\Sigma=\sigma^2=1$, $W_p(m,\Sigma)$ naturally reduce to χ^2_m.

$$\chi^2_n = W_1(n,1)$$

▷ `R.` **Code**

Distribution functions are in package `MCMCpack`, or use `rWishart()` function.

☐ Large sample $\bar{X}$ and S^2

- $\sqrt{n}(\bar{X}-\mu) \xrightarrow{\mathscr{L}} N_p(0,\Sigma)$;
- $n(\bar{X}-\mu)'S^{-1}(\bar{X}-\mu) \xrightarrow{\mathscr{L}} \chi^2_p$

4.2.4 Hypothesis Testing for Normal Population

- **One-Population Hypothesis Testing**:

Conduct hypothesis testing to μ:

$$H_0: \mu = \mu_0 \longleftrightarrow H_1: \mu \neq \mu_0$$

☐ Hotelling's T^2 test

– One-Dimensional case: t-test

$$T = \frac{\sqrt{n}(\bar{X}-\mu_0)}{S} \sim t_{n-1}$$

i.e.

$$T^2 = [\sqrt{n}(\bar{X}-\mu_0)]S^{-1}[\sqrt{n}(\bar{X}-\mu_0)] \sim t^2_{n-1} = F_{1,n-1}$$

[32] $W_p(m,\Sigma)$ is a distribution defined on $p \times p$ matrix space.

– Multi-Dimensional case: Hotelling's T^2

$$T^2 = [\sqrt{n}(\bar{X}-\mu_0)']S^{-1}[\sqrt{n}(\bar{X}-\mu_0)] \sim N_p(0,\Sigma)'\frac{W_p(n-1,\Sigma)}{n-1}N_p(0,\Sigma) = \frac{p}{n-p}(n-1)F_{p,n-p}$$

And we can get the distribution of **Hotelling's** T^2:

$$\frac{n-p}{p}\frac{T^2}{n-1} \sim F_{p,n-p}$$

Rejection Rule:

$$T^2 > \frac{p(n-1)}{n-p}F_{p,n-p,\alpha}$$

Property:

Invariant for X transform: For $Y = CX + d$, then

$$T_Y^2 = n(\bar{X}-\mu_0)'S^{-1}(\bar{X}-\mu_0) = T_X^2$$

☐ LRT of $\hat{\mu}$

Monovariate case see sec.2.4.3.

LRT uses the statistic:

$$\Lambda = \frac{\max_{H_0} L(\mu_0,\Sigma)}{\max_{H_0\cup H_1} L(\mu,\Sigma)} = (1+\frac{T^2}{n-1})^{-n/2}$$

where $T^2 = n(\bar{x}-\mu_0)'S^{-1}(\bar{x}-\mu_0)$

- **Two-Population Hypothesis Testing**:

Conduct hypothesis testing to $\delta = \mu_1 - \mu_2$:

$$H_0 : \delta = \delta_0 \longleftrightarrow H_1 : \delta \neq \delta_0$$

Notation: The two sample of size n_1, n_2, each denoted as

$$X_{1,ij} \qquad X_{2,ij}$$

with mean μ_1, μ_2 and covariance matrix Σ_1, Σ_2

– Paired Samples: $n_1 = n_2$

For two paires samples $\{X_{1,ij}\}$, $X_{2,ij}$, take subtraction as

$$D_{ij} = X_{1,ij} - X_{2,ij}$$

denote $\bar{D} = \frac{1}{n}\sum_{j=1}^n D_j$, $S_D^2 = \frac{1}{n-1}\sum_{j=1}^n(D_j - \bar{D}'(D_j - \bar{D}))$
and conduct test to

$$H_0 : \bar{D} = \delta_0 \longleftrightarrow H_1 : \bar{D} \neq \delta_0$$

And the folloeing steps are as in One-population testing, test

$$T^2 = n(\bar{D}-\delta)'(S_D^2)^{-1}(\bar{D}-\delta) \sim \frac{(n-1)p}{n-p}F_{p,n-p}$$

– Under Equal Unknown Variance: $\Sigma_1 = \Sigma_2$

$$\bar{X}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} X_{1,j} \qquad \bar{X}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} X_{1,j} \tag{4.49}$$

$$S_1 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(X_{1,j} - \bar{X}_1)(X_{1,j} - \bar{X}_1)' \qquad S_2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(X_{2,j} - \bar{X}_2)(X_{2,j} - \bar{X}_2)' \tag{4.50}$$

And denote pooled variance

$$S_{\text{pooled}} = \frac{1}{n_1 + n_2 - 2}\left((n_1 - 1)S_1 + (n_2 - 1)S_2\right) \sim \frac{W_p(n_1 + n_2 - 2, \Sigma)}{n_1 + n_2 - 2} \tag{4.51}$$

Under H_0, we have

$$T^2 = \frac{1}{\frac{1}{n_1} + \frac{1}{n_2}}(\bar{X}_1 - \bar{X}_2 - \delta_0)' S_{\text{pooled}}^{-1}(\bar{X}_1 - \bar{X}_2 - \delta_0) \sim \frac{p(n_1 + n_2 - 2)}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1}$$

4.2.5 Confidence Region

Estimate the confidence region for μ of $X \sim N_p(\mu, \Sigma)$, Monovariate case see sec.2.3.3

- Confidence Region:

  Also use Hotelling's T^2

  $$\frac{n-p}{p}\frac{T^2}{n-1} \sim F_{p,n-p}$$

  And take $100(1-\alpha)\%$ confidence region of μ as

  $$R(x) = \{x | T^2 \le c^2\} \qquad c^2 = \frac{p}{n-p}(n-1)F_{p,n-p,\frac{\alpha}{2}}$$

  The shape of $R(x)$ is an ellipsoid.

