Statistical Modelling Lecture Notes (2025/2026)

Griffin Reimerink

Contents

1 Introduction

1.1 Distributions related to the normal distribution

Density function of the normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

Chi-squared distribution

If $Z \sim \mathcal{N}(0,1)$, then $Z^2 \sim \chi^2(1)$ is **Chi-squared** distributed with 1 **degree of freedom**.
If $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ are independent, then

$$\sum_{i=1}^{n} \left(\frac{Y_1 - \mu_i}{\sigma_i}\right)^2 \sim \chi^2(n)$$

where n is the number of degrees of freedom.

t-distribution

If $Z \sim \mathcal{N}(0,1)$ is independent to $X \sim \chi^2(n)$, then

$$\frac{Z}{\sqrt{X/n}} \sim t(n)$$

is t-**distributed** with n degrees of freedom.

F-distribution

If $X_1 \sim \chi^2(n_1)$ and $X_2 \sim \chi^2(n_2)$ are independent, then

$$\frac{X_1/n_1}{X_2/n_2} \sim F(n_1, n_2)$$

is F-**distributed** with parameters n_1, n_2.

Definition *Variance matrix*

Let $\boldsymbol{y}$ be a random vector with expectation vector $\boldsymbol{\mu}$. Then the **variance matrix** of $\boldsymbol{y}$ is

$$\mathrm{Var}(\boldsymbol{y}) = \mathbb{E}\left[(\boldsymbol{y} - \mu)(\boldsymbol{y} - \mu)^\top\right]$$

Note that the variance matrix is symmetric and hence the eigenvalues are real.

Multivariate normal distribution

If $Z_1, \ldots, Z_n$ are i.i.d. $\mathcal{N}(0,1)$ random variables, and $\boldsymbol{z}^\top = (Z_1, \ldots, Z_n)$, then

$$f(Z_1, \ldots, Z_n) = \prod_{i=1}^{n} f(z_i) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{z}\right)$$

and $\boldsymbol{z} \sim \mathcal{N}(0, I_n)$ is **multivariate normal distributed**,
where $\mathbb{E}[\boldsymbol{z}] = \boldsymbol{0}$ is the zero vector and $\mathrm{Var}[\boldsymbol{z}] = I_n$ is the identity matrix.

Definition *Covariance matrix*

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be random vectors with expectation vectors $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ respectively.
The **covariance matrix** between $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}\left[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{y} - \boldsymbol{\mu}_y)^\top\right]$$

If $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent, then the covariance matrix is the zero matrix.

1.2 Generating functions

Definition *Generating functions*

Probability generating function:

$$G_Y(t) = \mathbb{E}[t^Y]$$

Moment generating function:

$$M_{\boldsymbol{y}}(\boldsymbol{t}) = \mathbb{E}[\exp(\boldsymbol{t}^\top \boldsymbol{y})]$$

Characteristic function:

$$\varphi_{\boldsymbol{y}}(\boldsymbol{t}) = \mathbb{E}[\exp(i\boldsymbol{t}^\top \boldsymbol{y})]$$

Moment generating property

If the n-th derivative of $M_{\boldsymbol{y}}$ is continuous around zero, then

$$M_Y^{(k)}(0) = \mathbb{E}[Y^k] \qquad \text{for all } k = 0, 1, \dots, n$$

If $\mathbb{E}[Y^n]$ exists, then

$$\mathbb{E}[Y^k] = (-i)^k \varphi_Y^{(k)}(0) \qquad \text{for all } k = 0, 1, \dots, n$$

Properties of the characteristic function

- If $\mathbb{E}[|\boldsymbol{y}|] < \infty$, then $\dot{\varphi}(\boldsymbol{t})$ exists and is continuous, and $\dot{\varphi}(\boldsymbol{0}) = -i\mathbb{E}[\boldsymbol{y}^\top]$
- If $\mathbb{E}[|\boldsymbol{y}|^2] < \infty$, then $\ddot{\varphi}(\boldsymbol{t})$ exists and is continuous, and $\ddot{\varphi}(\boldsymbol{0}) = -\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top]$
- If $\mathbb{P}(\boldsymbol{y} = \boldsymbol{c}) = 1$, then $\varphi_{\boldsymbol{y}}(\boldsymbol{t}) = \exp(i\boldsymbol{t}^\top \boldsymbol{c})$
- If $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, V)$, then $\varphi_{\boldsymbol{y}}(\boldsymbol{t}) = \exp\left(i\boldsymbol{t}^\top \mu - \frac{\boldsymbol{t}^\top V \boldsymbol{t}}{2}\right)$

