Statistical Modelling - Concise Notes

MATH50011

Arnav Singh

Colour Code - **Definitions** are **green** in these notes, **Consequences** are **red** and **Causes** are **blue**

Content from MATH40005 assumed to be known.

Mathematics
Imperial College London
United Kingdom
April 6, 2022

Contents

1 Statistical Models

1.2 Parametric Statistical Models

Definition 1.1 ***Statistical Model***

Statistical model; collection of probability distribution $\{P_\theta : \theta \in \Theta\}$ on a given sample space.
Set Θ - (**Parameter Space**) - set of all possible parametric values, $\Theta \subset \mathbb{R}^p$

Definition 1.2 ***Identifiable***

Statistical model is **identifiable** if map $\theta \mapsto P_\theta$, one-to-one, $P_{\theta_1} = P_{\theta_2} \implies \theta_1 = \theta_2 \quad \forall \theta_1, \theta_2 \in \Theta$

1.3 Using Models

Requirements for a model

1. Agree with observed data "reasonable" well
2. reasonably simple (no excess parameters)
3. easy to interpret (parameter have practical meaning)

2 Point Estimation

Definition 2.1 ***Statistic***

Statistic - function of observable random variable.

Definition 2.2 ***Estimate/Estimators***

t a statistic
$t(y_1, \ldots, y_n)$ called **estimate** of θ
$T(Y_1, \ldots, Y_n)$ an **estimator** of Θ

2.1 Properties of estimators

2.1.1 Bias

Definition 2.3 ***Bias***

T estimator for $\theta \in \Theta \subset \mathbb{R}$

$$bias_\theta(T) = E_\theta(T) - \theta$$

unbiased if $bias_\theta(T) = 0, \quad \forall \theta \in \Theta$

If $\Theta \subset \mathbb{R}^k$ often interested in $g(\theta),\ g : \theta \to \mathbb{R}$

$$\text{extend } bias_\theta(T) = E_\theta(T) - g(\theta)$$

2.1.2 Standard error

Definition 2.4

T estimator for $\theta \in \Theta \subset \mathbb{R}$

$$SE_\theta(T) = \sqrt{Var_\theta(T)}$$

Standard error, is standard deviation of sampling distribution of T

2.1.3 Mean Square Error

Definition 2.5

T estimator for $\theta \in \Theta \subset \mathbb{R}$
Mean square error of T

$$\begin{aligned} MSE_\theta(T) &= E_\theta(T - \theta)^2 \\ &= Var_\theta(T) + [bias_\theta(T)]^2 \end{aligned}$$

3 The Cramér-Rao Lower Bound

Theorem 3.1 *(Cramér-Rao Lower Bound)*

$T = T(X)$ unbiased estimator for $\theta \in \Theta \subset \mathbb{R}$ for $X = (X_1, \dots, X_n)$ with just pdf $f_\theta(x)$ under mild regularity conditions:

$$Var_\theta(T) \geq \frac{1}{I(\theta}$$

For I_θ the **Fisher information of sample**

$$\begin{aligned} I(\theta) &= E_\theta \left[\left\{ \frac{\partial}{\partial \theta} \log f_\theta(x) \right\}^2 \right] \\ &= -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(x) \right] \\ I_n(\theta) &= -nE_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(x) \right] \end{aligned}$$

Proposition.

For a random sample: Fisher info proportional to sample size

Jensen's inequality

For X a random variable with φ a convex function

$$\varphi(E[X]) \leq E\left[\varphi(X)\right]$$

Call $E\left[\varphi(X)\right] - \varphi\left(E[X]\right)$ the **Jensen gap**

4 Asymptotic Properties

Definition 4.1

Sequence of estimators $(T_n)_{n \in \mathbb{N}}$ for $g(\theta)$ called **(weakly) consistent** if $\forall \theta \in \Theta$

$$T_n \xrightarrow{P_\theta} g(\theta) \quad (n \to \infty)$$

Definition 4.2

Convergence in probability: $T_n \xrightarrow{P_\theta} g(\theta)$

$$\forall \epsilon > 0 : \lim_{n \to \infty} P_\theta(|T_n - g(\theta)| < \epsilon) = 1$$

Lemma - (Portmanteau Lemma)

X, X_n real valued random value.
Following are equivalent:

1. $X_n \to X$ as $n \to \infty$
2. $E[f(X_n)] \to E[f(X)] \quad n \to \infty$ for all bounded + continuous functions $f : \mathbb{R} \to \mathbb{R}$

Definition 4.3

Sequence of estimators $(T_n)_{n \in \mathbb{N}}$ for $g(\theta)$ **asymptotically unbiased** if $\forall \theta \in \Theta$

$$E_\theta \to g(\theta) \quad n \to \infty$$

Lemma.

