Part IB — Statistics

Theorems with proof

Based on lectures by D. Spiegelhalter

Notes taken by Dexter Chua

Lent 2015

Estimation
Review of distribution and density functions, parametric families. Examples: binomial, Poisson, gamma. Sufficiency, minimal sufficiency, the Rao-Blackwell theorem. Maximum likelihood estimation. Confidence intervals. Use of prior distributions and Bayesian inference. [5]

Hypothesis testing
Simple examples of hypothesis testing, null and alternative hypothesis, critical region, size, power, type I and type II errors, Neyman-Pearson lemma. Significance level of outcome. Uniformly most powerful tests. Likelihood ratio, and use of generalised likelihood ratio to construct test statistics for composite hypotheses. Examples, including t-tests and F-tests. Relationship with confidence intervals. Goodness-of-fit tests and contingency tables. [4]

Linear models
Derivation and joint distribution of maximum likelihood estimators, least squares, Gauss-Markov theorem. Testing hypotheses, geometric interpretation. Examples, including simple linear regression and one-way analysis of variance. Use of software. [7]

Contents

0 Introduction **3**

1 Estimation **4**
1.1 Estimators . . . 4
1.2 Mean squared error . . . 4
1.3 Sufficiency . . . 4
1.4 Likelihood . . . 5
1.5 Confidence intervals . . . 5
1.6 Bayesian estimation . . . 5

2 Hypothesis testing **6**
2.1 Simple hypotheses . . . 6
2.2 Composite hypotheses . . . 7
2.3 Tests of goodness-of-fit and independence . . . 7
2.3.1 Goodness-of-fit of a fully-specified null distribution . . . 7
2.3.2 Pearson's chi-squared test . . . 7
2.3.3 Testing independence in contingency tables . . . 7
2.4 Tests of homogeneity, and connections to confidence intervals . . . 7
2.4.1 Tests of homogeneity . . . 7
2.4.2 Confidence intervals and hypothesis tests . . . 7
2.5 Multivariate normal theory . . . 8
2.5.1 Multivariate normal distribution . . . 8
2.5.2 Normal random samples . . . 9
2.6 Student's t-distribution . . . 10

3 Linear models **11**
3.1 Linear models . . . 11
3.2 Simple linear regression . . . 11
3.3 Linear models with normal assumptions . . . 12
3.4 The F distribution . . . 14
3.5 Inference for $\boldsymbol{\beta}$ . . . 14
3.6 Simple linear regression . . . 14
3.7 Expected response at $\mathbf{x}^*$ . . . 14
3.8 Hypothesis testing . . . 14
3.8.1 Hypothesis testing . . . 14
3.8.2 Simple linear regression . . . 14
3.8.3 One way analysis of variance with equal numbers in each group . . . 14

0 Introduction

1 Estimation

1.1 Estimators

1.2 Mean squared error

1.3 Sufficiency

Theorem (The factorization criterion)**.** T is sufficient for θ if and only if

$$f_{\mathbf{X}}(\mathbf{x} \mid \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$$

for some functions g and h.

Proof. We first prove the discrete case.

Suppose $f_{\mathbf{X}}(\mathbf{x} \mid \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$. If $T(\mathbf{x}) = t$, then

$$\begin{aligned} f_{\mathbf{X}|T=t}(\mathbf{x}) &= \frac{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t)}{\mathbb{P}_\theta(T = t)} \\ &= \frac{g(T(\mathbf{x}), \theta)h(\mathbf{x})}{\sum_{\{\mathbf{y}:T(\mathbf{y})=t\}} g(T(\mathbf{y}), \theta)h(\mathbf{y})} \\ &= \frac{g(t, \theta)h(\mathbf{x})}{g(t, \theta)\sum h(\mathbf{y})} \\ &= \frac{h(\mathbf{x})}{\sum h(\mathbf{y})} \end{aligned}$$

which doesn't depend on θ. So T is sufficient.

The continuous case is similar. If $f_{\mathbf{X}}(\mathbf{x} \mid \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$, and $T(\mathbf{x}) = t$, then

$$\begin{aligned} f_{\mathbf{X}|T=t}(\mathbf{x}) &= \frac{g(T(\mathbf{x}), \theta)h(\mathbf{x})}{\int_{\mathbf{y}:T(\mathbf{y})=t} g(T(\mathbf{y}), \theta)h(\mathbf{y})\, \mathrm{d}\mathbf{y}} \\ &= \frac{g(t, \theta)h(\mathbf{x})}{g(t, \theta)\int h(\mathbf{y})\, \mathrm{d}\mathbf{y}} \\ &= \frac{h(\mathbf{x})}{\int h(\mathbf{y})\, \mathrm{d}\mathbf{y}}, \end{aligned}$$

which does not depend on θ.