1.3 Convergence

Definition *Convergence*

Convergence in distribution

$$\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y} \iff \mathbb{P}(\boldsymbol{y}_n \geq \boldsymbol{x}) \to \mathbb{P}(\boldsymbol{y} \leq \boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^p$$

Convergence in probability

$$\boldsymbol{y}_n \xrightarrow{P} \boldsymbol{y} \iff \lim_{n\to\infty} \mathbb{P}(\|\boldsymbol{y}_n - \boldsymbol{y}\| > \varepsilon) \quad \text{for all } \varepsilon > 0$$

Proposition *Some asymptotic properties*

$$\begin{aligned}
\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y} &\implies \boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y} \\
\boldsymbol{y}_n \xrightarrow{P} \boldsymbol{c} &\iff \boldsymbol{y}_n \xrightarrow{D} \boldsymbol{c} \\
\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y},\ d(\boldsymbol{y}_n, \boldsymbol{x}_n) \xrightarrow{P} 0 &\implies \boldsymbol{x}_n \xrightarrow{D} \boldsymbol{y} \\
\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y},\ \boldsymbol{x}_n \xrightarrow{P} \boldsymbol{c} &\implies (\boldsymbol{y}_n, \boldsymbol{x}_n) \xrightarrow{D} (\boldsymbol{y}, \boldsymbol{c}) \\
\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y},\ \boldsymbol{x}_n \xrightarrow{P} \boldsymbol{x} &\implies (\boldsymbol{y}_n, \boldsymbol{x}_n) \xrightarrow{P} (\boldsymbol{y}, \boldsymbol{x})
\end{aligned}$$

Lemma *Slutsky's lemma*

$$\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y},\ \boldsymbol{x}_n \xrightarrow{D} \boldsymbol{c} \implies \boldsymbol{y}_n + \boldsymbol{x}_n \xrightarrow{D} \boldsymbol{y} + \boldsymbol{c}$$
$$\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y},\ \boldsymbol{x}_n \xrightarrow{D} \boldsymbol{c} \implies \boldsymbol{x}_n \boldsymbol{y}_n \xrightarrow{D} \boldsymbol{c}\boldsymbol{y}$$
$$\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y},\ \boldsymbol{x}_n \xrightarrow{D} \boldsymbol{c} \implies \boldsymbol{x}_n^{-1} \boldsymbol{y}_n \xrightarrow{D} \boldsymbol{c}^{-1}\boldsymbol{y}$$

Proposition

$\boldsymbol{y}_n$ converges in distribution to $\boldsymbol{y}$ if and only if $\varphi_{\boldsymbol{y}_n}(\boldsymbol{t})$ converges to $\varphi_{\boldsymbol{y}}(\boldsymbol{t})$ for all $\boldsymbol{t}$.

Theorem *Continuity theorem*

Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $\mathbb{P}(\boldsymbol{y} \in C) = 1$. Then

1.4 Central limit theorem

Theorem *Central limit theorem*

If $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots$ are i.i.d. with mean vector $\boldsymbol{\mu}$ and positive definite variance matrix Σ, then

$$\sqrt{n}(\overline{\boldsymbol{y}_n} - \boldsymbol{\mu}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \Sigma))$$

Theorem *Lindeberg-Feller CLT*

Suppose that

- X_{nj} are independent for $j = 1, \ldots, n$ (per row)
- $\mathbb{E}[X_{nj}] = 0$ (often without loss of generality)
- $\mathrm{Var}(X_{nj}) = \sigma_{nj}$
- $Z_n = \sum_{j=1}^{n} X_{nj}$
- $B_n^2 = \mathrm{Var}(Z_n)$ (monotonically increases with n)

