STATISTICAL MODELLING

Statistical model is collection of probability distributions $\{P_\theta : \theta \in \Theta\}$ on a given sample space.

↳ statistical model called IDENTIFIABLE if $\theta \mapsto P_\theta$ one-to-one, i.e. $P_{\theta_1} = P_{\theta_2} \Rightarrow \theta_1 = \theta_2$, so distinct parameter values give rise to distinct distributions. (no 2 different θ lead to same observed distribution)

BIAS: given estimator T, $\text{bias}_\theta(T) = E_\theta(T) - \theta$

↳ if T estimator for $g(\theta)$, then $\text{bias}(T) = E(T) - g(\theta)$.

STANDARD ERROR: $SE_\theta(T) = \sqrt{\text{Var}_\theta(T)} = \sqrt{E_\theta(T^2) - E_\theta(T)^2}$

MEAN SQUARE ERROR: $MSE_\theta(T) = E_\theta\left[(T-\theta)^2\right]$.

$$\boxed{\Rightarrow MSE_\theta(T) = \text{Var}_\theta(T) + \text{bias}_\theta(T)^2.}$$

CRAMER-RAO LOWER BOUND: suppose $T = T(X)$ unbiased:

$$\text{Var}_\theta(T) \geq \frac{1}{I(\theta)}, \quad \text{where } I(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right].$$

N/B: $f_\theta(X)$ joint p.d.f. of $X = (X_1, \dots, X_n)$.

For $X_1, \dots, X_n$ and $f_\theta^{(1)}$ pdf of single observation,

$$\Rightarrow I_f(\theta) = n \cdot I_{f_\theta^{(1)}}(\theta).$$

$\Rightarrow I \propto$ sample size for random sample.

JENSEN'S: for convex function g and r.v. X,

$$g(E(X)) \leq E(g(X)).$$

- $X_n \xrightarrow{a.s.} X$ if $P(\lim_{n\to\infty} X_n = X) = 1$.
- $X_n \xrightarrow{P} X$ if $\forall \varepsilon > 0$, $P(|X_n - X| > \varepsilon) = 0$.
- $X_n \xrightarrow{D} X$ if $\lim_{n\to\infty} P(X_n \leq x) = F_X(x) = P(X \leq x)$ where X has c.d.f. F_X.

(a.s. $\Rightarrow$ P $\Rightarrow$ D.)

<u>CONSISTENT</u> : sequence of estimators T_n consistent if

$\forall \theta \in \Theta$, $\underline{T_n \xrightarrow{P_\theta} g(\theta)}$. → T_n converging in probability.

<u>PORTAMANTEAU LEMMA</u> : $\boxed{X_n \xrightarrow{D} X \iff E(f(X_n)) \longrightarrow E(f(X))}$.

for all <u>bounded</u> + <u>continuous</u> $f : \mathbb{R} \to \mathbb{R}$.

<u>ASYMPTOTICALLY UNBIASED</u> : T_n for $g(\theta)$ if $\underline{E(T_n) \to g(\theta)}$.

$$\boxed{\text{If } T_n \text{ asymptotically unbiased and } Var_\theta(T_n) \to 0, \Rightarrow T_n \text{ consistent for } g(\theta).}$$

proof by Markov's inequality: $P(|X| \geq a) \leq \frac{E(|X|)}{a}$.

<u>ASYMPTOTICALLY NORMAL</u> : T_n for θ if $\underline{\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2(\theta))}$

<u>CENTRAL LIMIT THM</u> : $Y_1, \dots, Y_n$ i.i.d., $E(Y_i) = \mu$, $Var(Y_i) = \sigma^2$

$$\Rightarrow \sqrt{n}(\bar{Y} - \mu) \xrightarrow{D} N(0, \sigma^2).$$

↳ i.e. sample averages are asymptotically normal.

<u>SLUTSKY'S</u> : if $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} c$ for constant c,

- $X_n + Y_n \xrightarrow{D} X + c$
- $Y_n X_n \xrightarrow{D} cX$, • $Y_n^{-1} X_n \xrightarrow{D} c^{-1} X$.

<u>DELTA METHOD</u> : T_n asymptotically normal, so $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2(\theta))$,

$$\Rightarrow \sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{D} \underline{N(0, g'(\theta)^2 \cdot \sigma^2(\theta))} \qquad (g' \neq 0)$$

(note: odds of event A happening defined as $P(A)/(1-P(A))$.)