Now suppose T is sufficient so that the conditional distribution of $\mathbf{X} \mid T = t$ does not depend on θ. Then

$$\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}, T = T(\mathbf{x})) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x} \mid T = T(\mathbf{x}))\mathbb{P}_\theta(T = T(\mathbf{x})).$$

The first factor does not depend on θ by assumption; call it $h(\mathbf{x})$. Let the second factor be $g(t, \theta)$, and so we have the required factorisation. □

Theorem. Suppose $T = T(\mathbf{X})$ is a statistic that satisfies

$$\frac{f_{\mathbf{X}}(\mathbf{x}; \theta)}{f_{\mathbf{X}}(\mathbf{y}; \theta)} \text{ does not depend on } \theta \text{ if and only if } T(\mathbf{x}) = T(\mathbf{y}).$$

Then T is minimal sufficient for θ.

Proof. First we have to show sufficiency. We will use the factorization criterion to do so.

Firstly, for each possible t, pick a favorite $\mathbf{x}_t$ such that $T(\mathbf{x}_t) = t$.

Now let $\mathbf{x} \in \mathcal{X}^N$ and let $T(\mathbf{x}) = t$. So $T(\mathbf{x}) = T(\mathbf{x}_t)$. By the hypothesis, $\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{x}_t;\theta)}$ does not depend on θ. Let this be $h(\mathbf{x})$. Let $g(t,\theta) = f_{\mathbf{X}}(\mathbf{x}_t,\theta)$. Then

$$f_{\mathbf{X}}(\mathbf{x};\theta) = f_{\mathbf{X}}(\mathbf{x}_t;\theta)\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{x}_t;\theta)} = g(t,\theta)h(\mathbf{x}).$$

So T is sufficient for θ.

To show that this is minimal, suppose that $S(\mathbf{X})$ is also sufficient. By the factorization criterion, there exist functions g_S and h_S such that

$$f_{\mathbf{X}}(\mathbf{x};\theta) = g_S(S(\mathbf{x}),\theta)h_S(\mathbf{x}).$$

Now suppose that $S(\mathbf{x}) = S(\mathbf{y})$. Then

$$\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{y};\theta)} = \frac{g_S(S(\mathbf{x}),\theta)h_S(\mathbf{x})}{g_S(S(\mathbf{y}),\theta)h_S(\mathbf{y})} = \frac{h_S(\mathbf{x})}{h_S(\mathbf{y})}.$$

This means that the ratio $\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{y};\theta)}$ does not depend on θ. By the hypothesis, this implies that $T(\mathbf{x}) = T(\mathbf{y})$. So we know that $S(\mathbf{x}) = S(\mathbf{y})$ implies $T(\mathbf{x}) = T(\mathbf{y})$. So T is a function of S. So T is minimal sufficient. □

Theorem (Rao-Blackwell Theorem)**.** Let T be a sufficient statistic for θ and let $\tilde{\theta}$ be an estimator for θ with $\mathbb{E}(\tilde{\theta}^2) < \infty$ for all θ. Let $\hat{\theta}(\mathbf{x}) = \mathbb{E}[\tilde{\theta}(\mathbf{X}) \mid T(\mathbf{X}) = T(\mathbf{x})]$. Then for all θ,

$$\mathbb{E}[(\hat{\theta} - \theta)^2] \leq \mathbb{E}[(\tilde{\theta} - \theta)^2].$$

The inequality is strict unless $\tilde{\theta}$ is a function of T.

Proof. By the conditional expectation formula, we have $\mathbb{E}(\hat{\theta}) = \mathbb{E}[\mathbb{E}(\tilde{\theta} \mid T)] = E(\tilde{\theta})$. So they have the same bias.

By the conditional variance formula,

$$\mathrm{var}(\tilde{\theta}) = \mathbb{E}[\mathrm{var}(\tilde{\theta} \mid T)] + \mathrm{var}[\mathbb{E}(\tilde{\theta} \mid T)] = \mathbb{E}[\mathrm{var}(\tilde{\theta} \mid T)] + \mathrm{var}(\hat{\theta}).$$

Hence $\mathrm{var}(\tilde{\theta}) \geq \mathrm{var}(\hat{\theta})$. So $\mathrm{mse}(\tilde{\theta}) \geq \mathrm{mse}(\hat{\theta})$, with equality only if $\mathrm{var}(\tilde{\theta} \mid T) = 0$. □

1.4 Likelihood

1.5 Confidence intervals

1.6 Bayesian estimation

2 Hypothesis testing

2.1 Simple hypotheses

Lemma (Neyman-Pearson lemma)**.** Suppose $H_0 : f = f_0$, $H_1 : f = f_1$, where f_0 and f_1 are continuous densities that are nonzero on the same regions. Then among all tests of size less than or equal to α, the test with the largest power is the likelihood ratio test of size α.