(T_n) asymptotically unbiased for $g(\theta)$ and $\forall \theta \in \Theta$

$$Var_\theta(T_n) \to 0 \quad n \to \infty$$

$\implies (T_n)$ consistent for $g(\theta)$

Definition 4.4

Sequence (T_n) of estimators for $\theta \in \mathbb{R}$ **asymptotically normal** if

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta))$$

for some $\sigma^2)(\theta)$

Theorem 4.1 *(Central Limit Theorem)*

$Y_1, \ldots, Y_n$ be iid random variable with $E(Y_i) = \mu,\ Var(Y_i) = \sigma^2$

$$\implies \text{sequence } \sqrt{n}(\bar{Y} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Remark.
Under mild regularity conditions for asymptotically normal estimators T_n

$$SE_\theta(T_n) \approx \frac{\sigma(T_n)}{\sqrt{n}}$$

Lemma. *(Slutsky)*

X_n, X, Y_n random variables

If $X_n \underset{d}{\rightarrow} X$ and $Y_n \underset{p}{\rightarrow} c$ for constant c

1. $X_n + Y_n \underset{d}{\rightarrow} X + c$
2. $Y_n X_n \underset{d}{\rightarrow} cX$
3. $Y_n^{-1} X_n \underset{d}{\rightarrow} c^{-1} X \quad$ provided $c \neq 0$

Theorem 4.2 *(Delta Method)*

Suppose T_n asymptotically normal estimator of θ with

$$\sqrt{n}(T_n - \theta) \underset{d}{\rightarrow} N(0, \sigma^2(\theta))$$

$g : \Theta \rightarrow \mathbb{R})$ differentiable function with $g'(\theta) \neq 0$. Then

$$\sqrt{n}[g(T_n) - g(\theta)] \underset{d}{\rightarrow} N(0, g'(\theta)^2 \sigma^2(\theta))$$

Theorem 4.3 *(Continuous Mapping Theorem)*

$k, m \in \mathbb{N}, X, X_n, \quad \mathbb{R}^k-$valued random variable.
$g : \mathbb{R}^k \rightarrow \mathbb{R}^m$ continuous function at every point of C s.t $P(X \in C) = 1$

- If $X_n \underset{d}{\rightarrow} X \implies g(X_n) \underset{d}{\rightarrow} g(x)$ as $n \rightarrow \infty$
- If $X_n \underset{p}{\rightarrow} X \implies g(X_n) \underset{p}{\rightarrow} g(X)$ as $n \rightarrow \infty$
- If $X_n \underset{a.s}{\longrightarrow} X \implies g(X_n) \underset{a.s}{\longrightarrow} g(X)$ as $n \rightarrow \infty$

5 Maximum Likelihood Estimation

Definition 5.1 *(Likelihood function)*

Suppose observer Y with realisation y
Likelihood function

$$L(\theta) = L(\theta : y) = \begin{cases} P(Y = y : \theta) & \text{discrete data} \\ f_Y(y : \theta) & \text{absolutely continuous data} \end{cases}$$

Likelihood function is the joint pdf/pmf or observed data as a function of unknown parameter.

Random sample $Y = (Y_1, \ldots, Y_n) \quad Y_i$ iid.
If Y_i has pdf $f(\cdot; \theta)$

$$\implies L(\theta) = \prod_{i=1}^{n} f(y_i : \theta)$$

Definition 5.2 *(Maximum Likelihood Estimator)*

MLE of θ is estimator $\hat{\theta}$ s.t

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

5.1 Properties of Maximum Likelihood estimators

5.1.1 MLEs functionally invariant

g bijective function
$\hat{\theta}$ MLE of $\theta \implies \hat{\phi} = g(\hat{\theta})$ a MLE of $\phi = g(\theta)$

5.1.2 Large Sample property

Theorem 5.1

$X_1, X_2, \ldots$ iid observations with pdf/pmf f_θ
$\theta \in \Theta,\ \Theta$ an open interval
$\theta_0 \in \Theta$ - true parameter.

Under regularity conditions ($\{x : f_\theta(x) > 0\}$ indpendent of θ). We have

1. $\exists$ consistent sequence $(\hat{\theta})_{n \in \mathbb{N}}$ of MLE
2. $(\hat{\theta})_{n \in \mathbb{N}}$ consistent sequence of MLEs $\implies \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, (I_f(\theta_0))^{-1})$ *(Asymptotic normality of MLE)*
   Where $I_f\theta$ Fisher information of sample size = 1

Remark: if MLE unique $(\forall n) \implies$ sequence of MLEs consistent

Remark
Limiting distribution depends on $I_f(\theta_0)$, which is often unknown in practical situations. $\implies$ need to estimate $I_f(\theta_0)$

iid sample; $I_f(\theta_0)$ estimated by

- $I_f(\hat{\theta})$
- $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\log(f(x_i : \theta))|_{\theta=\hat{\theta}}\right)^2$
- $-\frac{1}{n}\sum_{i=1}^{n}(\frac{\partial}{\partial\theta})^2\log(f(x_i : \theta))|_{\theta=\hat{\theta}}$

Often consistent $\implies$ converge to $I_f(\theta_0)$ in probability
Remark

Standard error of asymptotically normal MLE $\hat{\theta}_n$
Approximated by $SE(\hat{\theta}_n) = \sqrt{\hat{I}_n^{-1}}/\sqrt{n}\ \hat{I}_n$ estimator from above.

Remark - Multivariate version.
$\Theta \subset \mathbb{R}^k$ open set, $\hat{\theta}_n$ MLE based on n observation.

