Chapter 1

Preliminaries

1.1 Elements of Statistical Modelling

Probability and statistics can be characterised as the study of variability. In particular, statistical inference is the science of analysing statistical data, viewed as the outcome of some random process, in order to draw conclusions about that random process.

Statistical models help us to *understand* the random process by which observed data have been generated. This may be of interest in itself, but also allows us to make *predictions* and perhaps most importantly *decisions* contingent on our inferences concerning the process.

It is also important, as part of the modelling process, to acknowledge that our conclusions are only based on a (potentially small) sample of possible observations of the process and are therefore subject to error. The science of statistical inference therefore involves assessment of the uncertainties associated with the conclusions we draw.

Probability theory is the mathematics associated with randomness and uncertainty. We usually try to describe random processes using probability models. Then, statistical inference may involve estimating any unspecified features of a model, comparing competing models, and assessing the appropriateness of a model; all in light of observed data.

One (rather simplistic) view of the process of statistical analysis might be displayed as follows

Specify models

$\downarrow$

Estimate model parameters

↓

Compare models

↓

Assess chosen model

↓

Base predictions/decisions on chosen model

In order to identify 'good' statistical models, we require some principles on which to base our modelling procedures. In general, we have three requirements of a statistical model

Plausibility

Parsimony

Goodness of fit

The first of these is not a statistical consideration, and a subject-matter expert usually needs to be consulted about this. Parsimony and goodness of fit are statistical issues. Indeed, there is usually a trade-off between the two and our statistical modelling strategies will take account of this.

Almost all statistical models, and all the ones we shall deal with in MA325, can be formulated as *regression* models.

In practical applications, we often distinguish between a *response* variable and a group of *explanatory* variables. The aim is to determine the pattern of dependence of the response variable on the explanatory variables. We denote the n observations of the response variable by $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$. In a statistical model, these are assumed to be observations of *random variables* $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$. Usually, we assume that $Y_1, Y_2, \ldots, Y_n$ are *independent* random variables. Associated with each y_i is a vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ of values of p explanatory variables.

A regression model has the general form

$$\text{response} = \text{structure} * \text{randomness}.$$

In other words, the structural part of the model, which describes how the response depends on the explanatory variables, combines with the random part, which depends on the probability distribution of the response. The statistical modellers task is to 'separate' these.

Regression with a single explanatory variable

Associated with each observation of a continuous response is a single continuous explanatory variable, so the data may be represented as $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$.

Possible models include

$$y_i = \alpha + \epsilon_i \qquad i = 1, \ldots, n$$

$$y_i = \alpha + \beta x_i + \epsilon_i \qquad i = 1, \ldots, n$$

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \qquad i = 1, \ldots, n$$

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_k x_i^k + \epsilon_i \qquad i = 1, \ldots, n$$

In each case the random part of the model $\boldsymbol{\epsilon} \equiv (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$ has some probability distribution. This might be fully specified, partially specified, or even totally unspecified. In examples like this, it is common to assume that the ϵ_i are independent and normally distributed with zero mean, but unspecified variance.

For complex datasets, a wider class of models is required. In MA325, we will study the *generalised linear models*. These are a flexible family of models allowing fairly general patterns of dependence of the response variable on the explanatory variables, together with a wide range of probability distributions for the response.

1.2 Example datasets to be analysed

1.2.1 `weld`: Welding diameter

This dataset, obtained from the Welding Institute in Abingdon, represents 21 measurements of the current (in amps) and the resulting minimum diameter of the weld.

1.2.2 `nitric`: Nitric acid

This dataset represents 21 successive days of operation of a plant oxidising ammonia to nitric acid. The variables `x1`, `x2` and `x3` are respectively, flow of air to the plant, temperature of the cooling water entering the absorption tower, and concentration of nitric acid in the absorbing liquid. The response `y` is ten times the percentage of ingoing ammonia that is lost as unabsorbed nitric acid (an indirect measure of the yield).

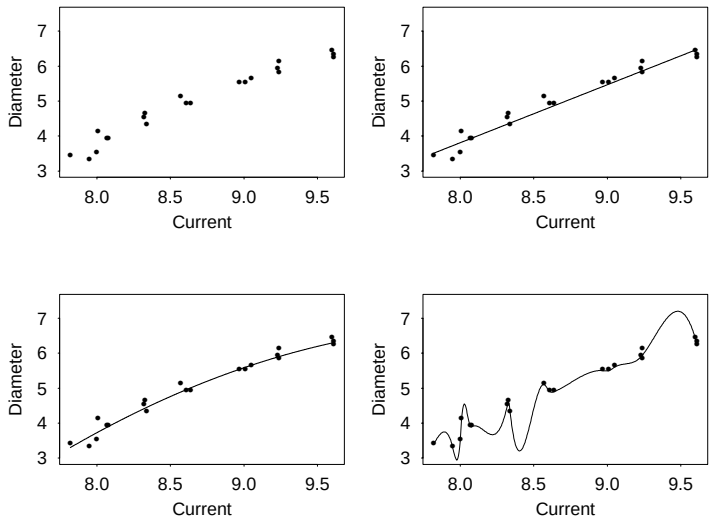

Figure 1.1: Possible models for the welding data

1.2.3 `survival`: Time to death

This dataset represents survival times (in 10 hour units) of 48 animals each allocated to one of 12 combinations of 4 treatments and 3 poisons. This will be taken up in Worksheet 3.

1.2.4 `birth`: Weight of newborn babies

This dataset contains data on the weight of 24 newborn babies. There are two explanatory variables; sex (S; a qualitative variable, coded "1=male" and "2=female") and gestational age (X; a quantitative variable, in weeks) together with the response variable, birth weight (Y; in grams).