- Individual Converage Interval

  Use the decomposition of S^2 as a positive finite matrix $S^2 = A^T A$, where A is some $p \times p$ matrix, then

  $$T^2 = [\sqrt{n}(\bar{X} - \mu_0)']S^{-1}[\sqrt{n}(\bar{X} - \mu_0)] = [A^{-1\prime}\sqrt{n}(\bar{X} - \mu_0)]'[A^{-1\prime}\sqrt{n}(\bar{X} - \mu_0)]$$

  Thus denote $Z = A^{-1\prime}(X - \mu_0) \sim N_p(0, A^{-1\prime}\Sigma A^{-1})$, the T^2 estimator of Z would be

  $$T_Z^2 = [\sqrt{n}\bar{Z}]' S_Z^{-1}[\sqrt{n}\bar{Z}] = n\bar{Z}'\bar{Z} = \frac{1}{n}\sum_{i=1}^{n}\bar{Z}_i^2 \sim F_{p,n-p}$$

  As a simplified case, we can take the **Individual Converage Interval** of Z_i, which is

  $$\frac{\sqrt{n}Z_i}{s_{Z_i}} \sim t_{n-1}$$

  And we can take the Confidence Region[33] as

  $$R(z) = \bigotimes_{i=1}^{n}(\bar{Z}_i \pm s_{Z_i} t_{n-1,\frac{\beta}{2}})$$

[33] The confidence region of Z can be transformed to that of X using $\hat{Z} = A^{-1\prime}(\hat{X} - \bar{X})$.

where β take

$$1 - p\beta = 1 - \alpha$$

Note: Consider that

$$P(\text{all } Z_i \text{ in } \text{CI}_i) \geq 1 - m\beta = 1 - \alpha$$

So the real CR for μ should be larger.

The shape of $R(x)$ is an oblique cubold.

4.2.6 Large Sample Multivariate Inference

Basic point:

$$\bar{X} \xrightarrow{\mathscr{L}} \mu \qquad S^2 \xrightarrow{\mathscr{L}} \Sigma$$

- One-sample Mean:

$$n(\bar{X} - \mu)S^{-1}(\bar{X} - \mu) \xrightarrow{\mathscr{L}} \chi_p^2$$

- Unequal Variance Two-sample Mean:

$$\bar{X}_1 - \bar{X}_2 \xrightarrow{\mathscr{L}} N(\mu_1 - \mu_2, \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2) \qquad \frac{1}{n_1}S_1 + \frac{1}{n_2}S_2 \xrightarrow{\mathscr{L}} \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2$$

Test:

$$T^2 = \left[(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)\right]' (\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2)^{-1} \left[(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)\right] \xrightarrow{\mathscr{L}} \chi_p^2$$

Section 4.3 Principal Component Analysis

PCA and next subsection FA focus on data dimension reduction. Why?

☐ ‘Curse of Dimesionality’

- Difficulty in computation complexity: Many algorithms has complexity $O(n^2)$ or more, high dimension n cause high complexity.
- Hughes Phenomenon: As the number of feature dimension increases, the classifier’s performance increases as well until an optimal dimension. Adding more features based on the same size as the training set will then degrade the classifier’s performance. [a]

[a] Example: Volumn of unit sphere in n-dim space

$$V_n = \pi^{n/2}\frac{1}{\Gamma(1 + n/2)} \to \left(\frac{2\pi e}{n}\right)^{n/2} \to 0$$

i.e. data will become ‘sparse’ in high dimension data -> difficult to extract information.

Key Idea of PCA: Find the components most powerful in explaining variance. (Similar to the idea of ANOVA)

4.3.1 Population Principal Component

For population $\vec{X} = (X_1, X_2, \dots, X_p) \sim (\mu, \Sigma)_p$, conduct spectrum decomposition to Σ such that

$$\Sigma P = P\Lambda \qquad P = \left[e_1, e_2, \dots, e_p\right] \qquad \Lambda = diag\{\lambda_1, \lambda_2, \dots, \lambda_p\},\ \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$$

where (λ_i, e_i) is the i^{th} eigenvalue-eigenvector pair of Σ, large λ_i suggests X is more 'extended' in e_i direction(large variance).

Then the **Principal Components** $Y = \{Y_i\}$

$$Y = P'X \sim (P'\mu, P'\Sigma P)_p = (P'\mu, \Lambda) \tag{4.52}$$

$$\begin{cases} Y_1 = e_1'X \sim (e_1'\mu, \lambda_1) \\ \vdots \\ Y_p = e_p'X \sim (e_p'\mu, \lambda_p) \end{cases} \tag{4.53}$$

Properties & Definitions:

- Trace of cov. matrix:
$$\sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} var(X_i) = \sum_{i=1}^{p} var(Y_i) = \sum_{i=1}^{p} \lambda_i$$
- $corr$ between Y_i, X_j:
$$\rho_{Y_i,X_j} = \frac{cov(Y_i, X_j)}{\sqrt{\lambda_i}\sqrt{\sigma_{jj}}} = \frac{(e_i)_j\sqrt{\lambda_i}}{\sqrt{\sigma_{jj}}}$$
- Factor Loading:
$$\text{FL}_{ij} = (e_i)_j\sqrt{\lambda_i}$$
- PC Score:
$$\text{PC Score}_i = Y_i = e_i'X \text{ or } Y_i = e_i'(X-\mu)$$

In practice, we pick the first several m PC such that

$$\sum_{i=1}^{m} \frac{\lambda_i}{\sum\limits_{k=1}^{p} \lambda_k} \text{ large enough}$$

Note: Another important point for PCA is the **interpretability** of principal components.

☐ Standardized Principal Component

To cancel out the influence due to scale, we can also obtain standardized PC from $Z = (V)^{-1/2}(X_\mu)$, where V is standard deviation matrix as def. in eqa(4.4).

And we have $\vec{Z} = (Z_1, Z_2, \dots, Z_p) \sim N_p(0, V^{-1/2}\Sigma V^{-1/2\prime}) = N_p(0, \rho)$. Then obtain (λ_i, e_i) pairs[34] from ρ to form PC.

$$\rho P = P\Lambda \qquad P = \left[e_1, e_2, \dots, e_p\right] \qquad \Lambda = diag\{\lambda_1, \lambda_2, \dots, \lambda_p\},\ \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$$

[34]The eigenvalue-eigenvector pairs obtained from ρ is generally **different** from Σ.

Then the Principal Components $W = \{W_i\}$

$$W = P'Z \sim (0, P'\rho P)_p = (0, \Lambda) \tag{4.54}$$

$$\begin{cases} W_1 = e_1'Z \sim (0, \lambda_1) \\ \vdots \\ W_p = e_p'Z \sim (0, \lambda_p) \end{cases} \tag{4.55}$$

Properties:

- Trace of cov. matrix:

$$\sum_{i=1}^{p} var(Z_i) = \sum_{i=1}^{p} var(W_i) = \sum_{i=1}^{p} \lambda_i = p$$

- $corr$ between Y_i, X_j:

$$\rho_{W_i,Z_j} = (e_i)_j\sqrt{\lambda_i}$$

4.3.2 Sample Principal Component

For sample matrix X denoted in eqa(4.32), with cov. matrix S in eqa(4.15). Then conduct the above spectrum decomposition to S to get sample PCs.