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, (I_f(\theta_{\mathbf{0}})^{-1})$$

$\theta_{\mathbf{0}}$ the true parameter, $I_f(\theta)$ **Fisher information matrix**

$$\begin{aligned} I_f(\theta) &:= E_\theta\left[(\nabla \log f(X;\theta))^T (\nabla \log f(X;\theta))\right] \\ &:= -E_\theta\left[\nabla^T \nabla \log f(X:\theta)\right] \end{aligned}$$

Definition 5.3

Converges in distribution for random vector
$\mathbf{X}, \mathbf{X_1}, \mathbf{X_2}$ random vectors of dimension k

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \quad (n \to \infty)$$

If $P(\mathbf{X}_n \leq z) \xrightarrow[n\to\infty]{} P(\mathbf{X} \leq z) \quad \forall z \in \mathbb{R}^k : z \mapsto P(X \leq Z)$ continuous

6 Confidence Regions

Definition 6.1 *(Confidence interval)*

$1-\alpha$ **confidence interval** for θ, a random interval I containing 'true' paramter with probability $\geq 1-\alpha$

$$P_{\theta \in I} \geq 1-\alpha \quad \forall \theta \in \Theta$$

6.1 Construction of confidence intervals

Definition 6.2

Pivotal Quantity for θ a function $t(Y,\theta)$ of data and θ
s.t distribution of $t(Y,\theta)$ known (no dependency on unknown parameters)
Know distribution of $t(Y,\theta) \implies$ can find constant a_1, a_2 s.t $P(a_1 \leq t(Y_1,\theta) \leq a_2) \geq 1-\alpha$
$\implies P(h_1(Y) \leq \theta \leq h_2(Y)) \geq 1-\alpha$

Call $[h_1(Y), h_2(Y)]$ a **random interval**
with observed interval $[h_1(y), h_2(y)]$ a $1-\alpha$ **confidence interval for** θ

6.2 Asymptotic confidence intervals

We often know

$$\begin{aligned} &\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta)) \\ \implies &\underbrace{\sqrt{n}(\frac{T_n - \theta}{\sigma(\theta)})}_{\text{use as pivotal quantity}} \xrightarrow{d} N(0,1) \end{aligned}$$

Definition 6.3

Sequence of random intervals I_n
an **asymptotic $1-\alpha$ Confidence Interval** if

$$\lim_{n\to\infty} P_\theta(\theta \in_n) \geq 1-\alpha \quad \theta$$

Simplification

Given consistent estimator $\hat{\sigma}_n$ for $\sigma(\theta)$ $\hat{\sigma}_n \xrightarrow{P_\theta} \sigma(\theta)$ $\forall\theta$

$$\sqrt{n}(\frac{T_n-\theta}{\sigma(\theta)}) \xrightarrow{d} N(0,1)$$

$$T_n \pm c_{\alpha/2} \times \underbrace{\frac{\hat{\sigma}_n}{\sqrt{n}}}_{\text{estimates } SE(T_n)}$$

$$T_n \pm c_{\alpha/2} SE(T_n)$$

Simplification.
$\hat{\sigma}^2 = \frac{Y}{n}(1-\frac{Y}{n})$ $\quad \hat{\sigma}^2 \xrightarrow{P} \theta(1-\theta)$

$$\underbrace{\sqrt{n}\frac{Y/n-\theta}{\sqrt{\frac{Y}{n}(1-\frac{Y}{n})}}}_{\text{pivotal quantity}} \implies \frac{y}{n} \pm \frac{c_{\alpha/2}}{\sqrt{n}}\sqrt{\frac{y}{n}(1-\frac{y}{n}}$$

6.3 Simultaneous Confidence Interval/Confidence regions.

Definition 6.4

$\theta = (\theta_1, \ldots, \theta_k)^T \in \Theta \in \mathbb{R}^k$
With random intervals $(L_i(\mathbf{Y}), U_i(\mathbf{Y}))$ s.t

$$\forall\theta : P_\theta(L_i(\mathbf{Y} < \theta_i < U_i(\mathbf{Y}), i \in \{1, \ldots, k\}) \geq 1-\alpha$$

$(L_i(\mathbf{y}, U_i(\mathbf{y}))$ $i \in \{1, \ldots, k\}$ a **$1-\alpha$ simultaneous confidence interval** for $\theta_1, \ldots, \theta_k$
Remark - (Bonferroni correction)
take $[L_i, U_i]$ a $1-\alpha$ confidence interval for $\theta_i, \ i \in \{1, \ldots, k\}$

7 Hypothesis Testing

7.1 Prelim

Definition 7.1 *(Hypotheses)*

We have 2 complementary hypothesis

- H_0 : Null hypothesis - consider to be the status quo
- H_1: Alternative hypothesis

Definition 7.2 *(Hypthesis Test)*

Hypothesis test a rule that specifies for which valus of a sample $x_1, \ldots, x_n$ a decision is to be made

- accept H_0 as true
- reject H_0 and accept H_1

Rejection region/Critical region - subset of sample space for which H_0 rejected

Definition 7.3 *(Types of error)*

	H_0 True	H_0 False
Don't reject H_0	✓	**Type II Error**
Reject H_0	**Type I Error**	✓

7.2 Power of a Test

Definition 7.4 *(Power function)*

Θ parameter space with $\Theta_0 \subset \Theta,\ \Theta_1 = \Theta \backslash \Theta_0$
Consider:

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

Given a test for this hypothsis, we have a **Power function**

$$\beta : \theta \to [0,1]$$
$$\beta(\theta) = P_\theta(\text{reject} H_0)$$

$\theta \in \Theta_0 \implies$ want $\beta(\theta)$ small
$\theta \in \Theta_1 \implies$ want $\beta(\theta)$ large