<u>CONTINUOUS MAPPING THM</u> : convergence in d, P, a.s. all preserved under continuous mappings.

<u>MLE</u> : parameter value $\theta \in \Theta$ for which observed data is <u>most likely</u>.

$L(\underline{\theta}) = \prod_{i=1}^{n} f(y_i; \underline{\theta})$ → product of pdfs → likelihood fun^n.

HYPOTHESIS TEST : H_0 : for which values of sample $X_1, \dots, X_n$ to accept H_0, or reject otherwise + accept H_1.

↳ subset of sample space to reject H_0 is called CRITICAL REGION.

TYPE I ERROR : false +ve → reject H_0 when H_0 true.

TYPE II ERROR : false −ve → accept / don't reject H_0 when H_0 false.

α-level test where $P_\theta(\text{reject } H_0) \leq \alpha$. (small α.)

POWER FUNCTION : $\beta(\theta) = P_\theta(\text{reject } H_0)$

↳ if $\theta \in \Theta_0$ then want $\beta(\theta)$ be small.

↳ if $\theta \in \Theta_1$ then want $\beta(\theta)$ be large.

p-value : reject H_0 iff $p \leq \alpha$ (p-value repeated from observed values)

construct test from confidence region : for region $A(Y)$ of $1-\alpha$ and test $H_0 : \theta \in \Theta_0$, $H_1 : \theta \notin \Theta_0$, reject H_0 if :

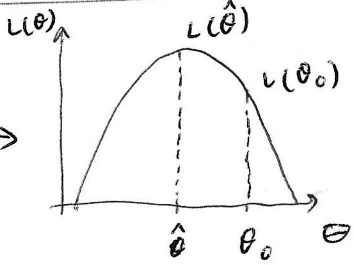
$\Theta_0 \cap A(y) = \phi$ i.e. if none of elements of H_0 are in confidence region.

LIKELIHOOD RATIO TEST : $t(\underline{y}) = \dfrac{\sup_{\theta \in \Theta} L(\theta; \underline{y})}{\sup_{\theta \in \Theta_0} L(\theta; \underline{y})} = \dfrac{\text{max lik. under } H_0 + H_1}{\text{max lik. under } H_0}$

for observed data $\underline{y}$.

↳ reject H_0 if $t(\underline{y})$ large i.e. $t(\underline{y}) \geq k$.

compare $L(\hat{\theta})$ to $L(\theta_0)$ and if $L(\theta_0) << L(\hat{\theta})$, θ_0 likely one.



DEF^N of χ_2 : For $X_1, X_2, \dots, X_n \sim N(0,1)$ i.i.d. $\Rightarrow \sum_{i=1}^n X_i^2 \sim \chi_n^2$.

LRT DISTRIBUTION : $2 \log t(\underline{Y}) \xrightarrow{D} \chi_r^2$ (LRT statistic asymptotically χ^2)

where r = # of indep params in full model − # of indep params under H_0.

(pf: ① Taylor's expansion of $\log L(\theta)$. ② Slutsky's + continuous mapping thm + MLE asymptotically normal + WLLN.)

MAXIMUM LIKELIHOOD ESTIMATOR : of θ is $\hat{\theta}$ such that

$$L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta).$$

e.g. for Bern(θ) and Pois(θ), MLE is $\bar{X}$ and for exp(θ), MLE is $\frac{1}{\bar{X}}$.

MLE not necessarily unbiased.

MLE functionally invariant : bijective g then $\hat{\phi} = g(\hat{\theta})$ MLE of $\phi = g(\theta)$.

if $\hat{\theta}_n$ MLE of $X_1, \dots, X_n$ sequence, then $\hat{\theta}_n$ asymptotically normal

$\Rightarrow \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N\left(0, \frac{1}{I_f(\theta)}\right)$ for $I_f(\theta) = E\left[\left(\frac{\partial}{\partial \theta} \log f_\theta(X)\right)^2\right]$.

↳ for $\hat{\theta}_n = \frac{1}{n} \sum X_i \Rightarrow \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, \sigma^2)$ where $\sigma^2 = Var(X_i)$

↳ CLT.

CONFIDENCE REGION : random interval containing true parameter with probability of $1-\alpha$.

↳ $P(\theta \in I) \geq 1-\alpha$.

construct confidence interval by pivotal quantity.

PIVOTAL QUANTITY : $t(\underline{Y}, \theta)$ s.t. distribution of t completely known + doesn't depend on any unknown params.