1.2.5 `beetle`: Mortality from carbon disulphide

This dataset represents the number of beetles exposed (N) and number killed (Y) in eight groups exposed to different doses (X) of a particular insecticide. Interest is focussed on how mortality is related to dose. It seems sensible to model the number of beetles killed in each group as the binomial random variable with probability of death depending on dose.

1.2.6 `shuttle`: Challenger disaster

This dataset concerns the 23 space shuttle flights before the Challenger disaster. The disaster is thought to have been caused by the failure of a number of O-rings, of which there are six in total. The data consist of four columns, the number Y of damaged O-rings for each pre-Challenger flight, together with the launch temperature T in degrees Fahrenheit, the pressure P at which the pre-launch test of O-ring leakage was carried out and the name of the orbiter (S; coded 1 = "Atlantis", 2 = "Challenger", 3 = "Columbia", 4 = "Discovery",). The Challenger launch temperature on 20th January 1986 was 31F. By fitting generalised linear models to this data, we can predict the probability of O-ring damage at the Challenger launch.

1.2.7 `heart`: Treatment for heart attack

This dataset represents the results of a clinical trial to assess the effectiveness of a thrombolytic (clot-busting) treatment for patients who have suffered an acute myocardial infarction (heart attack). The two columns represent the number of patients who did and did not survive for 35 days. There are four categorical explanatory variables, representing site of infarction (S: anterior, inferior or other); time between infarction and treatment (T: ≤ 12 or > 12 hours); whether the patient was already taking Beta-blocker medication prior to the infarction (B: yes or no); and the treatment the patient was given (R: active or placebo).

1.2.8 `hodk`: Treatment for Hodgkin's disease

This dataset is a cross-classification of 538 patients (with Hodgkin's disease) according to two factors, H, the histological type of their disease (4 levels) and R, their response to treatment (3 levels).

1.2.9 `accident`: Road traffic accident

This dataset concerns the number of road accidents and the volume of traffic observed on Mill Road and Trumptington Road in Cambridge during morning, midday and afternoon. By analysing this we should be able to answer questions like: (i) Is Mill Road more dangerous than Trumpington Road? (ii) How does time of day affect the rate of road accident?

1.3 Distribution theory (revision)

1.3.1 Probability Distributions

A random variable Y is described by its sample space S_Y, together with the probabilities assigned to subsets of the sample space. These define the *probability distribution* of the random variable.

We distinguish between *Discrete* probability distributions and *Continuous* probability distributions **Discrete** probability distributions are defined by a probability (mass) function (p.f.)

$$f_Y(y) \equiv P(Y = y) \qquad \text{for } y \in S_Y$$

where

$$\sum_{y \in S_Y} f_Y(y) = 1.$$

Continuous probability distributions are defined by a probability density function (p.d.f.) $f_Y(y)$ where

$$P(y_1 < Y \leq y_2) = \int_{y_1}^{y_2} f_Y(y)dy.$$

and hence

$$\int_{-\infty}^{\infty} f_Y(y)dy = 1.$$

The expectation of a random variable Y is given by.

$$E(Y) = \sum_{y \in S_Y} y f_Y(y)$$

if Y is discrete, and

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y)dy$$

if Y is continuous.

More generally, for any function g, we can write

$$E[g(Y)] = \sum_{y \in S_Y} g(y) f_Y(y)$$

if Y is discrete, and

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y) f_Y(y) dy$$

if Y is continuous.

Note that expectation is a *linear* operator, so

$$E(a + bY) = a + bE(Y)$$

for any $a, b \in \mathcal{R}$.

The *variance* of a random variable Y is defined by

$$Var(Y) \equiv \quad E\left([Y - E(Y)]^2\right) = \quad E\left(Y^2\right) - E(Y)^2.$$

1.4 Examples of discrete probability distributions

1.4.1 Bernoulli distribution

The Bernoulli distribution depends on one parameter, $p \in (0, 1)$. If random variable Y has a Bernoulli(p) distribution then $S_Y = \{0, 1\}$ and

$$f_Y(y) = \begin{cases} 1 - p & \text{if } y = 0 \\ p & \text{if } y = 1 \end{cases}$$

Thus $f_Y(y) = p^y (1 - p)^{1-y} \quad y \in \{0, 1\}$.

$$E(Y) = p, \quad Var(Y) = p(1 - p).$$

The Bernoulli distribution can be used to model any situation where the only two outcomes are possible, and can be coded as 0 and 1 (binary response).

1.4.2 Binomial distribution

The binomial distribution depends on two parameters, $n \in \mathcal{Z}_+$ and $p \in (0, 1)$. If random variable Y has a binomial(n, p) distribution then $S_Y = \{0, 1, \ldots, n\}$ and

$$f_Y(y) = \binom{n}{y} p^y (1 - p)^{n-y} \qquad y \in \{0, 1, \ldots, n\},$$

$$E(Y) = np, \quad Var(Y) = np(1-p).$$

If $Y_1, \ldots, Y_n$ are independent Bernoulli(p) random variables then $Y_1 + \ldots + Y_n$ is a binomial(n, p) random variable. Therefore, the binomial distribution is used to model any situation where a series of events have two possible outcomes, the chances of which remain constant.

The binomial($1, p$) distribution is the Bernoulli(p) distribution.

Applications include studies of mortality, or remission under treatment.