7.3 p-Value

Definition 7.5 *(p-value)*

$$p = \sup_{\theta \in \Theta_0} P_\theta(\text{observing something 'at least as extreme' as the observation})$$

reject $H_0 \iff p \leq \alpha$
For test based on statistic T with rejection for large value of T we have

$$p = \sup_{\theta \in \Theta_0} P_\theta(T \geq t)$$

for t our observed value

7.4 Connection between tests & confidence intervals

7.4.1 Constructing a test from confidence region

Y a random observation.
$A(Y)$ a $1-\alpha$ confidence region for θ

$$P(\theta \in A(Y)) \geq 1 - \alpha \quad \forall \theta \in \Theta$$

Define test for $\begin{matrix} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \notin \Theta_0 \end{matrix}$ for $\Theta_0 \subset \Theta$ a fixed subset with level α s.t

$$\text{Reject } H_0 \text{ if } \Theta_0 \cap A(Y) = \emptyset$$

$$\begin{aligned} P_\theta(\text{Type I error}) = P_\theta(\text{reject}) &= P_\theta(\Theta_0 \cap A(Y) = \emptyset) \\ &\leq P_\theta(\theta \notin A(Y)) \leq \alpha \end{aligned}$$

7.4.2 Constructing confidence region from tests

Suppose $\forall \theta_0 \in \Theta$ we have a level α test ϕ_{θ_0} for

$$H_0^{\theta_0} : \theta = \theta_0 \quad \text{vs.} \quad H_1^{\theta_0} : \theta \neq \theta_0$$

A decision rule ϕ_{θ_0} to reject/not reject $H_0^{\theta_0}$ satisfying:

$$P_{\theta_0}(\phi_{\theta_0} \text{ reject } H_0^{\theta_0}) \leq \alpha$$

Consider random set:

$$A := \left\{ \theta_0 \in \Theta : \phi_{\theta_0} \text{ doesn't reject } H_0^{\theta_0} \right\}$$

We see A a $1-\alpha$ confidence region for θ
$\forall \theta \in \Theta$ $P_\theta(\theta \in A) = P_\theta(\phi_\theta \text{ not rejects }) = 1 - P_\theta(\phi_\theta \text{ rejects }) \geq 1 - \alpha$

8 Likelihood Ratio Tests

(Numbers don't line up with official notes!!!)

Definition 8.1 *(Likelihood ratio statistic)*

$$t(\mathbf{y}) = \frac{sup_{\theta\in\Theta} L(\theta;\mathbf{y})}{sup_{\theta\in\Theta_0} L(\theta;\mathbf{y})} = \frac{\text{max likelihood under } H_0 + H_1}{\text{max likelihood under } H_0}$$

Theorem 8.1

$X_1, \dots, X_n \sim N(0,1),\ X_i$ independent

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

Theorem 8.2

Under regularity conditions

$$2\log t(\mathbf{Y}) \xrightarrow{D} \chi_r^2 \quad (n \to \infty)$$

under H_0 where r the number of independent restrictions on θ needed to define H_0

9 Linear models with 2nd order assumptions

9.1 Simple Linear Regression

Definition 9.1 *(Simple Linear Model)*

$$\underbrace{Y_i}_{\substack{\text{outcome} \\ \text{observable random var}}} = \underbrace{\beta_1 + \overbrace{a_i}^{\substack{\text{covariate} \\ \text{(observable constant)}}} \beta_2}_{\substack{\text{unknown} \\ \text{parameters}}} + \overbrace{\epsilon_i}^{\text{error (not observable)}}$$

Least Square Estimators
$\hat{\beta}_1, \hat{\beta}_2$ of β_1, β_2 defined as minimisers of

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - a_i\beta_2)^2$$

Remark

- $e_i = y_i = \hat{\beta}_1 - a_i\hat{\beta}_2$ - **residuals** are observable, not i.i.d
- unkown parameters β_1, β_2 and σ^2
- $Y_1, \dots, Y_n$ generally not i.i.d observations
  independence holds if $\epsilon_1, \dots, \epsilon_n$ independent
  Y_i not of same distribution, distribution depending on covariate a_i

9.2 Matrix Algebra

Lemma 5

(i) $A \in \mathbb{R}^{n\times m}, B \in \mathbb{R}^{m\times n}$
$(AB)^T = B^T A^T$

(ii) $A \in \mathbb{R}^{n\times n}$ invertible
$\implies (A^{-1})^T = (A^T)^{-1}$

(iii) $trace(AB) = trace(BA)$

(iv) $rank(X^T X) = rank(X)$

Lemma 8
$A \in \mathbb{R}^{n\times n}$ symmetric $\implies \exists$ orthogonal P s.t $P^T AP$ diagonal with diagonal entries = e.vals of A
A positive definite, symmetric $\implies \exists Q$ non-singular s.t $Q^T AQ = I_n$

9.3 Review of rules for E, cov for random vectors

Definition 9.2

$\mathbf{X} = (X_1, \dots, X_n)^T$ random vector

$$\implies E(\mathbf{X}) = (E(X_1), \dots, E(X_n))^T$$

Lemma 9
$\mathbf{X}, \mathbf{Y}$ random vector

(i) $E(\mathbf{X} + \mathbf{Y}) = E(\mathbf{X}) + E(\mathbf{Y})$

(ii) $E(a\mathbf{X}) = aE(\mathbf{X})$

(iii) AB deterministic matrices
$E(A\mathbf{X}) = AE(\mathbf{X})$, $E(\mathbf{X^T}B) = E(\mathbf{X})^T B$