$$\hat{Y} = \hat{P}\hat{\Lambda}\hat{P}' \qquad \hat{P} = \left[\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_p\right] \quad \hat{\Lambda} = diag\{\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p\},\ \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p$$

Properties and Definitions

- Trace of cov. matrix:

$$\sum_{i=1}^{p} s_{ii} = \sum_{i=1}^{p} \hat{\lambda}_i$$

- Sample corr & factor load:

$$\rho(\hat{y}_i, x_j) = \frac{(\hat{e}_i)_j\sqrt{\hat{\lambda}_j}}{\sqrt{s_{jj}}}$$

□ Large Sample & Normal PCA

Under normal assumption or large sample case, i.e.

$$X \sim N_p(\mu, \Sigma) \text{ or } X \xrightarrow{\mathscr{L}} N_p(\mu, \Sigma)$$

We can examine the (asymptotic) distribution of $(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p)$ and $(\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_p)$:

- $\hat{\lambda}$:

$$\sqrt{n}(\hat{\lambda} - \lambda) \sim N_p(0, 2\Lambda^2)$$

- $\hat{e}_i$:

$$\sqrt{n}(\hat{e}_i - e_i) \sim N_p(0, E_i), \quad E_i = \lambda_i \sum_{k\neq i} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} e_k e_k'$$

- Independence:

$$\hat{\lambda}_i \parallel \hat{e}_i$$

Section 4.4 Factor Analysis

Key idea of FA: For a model with p variable $X = (X_1, X_2, \ldots, X_p) \sim (\mu, \Sigma)_p$ (especially when p large and X_i interrelated), there would be some internal, latent **factors** F behind X determining the model structure.[35]

4.4.1 Orthogonal Factor Model

$$\underset{p\times 1}{X} - \mu = \underset{p\times m}{L}\underset{m\times 1}{F} + \underset{p\times 1}{\varepsilon},\ m < p$$

where L is the const **loading matrix** ; F is r.v. **factor**; and ε is r.v. error.

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} \quad L = \begin{bmatrix} \ell_{11} & \ell_{12} & \ldots & \ell_{1m} \\ \ell_{21} & \ell_{22} & \ldots & \ell_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{p1} & \ell_{p2} & \ldots & \ell_{pm} \end{bmatrix} \quad F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

Note: Intuitively, we cannot estimate $(m+p)$ （unobserable) r.v. from p r.v., so we need the following assumptions on F and ε

$$\begin{aligned} &E(F) = 0 && cov(F) = I_n \\ &E(\varepsilon) = 0 && cov(\varepsilon) = \Psi = diag\{\psi_1, \psi_2, \ldots, \psi_p\} \\ &\varepsilon \,\|F \Leftrightarrow cov(F, \varepsilon) = 0 \end{aligned} \tag{4.56}$$

Derived Conclusions:

- Representation of Σ:

$$cov(X) = \Sigma = LL' + \Psi$$

  - Diagonal Elements:

$$var(X_i) = \sum_{k=1}^{m} \ell_{ik}^2 + \psi_i = h_i^2 + \psi_i$$

    where h_i^2 is Communality, ψ_i is Specific variance.

  - NonDiagonal Elements:

$$cov(X_i, X_j) = \sum_{k=1}^{m} \ell_{im}\ell_{jm}$$

- relation bet. X and F:

$$cov(X, F) = L$$

☐ Factor Rotation For any orthonormal rotation/reflection matrix $\underset{m\times m}{T}$, $\tilde{L} = LT$ satisfies the same factor model (with a different $\tilde{F}$):

$$\begin{aligned} X =& LF + \varepsilon = LTT'F + \varepsilon = \tilde{L}\tilde{F} + \varepsilon \qquad \tilde{L} = LT,\ \tilde{F} = T'F \\ \Sigma =& LL' + \Psi = \tilde{L}\tilde{L}' + \Psi \end{aligned}$$

Comment: Factor rotation reflects the arbitrariness of selection of L, allowing us to choose an **interpretable** L for FA model.

[35] As the most simplified case, here only consider X linear dependent on F.

4.4.2 Principal Component Approach

Origin: when $m = p$, factor decomposition reduces to spectrum(PC) decomposition.(At the same time Ψ can be taken 0.)

$$\begin{aligned} X =& LF + \varepsilon = PY \qquad \Rightarrow \Psi = 0 \\ \Sigma =& LL' + \Psi = P\Lambda P' \quad \Rightarrow L = P\Lambda^{1/2} \end{aligned} \tag{4.57}$$

Then take the first m eigenvectors to form L, and use $\psi_i = \sigma_{ii} - \sum\limits_{k=1}^{m} \ell_{ik}^2$ as an approximation.

$$\Sigma = LL' + \Psi \qquad L = \left[\sqrt{\lambda_1}e_1, \sqrt{\lambda_2}e_2, \ldots, \sqrt{\lambda_m}e_m\right] \qquad \Psi = diag\{\psi_i\}$$

□ Sample Factor Decomposition

From sample cov. matrix S and eigenvalue-eigenvector pairs $(\hat{\lambda}_i, e_i)$, pick the first m paris to form $L = \{\ell_{ij}\}$:

$$\hat{L} = \{\hat{\ell}_{ij}\} = \left[\sqrt{\hat{\lambda}_1}\hat{e}_1, \sqrt{\hat{\lambda}_2}\hat{e}_2, \ldots, \sqrt{\hat{\lambda}_m}\hat{e}_m\right] \qquad \hat{\Psi} = diag\{s_{ii} - \sum_{k=1}^{m} \hat{\ell}_{ik}^2\}$$

- Selection of m: Construct Residual Matrix

$$\hat{E} = S - (\hat{L}\hat{L}' + \hat{\Psi})$$

Residual matrix is trace 0, pick m such that

$$\text{Sum of All Elements in } \hat{E} < \sum_{k=m+1}^{p} \hat{\lambda}_k^2 \text{ small enough}$$

4.4.3 MLE Method

Assumption: Factor F and error ε are normal.(Then also $X \sim N_p(\mu, \Sigma)$ is normal)

$$F \sim N_m(0, I_m) \quad \varepsilon \sim N_p(0, \Psi) \quad X \sim N_p(\mu, \Sigma)$$

Likelihood Function:

$$L(\mu, \Sigma) = (2\pi)^{-np/2}|\Sigma|^{-n/2} \exp\left(-\frac{1}{2}tr\left[\Sigma^{-1}\left(\sum_{k=1}^{n}(x_k - \bar{x})(x_k - \bar{x})' + n(\bar{x} - \mu)(\bar{x} - \mu)'\right)\right]\right)$$

Maximize L to get $\hat{L}$ and $\hat{\Psi}$, usually for convenient (and to counteract the arbitrariness of factor rotation) we further assume

$$L'\Psi^{-1}L = \varXi \text{ (diagonal matrix)}$$

- Estimtor of communality variance h_i^2:

$$\hat{h}_i^2 = \sum_{k=1}^{m} \hat{l}_{ik}^2$$

Section 4.5 Canonical Correction Analysis

Key idea of CCA: For a model with two multivariate population $X^{(1)} = (X_1^{(1)}, X_2^{(1)}, \ldots, X_p^{(1)})$, $X^{(2)} = (X_1^{(2)}, X_2^{(2)}, \ldots, X_q^{(2)})$ with covariance

$$\underset{(p+q)\times(p+q)}{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

find a few condensed variable to measure their similarity.