Proof. Under the likelihood ratio test, our critical region is

$$C = \left\{\mathbf{x} : \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} > k\right\},$$

where k is chosen such that $\alpha = \mathbb{P}(\text{reject } H_0 \mid H_0) = \mathbb{P}(\mathbf{X} \in C \mid H_0) = \int_C f_0(\mathbf{x})\, \mathrm{d}\mathbf{x}$. The probability of Type II error is given by

$$\beta = \mathbb{P}(\mathbf{X} \notin C \mid f_1) = \int_{\bar{C}} f_1(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

Let C^* be the critical region of any other test with size less than or equal to α. Let $\alpha^* = \mathbb{P}(\mathbf{X} \in C^* \mid H_0)$ and $\beta^* = \mathbb{P}(\mathbf{X} \notin C^* \mid H_1)$. We want to show $\beta \leq \beta^*$.

We know $\alpha^* \leq \alpha$, i.e

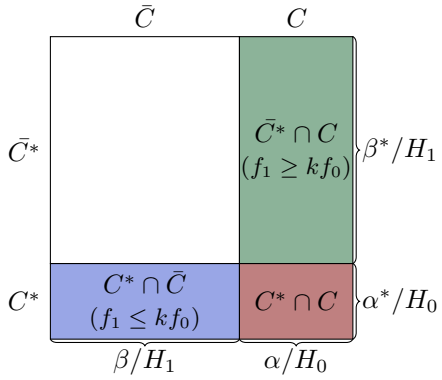
$$\int_{C^*} f_0(\mathbf{x})\, \mathrm{d}\mathbf{x} \leq \int_C f_0(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

Also, on C, we have $f_1(\mathbf{x}) > k f_0(\mathbf{x})$, while on $\bar{C}$ we have $f_1(\mathbf{x}) \leq k f_0(\mathbf{x})$. So

$$\begin{aligned}
\int_{\bar{C}^* \cap C} f_1(\mathbf{x})\, \mathrm{d}\mathbf{x} &\geq k \int_{\bar{C}^* \cap C} f_0(\mathbf{x})\, \mathrm{d}\mathbf{x} \\
\int_{\bar{C} \cap C^*} f_1(\mathbf{x})\, \mathrm{d}\mathbf{x} &\leq k \int_{\bar{C} \cap C^*} f_0(\mathbf{x})\, \mathrm{d}\mathbf{x}.
\end{aligned}$$

Hence

$$\begin{aligned}
\beta - \beta^* &= \int_{\bar{C}} f_1(\mathbf{x})\, \mathrm{d}\mathbf{x} - \int_{\bar{C}^*} f_1(\mathbf{x})\, \mathrm{d}\mathbf{x} \\
&= \int_{\bar{C} \cap C^*} f_1(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int_{\bar{C} \cap \bar{C}^*} f_1(\mathbf{x})\, \mathrm{d}\mathbf{x} \\
&\quad - \int_{\bar{C}^* \cap C} f_1(\mathbf{x})\, \mathrm{d}\mathbf{x} - \int_{\bar{C} \cap \bar{C}^*} f_1(\mathbf{x})\, \mathrm{d}\mathbf{x} \\
&= \int_{\bar{C} \cap C^*} f_1(\mathbf{x})\, \mathrm{d}\mathbf{x} - \int_{\bar{C}^* \cap C} f_1(\mathbf{x})\, \mathrm{d}\mathbf{x} \\
&\leq k \int_{\bar{C} \cap C^*} f_0(\mathbf{x})\, \mathrm{d}\mathbf{x} - k \int_{\bar{C}^* \cap C} f_0(\mathbf{x})\, \mathrm{d}\mathbf{x} \\
&= k \left\{ \int_{\bar{C} \cap C^*} f_0(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int_{C \cap C^*} f_0(\mathbf{x})\, \mathrm{d}\mathbf{x} \right\} \\
&\quad - k \left\{ \int_{\bar{C}^* \cap C} f_0(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int_{C \cap C^*} f_0(\mathbf{x})\, \mathrm{d}\mathbf{x} \right\} \\
&= k(\alpha^* - \alpha) \\
&\leq 0.
\end{aligned}$$