- by asymptotic normality of MLE, $\sqrt{n} \frac{(T_n - \theta)}{\sigma(\theta)} \xrightarrow{D} N(0,1)$

$\Rightarrow \sqrt{n} \frac{(T_n - \theta)}{\sigma(\theta)} \sim N(0,1)$ approx. can be used as pivotal quantity.

BONFERRONI CORRECTION : for confidence intervals $[L_i, U_i]$ of θ_i $(i = 1, \dots, k)$ of $1 - \alpha/k$. $\Rightarrow (L_1, U_1) \times \dots \times (L_k, U_k)$ $1-\alpha$ CI for $(\theta_1, \dots, \theta_k)^T$.

(e.g. if $[L_1, U_1]$ is $0.99 = 1 - 0.01$ and $[L_2, U_2]$ is $0.97 = 1 - 0.03$ then $[L_1, U_1] \times [L_2, U_2]$ is $1 - 0.01 - 0.03 = 0.96$. $(1 - \alpha_1 - \alpha_2)$)

(simple)
LINEAR MODEL : $Y_i = \beta_0 + a_i \beta_1 + \varepsilon_i$, $i = 1 \dots n$

and goal how to "estimate" linear parameters β_1, β_2 (and also $Var(\varepsilon_i) = \sigma^2$) ?

STATISTICAL ERROR ε_i : amount by which observation differs from its expected value $\rightarrow$ $E(Y_i) = \beta_0 + a_1 \beta_1$.

$\boxed{\text{LINEAR ALGEBRA lemma : } X_{n \times p} \text{ matrix, } rank(X^T X) = rank(X).}$

for $\underline{X} = (X_1 \dots X_n)^T$, $E(\underline{X}) = (EX_1, \dots, EX_n)^T$.
for deterministic $\underline{A}, \underline{B}$: $E(\underline{A}\,\underline{X}) = \underline{A}\, E(\underline{X})$ and $E(\underline{X}^T \underline{B}) = E(\underline{X})^T \underline{B}$.

$cov(\underline{X}, \underline{Y}) = (cov(X_i, Y_j))_{i,j} = E(\underline{X}\,\underline{Y}^T) - E(\underline{X}) \cdot E(\underline{Y})^T$ $(cov(\underline{X}) = cov(\underline{X}, \underline{X}))$ $(cov(\underline{X}) = Var(\underline{X}))$

- $cov(X, Y) = cov(Y, X)^T$
- $cov(aX + bY, Z) = a\, cov(X, Z) + b\, cov(Y, Z)$
- $cov(AX, BY) = A\, cov(X, Y)\, B^T$
- $cov(\underline{X})$ +ve definite + symmetric.

if $cov(X, Y) = 0$, X, Y same distribution $\Rightarrow$ X, Y independent.
X, Y independent $\Rightarrow$ $cov(X, Y) = 0$.

(general)
LINEAR MODEL : $\underline{Y} = \underline{X}\, \underline{\beta} + \underline{\varepsilon}$ $\quad \underline{X} \in \mathbb{R}^{n \times p}$, $\underline{\beta} \in \mathbb{R}^p$, $\underline{Y} \in \mathbb{R}^n$.

$E(\underline{\varepsilon}) = \underline{0}$ $\Rightarrow$ $E(\underline{Y}) = \underline{X}\, \underline{\beta}$. ($\underline{X}$ design matrix)

SECOND ORDER ASSUMPTION : $cov(\underline{\varepsilon}) = \sigma^2 I_n$ $\rightarrow$ errors of 2 different observat^{n}s independent + variance of all errors identical.

NORMAL THEORY ASSUMPTION : $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I_n)$. $\rightarrow$ used to construct tests.

FULL RANK : X has full rank of p (since assuming $n > p$).

LEAST SQUARES : $S(\underline{\beta}) = (\underline{Y} - \underline{X}\,\underline{\beta})^T (\underline{Y} - \underline{X}\,\underline{\beta})$ $\leftarrow$ want to minimise w.r.t. β.

$$\boxed{\Rightarrow \text{LS eq}^n : X^T X \hat{\beta} = X^T \underline{Y}}$$

Solution exists iff $(X^T X)^{-1}$ exists iff $rank(X^T X) = p = rank(X)$.
$\hookrightarrow$ $cov(\hat{\underline{\beta}}) = \sigma^2 (X^T X)^{-1}$.