1.4.3 Poisson distribution

The Poisson distribution depends on one parameter, $\lambda \in \mathcal{R}_+$. If random variable Y has a Poisson(λ) distribution then $S_Y = \{0, 1, \ldots\}$ and

$$f_Y(y) = \frac{\exp(-\lambda)\lambda^y}{y!} \qquad y \in \{0, 1, \ldots\},$$

$$E(Y) = \lambda, \quad Var(Y) = \lambda.$$

If $Y_1, \ldots, Y_n$ are independent Poisson random variables with parameters $\lambda_1, \ldots, \lambda_n$, then $Y_1 + \ldots + Y_n$ is a Poisson random variable with parameter $\lambda_1 + \ldots + \lambda_n$.

A binomial(n, p) distribution with large n may be approximated by a Poisson distribution with mean np.

The Poisson distribution is used to model 'counts', particularly counts of a phenomenon over a particular unit of time, or in a particular spatial region. Applications include counts of cases of a particular disease, or number of arrivals in a queue.

1.5 Examples of continuous probability distributions

1.5.1 Normal distribution

The normal (or Gaussian) distribution depends on two parameters, the *mean* $\mu \in \mathcal{R}$ and the *variance* $\sigma^2 \in \mathcal{R}_+$, and is usually denoted $N(\mu, \sigma^2)$. If random variable Y has a $N(\mu, \sigma^2)$

distribution then $S_Y = \mathcal{R}$ and

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \qquad y \in \mathcal{R},$$

$$E(Y) = \mu, \quad Var(Y) = \sigma^2.$$

If $Y_1, \ldots, Y_n$ are independent normal random variables with means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$, then $Y_1 + \ldots + Y_n$ is a normal random variable with mean $\mu_1 + \ldots + \mu_n$ and variance $\sigma_1^2 + \ldots + \sigma_n^2$

If Y is a $\mathrm{N}(\mu, \sigma^2)$ random variable then $a + bY$ is a $\mathrm{N}(a + b\mu, b^2\sigma^2)$ random variable.

The *central limit theorem* states that if $Y_1, Y_2, \ldots, Y_n$ are independent identically distributed random variables with $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2 < \infty$ for $i = 1, \ldots, n$, then as $n \to \infty$, the distribution of $\overline{Y} \equiv \frac{1}{n}\sum_{i=1}^n Y_i$ tends to $N(\mu, \sigma^2/n)$. Similar central limit theorems for the distribution of $\overline{Y}$ exist when $Y_1, Y_2, \ldots, Y_n$ are not identically distributed, or even independent, although these theorems require additional conditions to be satisfied.

1.5.2 Gamma distribution

The gamma distribution depends on two parameters, the *shape* parameter $\alpha \in \mathcal{R}_+$, and the *scale* parameter $\beta \in \mathcal{R}_+$. If random variable Y has a gamma(α, β) distribution then $S_Y = \mathcal{R}_+$ and

$$f_Y(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y) \qquad y \in \mathcal{R}_+,$$

$$E(Y) = \frac{\alpha}{\beta}, \quad Var(Y) = \frac{\alpha}{\beta^2}.$$

The gamma function $\Gamma(t)$ is defined by

$$\Gamma(t) = \int_0^\infty x^{t-1} \exp(-x) dx.$$

If $Y_1, \ldots, Y_n$ are independent Gamma random variables with common scale parameter β and shape parameters $\alpha_1, \ldots, \alpha_n$, then $Y_1 + \ldots + Y_n$ is a Gamma random variable with scale parameter β and shape parameter $\alpha_1 + \ldots + \alpha_n$.

If Y is a Gamma(α, β) random variable then bY is a Gamma$(\alpha, \beta/b)$ random variable.

The gamma distribution is a flexible model for many positive variables, including those arising in meteorology, lifetime testing, demography and economics.

There are two important special cases of the gamma distribution.

1. A gamma distribution with shape parameter $\alpha = 1$ is an exponential distribution, *i.e.* gamma$(1, \beta) \equiv$ exponential(β).

2. A gamma distribution with shape parameter $\alpha = \frac{k}{2}$ and scale parameter $\beta = \frac{1}{2}$ is called a *chi-squared* distribution with k degrees of freedom, usually denoted χ^2_k.

3. If random variable Y has a χ^2_k distribution then $S_Y = \mathcal{R}_+$ and

$$f_Y(y) = \frac{1}{\Gamma(k/2)2^{k/2}} y^{k/2-1} \exp(-y/2) \qquad y \in \mathcal{R}_+,$$

$$E(Y) = k, \quad Var(Y) = 2k.$$

1.5.3 Derived distributions

If Y_1 is a $\chi^2_{k_1}$ random variable, Y_2 is a $\chi^2_{k_2}$ random variable and Z is a $\mathrm{N}(0,1)$ random variable, and Y_1, Y_2 and Z are independent then

1. Z^2 has a chi-squared distribution with 1 degree of freedom.

2. $Y_1 + Y_2$ has a chi-squared distribution with $k_1 + k_2$ degrees of freedom.

3. $Z/\sqrt{Y_1/k_1}$ has a t distribution with k_1 degrees of freedom.

4. $(Y_1/k_1)/(Y_2/k_2)$ has an F distribution with k_1 degrees of freedom in the numerator and k_2 degrees of freedom in the denominator.

1.6 Multivariate Probability Distribution

A multivariate probability distribution describes the joint variation of a collection of random variables.