4.5.1 Canonical Variate Pair

By using the linear combination, we can construct a pair of vector $\underset{p\times 1}{a}$ and $\underset{q\times 1}{b}$ such that $corr(a'X^{(1)}, b'X^{(2)})$ large, i.e.

$$\{a, b\} = \arg\max_{a,b\neq 0} \frac{a'\Sigma_{12}b}{\sqrt{a'\Sigma_{11}a}\sqrt{b'\Sigma_{22}b}} \tag{4.58}$$

where $U_1 = a'X^{(1)}$, $V_1 = b'X^{(2)}$ with $var(U_1) = var(V_1) = 1$ are the **(first) canonical variate pair**, and $\rho_1^* = corr(U_1, V_1)$ is the **(first) canonical correlation**.

Similarly, the k^{th} canonical pair (U_k, V_k) satisfy the same criterion as eqa(4.58) but with $a_k \in \text{span}\{a_1, \ldots, a_{k-1}\}^{\perp}$, $b_k \in \text{span}\{b_1, \ldots, b_{k-1}\}^{\perp}$, $k \leq \min\{p, q\}$.

Result: U_k, V_k can be expressed as

$$U_k = a_k'X^{(1)} = e_k'\Sigma_{11}^{-1/2}X^{(1)} \qquad V_k = b_k'X^{(2)} = f_k'\Sigma_{22}^{-1/2}X^{(2)}$$

where e_k is the k^{th} eigen vector of $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$, f_k is the k^{th} eigenvector of $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$. e_k and f_k satisfies:

$$f_k = \frac{1}{\rho_k^*}\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}e_k \quad e_k = \frac{1}{\rho_k^*}\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}f_k$$

4.5.2 Canonical Correlation based on Standardized Variables

Using standardized variable of X:

$$Z_k^{(\nu)} = \frac{X_k^{(\nu)} - \mu_k^{(\nu)}}{\sqrt{\sigma_{kk}^{(\nu)}}}, \ k = 1, 2, \ldots p \text{ or } q, \ \nu = 1, 2$$

with covariance

$$\underset{(p+q)\times(p+q)}{\rho} = V^{-1/2}\Sigma V^{-1/2} = \begin{bmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{bmatrix}$$

And similarly, the CCA pair is

$$U_k = a_k'Z^{(1)} = e_k'\rho_{11}^{-1/2}Z^{(1)} \qquad V_k = b_k'Z^{(2)} = f_k'\rho_{22}^{-1/2}Z^{(2)}$$

with e_k is the k^{th} eigenvector of $\rho_1 1^{-1/2}\rho_{12}\rho_{22}^{-1}\rho_{21}\rho_{11}^{-1/2}$, $f_[k]$ is the k^{th} eigenvector of $\rho_{22}^{-1/2}\rho_{21}\rho_{11}^{-1}\rho_{12}\rho_{22}^{-1/2}$, and

$$f_k = \frac{1}{\rho_k^*}\rho_{22}^{-1/2}\rho_{21}\rho_{11}^{-1/2}e_k \quad e_k = \frac{1}{\rho_k^*}\rho_{11}^{-1/2}\rho_{12}\rho_{22}^{-1/2}f_k$$

4.5.3 Sample Canonical Correlation

Replacement:

$$\Sigma \longrightarrow S \qquad \rho \longrightarrow R$$

to get

$$\hat{U} = \hat{A}x^{(1)} \qquad \hat{V} = \hat{B}x^{(2)}$$

and we can use $\hat{U}, \hat{V},\ \hat{A}, \hat{B}$ to express S_{12} as

$$S_{12} = \hat{A}^{-1} \begin{bmatrix} \hat{\rho}_1^* & 0 & \dots & 0 \\ 0 & \hat{\rho}_2^* & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\rho}_p^* \end{bmatrix} (\hat{B}^{-1})'$$

When applying CCA, we pick the first r canonical variable, thus some infomation is lost. But we hope the first r canonical variables can contain enough information of $X^{(1)}$ and $X^{(2)}$.

Determine of r: consider the error if approximation by expressing

$$\hat{A}^{-1} = [\alpha_1, \alpha_2, \dots, \alpha_p] \qquad \hat{B}^{-1} = [\beta_1, \beta_2, \dots, \beta_p]$$

and

$$\begin{aligned} S_{12} &= \sum_{i=1}^{p} \hat{\rho}_i^* \alpha_i \beta_i' \\ S_{11} &= \hat{A}^{-1}(\hat{A}^{-1})' = \sum_{i=1}^{p} \alpha_i \alpha_i' \\ S_{22} &= \hat{B}^{-1}(\hat{B}^{-1})' = \sum_{i=1}^{p} \beta_i \beta_i' \end{aligned}$$

Total sample variance explained by the first r canonical variables:

$$\frac{\sum_{i=1}^{r} \alpha_i' \alpha_i}{tr(S_{11})} \qquad \frac{\sum_{i=1}^{r} \beta_i' \beta_i}{tr(S_{22})}$$

Section 4.6 Discriminant Analysis

Key idea of DA: for $\underset{n\times p}{X}$ with an extra column labeling the classification, we want to determine a rule to assign new objects. More specifically, determine the classification region R_i for each class π_i.

4.6.1 Classification Criterion

- Two-category classification case:Each row of X is labeled in π_1 or π_2, for two-category, only one of R_1, R_2 is needed.