GAUSS-MARKOV THM: under FR, SOA, $\forall \underline{c} \in \mathbb{R}^p$:

estimator $\underline{c}^T\hat{\underline{\beta}}$ has smallest variance out of all unbiased estimators $\underline{c}^T\underline{\beta}$

P projection matrix iff $P^T = P$ and $P^2 = P$.

- P projection onto L given by: $P = X(X^TX)^{-1}X^T$

with $X = (\underline{x}_1 \dots \underline{x}_r)$ for $\underline{x}_1, \dots, \underline{x}_r$ basis of L.

↳ $I - P$ projects to $L^\perp$.

LEMMA: for $n \times n$ proj. matrix $\underline{P}$ w/ rank r

↳ $\underline{P}$ has r evalues of 1 and $n-r$ evalues of 0.

↳ rank $\underline{P}$ = trace $\underline{P}$.

$$\hat{\underline{Y}} = X\overbrace{(X^TX)^{-1}X^T\underline{Y}}^{\hat{\underline{\beta}}} = P\underline{Y} \quad ; \quad \underline{e} = \underline{Y} - \hat{\underline{Y}} = \text{vector of residuals.}$$

RESIDUAL SUM of SQUARES: $RSS = \underline{e}^T\underline{e} = \sum_{i=1}^n e_i^2 = \underline{Y}^TQ\underline{Y}$ for $Q = I - P$.

↳ it is minimum value of $S(\underline{\beta})$.

$\hat{\sigma}^2 = \dfrac{RSS}{n-p}$ ← $p = \text{rank } X$. is unbiased estimator for σ^2, where $\text{cov}(\underline{\varepsilon}) = \sigma^2 I_n$

COEFFICIENT of DETERMINATION: $R^2 = 1 - \dfrac{RSS}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \dfrac{RSS}{\text{RSS in intercept-only model}}$

↳ more parameters, lower RSS.

smaller RSS ⇒ larger R^2, $0 \le R^2 \le 1$ shows "fit" of model.

MULTIVARIATE NORMAL: for $X_1, \dots, X_r \sim N(0,1)$ i.i.d. and $\underline{\mu} \in \mathbb{R}^n$, $\underline{A} \in \mathbb{R}^{n \times r}$

$$\Rightarrow \underline{Z} = \underline{A}\underline{X} + \underline{\mu} \sim N(\underline{\mu}, \underline{A}\underline{A}^T).$$

if $\underline{Z} \sim N(\underline{\mu}, \Sigma)$, then $\underline{A}\underline{Z} + \underline{b} \sim N(\underline{A}\underline{\mu} + \underline{b}, \underline{A}\Sigma\underline{A}^T)$

LEMMA: if $\underline{Z} = \begin{pmatrix} \underline{Z}_1 \\ \vdots \\ \underline{Z}_k \end{pmatrix} \sim N(\underline{\mu}, \Sigma)$ and $\Sigma = \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_k \end{pmatrix}$

$\Rightarrow \underline{Z}_1, \dots, \underline{Z}_k$ independent.

uncorrelated + jointly normal ⇓ independent

NON-CENTRAL χ^2 : $\underline{Z} \sim N(\underline{\mu}, I_n) \Rightarrow \underline{U} = \underline{Z}^T \underline{Z} = \sum Z_i^2 \sim \chi_n^2(\delta)$

where $\delta = \sqrt{\underline{\mu}^T \underline{\mu}}$ and $\underline{\mu} \in \mathbb{R}^n$.

$\boxed{\text{if } U \sim \chi_n^2(\delta), \text{ then } E(U) = n + \delta^2 \text{ and } Var(U) = 2n + 4\delta^2}$

NON-CENTRAL t : $X \sim N(\delta, 1)$ and $U \sim \chi_n^2 \Rightarrow \dfrac{X}{\sqrt{U/n}} \sim t_n(\delta)$

F-DISTRIBUTION : $W_1 \sim \chi_{n_1}^2$ and $W_2 \sim \chi_{n_2}^2$

$\Rightarrow F = \dfrac{W_1/n_1}{W_2/n_2} \sim F_{n_1,n_2}(0)$ ($\uparrow$ or δ)

ratio of 2 chi-squared ↓ f cannot take -ve values.

for +ve semidefinite, symmetric $\underline{A} \in \mathbb{R}^{n\times n}$, $A = LL^T$ and $L^TL = \text{diag}(\text{non-zero evalues of } \underline{A})$.

if $\underline{X} \sim N(\underline{\mu}, I)$, $\underline{A}$ +ve semidefinite, symmetric and $\underline{B}$ s.t. $\underline{B}\underline{A} = 0$

$\Rightarrow \underline{X}^T \underline{A} \underline{X}$ and $\underline{B}\underline{X}$ are independent.

pf: since if r.v. X indep of Y then $g(X)$ indep of $f(Y)$.

if $\underline{Z} \sim N(\underline{\mu}, I_n)$ and A_1, A_2 proj matrices with $A_1 A_2 = 0$

$\Rightarrow \underline{Z}^T A_1 \underline{Z}$ and $\underline{Z}^T A_2 \underline{Z}$ independent.