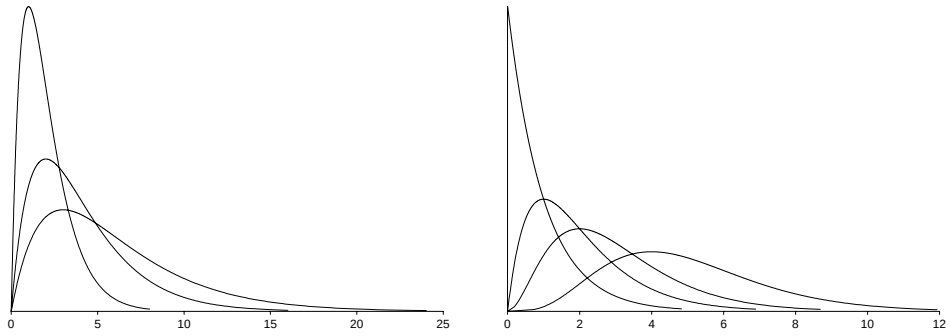

Figure 1.2: Density functions for some gamma distributions

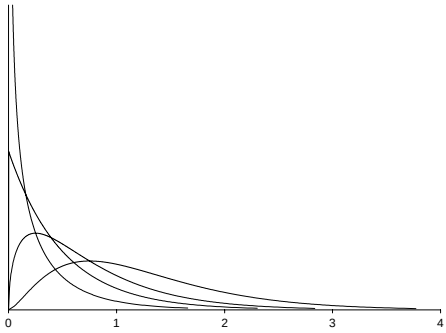

Figure 1.3: Density functions for some gamma distributions

Suppose that $Y_1, Y_2, \ldots, Y_n$ are random variables. Then we write $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ as the *vector* of random variables, and $S_{\mathbf{Y}}$ as the sample space for $\mathbf{Y}$. Usually we assume that $S_{\mathbf{Y}} = \mathcal{R}^n$.

Discrete multivariate probability distributions are defined by a probability (mass) function (p.f.)

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &\equiv P(\mathbf{Y} = \mathbf{y}) \\ &= P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) \qquad \text{for } \mathbf{y} \in S_{\mathbf{Y}} \end{aligned}$$

where

$$\sum_{\mathbf{y} \in S_{\mathbf{Y}}} f_{\mathbf{Y}}(\mathbf{y}) = 1.$$

Continuous multivariate probability distributions are defined by a *joint* probability

density function (p.d.f.) $f_{\mathbf{Y}}(\mathbf{y})$ where

$$P(\mathbf{Y} \in A) = \int_A f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} \qquad A \subseteq S_{\mathbf{Y}}.$$

and hence

$$\int_{\mathcal{R}^n} f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} = 1.$$

Henceforth, we shall restrict attention to continuous distributions. All of the properties we describe, can be applied to discrete distributions by replacing integration by summation. If $\mathbf{Y}$ has p.d.f. $f_{\mathbf{Y}}(\mathbf{y})$, then the p.d.f of random variable Y_1, a component of $\mathbf{Y}$ is obtained from the joint p.d.f. $f_{\mathbf{Y}}$ by integrating out all the other variables (or for a discrete distribution by summing the p.f.), for example

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{Y}}(y_1, y_2, y_3, \ldots, y_n) dy_2 dy_3 \cdots dy_n$$

The distribution of a component (or a set of components, or a function of components) of a jointly distributed collection of variables is called the *marginal* distribution.

The expectation (mean) $E(\mathbf{Y})$ of $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ is defined to be

$$E(\mathbf{Y}) = [E(Y_1), E(Y_2), \ldots E(Y_n)]^T,$$

the vector containing the marginal expectations of each of the random variables. It is clear that, for example,

$$E(Y_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} y_1 f_{\mathbf{Y}}(y_1, y_2, \ldots, y_n) dy_1 dy_2 \cdots dy_n.$$

More generally, for any function $g(\mathbf{Y})$ of $\mathbf{Y}$, we define

$$E[g(\mathbf{Y})] \equiv \int_{\mathcal{R}^n} g(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}.$$

It is immediately clear that

$$E(Y_1 + Y_2 + \ldots + Y_n) = E(Y_1) + E(Y_2) + \ldots + E(Y_n).$$

A more general and useful result is

$$E(\mathbf{a} + \mathbf{B}\mathbf{Y}) = \mathbf{a} + \mathbf{B}E(\mathbf{Y})$$

where $\mathbf{B}$ is any $p \times n$ matrix of constants, and $\mathbf{a}$ is any vector of p constants. The covariance of a pair of jointly distributed random variables, Y_1 and Y_2 is defined to be

$$\begin{aligned} Cov(Y_1, Y_2) &\equiv E([Y_1 - E(Y_1)][Y_2 - E(Y_2)]) \\ &= E(Y_1 Y_2) - E(Y_1)E(Y_2) \end{aligned}$$

Note that $Cov(Y_1, Y_1) = Var(Y_1)$ and $Cov(Y_i, Y_j) = Cov(Y_j, Y_i)$, and

$$\begin{aligned} Var(Y_1 + Y_2 + \ldots + Y_n) &= \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(Y_i, Y_j) \\ &= \sum_{i=1}^{n} Var(Y_i) + \sum_{i \neq j} Cov(Y_i, Y_j) \\ &= \sum_{i=1}^{n} Var(Y_i) + \sum_{i<j} 2Cov(Y_i, Y_j) \end{aligned}$$

A more general and useful result is

$$Var(\mathbf{a} + \mathbf{B}\mathbf{Y}) = \mathbf{B}Var(\mathbf{Y})\mathbf{B}^T$$

where $\mathbf{B}$ is any $p \times n$ matrix of constants, and $\mathbf{a}$ is any vector of p constants.

The correlation of random variables Y_1 and Y_2 is defined as

$$Corr(Y_1, Y_2) = \frac{Cov(Y_1, Y_2)}{\sqrt{Var(Y_1)Var(Y_2)}}.$$

Note that $-1 \leq Corr(Y_1, Y_2) \leq 1$.