If the **Lindeberg condition** holds:

$$\lim_{n\to\infty} \left[\frac{1}{B_n^2} \sum_{j=1}^{n} \mathbb{E}\left[X_{nj}^2 \mathbb{1}(|X_{nj}| \geq \varepsilon B_n) \right] \right] = 0 \qquad \text{for all } \varepsilon > 0$$

Then

$$\frac{Z_n}{B_n} \xrightarrow{D} \mathcal{N}(0, 1)$$

Conversely, if

$$\lim_{n\to\infty} \frac{1}{B_n^2} \max_{j \leq n} \sigma_{nj} = 0 \quad \text{and} \quad \frac{Z_n}{B_n} \xrightarrow{D} \mathcal{N}(0, 1)$$

then the Lindeberg condition holds.

2 Estimation

2.1 Exponential family

Definition *Exponential family*

A probability distribution $f_\theta(y)$ belongs to the **exponential family** if it can be written as

$$\begin{aligned} f_\theta(y) &= s(y)t(\theta)e^{a(y)b(\theta)} \\ &= \exp[a(y)b(\theta) + c(\theta) + d(y)] \end{aligned}$$

If $a(y) = y$, then the distribution is **canonical**. $b(\theta)$ is the **natural parameter** of the distribution.

Lemma

Univariate distributions in the exponential family are concave, i.e. they have a unique maximum.

Lemma

For a density function $f_\theta(y) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$ in the exponential family, if $b'(\theta) \neq 0$ then

$$\mathbb{E}[a(Y)] = -\frac{c'(\theta)}{b'(\theta)} \qquad \mathrm{Var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{(b'(\theta))^3}$$

Definition *Multivariate likelihood*

$$L(\theta, \boldsymbol{y}) = \prod_{i=1}^{n} f(y_i, \theta)$$

Notation

Log-likelihood:

$$\ell(\theta; \boldsymbol{y}) = \log L(\theta; \boldsymbol{y})$$

Score function:

$$U(\theta; \boldsymbol{y}) = \frac{\mathrm{d}}{\mathrm{d}\theta}\ell(\theta; \boldsymbol{y})$$

Information number:

$$J = \mathrm{Var}[U(\theta; \boldsymbol{y})]$$

Lemma

For a distribution in the exponential family, we have $\mathbb{E}[U] = 0$ and $J > 0$.

Lemma

For a distribution in the exponential family, we have

$$\mathbb{E}[-U'] = \mathrm{Var}[U] = b''(\theta)\frac{c'(\theta)}{b'(\theta)} - c''(\theta) = J$$

2.2 Maximum likelihood estimators

Definition *Maximum likelihood estimator*

The **maximum likelihood estimator** $\hat{\theta}_n$ is the value of θ which maximizes the likelihood.

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\mathrm{argmax}}\, L(\theta, \boldsymbol{y}) = \underset{\theta \in \Theta}{\mathrm{argmax}}\, \ell(\theta, \boldsymbol{y})$$

Theorem *Strong consistency of MLE*

If $\{Y_n\}$ are i.i.d. with density $f(y;\theta)$ and true parameter θ_0 and

- $\theta \in \Theta$ with Θ compact
- $f(y;\theta)$ is continous in θ for all y
- there exists a dominating function $K(y)$ such that $\mathbb{E}_{\theta_0}|K(Y)| < \infty$ and $U(y;\theta) = \log f(y;\theta) - \log f(y;\theta_0) \leq K(y)$ for all y, θ
- for all $\theta \in \Theta$, there exists $\rho > 0$ such that $\sup_{|\theta' - \theta| < \rho} f(y;\theta')$ is measurable in y
- if $f(y;\theta) = f(y;\theta_0)$ almost everywhere, then $\theta = \theta_0$

then for any sequence of ML estimates we have $\{\hat{\theta}_n\} \to \theta_0$:

$$\lim_{n\to\infty} \mathbb{P}(\|\hat{\theta}_n - \theta_0\| \leq \varepsilon) = 1$$

Note: all conditions are necessary.