  Some basic concept in classification model:

  - Prior Possibility p_i, $i = 1, 2$;

- Penalty for misclassification $c(i|j)$, $i,j=1,2$: cost if a π_j object is classified in R_i.
- Conditional Probability $P(i|j)$, $i,j=1,2$: probability that a π_j object falls in region R_i

☐ Determination Criterion:

- Expected Cost of Misclassification (ECM) Criterion: Minimizing ECM,

$$\text{ECM} = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2$$

For two-category problem, R_1, R_2 can be determined as

$$\begin{aligned} R_1 &= \arg \frac{f_{\pi_1}(x)}{f_{\pi_2}(x)} \geq \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1} \\ R_2 &= \complement_{R_x}^{R_1} = \arg \frac{f_{\pi_1}(x)}{f_{\pi_2}(x)} < \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1} \end{aligned}$$

- Total Probability of Misclassification (TPM) Criterion: Minimizing TPM,

$$\text{TPM} = P(\text{misclass}) = P(2|1)p_1 + P(1|2)p_2$$

actually $\arg\min \text{TPM} = \underset{c(1|2)=c(2|1)}{\arg\min} \text{ECM}$

- Posterior Probability Criterion: Maximize posterior probability $P(\pi_i|x_0)$,

$$P(X\in\pi_i|X=x_0) = \frac{p_i f_{\pi_i(x_0)}}{p_1 f_{\pi_1}(x_0) + p_2 f_{\pi_2}(x_0)},\ i=1,2$$

Also equivalent to ECM for $c(1|2)=c(2|1)$

- Here only introduce ECM: $\{R_i\} = \arg\min \text{ECM}$

$$\begin{aligned} \text{ECM}(i) &= \sum_{j\neq i} c(j|i)P(j|i) \\ \text{ECM} &= \sum_{i=1}^{g} p_i \text{ECM}_i = \sum_{i=1}^{g}\sum_{j\neq i} c(j|i)p(j|i)p_i \end{aligned}$$

4.6.2 Linear & Quadratic Discriminant Analysis

Now take two-category ECM criterion as example. An estimation to $P(1|2)$, $P(2|1)$, i.e. to f_{π_1}, f_{π_2} is needed.

Assumption: for $\pi_1 : X\sim N(\mu_1,\Sigma_1)$,$\pi_2 : X\sim N(\mu_2,\Sigma_2)$, further for

- $\Sigma_1=\Sigma_2=\Sigma$: Linear Discriminant Analysis (LDA).

$$f_{\pi_i}(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)\right),\ i=1,2$$

then

$$\begin{aligned} R_1 &= \arg(\mu_1-\mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) \geq \ln\left(\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right) \\ R_2 &= \arg(\mu_1-\mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) < \ln\left(\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right) \end{aligned}$$

Note that L.H.S. is a linear combination of x, thus called LinearDA.

Sample estimation to Σ: use pooled variance in eqa(**??**).

- $\Sigma_1 \neq \Sigma_2$: Quadratic Discriminant Analysis (QDA).

$$f_{\pi_i}(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right),\ i=1,2$$

then

$$R_1 = -\frac{1}{2}x'(\Sigma_1^{-1}-\Sigma_2^{-1})x + (\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})x - \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2) \geq \ln\left(\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right)$$
$$R_2 = -\frac{1}{2}x'(\Sigma_1^{-1}-\Sigma_2^{-1})x + (\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})x - \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2) < \ln\left(\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right)$$

Note that L.H.S. is a quadric form of x, thus called QuadraticDA.

- Two extension: allow more flexible estimation to variance:
  - $\hat{\Sigma}_i(\alpha) = \alpha\hat{\Sigma}_i + (1-\alpha)\hat{\Sigma}$, shrink between QDA and LDA;
  - $\hat{\Sigma}_i(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\hat{\sigma}^2 I$, shrink toward scalar cov.

4.6.3 Fisher's Discriminant Analysis

Project X onto some hyperplane and conduct low-dimensional classification.

Project x onto some hyperplane by $y = a'x$, then we maximize $\psi = \frac{\text{mean of treatment}^2}{\text{variance}}$[36]. i.e.

$$\psi = \frac{\sum_{i=1}^{g}(\mu_{iY}-\mu_Y)^2}{\sigma_Y^2} = \frac{a'\left(\sum_{i=1}^{g}(\mu_i-\mu)(\mu_i-\mu)'\right)a}{a'\Sigma a}$$
$$= \frac{a'B_\mu a}{a'\Sigma a}$$

Result: a is the largest eigen vector of $W^{-1}B$.

Relation between FDA and LDA: in FDA, take the first ξ eigenvectors to conduct classification, thus loses more information. But when $\xi = g-1$, FDA $\equiv$ LDA.[37]

4.6.4 Evaluation of Discriminant Model

☐ Judging Index:

---

[36] MANOVA Model: For g groups with same Σ, consider an MANOVA model: $X_{ij} = \mu + \tau_i + e_{ij}$. Then MANOVA table gives Sum of Squares and cross Products (SSP):

$$\text{Treatment:} B = \sum_{i=1}^{g} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'$$
$$\text{Residual:} W = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)'$$
$$\text{Total:} T = B + W = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})(x_{ij}-\bar{x})'$$

use B and W to measure the variance of sample.

[37] Because a is eigenvector of $W^{-1}B$, while $\text{rk}(B) = g-1$, thus there are $g-1$ non-zero eigenvalues at most.

- Total Probability of Misclassification (TPM):

$$\mathrm{TPM} = p_1 P(2|1) + p_2 P(1|2) = p_1 \int_{R_2} f_{\pi_1}(x)\,\mathrm{d}x + p_2 \int_{R_1} f_{\pi_2}(x)\,\mathrm{d}x$$

- APparent Error Rate (APER): used with cross validation (CV). The fraction of misclassification in training set.

Section 4.7 Clustering Analysis

Key idea of CA: Group a collection of data according to similarity and relation of objects.

4.7.1 Agglomerative Clustering Algorithm

Hierarchical clustering: start with individual points and combine them to form groups. Agglomerative clustering algorithm is conducted by repeating following steps:

1. Start: All $k = n$ points are individual clusters;
2. Use a distance/dissimilarity matrix $\underset{k\times k}{D}$ to express distances between clusters; the 'distance' between clusters is diversified, choice of which see the following part;
3. merge closest pairs of clusters(or points) to form larger clusters, and now number of clusters $k <-k-1$;
4. Go back to step.2 and repeat, until there is only one cluster left.

□ Choice of between-cluster distance: To express distance between two clusters A and B,

Choice of distance functional $D(\cdot,\cdot)$:

- Euclidean Distance D_{E};
- Mahalanobis Distance D_{M};
- Jaccard Distance $D_{\mathrm{J}} = 1 - \dfrac{|A\cap B|}{|A\cup B}$;
- etc.