For a jointly distributed collection of random variables $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$, $Var(\mathbf{Y})$ is the *variance-covariance matrix*, a $n \times n$ matrix whose entries are given by

$$Var(\mathbf{Y})_{ij} \equiv Cov(Y_i, Y_j).$$

1.6.1 Independent random variables

A collection of continuous random variables $Y_1, Y_2, \ldots, Y_n$ are said to be jointly *independent* if and only if their joint density is the product of their marginal densities *i.e.*

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{Y_1}(y_1) f_{Y_2}(y_2) \cdots f_{Y_n}(y_n).$$

Furthermore, if $Y_1, Y_2, \ldots, Y_n$ are independent then $E(Y_1 Y_2 \cdots Y_n) = E(Y_1)E(Y_2)\cdots E(Y_n)$, and hence for any pair of independent random variables

$$Y_1 \text{ and } Y_2 \text{ are independent} \Rightarrow Cov(Y_1, Y_2) = 0.$$

Independent random variables are uncorrelated, but uncorrelated random variables are not necessarily independent.

It follows that if $Y_1, Y_2, \ldots, Y_n$ are independent random variables then

$$Var(Y_1 + Y_2 + \ldots + Y_n) = Var(Y_1) + Var(Y_2) + \ldots + Var(Y_n).$$

1.6.2 Conditional distributions

If Y_1 and Y_2 are jointly distributed random variables with p.d.f. $f_{\mathbf{Y}}(y_1, y_2)$, then the distribution of Y_1 *conditional on* $Y_2 = y_2$, for any value y_2 for which $f_{Y_2}(y_2) > 0$ is determined by the conditional p.d.f

$$f_{Y_1|Y_2}(y_1|Y_2 = y_2) = \frac{f_{\mathbf{Y}}(y_1, y_2)}{f_{Y_2}(y_2)}.$$

Hence, if Y_1 and Y_2 are independent, $f_{Y_1|Y_2}(y_1|Y_2 = y_2) = f_{Y_1}(y_1)$, and 'knowledge of Y_2 does not influence the distribution of Y_1.'

More generally, Y_1 and Y_2 can be replaced in the definition above by collections of jointly distributed random variables.

1.6.3 The multivariate normal distribution

Suppose that $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ is a collection of jointly distributed random variables, with $E(\mathbf{Y}) = \boldsymbol{\mu}$ and $Var(\mathbf{Y}) = \boldsymbol{\Sigma}$. Then $\mathbf{Y}$ is said to have a *multivariate normal distribution*, denoted $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if the p.d.f. of $\mathbf{Y}$ is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})\right].$$

The multivariate normal distribution has several appealing properties. If $\mathbf{Y}$ has the multivariate normal distribution denoted $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then:

1. The marginal distribution of any component of $\mathbf{Y}$ is univariate normal. For example, Y_1 is a $N(\mu_1, \Sigma_{11})$ random variable.

2. The conditional distribution of any component given the other components is univariate normal.

3. If $\mathbf{X} = \mathbf{a} + \mathbf{BY}$, where $\mathbf{B}$ is any $p \times n$ matrix of constants, and $\mathbf{a}$ is any vector of p constants, then $\mathbf{X}$ is a multivariate normal, with $E(\mathbf{X}) = \mathbf{a} + \mathbf{B}\boldsymbol{\mu}$ and $Var(\mathbf{X}) = \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T$.

4. If $\mathbf{Y}$ has a multivariate normal distribution then $Cov(Y_i, Y_j) = 0 \Rightarrow Y_i$ and Y_j are independent.

1.7 Likelihood based statistical theory

1.7.1 The likelihood function

The probability distribution theory discussed previously enables us to calculate probabilities, and other quantities of interest (*e.g.* expectations) for a probability model of a random process. Therefore, given the model, we can make statements about possible outcomes of the process.

Statistical inference is concerned with the inverse problem. Given outcomes of a random process (observed data), what conclusions (inferences) can we draw about the process itself?

We assume that the n observations of the response $\mathbf{y} = (y_1, \ldots, y_n)^T$ are observations of random variables $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, which have joint p.d.f. $f_{\mathbf{Y}}$ (joint p.f. for discrete variables). We use the observed data $\mathbf{y}$ to make inferences about $f_{\mathbf{Y}}$.

We usually make certain assumptions about $f_{\mathbf{Y}}$. In particular, we usually assume that $y_1, \ldots, y_n$ are observations of *independent* random variables. Hence

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{Y_1}(y_1) f_{Y_2}(y_2) \cdots f_{Y_n}(y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i).$$

In parametric statistical inference, we specify a joint distribution $f_{\mathbf{Y}}$, for $\mathbf{Y}$, which is known, except for the values of parameters $\theta_1, \theta_2, \ldots, \theta_p$ (sometimes denoted $\boldsymbol{\theta}$). Then we use the observed data $\mathbf{y}$ to make inferences about $\theta_1, \theta_2, \ldots, \theta_p$. In this case, we usually write $f_{\mathbf{Y}}$ as $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$, to make explicit the dependence on the unknown $\boldsymbol{\theta}$.

An important concept for parametric statistical inference, particularly for complex statistical models is *likelihood*.

Until now, we have thought of the joint density $f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$ as a function of $\mathbf{y}$ for fixed $\boldsymbol{\theta}$, which describes the relative probabilities of different possible (sets of) $\mathbf{y}$, given a particular set of parameters $\boldsymbol{\theta}$. However, in statistical inference, we have observed $y_1, \ldots, y_n$ (values of $Y_1, \ldots, Y_n$). Knowledge of the probability of alternative possible realisations of $\mathbf{Y}$ is largely irrelevant. What we want to know about is $\boldsymbol{\theta}$.