Definition 9.3 *(Covariance)*

X,Y random vectors

$$cov(\mathbf{X}, \mathbf{Y}) = E(\mathbf{X}\mathbf{Y^T}) - E(\mathbf{X})E(\mathbf{Y})^T$$
$$cov(\mathbf{X}) = cov(\mathbf{X}, \mathbf{X})$$

Lemma 10
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ random vector
A, B deterministic matrices, $a, b \in \mathbb{R}$

(i) $cov(\mathbf{X}, \mathbf{Y}) = cov(\mathbf{Y}, \mathbf{X})^T$

(ii) $cov(a\mathbf{X} + b\mathbf{Y}, Z) = a \cdot cov(\mathbf{X}, \mathbf{Z}) + b \cdot cov(\mathbf{Y}, \mathbf{Z})$

(iii) $cov(A\mathbf{X}, B\mathbf{Y}) = Acov(\mathbf{X}, \mathbf{Y})B^T$

(iv) $cov(A\mathbf{X}) = Acov(\mathbf{X})A^T$
$cov(\mathbf{X})$ positive semidefinite and symmetric
i.e. $\mathbf{c}^T cov(\mathbf{X})\mathbf{c} \geq 0 \ \forall \mathbf{c}$
All e.val. of $cov(\mathbf{X})$ non-negative

(v) $\mathbf{c}, \mathbf{Y}$ independent $\implies cov(\mathbf{X}, \mathbf{Y}) = 0$

9.4 Linear Model

Definition 9.4

In a **linear model**

$$\mathbf{Y} = X\beta + \epsilon$$

- $\mathbf{Y}$ - n. dimensional random vector (observable)
- $X \in \mathbb{R}^{n \times p}$ known matrix - **design matrix**
- $\beta \in \mathbb{R}^p$
- ϵ n-variate random vector (not observable); $E(\epsilon) = 0$

Assumptions
2nd order assumptions (SOA)

$$cov(\epsilon) = (cov(\epsilon_i, \epsilon_j))_{\substack{i=1,\dots,n \\ j=1,\dots,n}} = \sigma^2 I_n \quad \sigma^2 > 0$$

Normal theory assumptions (NTA)
$\epsilon \sim N(0, \sigma^2 I_n)$, some $\sigma^2 > 0$
$N-$multivariate n-dimensional normal multivariate distribution

$$NTA \implies SOA$$

Full rank (FR)
X has full rank $rank(X) = r$

9.5 Identifiability

Definition 9.5

Suppose statistical model with unkown parameter θ
θ **identifiable** if no 2 different values of θ yield same distribution of observed data.

9.6 Least Square estimation

Estimate β by least squares.
Least squares: choose β to minimise

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 \\ &= (Y - X\beta)^T(Y - X\beta) \\ &= Y^TY - 2Y^TX\beta + \beta^TX^TX\beta \\ \frac{\partial S(\beta)}{\partial \beta} &= \frac{\partial S(\beta)}{\partial \beta_i}_{\,i=1,\dots,p} = -2X^TY + 2X^TX\beta \end{aligned}$$

$$\begin{aligned} \text{Unique solution} &\iff X^TX \text{ invertible } (rank = p) \quad rank(X^TX) = rank(X) \\ &\iff \text{linear model of full rank} \end{aligned}$$

$\hat{\beta}$ satisfies LSE $\implies$ minimise $S(\beta)$

9.7 Properties of LSE

Assume (FR) and (SOA) $\implies \hat{\beta} = (X^TX)^{-1}X^TY$

- $\hat{\beta}$ linear in $\mathbf{X}$
  i.e. $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}^p, y \mapsto (X^TX)^{-1}X^T\mathbf{y}$ linear mapping
- $\hat{\beta}$ unbiased for β
  $\forall \beta\ E(\hat{\beta}) = (X^TX)^{-1}X^TE(\mathbf{Y}) = (X^TX)^{-1}X^TX\beta = \beta$
- $cov(\hat{\beta}) = \sigma^2(X^XX)^{-1}$

Definition 9.6

Estimator $\hat{\gamma}$ linear if $\exists L \in \mathbb{R}^n$ s.t $\hat{\gamma} = L^TY$

Theorem 9.1 *(Gauss-Markov Theorem for FR linear models)*

Assume (FR),(SOA)
Let $\mathbf{c} \in \mathbb{R}^p, \hat{\beta}$ a least square estimator of β in a linear model.
$\implies$ estimator $c^T\beta$ has smallest variance among all linear unbiased estimators for $c^T\beta$

9.8 Projection Matrices

Definition 9.7

L a linear subspace of $\mathbb{R}^n, dim(L) = r \leq n$
$P \in \mathbb{R}^{n\times n}$ a projection matrix onto L if

(i) $P\mathbf{x} = \mathbf{x} \quad \forall \mathbf{x} \in L$

(ii) $P\mathbf{x} = \mathbf{0} \quad \forall \mathbf{x} \in L^{\perp} = \{\mathbf{z} \in \mathbb{R}^n : \mathbf{z}^T\mathbf{y} = 0\ \forall \mathbf{y} \in L\}$

Lemma 11
P a projection matrix $\iff \underbrace{P^T = P}_{P \text{ symmetric}}$ and $\underbrace{P^2 = P}_{P \text{ independent}}$