Theorem *Asymptotic normality of MLE*

If $\{Y_n\}$ are i.i.d with density $f(y;\theta)$ and true parameter θ_0 and

- Θ is an open subset of $\mathbb{R}^p$ and θ_0 is an interior point of the confidence interval around $\hat{\theta}_n$
- 2nd partial derivatives of $f(y;\theta)$ with respect to θ are continuous for all y
- There exists a dominating function $K(y)$ with $\mathbb{E}_{\theta_0}|K(Y)| < \infty$ such that the absolute value of each element of
$$\dot{\boldsymbol{\psi}}(y;\theta) = \frac{\partial^2}{\partial\theta\partial\theta^\top} \log g(y;\theta)$$
is integrable and bounded by $K(y)$ uniformly in a neighborhood of θ_0
- the Fisher information matrix $J(\theta_0) = -\mathbb{E}_{\theta_0}\dot{\boldsymbol{\psi}}(y,\theta)$ is positive definite
- if $f(y;\theta) = f(y;\theta_0)$ almost everywhere, then $\theta = \theta_0$

then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathscr{N}(\mathbf{0}, J(\theta_0)^{-1}$$

Theorem *Cramer-Rao lower bound*

If

- $g(\theta) = \mathbb{E}_\theta[\hat{\theta}(Y)]$
- $\frac{\partial}{\partial\theta} f(y;\theta)$ exists and passes the integral sign in both $\int f(y;\theta) = 1$ and $\int \hat{\theta}(y) f(y;\theta)\,\mathrm{d}y = g(\theta)$
- $0 < J(\theta)$

then

$$\mathrm{Var}[\hat{\theta}(Y)] \geq \frac{g'(\theta)^2}{J(\theta)} \qquad \text{for all } \theta \in \Omega$$

Corollary

If the conditions for asymptotic normality hold, then

$$J(\theta_0)^{1/2}\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I)$$

If $J(\theta)$ is continuous, then

$$J(\hat{\theta}_n) \xrightarrow{\mathbb{P}} J(\theta_0) \qquad J(\hat{\theta}_n)^{-1} \xrightarrow{\mathbb{P}} J(\theta_0)^{-1}$$

Corollary

The maximum likelihood estimator attains the Cramer-Rao lower bound if and only if $\hat{\theta}(y)$ is a sufficient statistic for θ from the exponential family.

Likelihood ratio test

Assume the maximum likelihood estimator exists and is asymptotically normal.
Suppose that $H_0 : \theta_1 = \cdots = \theta_r = 0$ where $1 \leq r \leq k$, and θ_0 satisfies H_0. Then

$$-2\log\left(\frac{\sup_{\Theta_0}\prod_{i=1}^n f(x_i|\boldsymbol{\theta})}{\sup_{\Theta}\prod_{i=1}^n f(x_i|\boldsymbol{\theta})}\right) = -2\log\left(\frac{L(\boldsymbol{\theta}_n^*)}{L(\hat{\boldsymbol{\theta}}_n)}\right) \xrightarrow{d} \chi_r^2$$

3 Linear models

Linear models

Linear models have the form

$$E(Y_i) = \mu_i = \boldsymbol{x}_i^\top \beta \qquad Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

In **generalized models** we have

$$g(\mu_i) = \boldsymbol{x}_i^\top \beta$$

for a non-linear **link function** g which is injective and continuous.

Weibull distribution

The density function of the **Weibull distribution** is

$$f(y,\theta) = \frac{\lambda y^{\lambda-1}}{\theta^\lambda}\exp\left[-\left(\frac{y}{\theta}\right)^\lambda\right]$$

It belongs to the exponential family: $f(y,\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$ with

$$a(y) = y^\lambda \qquad b(\theta) = -\theta^{-\lambda} \qquad c(\theta) = \log(\lambda) - \lambda\log(\theta) \qquad d(y) = (\lambda-1)\log y$$