Our only link between the observed data $y_1, \ldots, y_n$ and $\boldsymbol{\theta}$ is through the function $f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$. Therefore, it seems sensible that parametric statistical inference should be based on this function. We can think of $f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ for fixed $\mathbf{y}$, which describes the relative *likelihoods* of different possible (sets of) $\boldsymbol{\theta}$, given observed data $y_1, \ldots, y_n$.

> When $f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$ is considered as a function of $\boldsymbol{\theta}$ for fixed (observed) $y_1, \ldots, y_n$, we call it the *likelihood function*.

The likelihood function is of central importance in parametric statistical inference. It provides a means for comparing different possible values of $\boldsymbol{\theta}$, based on the probabilities (densities) that they assign to the observed data $y_1, \ldots, y_n$.

1. Often the likelihood function is written $L(\boldsymbol{\theta})$ or $L(\boldsymbol{\theta};\mathbf{y})$. We shall continue to use $f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$. Wherever parametric statistical inference is our concern, we treat $f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ and call it the likelihood.

2. Frequently it is more convenient to consider the log-likelihood function $\log f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$, This is often denoted $l(\boldsymbol{\theta})$ or $l(\boldsymbol{\theta};\mathbf{y})$.

3. Nothing in the definition of the likelihood requires $y_1, \ldots, y_n$ to be observations of independent random variables, although we shall frequently make this assumption.

4. Any factors which depend on $y_1, \ldots, y_n$ alone (and not on $\boldsymbol{\theta}$) can be ignored when writing down the likelihood. Such factors give no information about the relative likelihoods of different possible values of $\boldsymbol{\theta}$.

♡ **Example** 1.1. $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, independent identically distributed (i.i.d.) Bernoulli(p) random variables. Here $\boldsymbol{\theta} = (p)$ and

$$f_{\mathbf{Y}}(\mathbf{y};p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{\Sigma y_i}(1-p)^{n-\Sigma y_i}$$

Note that as $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, the m.l.e. for any component of $\boldsymbol{\theta}$ is given by the corresponding component of $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^T$. Similarly, the m.l.e. for any function of parameters $g(\boldsymbol{\theta})$ is given by $g(\hat{\boldsymbol{\theta}})$.

As log is a strictly increasing function, the value of $\boldsymbol{\theta}$ which maximises $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$ also maximises $\log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$. It is almost always easier to maximise $\log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$. This is achieved in the usual way; finding a stationary point by differentiating $\log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$ with respect to $\theta_1, \ldots, \theta_p$ and solving the resulting p simultaneous equations. It should also be checked that the stationary point is a maximum.

♡ **Example** 1.3. $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. Bernoulli(p) random variables. Here $\boldsymbol{\theta} = (p)$ and

$$\begin{aligned} \log f_{\mathbf{Y}}(\mathbf{y}; p) &= n\overline{y} \log p + n(1-\overline{y}) \log(1-p) \\ \tfrac{\partial}{\partial p} \log f_{\mathbf{Y}}(\mathbf{y}; p) &= \tfrac{n\overline{y}}{p} - \tfrac{n(1-\overline{y})}{1-p} \\ \Rightarrow \quad 0 &= \tfrac{n\overline{y}}{\hat{p}} - \tfrac{n(1-\overline{y})}{1-\hat{p}} \\ \Rightarrow \quad \hat{p} &= \overline{y}. \end{aligned}$$

Note that $\frac{\partial^2}{\partial p^2} \log f_{\mathbf{Y}}(\mathbf{y}; p) = -n\overline{y}/p^2 - n(1-\overline{y})/(1-p)^2 < 0$ everywhere, so the stationary point is clearly a maximum.

♡ **Example** 1.4. $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. N(μ, σ^2) random variables. Here $\boldsymbol{\theta} = (\mu, \sigma^2)$ and

$$\begin{aligned} \log f_{\mathbf{Y}}(\mathbf{y}; \mu, \sigma^2) &= -\tfrac{n}{2} \log(2\pi) - \tfrac{n}{2} \log(\sigma^2) - \tfrac{1}{2\sigma^2} \textstyle\sum (y_i - \mu)^2 \\ \tfrac{\partial}{\partial \mu} \log f_{\mathbf{Y}}(\mathbf{y}; \mu, \sigma^2) &= \tfrac{1}{\sigma^2} \textstyle\sum (y_i - \mu) = \tfrac{n(\overline{y}-\mu)}{\sigma^2} \\ \Rightarrow \quad 0 &= \tfrac{n(\overline{y}-\hat{\mu})}{\hat{\sigma}^2} \qquad (1) \\ \tfrac{\partial}{\partial \sigma^2} \log f_{\mathbf{Y}}(\mathbf{y}; \mu, \sigma^2) &= -\tfrac{n}{2\sigma^2} + \tfrac{1}{2(\sigma^2)^2} \textstyle\sum (y_i - \mu)^2 \\ \Rightarrow \quad 0 &= -\tfrac{n}{2\hat{\sigma}^2} + \tfrac{1}{2(\hat{\sigma}^2)^2} \textstyle\sum (y_i - \hat{\mu})^2 \qquad (2) \end{aligned}$$

Solving (1) and (2), we obtain

$$\begin{aligned} \hat{\mu} &= \overline{y} \\ \hat{\sigma}^2 &= \frac{1}{n} \textstyle\sum (y_i - \hat{\mu})^2 = \frac{1}{n} \sum (y_i - \overline{y})^2. \end{aligned}$$

Strictly, to show that this stationary point is a maximum, we need to show that the Hessian matrix (the matrix of second derivatives with elements $[\mathbf{H}(\boldsymbol{\theta})]_{ij} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$) is negative definite at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, that is $\mathbf{a}^T \mathbf{H}(\hat{\boldsymbol{\theta}}) \mathbf{a} < 0$ for every $\mathbf{a} \neq \mathbf{0}$. Here

$$\mathbf{H}(\hat{\mu}, \hat{\sigma}^2) = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2(\hat{\sigma}^2)^2} \end{pmatrix}$$

which is clearly negative definite.