Lemma 12
A a $n \times n$ projection matrix $(A = A^T, A^2 = A)$ of $rank(r)$

(i) r of e.val of A are 1 and $n - r$ are 0

(ii) $rank(A) = trace(A)$

9.9 Residuals, Estimation of the variance

Definition 9.8

$\hat{Y} = X\hat{\beta}$, $\hat{\beta}$ a least squares estimator, called vector of fitted values.
Lemma 13
$\hat{Y}$ unique and

$$\hat{Y} = PY$$

P the projection matrix onto column space of X

Definition 9.9

Vector of residuals.

$$\begin{aligned}\mathbf{e} &= Y - \hat{Y} : \text{ vector of residuals}\\ &= Y - PY = QY, Q = I - P : \text{ the projection of matrix onto } span(X)^{\perp}\end{aligned}$$

$$E(\mathbf{e}) = E(QY) = QE(Y) = \underbrace{QX}_{=0}\beta = 0$$

Diagnostic plots
Suppose data comes from model

$$Y = X\beta + Z\gamma + \epsilon \quad E(\epsilon) = 0$$

$z \in \mathbb{R}^n \backslash span(X), \gamma \in \mathbb{R}$ deterministic
But analyst works with

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

$\implies$ if $\gamma \neq 0$, used wrong model

$$\implies E(\epsilon) = E(QY) = E(Q(X\beta + Z\gamma + \epsilon)) = QZ\gamma$$

$\implies$ plot QZ against residuals yields line through the origin.
if non-zero slope $\implies$ consider including Z

Residual sum of squares

Definition 9.10 *(Residual sum of squares)*

$$RSS = e^T e$$

Other forms

- RSS $= \sum_{i=1}^{n} e_i^2$
- RSS $= S(\hat{\beta}) = \|Y - X\hat{\beta}\|^2$
- RSS $= Y^T Y - \hat{Y}^T \hat{Y}$
- RSS $= (Y - \hat{Y})^T (Y - \hat{Y})$
- RSS $= (QY)^T QY$
- RSS $= Y^T QY$

Theorem 9.2

$$\hat{\sigma}^2 = \frac{RSS}{n - r}$$

An unbiased estimator of σ^2, $r = rank(X)$

Coefficient of determination - $(\mathbb{R}^2)$

For models containing intercept term (X has column of 1s or other constants)

$$R^2 = 1 - \frac{RSS}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Small RSS 'better' $\implies$ want large R^2
$0 \leq R^2 \leq 1 \implies R^2 = 1$ for perfect model.

Remark
$\frac{RSS}{n}$ an estimator of σ^2

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

estimator of σ^2 in model with only intercept term.

$$\implies \frac{RSS/n}{\frac{1}{n}\sum(Y_i - \bar{Y})^2} \approx \frac{\text{Var. in model}}{\text{Total variance}} \implies R^2 \approx \frac{\text{Total var. - Var. in Model}}{\text{Total var.}}$$

10 Linear Models with Normal theory Assumptions

10.1 Distributional Results

10.1.1 Multivariate Normal Distribution

Denoted $N(\underbrace{\mu}_{\in\mathbb{R}^n}, \underbrace{\Sigma}_{\in\mathbb{R}^{n\times n}})$, distribution of random vec. μ - Expectation, Σ - Covariance

Definition 10.1

Σ - positive definite
$Z \sim N(\mu, \Sigma)$ if Z has pdf of form

$$f(z) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1.2}} \exp\left(-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right)$$

n-variate random vector Z follows MVN distribution if

- $\forall a \in \mathbb{R}^n$ random variable a^TZ follows univariate normal distribution
- $X_1, \dots, X_n \sim N(0,1)$ iid, let $\mu \in \mathbb{R}^d, A \in \mathbb{R}^{n\times r}$
  $\implies Z = AX + \mu \sim N(\mu, AA^T)$
- $Z \sim N(\mu, \Sigma)$ if its characteristic function $\phi : \mathbb{R}^n \to \mathbb{C}, \phi(t) = E(\exp(iZ^Tt)$ satisfies

$$\phi(t) = \exp\left(i\mu^Tt - \frac{1}{2}t^T\Sigma t\right) \quad \forall\ t \in \mathbb{R}^n, \mu \in \mathbb{R}^n, \Sigma \in \mathbb{R}^{n\times n} \text{ symm. pos. def}$$

Remark
$Z \sim N(\mu, \Sigma) \implies$

- $E(Z) = \mu$
- $cov(Z) = \Sigma$
- A deterministic matrix, b deterministic vector
  $AZ + b \sim N(A\mu + b, A\Sigma A^T)$

Remark
X, Y random variables
$cov(X,Y) \neq= 0 \not\Longrightarrow X, Y$ independent
Lemma 14
$i = 1, \dots, k$ let $A_i \in \mathbb{R}^{n_i \times n_i}$ positive semidefinite and symmetric
Z_i a n_i-variate random vector
if $Z = \begin{pmatrix} Z_1 \\ \dots \\ Z_k \end{pmatrix} \sim N(\mu, \Sigma)$ for some $\mu \in \mathbb{R}^{\sum_{i=1}^k n_i}$ and $\Sigma = diag(A_1, \dots, A_n) \implies Z_1, \dots, Z_k$ independent.