3.1 The IWLS algorithm

Algorithm *Newton-Raphson algorithm for finding MLE*

We repeatedly apply the following step to find the MLE θ:

$$\theta_{m+1} = \theta_m - \frac{U(\theta_m)}{U'(\theta_m)}$$

Algorithm *Modified Newton-Raphson (Nelder & Wedderburn)*

We repeatedly apply

$$\theta_{m+1} = \theta_m + \frac{U(\theta_m)}{\mathbb{E}[-U'(\theta_m)]} = \theta_m + \frac{U(\theta_m)}{J(\theta_m)}$$

Assuming convergence, we end up with the MLE (the zero of the score function).

$$\theta = \theta + \frac{U(\theta)}{J(\theta)} \implies U(\theta) = 0$$

Preparations for IWLS

Suppose $Y_1, \dots, Y_n$ are independently distributed with densities in the exponential family:

$$f(Y_i, \theta_i) = \exp[a(Y_i)b(\theta_i) + c(\theta_i) + d(Y_i)]$$

Define

$$\mu_i = -\frac{c'(\theta_i)}{b'(\theta_i)} \qquad \eta_i = g(\mu_i) = \boldsymbol{x}_i^T \beta$$

Let W be the $n \times n$ diagonal matrix with elements

$$w_i i = \frac{1}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

and let X be the matrix such that the i-th row is $\boldsymbol{x}_i^T$. Finally, define

$$z_i^{(m)} = \sum_{k=1}^{p} x_{ik} b_k^{(m)} + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i}$$

Proposition

The vector

$$\boldsymbol{b}^{(m+1)} = (X^T W^{(m)} X)^{-1} X^T W^{(m)} \boldsymbol{z}^{(m)}$$

is the global minimizer of

$$\left\| (W^{(m)})^{\frac{1}{2}} \boldsymbol{z}^{(m)} - (W^{(m)})^{\frac{1}{2}} X \boldsymbol{b} \right\|_2^2$$

over any $\boldsymbol{b} \in \mathbb{R}^{p+1}$

Algorithm *IWLS (Iterative Weighted Least Squares)*

1. Start with $\boldsymbol{b}^{(1)}$, possibly rational.
2. Set $m = 1$.
3. Compute $W^{(m)}$ and $W^{(m+1)}$ and $\boldsymbol{b}^{m+1}$.
4. While $m < 100$ and $\|\boldsymbol{b}^{(m)} - \boldsymbol{b}^{(m+1)}\| > 0.0001$, repeat the following steps:
   (a) Compute $W^{(m+1)}$ and $\boldsymbol{z}^{(m+1)}$ from $\boldsymbol{b}^{(m+1)}$
   (b) Compute the update
   $$\boldsymbol{b}^{(m+1)} = (X^T W^{(m+1)} X)^{-1} W^{(m+1)} \boldsymbol{z}^{(m+1)}$$
   (c) $m \leftarrow m + 1$

If the IWLS algorithm generates a converging sequence, then

$$\boldsymbol{b}^{(m)} \to b = \hat{\beta} = \hat{\beta}_{MLE} = (X^T \hat{W} X)^{-1} X^T \hat{W} \hat{\boldsymbol{z}}$$

Proposition

Under regularity assumptions the estimator obtained from IWLS is consistent

$$\hat{\beta} \xrightarrow{P} \beta$$

and asymptotically normal

$$(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, (X^T \hat{W} X)^{-1}\right)$$

3.2 Inference

Proposition

The score function is asymptotically normal:

$$\boldsymbol{U}(\beta) = \boldsymbol{U} \xrightarrow{d} \mathcal{N}(\mathbf{0}, J) \implies \boldsymbol{U}^T J^{-1} \boldsymbol{U} \xrightarrow{d} \chi^2(p)$$

Definition *Asymptotic chi-square test*

We have

$$(\beta_0 - \hat{\beta})^T J(\hat{\beta})(\beta_0 - \hat{\beta}) \xrightarrow{d} \chi^2(p)$$

This is known as the **Wald statistic**.
We reject H_0 by the **asymptotic χ^2 test** if

$$(\beta_0 - \hat{\beta})^T J(\hat{\beta})(\beta_0 - \hat{\beta}) > \chi^2_{p,1-\alpha}$$