□

2.2 Composite hypotheses

Theorem (Generalized likelihood ratio theorem)**.** Suppose $\Theta_0 \subseteq \Theta_1$ and $|\Theta_1| - |\Theta_0| = p$. Let $\mathbf{X} = (X_1, \cdots, X_n)$ with all X_i iid. If H_0 is true, then as $n \to \infty$,

$$2 \log \Lambda_{\mathbf{X}}(H_0; H_1) \sim \chi^2_p.$$

If H_0 is not true, then $2 \log \Lambda$ tends to be larger. We reject H_0 if $2 \log \Lambda > c$, where $c = \chi^2_p(\alpha)$ for a test of approximately size α.

2.3 Tests of goodness-of-fit and independence

2.3.1 Goodness-of-fit of a fully-specified null distribution

2.3.2 Pearson's chi-squared test

2.3.3 Testing independence in contingency tables

2.4 Tests of homogeneity, and connections to confidence intervals

2.4.1 Tests of homogeneity

2.4.2 Confidence intervals and hypothesis tests

Theorem (Duality of hypothesis tests and confidence intervals)**.** Suppose $X_1, \cdots, X_n$ have joint pdf $f_{\mathbf{X}}(\mathbf{x} \mid \theta)$ for $\theta \in \Theta$.

(i) Suppose that for every $\theta_0 \in \Theta$ there is a size α test of $H_0 : \theta = \theta_0$. Denote the acceptance region by $A(\theta_0)$. Then the set $I(\mathbf{X}) = \{\theta : \mathbf{X} \in A(\theta)\}$ is a $100(1-\alpha)\%$ confidence set for θ.

(ii) Suppose $I(\mathbf{X})$ is a $100(1-\alpha)\%$ confidence set for θ. Then $A(\theta_0) = \{\mathbf{X} : \theta_0 \in I(\mathbf{X})\}$ is an acceptance region for a size α test of $H_0 : \theta = \theta_0$.

Proof. First note that $\theta_0 \in I(\mathbf{X})$ iff $\mathbf{X} \in A(\theta_0)$.

For (i), since the test is size α, we have

$$\mathbb{P}(\text{accept } H_0 \mid H_0 \text{ is true}) = \mathbb{P}(\mathbf{X} \in A(\theta_0) \mid \theta = \theta_0) = 1 - \alpha.$$

And so

$$\mathbb{P}(\theta_0 \in I(\mathbf{X}) \mid \theta = \theta_0) = \mathbb{P}(\mathbf{X} \in A(\theta_0) \mid \theta = \theta_0) = 1 - \alpha.$$

For (ii), since $I(\mathbf{X})$ is a $100(1-\alpha)\%$ confidence set, we have

$$P(\theta_0 \in I(\mathbf{X}) \mid \theta = \theta_0) = 1 - \alpha.$$

So

$$\mathbb{P}(\mathbf{X} \in A(\theta_0) \mid \theta = \theta_0) = \mathbb{P}(\theta \in I(\mathbf{X}) \mid \theta = \theta_0) = 1 - \alpha. \qquad \square$$

2.5 Multivariate normal theory

2.5.1 Multivariate normal distribution

Proposition.

(i) If $\mathbf{X} \sim N_n(\mu, \Sigma)$, and A is an $m \times n$ matrix, then $A\mathbf{X} \sim N_m(A\boldsymbol{\mu}, A\Sigma A^T)$.

(ii) If $\mathbf{X} \sim N_n(\mathbf{0}, \sigma^2 I)$, then

$$\frac{|\mathbf{X}|^2}{\sigma^2} = \frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} = \sum \frac{X_i^2}{\sigma^2} \sim \chi_n^2.$$

Instead of writing $|\mathbf{X}|^2/\sigma^2 \sim \chi_n^2$, we often just say $|\mathbf{X}|^2 \sim \sigma^2\chi_n^2$.

Proof.

(i) See example sheet 3.

(ii) Immediate from definition of χ_n^2. $\square$

Proposition. Let $\mathbf{X} \sim N_n(\mu, \Sigma)$. We split $\mathbf{X}$ up into two parts: $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$, where $\mathbf{X}_i$ is a $n_i \times 1$ column vector and $n_1 + n_2 = n$.

Similarly write

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where Σ_{ij} is an $n_i \times n_j$ matrix.

Then

(i) $\mathbf{X}_i \sim N_{n_i}(\boldsymbol{\mu}_i, \Sigma_{ii})$

(ii) $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent iff $\Sigma_{12} = 0$.