10.1.2 Distributions derived from MVN

Definition 10.2 χ^2 *(Chi squared distribution)*

$Z \sim N(\mu, I_n),\ \mu \in \mathbb{R}^n$
$U = Z^T Z = \sum_{i=1}^n z_i^2$ has non-central χ^2 distribution with n degrees of freedom and non-centrality parameter; $\delta = \sqrt{\mu^T \mu}$

$$U \sim \chi_n^2(\delta), \quad \chi_n^2 = \chi_n^2(0)$$

Lemma
$U \sim \chi_n^2(\delta) \implies E(U) = n + \delta^2,\ Var(U) = 2n + 4\delta^2$
$U_i \sim \chi_{n_i}^2(\delta_i), i = 1, \ldots, k$ and U_i independent

$$\implies \sum_{i=1}^k U_i \sim \chi^2_{\sum_{n_i} \sqrt{\Sigma \delta_i^2}}$$

Definition 10.3

X, U independent random variables, $X \sim N(\delta, 1),\ U \sim \chi_n^2$

$$Y = \frac{X}{\sqrt{U/n}} \sim t_n(\delta)$$

Non-central t-distribution with n degrees of freedom and centrality parameter δ
$t_n = t_n(0)$
Remark
$Y_n \sim t_n\ \forall n \in \mathbb{N}$

$$Y_n \xrightarrow[n \to \infty]{d} N(0,1)$$

Definition 10.4

$W_1 \sim \chi_{n_1}^2(\delta), W_2 \sim \chi_{n_2}^2$ independently

$$F = \frac{W_1/n_1}{W_2/n_2} \sim F_{n_1,n_2}(\delta)$$

Non-central F distribution with (n_1, n_2) degrees of freedom and non-centrality parameter $= \delta$
$F_{n_1,n_2} = F_{n_1,n_2}(0)$

10.1.3 Some independence results

Lemma 16
$A \in \mathbb{R}^{n \times n}$ positive semidefinite and symmetric matrix of rank r

$$\implies \exists L \in \mathbb{R}^{n \times r} \text{ s.t } rank(L) = r, A = LL^T\ L^T L = diag(\text{non-zero evals of } A)$$

Lemma 17
$X \sim N(\mu, I),\ A \in \mathbb{R}^{n \times n}$ positive semidefinite symmetric, B s.t $BA = 0$

$$\implies X^T AX, BX \text{ independent}$$

Lemma 18
$Z \sim N(\mu, I_n),\ A$ a $n \times n$ projection matrix of rank r

$$\implies Z^T AZ \sim \chi_r^2(\delta) \quad \delta^2 = \mu^T A \mu$$

Lemma 19
$Z \sim N(\mu, I_n), A_1, A_2 \in \mathbb{R}^{n \times n}$ prejocetion matrix s.t $A_1 A_2 = 0$

$$\implies Z^T A_1 Z \& Z^T A_2 Z \text{ independent}$$

Lemma 20
$A_1, \ldots, A_k$ symmetric $n \times n$ matrices s.t $\Sigma(A_i) = I_n$ if rank$A_i = r_i$
Following equivalent

(i) $\Sigma r_i = n$

(ii) $A_i A_j = 0 \quad \forall i \neq j$

(iii) A_i independent $\forall i = 1, \ldots, k$

Theorem 10.1 *(The Fisher-Cochran Theorem)*

Consider linear model $Y = X\beta + \epsilon,\ E(\epsilon) = 0$ with (NTA)
(NTA): $\epsilon \sim N(0, \sigma^2 I_n) \implies Y \sim N(X\beta, \sigma^2 I_n)$

$$f(y) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right)$$

Estimation using maximum likelihood approach:

- Log-likelihood of data is
$$L(\beta, \mu^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\underbrace{(Y - X\beta)^T(Y - X\beta)}_{S(\beta}$$
- Maximising L w.r.t β (for fixed σ^2) equivalent to minimising $S(\beta) = (Y - X\beta)^T(Y - X\beta)$
Max likelihood equivalent to least squares for estimating β
- MLE for σ^2 is $\frac{RSS}{n}$
$$L(\hat{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}RSS \quad \text{w.r.t } \sigma^2$$

10.1.4 Confidence intervals, tests for one dimensional quantities.

Lemma 21 - *(*Distribution of RSS)
Assume (NTA) $\implies \frac{RSS}{\sigma^2} \sim \chi^2_{n-r}\ r = rank(X)$

Lemma 22
Assume (FR),(NTA) in linear model.
Let $c \in \mathbb{R}^p$

$$\frac{c^T\hat{\beta} - c^T\beta}{\sqrt{c^T(X^TX)^{-1}c\frac{RSS}{n-p}}} \sim t_{n-p}$$

10.2 The F-test

Lemma 23
Under $H_0 : E(Y) \in Span(X_0)$

$$F = \frac{RSS_0 - RSS}{RSS} \cdot \frac{n-r}{r-s} \sim F_{r-s,n-r}$$

$r = rank(X), s = rank(X_0)$
NEED EXPLAINING AND TYPING UP STILL

10.3 Confidence regions

Suppose $E(Y) = X\beta$ a linear model satisfying (FR),(NTA)
Want to find random set D s,t $P(\beta \in D) \geq 1 - \alpha\ \forall\beta, \sigma^2$

$$A = \frac{(\hat{\beta} - \beta)^T X^T X(\hat{\beta} - \beta)}{RSS} \cdot \frac{n-p}{p}$$

Find distribution of $A \implies$ use A as pivotal quantity for β
Numerator of first fraction re-written as

$$(Y - X\beta)^T P(Y - X\beta)$$

P, projection onto space span(cols. of X)

$$(Y - X\beta)^T P(Y - X\beta) = (Y - X\beta)^T PP(Y - X\beta) = [P(Y - X\beta)]^T[P(Y - X\beta)]$$