Definition *Deviance*

$$D = 2(\ell(\beta) - \ell(\hat{\beta}))$$

4 Normal linear models

Definition *Normal linear model*

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Model equation

Let $x_i^T = (1, x_{i1}, x_{i2}, \cdots, x_{ip})$ be row i of the $n \times p$ **design matrix** X.
Model equation:

$$\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I) \implies \boldsymbol{y} \sim \mathcal{N}(X\beta, \sigma^2 I)$$

Likelihood for the normal linear model

$$f(\boldsymbol{y}, \beta, \sigma^2) = (\sqrt{2\pi\sigma^2})^{-n} \exp\left[-\frac{1}{2\sigma^2}\|\boldsymbol{y} - X\beta\|_2^2\right]$$

The estimates of β and σ^2 are independent, hence we can first estimate β and use it for estimating σ^2.

4.1 Estimation of parameters

Definition *Residual sum of squares*

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N}(y_i - x_i^T\beta)^2 = (\boldsymbol{y} - X\beta)^T(\boldsymbol{y} - X\beta) = \boldsymbol{y}^T\boldsymbol{y} - 2\beta^T X^T \boldsymbol{y} + \beta^T X^T X\beta$$

Estimator of coefficients

The MLE of β is equal to the argmin of the residual sum of squares. By convexity,

$$\hat{\beta} = (X^T X)^{-1} X^T \boldsymbol{y}$$

globally minimizes the RSS function.

Estimator of the error variance

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-p}\|\underline{y} - X\hat{\beta}\|^2$$

We obtain the estimator of the residual standard error $\hat{\sigma}$ by taking the square root.
If $\hat{\sigma}$ is smaller then there is less error in $\boldsymbol{y}$ and hence greater estimation precision.

Definition R^2

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(\bar{y} - y_i)^2} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Generally, R^2 is squared correlation between prediction $\hat{\boldsymbol{y}} = X\hat{\beta}$ and response $\boldsymbol{y}$.

Definition *Prediction interval*

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0))} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0))}$$

Remark: the prediction interval is larger than a confidence interval, and prediction always involves some variation due to the term $\hat{\sigma}^2$

4.2 Detecting influential observations

Definition *Marginal testing parameters*

C_{jj} is element jj of $(X^TX)^{-1}$, and we define the **standard error** as

$$\text{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$$

Definition *Hat matrix*

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

The hat matrix projects outcomes $\boldsymbol{y}$ onto space spanned by the columns of predictor matrix X.

Predicted outcomes

The predicted outcomes are

$$\hat{\boldsymbol{y}} = X\hat{\beta} = C(X^TX)^{-1}X^T\boldsymbol{y} = H\boldsymbol{y}$$

Variance of predicted outcomes:

$$\text{Var}(\hat{\boldsymbol{y}} = \sigma^2 H$$

Variance of residuals:

$$\text{Var}(\boldsymbol{y} - \hat{\boldsymbol{y}} = \sigma^2(I - H)$$

Leverages

The **leverage** is the i-th diagonal value h_{ii} of H, which is in the interval $[0, 1]$ for all i.
The sum of leverages equals p, and the observation i is **influential** if

$$h_{ii} > 2 \cdot \bar{h} = \frac{2p}{n}$$

4.3 Residuals

Definition *Standardized residuals*

Size of residuals: $\hat{\varepsilon}_i = y_i - \hat{y}_i$
Standardized residuals with mean zero and approximate unit variance:

$$d_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}^2}}$$

An observation is an **outlier** if $|d_i| > 3$.