Proof.

(i) See example sheet 3.

(ii) Note that by symmetry of Σ, $\Sigma_{12} = 0$ if and only if $\Sigma_{21} = 0$.

From (†), $M_{\mathbf{X}}(\mathbf{t}) = \exp(\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t})$ for each $\mathbf{t} \in \mathbb{R}^n$. We write $\mathbf{t} = \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \end{pmatrix}$.

Then the mgf is equal to

$$M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\mathbf{t}_1^T\boldsymbol{\mu}_1 + \mathbf{t}_2^T\boldsymbol{\mu}_2 + \mathbf{t}_1^T\Sigma_{11}\mathbf{t}_1 + \frac{1}{2}\mathbf{t}_2^T\Sigma_{22}\mathbf{t}_2 + \frac{1}{2}\mathbf{t}_1^T\Sigma_{12}\mathbf{t}_2 + \frac{1}{2}\mathbf{t}_2^T\Sigma_{21}\mathbf{t}_1\right).$$

From (i), we know that $M_{\mathbf{X}_i}(\mathbf{t}_i) = \exp(\mathbf{t}_i^T\boldsymbol{\mu}_i + \frac{1}{2}\mathbf{t}_i^T\Sigma_{ii}\mathbf{t}_i)$. So $M_{\mathbf{X}}(\mathbf{t}) = M_{\mathbf{X}_1}(\mathbf{t}_1)M_{\mathbf{X}_2}(\mathbf{t}_2)$ for all $\mathbf{t}$ if and only if $\Sigma_{12} = 0$. $\square$

Proposition. When Σ is a positive definite, then $\mathbf{X}$ has pdf

$$f_{\mathbf{X}}(\mathbf{x};\mu,\Sigma) = \frac{1}{|\Sigma|^2}\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right].$$

2.5.2 Normal random samples

Theorem (Joint distribution of $\bar{X}$ and S_{XX})**.** Suppose $X_1,\cdots,X_n$ are iid $N(\mu,\sigma^2)$ and $\bar{X} = \frac{1}{n}\sum X_i$, and $S_{XX} = \sum(X_i-\bar{X})^2$. Then

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

(ii) $S_{XX}/\sigma^2 \sim \chi^2_{n-1}$.

(iii) $\bar{X}$ and S_{XX} are independent.

Proof. We can write the joint density as $\mathbf{X} \sim N_n(\boldsymbol{\mu},\sigma^2 I)$, where $\boldsymbol{\mu} = (\mu,\mu,\cdots,\mu)$.

Let A be an $n\times n$ orthogonal matrix with the first row all $1/\sqrt{n}$ (the other rows are not important). One possible such matrix is

$$A = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2\times 1}} & \frac{-1}{\sqrt{2\times 1}} & 0 & 0 & \cdots & 0 \\ \frac{1}{\sqrt{3\times 2}} & \frac{1}{\sqrt{3\times 2}} & \frac{-2}{\sqrt{3\times 2}} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{-(n-1)}{\sqrt{n(n-1)}} \end{pmatrix}$$

Now define $\mathbf{Y} = A\mathbf{X}$. Then

$$\mathbf{Y} \sim N_n(A\boldsymbol{\mu}, A\sigma^2 I A^T) = N_n(A\boldsymbol{\mu},\sigma^2 I).$$

We have

$$A\boldsymbol{\mu} = (\sqrt{n}\mu, 0, \cdots, 0)^T.$$

So $Y_1 \sim N(\sqrt{n}\mu,\sigma^2)$ and $Y_i \sim N(0,\sigma^2)$ for $i = 2,\cdots,n$. Also, $Y_1,\cdots,Y_n$ are independent, since the covariance matrix is every non-diagonal term 0.

But from the definition of A, we have

$$Y_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i = \sqrt{n}\bar{X}.$$

So $\sqrt{n}\bar{X} \sim N(\sqrt{n}\mu,\sigma^2)$, or $\bar{X} \sim N(\mu,\sigma^2/n)$. Also

$$\begin{aligned} Y_2^2 + \cdots + Y_n^2 &= \mathbf{Y}^T\mathbf{Y} - Y_1^2 \\ &= \mathbf{X}^T A^T A\mathbf{X} - Y_1^2 \\ &= \mathbf{X}^T\mathbf{X} - n\bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - n\bar{X}^2 \\ &= \sum_{i=1}^n (X_i-\bar{X})^2 \\ &= S_{XX}. \end{aligned}$$

So $S_{XX} = Y_2^2 + \cdots + Y_n^2 \sim \sigma^2\chi^2_{n-1}$.