1.7.3 Score

Suppose that $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, whose joint p.d.f. $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$ is completely specified except for the values of p unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$. Let

$$u_i(\boldsymbol{\theta}) \equiv \frac{\partial}{\partial \theta_i} \log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) \qquad i = 1, \ldots, p$$

and $\mathbf{u}(\boldsymbol{\theta}) \equiv [u_1(\boldsymbol{\theta}), \ldots, u_p(\boldsymbol{\theta})]^T$. Then we call $\mathbf{u}(\boldsymbol{\theta})$ the *vector of scores* or *score vector*. Where $p = 1$ and $\boldsymbol{\theta} = (\theta)$, the *score* is the scalar defined as

$$u(\theta) \equiv \frac{\partial}{\partial \theta} \log f_{\mathbf{Y}}(\mathbf{y}; \theta).$$

The maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ satisfies

$$u(\hat{\boldsymbol{\theta}}) = \mathbf{0} \qquad \Leftrightarrow \qquad u_i(\hat{\boldsymbol{\theta}}) = 0 \quad i = 1, \ldots, p.$$

Note that $\mathbf{u}(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$ for fixed (observed) $\mathbf{y}$. However, if we replace $y_1, \ldots, y_n$ in $\mathbf{u}(\boldsymbol{\theta})$, by the corresponding random variables $Y_1, \ldots, Y_n$ then we obtain a vector of random variables $\mathbf{U}(\boldsymbol{\theta}) \equiv [U_1(\boldsymbol{\theta}), \ldots, U_p(\boldsymbol{\theta})]^T$.

An important result in likelihood theory is that the expected score at the true (but unknown) value of $\boldsymbol{\theta}$ is zero, *i.e.*

$$E[\mathbf{U}(\boldsymbol{\theta})] = \mathbf{0} \qquad \Leftrightarrow \qquad E[U_i(\boldsymbol{\theta})] = 0 \quad i = 1, \ldots, p,$$

provided that

1. The expectation exists.
2. The sample space for $\mathbf{Y}$ does not depend on $\boldsymbol{\theta}$.

Proof (continuous $\mathbf{y}$ – in discrete case replace $\int$ by $\sum$)

$$\begin{aligned} E[U_i(\boldsymbol{\theta})] &= \int U_i(\boldsymbol{\theta}) f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) d\mathbf{y} \\ &= \int \frac{\partial}{\partial \theta_i} \log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) d\mathbf{y} \end{aligned}$$

$$
\begin{aligned}
&= \int \frac{\frac{\partial}{\partial\theta_i} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) d\mathbf{y} \\
&= \int \frac{\partial}{\partial\theta_i} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) d\mathbf{y} \\
&= \frac{\partial}{\partial\theta_i} \int f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) d\mathbf{y} \\
&= \frac{\partial}{\partial\theta_i} 1 = 0 \qquad i = 1, \ldots p.
\end{aligned}
$$

♡ **Example** 1.5. $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. Bernoulli(p) random variables. Here $\boldsymbol{\theta} = (p)$ and, from §2.4.2,

$$
\begin{aligned}
u(p) &= n\overline{y}/p - n(1-\overline{y})/(1-p) \\
E[U(p)] = 0 \quad &\Rightarrow \quad E[\overline{Y}] = p.
\end{aligned}
$$

♡ **Example** 1.6. $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. N(μ, σ^2) random variables. Here $\boldsymbol{\theta} = (\mu, \sigma^2)$ and, from §2.4.2,

$$
\begin{aligned}
u_1(\mu,\sigma^2) &= n(\overline{y} - \mu)/\sigma^2 \\
u_2(\mu,\sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum(y_i - \mu)^2 \\
E[\mathbf{U}(\mu,\sigma^2)] = \mathbf{0} \quad &\Rightarrow \quad E[\overline{Y}] = \mu \quad \text{and} \quad E[\tfrac{1}{n}\textstyle\sum(y_i - \mu)^2] = \sigma^2.
\end{aligned}
$$

1.7.4 Information

Suppose that $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, whose joint p.d.f. $f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$ is completely specified except for the values of p unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$. Previously, we defined the Hessian matrix $\mathbf{H}(\boldsymbol{\theta})$ to be the matrix with components

$$
[\mathbf{H}(\boldsymbol{\theta})]_{ij} \equiv \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) \qquad i = 1, \ldots, p;\ j = 1, \ldots, p.
$$

We call the matrix $-\mathbf{H}(\boldsymbol{\theta})$ the *observed information matrix.* Where $p = 1$ and $\boldsymbol{\theta} = (\theta)$, the *observed information* is a scalar defined as

$$
-H(\theta) \equiv -\frac{\partial}{\partial\theta^2} \log f_{\mathbf{Y}}(\mathbf{y};\theta).
$$