Location choice of cluster:

- Complete link: max $D(a\in A, b\in B)$;
- Single link: min $D(a\in A, b\in B)$;
- Centroid distance: $D(A\,\text{centroid}, B\,\text{centroid})$;
- Group average: $< D(a\in A, b\in B) >$

Note: pros-and-cons of agglomerative clustering algorithm

- No assumptions for final k needed;
- Intuitive display of relations;
- Large computational requirement: $\sim O(n^3)$;
- Sensitive to noise and outliers.

4.7.2 k-Means Clustering Algorithm

Assume we have a preset number of clusters k, we can use k-means clustering.

1. Choose number of clusters k;
2. Select k points as initial centroids, useful methods:
   - Randomly select;
   - Use Centroid of agglomerative algorithm;
   - Successively pick the farthest point from others.
3. For all points i, calculate its distance from the l^{th} centroid $D(i,l)$, and classify each i point to the nearest centroid cluster;
4. Re-calculate the centroid of new k clusters;
5. Repeat step 4,5 until convergence.(Convergence criterion can be e.g. $< \sum_i D(i \in g_l, l) > \to$ const)

Note: prosandcons of k-Means clustering algorithm

- Efficient:$\sim O(n)$;
- Sensitive to outliers;
- Ineffective for non-convex shapes.

4.7.3

Assumption and notation: X is generated from a mixed distribution of k normal:

$$X \sim \sum_{l=1}^{k} \pi_l N(\mu_l, \Sigma_l) = \sum_{l=1}^{k} \pi_l N(\theta_l),\ \sum_{l=1}^{k} \pi_l = 1,\ \pi_l \geq 0.$$

Use its likelihood function $L(\theta; x)$ and maximize posterior probability by $\dfrac{\partial \ell}{\partial \theta}$:

$$L(\{\theta_l\}; x) = \prod_{i=1}^{N} \sum_{l=1}^{k} \pi_l \frac{1}{(2\pi)^{p/2} |\Sigma_l|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_l)' \Sigma_l (x-\mu_l) \right)$$

Extreme value with constraint $\sum_{l=1}^{k} \pi_l = 1$ requires

$$\frac{\partial \ln L}{\partial \mu_l} = 0 \quad \frac{\partial \ln L}{\partial \Sigma_l} = 0 \quad \frac{\partial L + \lambda(\sum\limits_{l=1}^{k} \pi_l - 1)}{\partial \pi_l} = 0$$

Result:

$$\text{denote:}\gamma_{il} := \frac{\pi_i \phi_{\theta_l}(x_i)}{\sum\limits_{j=1}^{k} \pi_j \phi_{\theta_j}(x_i)}$$

$$\begin{cases} \mu_l = & \frac{\sum\limits_{i=1}^{N} \gamma_{il} x_i}{\sum\limits_{i=1}^{N} \gamma_{il}} \\ \Sigma_l = & \frac{\sum\limits_{i=1}^{N} \gamma_{il}(x_i - \mu_l)(x_i - \mu_l)'}{\sum\limits_{i=1}^{N} \gamma_{il}} \\ \pi_l = & \frac{1}{N} \sum_{i=1}^{N} \gamma_{il} \end{cases}$$

where γ_{il} is the posterior probability that the i^{th} object belongs to the l^{th} group.

The above constraint equations are difficult to solve, use iteration algorithm:

1. Use e.g. k-means method to get an initial estimation to $\hat{\theta}_l = (\hat{\mu}_l, \hat{\Sigma}_l)$, $\hat{\pi}_l = 1/k$;
2. E$_{\text{xpectation}}$-Step: Compute weight(responsibility) of each point;

$$\hat{\gamma}_i(x_i \in G_l) = \frac{\hat{\pi}_l \phi_{\hat{\theta}_l}}{\sum\limits_{j=1}^{l} \hat{\pi}_j \phi_{\hat{\theta}_j}}, \; 1 \le i \le N$$

3. M$_{\text{aximize}}$-Step: Re-calculate parameters $\{\mu_l, \Sigma_l, \pi_l\}$;
4. Repeat step 2,3 until convergence.

Note: EM method for Gaussion Mixture Model is a greedy algorithm -> local maximum.

4.7.4 DBSCAN & OPTICS Clustering Algorithm

DBSCAN algorithm (Density-Based Spatial Clustering of Application with Noise) is a kind of density clustering algorithm. OPTICS algorithm (Ordering Point To Indentify the Cluster Structure) is its improved version.

□ DBSCAN Algorithm

Key (preset) index in DBSCAN:

- Eps ε: Radius of neighbourhood of a point;
- MinPts M: Minimum number of points to be indentified as cluster core point, usually choose $M \ge \text{dim} + 1$;
- (Also, a distance norm is needed, e.g. Euclidean D).

Notation:

- Neighbourhood of point x_i:

$$\mathcal{N}_\varepsilon(x_i) \equiv \{y \in \mathbb{R}^n : 0 < D(y, x) < \varepsilon\}$$

- 'Density' (is actually an integer):

$$\rho_\varepsilon(x_i) \equiv \# x_j \in \mathcal{N}_\varepsilon(x_i)$$

- Three types of Points: X_c, X_{bd}, X_{noi}.
  - Core Point: label an x_i as core point if
$$\rho_\varepsilon(x_i) \geq M$$
    Denote the set of core point as X_c, and set of non-core point as X_{nc}
  - Border Point: label an $x_j \in X_{nc}$ as border point if
$$\exists (x_i \in X_c) \in \mathcal{N}_\varepsilon(x_j) \& x_j \in X_{nc}$$
    Denote the set of border point as X_{bd}
  - Noise Point: the set of noise point is
$$X_{noi} \equiv \complement_X^{X_c \cup X_{bd}}$$
- Point Relations: DDR, DR, DC
  - Directly Density Reachable: For $x_i, x_j \in X$, if $x_i \in X_c$, $x_j \in \mathcal{N}_\varepsilon(x_i)$, then say x_j is DDR from x_i;
  - Density Reachable: For point chain $x_{i_1}, x_{i_2}, \ldots, x_{i_m}$, $m \geq 2$. If $x_{i_{\kappa+1}}$ is DDR from x_{i_κ}, $\forall 1 \leq \kappa \leq m-1$, then say x_{i_m} is DR from x_{i_1}.
  - Density Connected: For point $x_{i_1}, x_{i_2}, x_{i_3}$, if x_{i_2} and x_{i_3} are both DR from x_{i_1}, then say x_{i_2} and x_{i_3} are DC.

  Note: DR is not symmetric for x_{i_1} and x_{i_m}; while DC is.