Finally, since Y_1 and $Y_2,\cdots,Y_n$ are independent, so are $\bar{X}$ and S_{XX}. □

2.6 Student's t-distribution

Proposition. If $k > 1$, then $\mathbb{E}_k(T) = 0$.

If $k > 2$, then $\mathrm{var}_k(T) = \frac{k}{k-2}$.

If $k = 2$, then $\mathrm{var}_k(T) = \infty$.

In all other cases, the values are undefined. In particular, the $k = 1$ case has undefined mean and variance. This is known as the *Cauchy distribution*.

3 Linear models

3.1 Linear models

Proposition. The least squares estimator satisfies

$$X^T X\hat{\boldsymbol{\beta}} = X^T\mathbf{Y}. \tag{3}$$

3.2 Simple linear regression

Theorem (Gauss Markov theorem)**.** In a full rank linear model, let $\hat{\boldsymbol{\beta}}$ be the least squares estimator of $\boldsymbol{\beta}$ and let $\boldsymbol{\beta}^*$ be any other unbiased estimator for $\boldsymbol{\beta}$ which is linear in the Y_i's. Then

$$\operatorname{var}(\mathbf{t}^T\hat{\boldsymbol{\beta}}) \leq \operatorname{var}(\mathbf{t}^T\boldsymbol{\beta}^*).$$

for all $\mathbf{t} \in \mathbb{R}^p$. We say that $\hat{\boldsymbol{\beta}}$ is the *best linear unbiased estimator* of $\boldsymbol{\beta}$ (BLUE).

Proof. Since $\boldsymbol{\beta}^*$ is linear in the Y_i's, $\boldsymbol{\beta}^* = A\mathbf{Y}$ for some $p \times n$ matrix A.

Since $\boldsymbol{\beta}^*$ is an unbiased estimator, we must have $\mathbb{E}[\boldsymbol{\beta}^*] = \boldsymbol{\beta}$. However, since $\boldsymbol{\beta}^* = A\mathbf{Y}$, $\mathbb{E}[\boldsymbol{\beta}^*] = A\mathbb{E}[\mathbf{Y}] = AX\boldsymbol{\beta}$. So we must have $\boldsymbol{\beta} = AX\boldsymbol{\beta}$. Since this holds for any $\boldsymbol{\beta}$, we must have $AX = I_p$. Now

$$\begin{aligned}
\operatorname{cov}(\boldsymbol{\beta}^*) &= \mathbb{E}[(\boldsymbol{\beta}^* - \boldsymbol{\beta})(\boldsymbol{\beta}^* - \boldsymbol{\beta})^T] \\
&= \mathbb{E}[(A\mathbf{Y} - \boldsymbol{\beta})(A\mathbf{Y} - \boldsymbol{\beta})^T] \\
&= \mathbb{E}[(AX\boldsymbol{\beta} + A\boldsymbol{\varepsilon} - \boldsymbol{\beta})(AX\boldsymbol{\beta} + A\boldsymbol{\varepsilon} - \boldsymbol{\beta})^T]
\end{aligned}$$

Since $AX\boldsymbol{\beta} = \boldsymbol{\beta}$, this is equal to

$$\begin{aligned}
&= \mathbb{E}[A\boldsymbol{\varepsilon}(A\boldsymbol{\varepsilon})^T] \\
&= A(\sigma^2 I)A^T \\
&= \sigma^2 AA^T.
\end{aligned}$$

Now let $\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} = (A - (X^TX)^{-1}X^T)\mathbf{Y} = B\mathbf{Y}$, for some B. Then

$$BX = AX - (X^TX)^{-1}X^TX = I_p - I_p = 0.$$

By definition, we have $A\mathbf{Y} = B\mathbf{Y} + (X^TX)^{-1}X^T\mathbf{Y}$, and this is true for all $\mathbf{Y}$. So $A = B + (X^TX)^{-1}X^T$. Hence

$$\begin{aligned}
\operatorname{cov}(\boldsymbol{\beta}^*) &= \sigma^2 AA^T \\
&= \sigma^2(B + (X^TX)^{-1}X^T)(B + (X^TX)^{-1}X^T)^T \\
&= \sigma^2(BB^T + (X^TX)^{-1}) \\
&= \sigma^2 BB^T + \operatorname{cov}(\hat{\boldsymbol{\beta}}).
\end{aligned}$$

Note that in the second line, the cross-terms disappear since $BX = 0$.