FISHER-COCHRAN THM : $A_1, \dots, A_n$ proj matrices s.t. $\sum A_i = I_n$ and $\underline{Z} \sim N(\underline{\mu}, I_n)$, then $\underline{Z}^T A_1 \underline{Z}, \dots, \underline{Z}^T A_k \underline{Z}$ all independent

AND $\underline{Z}^T A_i \underline{Z} \sim \chi_{r_i}^2(\delta_i)$ where $r_i = \text{rank } A_i$ and $\delta_i^2 = \underline{\mu}^T A_i \underline{\mu}$

($\delta_i = \sqrt{\underline{\mu}^T A_i \underline{\mu}}$)

- under NTA, $\underline{Y} \sim N(\underline{X}\underline{\beta}, \sigma^2 I_n)$.
- MLE for σ^2 is $\hat{\sigma}^2 = \dfrac{RSS}{n} \rightarrow$ biased since $\dfrac{RSS}{n-p}$ unbiased.

$$\boxed{\frac{RSS}{\sigma^2} \sim \chi_{n-r}^2 \quad \text{where } r = \text{rank } X}$$

$\leftarrow$ PIVOTAL QUANTITY for σ^2

TEST for one component of $\underline{\beta}$: $\dfrac{\underline{c}^T\hat{\beta} - \underline{c}^T\beta}{\sqrt{\underline{c}^T(X^TX)^{-1}\underline{c}\cdot\frac{RSS}{n-p}}} \sim t_{n-p}$

(under FR, NTA.)

↳ where $p = \text{rank } X$

e.g. if we want to test for $\beta_3 = 0$, then $\underline{c} = \begin{pmatrix}0\\0\\1\\0\\\vdots\end{pmatrix}$.

equally, $\dfrac{c^T\hat{\beta} - c^T\beta}{\sqrt{c^T(X^TX)^{-1}c\,\sigma^2}} \sim N(0,1)$ if σ^2 known.

$$\boxed{\begin{array}{l}\text{F-TEST : testing for more than one component of } \underline{\beta}.\\[1em] F = \dfrac{RSS_0 - RSS}{RSS}\cdot\dfrac{(n-r)}{(r-s)} \sim F_{r-s,\,n-r} \quad \text{where } r = \text{rank } X,\ s = \text{rank } X_0.\end{array}}$$

↳ i.e. RSS of reduced model − RSS of full model / RSS of full model.

we use $RSS = \underline{Y}^T Q \underline{Y}$ and $RSS_0 = \underline{Y}^T Q_0 \underline{Y}$ $(Q_0 = I - P_0)$ onto span $(X_0)^\perp$

(↳ reject if $F > c$ at α-sig level where $P(X \geq c) = \alpha$ for $X \sim F_{r-s,\,n-r}$.)

OUTLIERS : look for residuals that are too large:

$\boxed{\dfrac{e_i}{\sqrt{(1-P_{ii})\sigma^2}} \sim N(0,1)}$ where P is proj onto span X

↖ if dk σ^2, then plug in unbiased $\sigma^2 = \dfrac{RSS}{n-p}$.

$\text{Cov}(\underline{e}) = \sigma^2(I_n - P)$ and $\text{Var}(e_i) = \sigma^2(1 - P_{ii})$

↖ known as leverage.

COOK'S DISTANCE : $\boxed{D_i = \dfrac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X(\beta_{(i)} - \beta)}{p\,RSS/(n-p)}}$ where $\beta_{(i)}$ is LSE with ith observation removed.

↳ how much $\hat{\beta}$ changes if I remove observation i.

OR

$\boxed{D_i = r_i^2 \cdot \dfrac{P_{ii}}{(1-P_{ii})r}}$ $r = \text{rank } X$ $\quad r_i^2 = \dfrac{e_i^2}{(1-P_{ii})\sigma^2}$ standardised residual.