DBSCAN algorithm classify all points that are Density Connected to each other into a cluster $C \subset X$, i.e.

$$\text{Maximality:} x \in C \&\& y \text{ DR from } x \Rightarrow y \in C$$
$$\text{Connectivity:} x, y \in C \Rightarrow x, y \text{ DC.}$$

Pros and cons of DBSCAN:

- Insensitive to noise;
- Based on density, with no constraint on the shape of cluster;
- Suitable for clusters with uniformly densed data, otherwise difficult to choose proper Eps ε;
- Complexity $\sim O(n^2)$, at least $O(n \log n)$.

□ OPTICS Algorithm

OPTICS is based on DBSCAN and shares most of the basic concepts and ideas. Further define the following distance (preset ε and M):

- Core Distance: For $x_i \in X_c$, the smallest distance allowing x_i to become core point.
$$\text{CD}(x_i) = D(x_i, N_\varepsilon^M(x_i)),\ \rho_\varepsilon(x_i) \geq M$$
  where $N_\varepsilon^M(x_i)$ is the M^{th} closest point from x_i;
- Reachablity Distance: For $y \in X$, $x_i \in X_c \subset X$,
$$\text{RD}(y, x_i) = \max\{CD(x_i, D(y, x_i))\}$$

Or equivlantly

$$\mathrm{RD}(y, x_i) = \underset{\rho_d(x_i) \geq M, y \in \mathcal{N}_d(x_i)}{\arg\min} \quad d$$

Algorithm flow:

1. Construct X_c based on preset M, ε;
2. Pick an 'unprocessed' point $x_{n_i} \in X_c$ and calculate $\mathrm{RD}(x_j, x_{n_i})$, $\forall$'unprocessed'$x_j \in \mathcal{N}_\varepsilon(x_{n_i}) \cap X_c$. Pick the $x_j \in X_c$ with smallest RD and label as $x_{n_{i+1}}$ processed;
3. Repeat step 2 until all points are processed. Output $\{x_{n_i}\} = (x_{n_1}, x_{n_2}, \ldots, x_{n_{|X_c|}})$. Each x_{n_i} is attached with a $\mathrm{CD}(x_{n_i})$ and a $r(x_{n_i}) := \mathrm{RD}(x_{n_{i-1}}, x_{n_i})$[38].

Then break the ordering sequence n_i according to $r(x_{n_i})$, .e.g. break n_i if $r(x_{n_i}) \geq \tilde{\varepsilon}$

Comment: OPTIC is more stable than DBSCAN, capable of dealing with multi-density clustering.

[38] For $i = 1$, just define as 0

Chapter. V 数据科学导论部分

Instructor: Sheng Yu

☐ Road to Data Scientist

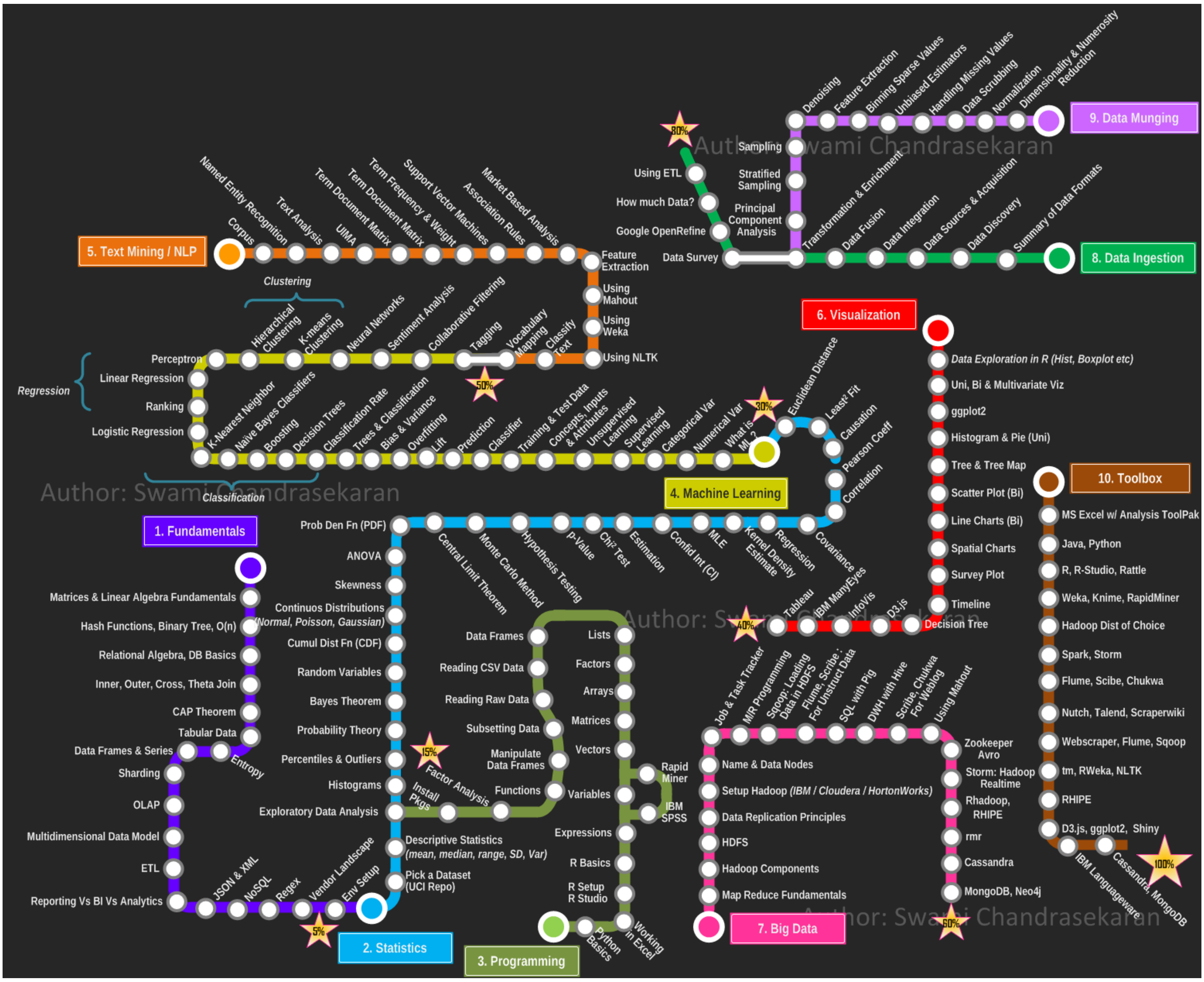


图 2: Road to Data Scientist

Comparison of `R`, `python`:focus on different aspects of 'Statistics':

- Differnece in programming philosophy: `R` for data analysis and `python` for data processing
- Difference in operating domain: `R` for statistical programming while `python` for general programming.