So for any $\mathbf{t} \in \mathbb{R}^p$, we have

$$\begin{aligned}
\operatorname{var}(\mathbf{t}^T\boldsymbol{\beta}^*) &= \mathbf{t}^T \operatorname{cov}(\boldsymbol{\beta}^*)\mathbf{t} \\
&= \mathbf{t}^T \operatorname{cov}(\hat{\boldsymbol{\beta}})\mathbf{t} + \mathbf{t}^T BB^T\mathbf{t}\sigma^2 \\
&= \operatorname{var}(\mathbf{t}^T\hat{\boldsymbol{\beta}}) + \sigma^2\|B^T\mathbf{t}\|^2 \\
&\geq \operatorname{var}(\mathbf{t}^T\hat{\boldsymbol{\beta}}).
\end{aligned}$$

Taking $\mathbf{t} = (0, \cdots, 1, 0, \cdots, 0)^T$ with a 1 in the ith position, we have

$$\text{var}(\hat{\beta}_i) \leq \text{var}(\beta_i^*). \qquad \square$$

3.3 Linear models with normal assumptions

Proposition. Under normal assumptions the maximum likelihood estimator for a linear model is

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y},$$

which is the same as the least squares estimator.

Lemma.

(i) If $\mathbf{Z} \sim N_n(\mathbf{0}, \sigma^2 I)$ and A is $n \times n$, symmetric, idempotent with rank r, then $\mathbf{Z}^T A \mathbf{Z} \sim \sigma^2 \chi_r^2$.

(ii) For a symmetric idempotent matrix A, $\text{rank}(A) = \text{tr}(A)$.

Proof.

(i) Since A is idempotent, $A^2 = A$ by definition. So eigenvalues of A are either 0 or 1 (since $\lambda \mathbf{x} = A\mathbf{x} = A^2\mathbf{x} = \lambda^2 \mathbf{x}$).

Since A is also symmetric, it is diagonalizable. So there exists an orthogonal Q such that

$$\Lambda = Q^T A Q = \text{diag}(\lambda_1, \cdots, \lambda_n) = \text{diag}(1, \cdots, 1, 0, \cdots, 0)$$

with r copies of 1 and $n - r$ copies of 0.

Let $\mathbf{W} = Q^T \mathbf{Z}$. So $\mathbf{Z} = Q\mathbf{W}$. Then $\mathbf{W} \sim N_n(\mathbf{0}, \sigma^2 I)$, since $\text{cov}(\mathbf{W}) = Q^T \sigma^2 I Q = \sigma^2 I$. Then

$$\mathbf{Z}^T A \mathbf{Z} = \mathbf{W}^T Q^T A Q \mathbf{W} = \mathbf{W}^T \Lambda \mathbf{W} = \sum_{i=1}^{r} w_i^2 \sim \chi_r^2.$$

(ii)

$$\begin{aligned} \text{rank}(A) &= \text{rank}(\Lambda) \\ &= \text{tr}(\Lambda) \\ &= \text{tr}(Q^T A Q) \\ &= \text{tr}(A Q^T Q) \\ &= \text{tr}\, A \end{aligned} \qquad \square$$

Theorem. For the normal linear model $\mathbf{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I)$,

(i) $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$

(ii) $\text{RSS} \sim \sigma^2 \chi_{n-p}^2$, and so $\hat{\sigma}^2 \sim \frac{\sigma^2}{n} \chi_{n-p}^2$.

(iii) $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent.

Proof.

– We have $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{Y}$. Call this $C\mathbf{Y}$ for later use. Then $\hat{\boldsymbol{\beta}}$ has a normal distribution with mean

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

and covariance

$$(X^TX)^{-1}X^T(\sigma^2 I)[(X^TX)^{-1}X^T]^T = \sigma^2(X^TX)^{-1}.$$

So

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1}).$$

– Our previous lemma says that $\mathbf{Z}^TA\mathbf{Z} \sim \sigma^2\chi^2_r$. So we pick our $\mathbf{Z}$ and A so that $\mathbf{Z}^TA\mathbf{Z} = \text{RSS}$, and r, the degrees of freedom of A, is $n - p$.

Let $\mathbf{Z} = \mathbf{Y} - X\boldsymbol{\beta}$ and $A = (I_n - P)$, where $P = X(X^TX)^{-1}X^T$. We first check that the conditions of the lemma hold:

Since $\mathbf{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I)$, $\mathbf{Z} = \mathbf{Y} - X\boldsymbol{\beta} \sim N_n(\mathbf{0}, \sigma^2 I)$.