Taking $P = X(X^TX)^{-1}X^T$

$$\implies [X(\hat{\beta} - \beta)]^T[X(\hat{\beta} - \beta)]$$

With

$$RSS = Y^TQY = (Y - X\beta)^TQ(Y - X\beta), \quad Q = I_P \implies Z = \frac{1}{\sigma}(Y - X\beta)$$

$$A = \frac{Z^TPZ}{Z^TQZ} \cdot \frac{n-p}{p} \quad Z \sim N(0,1), P + Q = I, rank(P) = p, P\&Q \text{ proj. mat.}$$

$\implies$ by Fisher-Cochran Theorem $A \sim F_{p,n-p}$
$1 - \alpha$ confidence region R for β defined by all $\gamma \in \mathbb{R}^p$ s.t

$$\frac{(\hat{\beta} - \gamma)^T X^T X(\hat{\beta} - \gamma)}{RSS} \cdot \frac{n-p}{p} \leq F_{p,n-p,\alpha}$$

$P(Z \geq F_{p,n-p,\alpha}) = \alpha$ for $Z \sim F_{p,n-p}$
R an ellipsoid central at $\hat{\beta}$

Remark
General definition of ellipsoid

$$\{z \in \mathbb{R}^p : (z - z_0)^T A^{-1}(z - z_0) \leq 1\} \quad A \text{ pos. semi def.}, z_0 \in \mathbb{R}^p$$

11 Diagnostics,Model selection, Extensions

11.1 Outliers

Definition 11.1 *(Outlier)*

Outlier - an obseravtion that does not conform to general pattern of the rest of the data.

Potential causes

- error in data recording mechanism
- Data set may be 'contaminated (e.g. mix of 2 or more populations)
- Indication that model/underlying theory needs improvement

Spot outliers $\implies$ look for residuals that are 'too large'

$$\mathbf{e} = (I - P); \quad P - \text{ projects onto } span(X)$$

X full rank $\implies P = X(X^TX)^{-1}X^T$

$$cov(\mathbf{e}) = (I - P)cov(Y)(I - P)^T = \sigma^2(I - P) \quad E(\mathbf{e}) = 0$$

$\implies$ under (NTA) $e_i \sim N(0, \sigma^2(1 - P_{ii}))$ $\quad P_ii$ the i^{th} diagonal of P

$$\implies \frac{e_i}{\sqrt{(1 - P_{ii}\sigma^2}} \sim N(0,1)$$

σ^2 unknown $\implies$ use unbiased estimator $\hat{\sigma}^2 = \frac{RSS}{n-p}$

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - P_{ii}}}$$

r_i not necessarily $\sim N(0,1)$ but distribution is close to it.
Remark
$r_i \not\sim t$; $\hat{\sigma}^2, e_i$ not independent
Remark
$X \sim N(0,1) \implies$ probability for large X v. rapidly decreasing
if (NTA) holds $\implies$ standardised residuals should be relatively small

11.2 Leverage

Definition 11.2

Leverage of i^{th} observation in linear model is P_{ii}
i^{th} diagonal matrices of hat matrix P

11.3 Cook's Distance

Definition 11.3 *(Cook's Distance)*

Measure how much i^{th} observation changes estimator $\hat{\beta}$

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})}{pRSS/(n-p)}$$

$\hat{\beta}_{(i)}$ - least squares estimator with i^{th} observation removed

Alternatively

$$\begin{aligned} D_i &= \frac{(\hat{Y} - Y_{(i)})^T (\hat{Y} - Y_{(i)})}{pRSS/(n-p)} \quad \hat{Y}_{(i)} = X\hat{\beta}_{(i)} \\ &= r_i^2 \frac{P_{ii}}{(1-P_{ii})r} \quad r_i \text{ standardised residuals}, r = rank(X) \end{aligned}$$

11.4 Under/Overfitting

Definition 11.4

1. Underfitting - necessary predictors left out
2. Overfitting - unnecessary predictors included

11.5 Weighted Least Squares

$cov(Y) = \sigma^2 I_n$ but now we take $cov(Y) = \sigma^2 V$ instead for V symmetric, positive definite.
Transform model s.t $cov(\epsilon) = \sigma^2 I$ to estimate β

V symmetric, positive definite $\implies \exists$ non-singular T s.t $T^T V T = I_n$ $TT^T = V^{-1}$
$\implies \exists$ orthogonal P, diagonal of e.vals of $V; D$ s.t $P^T V P = D$
Take $T = PD^{-1/2}P^T \implies V = PDP^T \implies T^T V T = PD^{-1/2}P^T PDP^T PD^{-1/2}P^T = I_n$
$TT^T = PD^{-1}P^T = V^{-1}$

Take $Z = T^T Y \implies$

$$E(Z) = \underbrace{T^T X}_{=\tilde{X}} \beta \quad cov(Z) = T^T V T \sigma^2 = \sigma^2 I_n$$

$\implies E(Z) = \tilde{X}\beta$ satisfies (SOA)
Assuming (FR);

$$\begin{aligned} \hat{\beta} &= [\tilde{X}^T X]^{-1} \tilde{X}^T Z \\ &= [X^T (TT^T) X]^{-1} X^T (TT^T) Y \\ &= (X^T V^{-1} X)^{-1} X^T V^{-1} Y \end{aligned}$$

$\hat{\beta}$; optimal estimator in sense of Gauss-Markov Theorem.