Definition *Studentized residuals*

Adapting for high leverage gives the **Studentized residuals**:

$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}^2 \cdot (1 - h_{ii})}}$$

Definition *Externally Studentized residual*

The **externally Studentized residual**

$$t_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}^2_{(i)} \cdot (1 - h_{ii})}} \sim t_{n-p-1}$$

is the basis for a statistical test with $H_0 : \hat{\varepsilon}_i$ is not an outlier

Definition *DFBETAS*

Let $\hat{\beta}_{j(i)}$ be the coefficient $\hat{\beta}_j$ computed without observation i. Then

$$\mathrm{DFBETAS}_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{\hat{\sigma}^2_{(i)} \cdot C_{jj}}}$$

Definition *Cook's distance*

$$C_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T X^T X (\hat{\beta} - \hat{\beta}_{(i)}}{p\hat{\sigma}^2}$$

5 Survival analysis

Definition *Survival function*

Let the random variable $Y \geq 0$ be the survival time.
The **survival function** gives probability of survival beyond y, that is

$$S(Y) = \mathbb{P}(Y \geq y) = 1 - F(y)$$

Definition *Hazard function*

The **hazard function** h is the probability of death in $[y, y + \delta_y]$ given survival up to y relatively to an infinitely small interval

$$h(y) = \lim_{\delta_y \to 0} \frac{\mathbb{P}(Y \in [y, y + \delta_y] \mid Y \geq y}{\delta_y} = -\frac{\mathrm{d}}{\mathrm{d}y} \log(S(y))$$

Definition *Cumulative hazard function*

For the **cumulative hazard function** we have

$$H(y) = \int_0^y h(t)\,\mathrm{d}t = -\log(1 - F(y))$$

The **median** $y_{0.50}$ is the solution of

$$\frac{1}{2} = \mathbb{P}(Y \leq y) = F(y)$$

Lack of memory

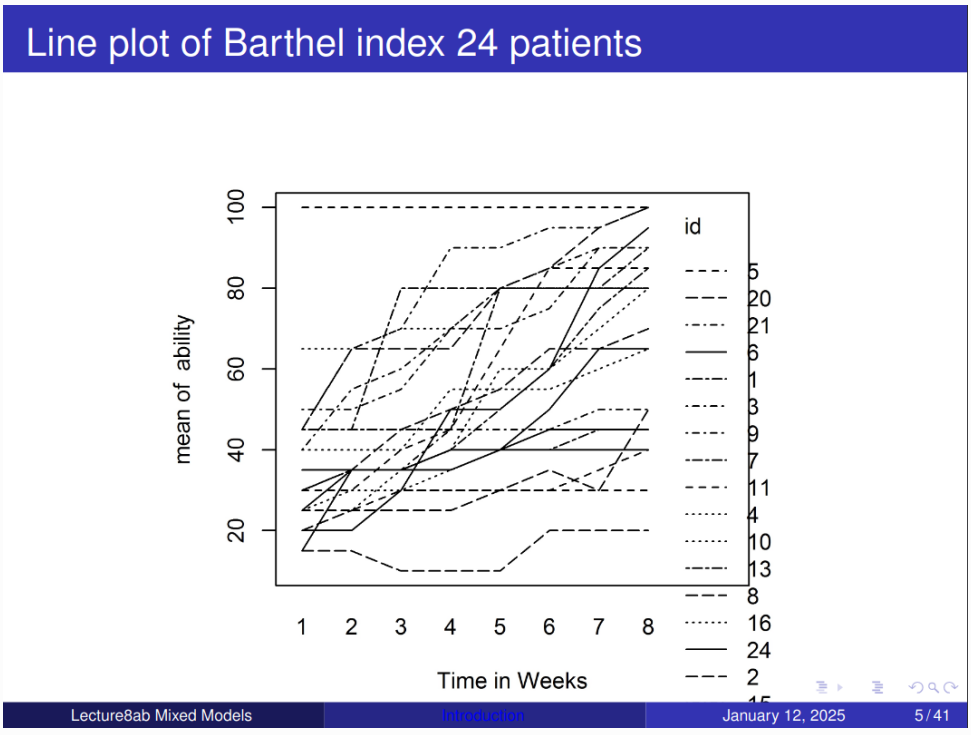
If the system lacks memory of survival beyond x, then

$$\mathbb{P}(X > x + y \mid X > x) = \mathbb{P}(X > y)$$

The concept of lack of memory leads to the exponential distribution.

$$F(x) = (1 - e^{-\theta x})\mathbf{1}_{[0,\infty)}$$

A work of art



Index