Since P is idempotent, $I_n - P$ also is (check!). We also have

$$\text{rank}(I_n - P) = \text{tr}(I_n - P) = n - p.$$

Therefore the conditions of the lemma hold.

To get the final useful result, we want to show that the RSS is indeed $\mathbf{Z}^TA\mathbf{Z}$. We simplify the expressions of RSS and $\mathbf{Z}^TA\mathbf{Z}$ and show that they are equal:

$$\mathbf{Z}^TA\mathbf{Z} = (\mathbf{Y} - X\boldsymbol{\beta})^T(I_n - P)(\mathbf{Y} - X\boldsymbol{\beta}) = \mathbf{Y}^T(I_n - P)\mathbf{Y}.$$

Noting the fact that $(I_n - P)X = \mathbf{0}$.

Writing $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = (I_n - P)\mathbf{Y}$, we have

$$\text{RSS} = \mathbf{R}^T\mathbf{R} = \mathbf{Y}^T(I_n - P)\mathbf{Y},$$

using the symmetry and idempotence of $I_n - P$.

Hence $\text{RSS} = \mathbf{Z}^TA\mathbf{Z} \sim \sigma^2\chi^2_{n-p}$. Then

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n} \sim \frac{\sigma^2}{n}\chi^2_{n-p}.$$

– Let $V = \begin{pmatrix}\hat{\boldsymbol{\beta}} \\ \mathbf{R}\end{pmatrix} = D\mathbf{Y}$, where $D = \begin{pmatrix}C \\ I_n - P\end{pmatrix}$ is a $(p+n) \times n$ matrix.

Since $\mathbf{Y}$ is multivariate, V is multivariate with

$$\begin{aligned}
\text{cov}(V) &= D\sigma^2 I D^T \\
&= \sigma^2 \begin{pmatrix} CC^T & C(I_n - P)^T \\ (I_n - P)C^T & (I_n - P)(I_n - P)^T \end{pmatrix} \\
&= \sigma^2 \begin{pmatrix} CC^T & C(I_n - P) \\ (I_n - P)C^T & (I_n - P) \end{pmatrix} \\
&= \sigma^2 \begin{pmatrix} CC^T & 0 \\ 0 & I_n - P \end{pmatrix}
\end{aligned}$$

Using $C(I_n - P) = 0$ (since $(X^TX)^{-1}X^T(I_n - P) = 0$ since $(I_n - P)X = 0$ — check!).

Hence $\hat{\boldsymbol{\beta}}$ and $\mathbf{R}$ are independent since the off-diagonal covariant terms are 0. So $\hat{\boldsymbol{\beta}}$ and $\text{RSS} = \mathbf{R}^T\mathbf{R}$ are independent. So $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent. □

3.4 The F distribution

Proposition. If $X \sim F_{m,n}$, then $1/X \sim F_{n,m}$.

3.5 Inference for $\boldsymbol{\beta}$

3.6 Simple linear regression

3.7 Expected response at $\mathbf{x}^*$

3.8 Hypothesis testing

3.8.1 Hypothesis testing

Lemma. Suppose $\mathbf{Z} \sim N_n(\mathbf{0}, \sigma^2 I_n)$, and A_1 and A_2 are symmetric, idempotent $n \times n$ matrices with $A_1 A_2 = 0$ (i.e. they are orthogonal). Then $\mathbf{Z}^T A_1 \mathbf{Z}$ and $\mathbf{Z}^T A_2 \mathbf{Z}$ are independent.

Proof. Let $\mathbf{X}_i = A_i \mathbf{Z}$, $i = 1, 2$ and

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix} = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \mathbf{Z}.$$

Then

$$\mathbf{W} \sim N_{2n}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \sigma^2 \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} \right)$$

since the off diagonal matrices are $\sigma^2 A_1^T A_2 = A_1 A_2 = 0$.

So $\mathbf{W}_1$ and $\mathbf{W}_2$ are independent, which implies

$$\mathbf{W}_1^T \mathbf{W}_1 = \mathbf{Z}^T A_1^T A_1 \mathbf{Z} = \mathbf{Z}^T A_1 A_1 \mathbf{Z} = \mathbf{Z}^T A_1 \mathbf{Z}$$

and

$$\mathbf{W}_2^T \mathbf{W}_2 = \mathbf{Z}^T A_2^T A_2 \mathbf{Z} = \mathbf{Z}^T A_2 A_2 \mathbf{Z} = \mathbf{Z}^T A_2 \mathbf{Z}$$

are independent □

3.8.2 Simple linear regression

3.8.3 One way analysis of variance with equal numbers in each group