Section 5.1 Basic R. Manipulation

5.1.1 Installation and Maintenance of R.

Installing and updating `R.`: update by delete old version and install new version.

- In CRAN (The Comprehensive R Archive Network): https://cran.r-project.org
- In Mirror@TUNA: https://mirrors.tuna.tsinghua.edu.cn/CRAN

Installing and updating RStudio: https://www.rstudio.com

Running `R.` command:

- In `R.` GUI;
- In `R.` command line terminal;
- `R. CMD BATCH`;
- `Rscript`;
  - Use `>` to redirect output(overwrite);
  - Use `>>` to append output.

`R.` package library: packages are collection of `R.` functions (as well as test data and sample code).

- `.libPaths()` show package library location[39] ;
- `library('PACKAGE_NAME1','PACKAGE_NAME2',...)` load packages.
- `install.packages('PACKAGE_NAME1','PACKAGE_NAME2',...)` install package from CRAN/mirrors;
- `installed.packages()` show all installed packages;
- `updata.packages(checkBuilt = TRUE, ask = FALSE)` update installed packages;

Working Directory manipulation:

- `getwd()` get current working directory;
- `setwd('TARGET_PATH')` set working directory as an existing path.

Recommended `R.` project organization: working directory organized like

- `data/` folder for structured original dataset;
- `result/` folder for output result;
- `presentation/` folder for result representing slides/reports/etc.;
- `.r` project file $\times n$.

5.1.2 Data Manipulation in R.

☐ Looking for help/example of function:

- `?FUN_NAME()`;
- `help('FUN_NAME')`;

☐ Atomic Classes

- `'abc'` Character;

[39] Unlike in `C` or `python` where `.` is an operator, `.` in `R.` is just a common character, without special meaning. This feature can be used in naming self-defined functions: use `.FUN_NAME1` for within-project function while `FUN_NAME2` for external interface.

- `3L` Integer;
- `2.4` Numeric;
- `TRUE`,`FALSE`,`T`,`F` Logical;
- Special types: `NA`, `NaN`, `NULL`, `Inf`

☐ Operators

- Numerical Operators: +,−,∗(multiply by column),/,%∗%(matrix multiply),^;
- Logical Operators: ==,etc.; & and | for common operator, && and || for comparing the first element;
- Round a numeric:
  - `as.integer()`, round towards 0
  - `trunc()`
  - `ceiling()`
  - `floor()`
  - `round(NUMBER_TO_ROUND,digits = DIGITS)`

☐ Data Structure

- Vector: Column vector is the **basic** data structure in `R.` (scalar is length=1 vector).

  Only data of the same class can be held in one vector.

  Initialization:

  - Ordinary way:
    - `c(1,2,3)`, `c(T,FALSE,TRUE)`, `c('a',NA,'b')`
    - `vector(mode = MODE,length = LENGTH)`

    where `c()` for 'combine';

    `c()` combines all things into one vector, e.g. `c(c(1,2,3),c(1,2))=(1,2,3,4,5)`.
  - Sequence vector:
    - `1:3.5=c(1,2,3)`, `3:1=c(3,2,1)`
    - `seq(from, to ,by, length.out)`, `length.out` for total vector length;
    - `rep(SEQ_TO_REP, times, lenght.out ,each)`, used in k-fold cross validation labelling.

  Operations:

  - between vectors of different length `SHORT` and `LONG`: First `SHORT <-rep(SHORT, length.out=length(LONG))`. Then operate `SHORT` and `LONG`.
  - Element access: `a[i]`

> **Vectorized Operation**: All operation in `R.` are based on vector, and vectorized operation is **Parallel Arithmetic**, which is **much faster** than loop such as `for` ⟶ Consider using vectorized opertion when writing code for **Speed**!

- Factor: A special kind of 'vector' in `R.`, used to label discrete categorical data.[40]

  Initialization:

  `factor(FACTOR_SEQ, levels = FACTOR_LEVEL, labels = ...)`, `FACTOR_LEVEL` is the 'rank' of each factor, `labels` is the 'tag' of levels.

- Matrix: Only data of the same class can be held in one matrix.

  Initilaization:

  `matrix(DATA_SEQ, nrow, ncol, byrow = FALSE)`

  If `length(DATA_SEQ) < nrow*ncol`, then `DATA_SEQ` is repeated. Default: fill by column (because matrix is stored by column).

  Operation:

  – Common operators `+-*/^` etc. operate in column-by-column mode (vectorized operation).
  – Binding matrix: `cbind` for `[A,B]` and `rbind` for `[A;B]`
  – Transpose: `t()`
  – Matrix multiplication: `%*%`
  – Inverse matrix: `solve()` (The essence of inversion is solving linear equations)
  – Diagonal matrix:
    * `diag(VECTOR)` returns a matrix diag{`VECTOR`}
    * `diag(MATRIX)` returns the diagonal element vector
  – Element access: `a[i,j]`, `a$OBJECT_NAME`
  – Dimension of matrix: `dim()`, `nrow()`, `ncol()`

- List: A pack containing various datatype

  Initialization: `list(OBJECT1,OBJECT2,...)`

  Element access: `a[[i]]`

- `data.frame`: 'Mixture' of matrix and list. `data.frame` is actually a kind of list(with some constraint), organized in the shape of matrix (but allowing different datatype for different columns, each column is a list object).

  Each column of `data.frame` has name: `names(DATA_FRAME)`, `colnames(DATA_FRAME)`

- Element access: `a[i,j]`, `a[[i]]`, `a$COL_NAME`

[40]Factor vector is stored as integer vector.

参考文献

[1] 概率导论 (第二版・修订版). Dimitri P. Bertsekas, John N. Tsitsiklis. 人民邮电出版社.

[2] 数理统计（第二版）. 韦来生. 科学出版社.

[3] Statistical Inference(2nd Edition). George Casella, Roger L. Berger. Duxbury Press.

[4] Applied Linear Statistical Models(5th Edition). Michael H. Kutner, Christopher J. Nachtsheim, John Neter, William Li. McGraw-Hill Compaines, Inc.

[5] 线性模型引论. 王松桂 et. al. 科学出版社.

[6] Linear Models with R(2nd Edition). Julian J. Faraway. CRC Press.

[7] 实用多元统计分析（第六版）. Richard A. Johnson, Dean W. Wichern. 清华大学出版社.

[8] R In Action: Data Analysis and Graphics with R(2nd Edition). Robert I. Kabacoff. Manning Publications Co.

[9] R Programming For Data Science. Roger D. Peng. Lean Publishing.

索引