Here, we are interpreting $\boldsymbol{\theta}$ as the true (but unknown) value of the parameter. As with the score, if we replace $y_1, \ldots, y_n$ in $\mathbf{H}(\boldsymbol{\theta})$, by the corresponding random variables $Y_1, \ldots, Y_n$,

we obtain a matrix of random variables. Then, we define the *expected information matrix* or *Fisher information matrix to be* $\mathcal{I}(\boldsymbol{\theta})$, where

$$[\mathcal{I}(\boldsymbol{\theta})]_{ij} = E(-[\mathbf{H}(\boldsymbol{\theta})]_{ij}) \qquad i = 1, \ldots, p;\ j = 1, \ldots, p.$$

An important result in likelihood theory is that the variance-covariance matrix of the score vector is equal to the expected information matrix *i.e.*

$$Var[\mathbf{U}(\boldsymbol{\theta})] = \mathcal{I}(\boldsymbol{\theta}) \quad \Leftrightarrow \quad Var[\mathbf{U}(\boldsymbol{\theta})]_{ij} = [\mathcal{I}(\boldsymbol{\theta})]_{ij} \qquad i = 1, \ldots, p;\ j = 1, \ldots, p,$$

provided that

1. The variance exists.
2. The sample space for $\mathbf{Y}$ does not depend on $\boldsymbol{\theta}$.

Proof (continuous $\mathbf{y}$ – in discrete case replace $\int$ by $\sum$)

$$\begin{aligned}
Var[\mathbf{U}(\boldsymbol{\theta})]_{ij} &= E[U_i(\boldsymbol{\theta})U_j(\boldsymbol{\theta})] \\
&= \int \frac{\partial}{\partial\theta_i}\log f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}\log f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})d\mathbf{y} \\
&= \int \frac{\frac{\partial}{\partial\theta_i} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})}\frac{\frac{\partial}{\partial\theta_j} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})d\mathbf{y} \\
&= \int \frac{1}{f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})}\frac{\partial}{\partial\theta_i} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})\frac{\partial}{\partial\theta_j} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})d\mathbf{y} \\
&\qquad\qquad i = 1, \ldots, p;\ j = 1, \ldots, p.
\end{aligned}$$

Now

$$\begin{aligned}
[\mathcal{I}(\boldsymbol{\theta})]_{ij} &= E\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})\right] \\
&= \int -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})d\mathbf{y} \\
&= \int -\frac{\partial}{\partial\theta_i}\left[\frac{\frac{\partial}{\partial\theta_j} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})}\right] f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})d\mathbf{y} \\
&= \int \left[-\frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})} + \frac{\frac{\partial}{\partial\theta_i} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})\frac{\partial}{\partial\theta_j} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})^2}\right] f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})d\mathbf{y}
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\int f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})d\mathbf{y} + \int \frac{1}{f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})}\frac{\partial}{\partial\theta_i}f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})d\mathbf{y} \\
&= Var[\mathbf{U}(\boldsymbol{\theta})]_{ij} \qquad i = 1,\ldots,p;\ j = 1,\ldots,p.
\end{aligned}
$$

♡ **Example** 1.7. $y_1,\ldots,y_n$ are observations of $Y_1,\ldots,Y_n$, i.i.d. Bernoulli(p) random variables. Here $\boldsymbol{\theta} = (p)$ and

$$
\begin{aligned}
u(p) &= \frac{n\overline{y}}{p} - \frac{n(1-\overline{y})}{(1-p)} \\
-H(p) &= \frac{n\overline{y}}{p^2} + \frac{n(1-\overline{y})}{(1-p)^2} \\
\mathcal{I}(p) &= \frac{n}{p} + \frac{n}{(1-p)} = \frac{n}{p(1-p)}.
\end{aligned}
$$

♡ **Example** 1.8. $y_1,\ldots,y_n$ are observations of $Y_1,\ldots,Y_n$, i.i.d. N(μ,σ^2) random variables. Here $\boldsymbol{\theta} = (\mu,\sigma^2)$ and,

$$
\begin{aligned}
u_1(\mu,\sigma^2) &= \frac{n(\overline{y}-\mu)}{\sigma^2} \\
u_2(\mu,\sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum(y_i-\mu)^2
\end{aligned}
$$

$$
-\mathbf{H}(\mu,\sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{n(\overline{y}-\mu)}{(\sigma^2)^2} \\ \frac{n(\overline{y}-\mu)}{(\sigma^2)^2} & \frac{1}{(\sigma^2)^3}\sum(y_i-\mu)^2 - \frac{n}{2(\sigma^2)^2} \end{pmatrix}
$$

$$
\mathcal{I}(\mu,\sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2(\sigma^2)^2} \end{pmatrix}.
$$

1.7.5 Asymptotic distribution of the m.l.e.

Maximum likelihood estimation is an attractive method of estimation for a number of reasons. It is intuitively sensible (choosing $\boldsymbol{\theta}$ which makes the observed data most probable) and usually reasonably straightforward to carry out. Even when the simultaneous equations we obtain by differentiating the log likelihood function are impossible to solve directly, solution by numerical methods is usually feasible.

Perhaps the most compelling reason for considering maximum likelihood estimation is the asymptotic behaviour of maximum likelihood estimators.

Suppose that $y_1, \ldots, y_n$ are observations of independent random variables $Y_1, \ldots, Y_n$, whose joint p.d.f. $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_{Y_i}(y_i; \boldsymbol{\theta})$ is completely specified except for the values of an unknown parameter vector $\boldsymbol{\theta}$, and that $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$.

Then, as $n \to \infty$, the distribution of $\hat{\boldsymbol{\theta}}$ tends to a multivariate normal distribution with mean vector $\boldsymbol{\theta}$ and variance covariance matrix $\mathcal{I}(\boldsymbol{\theta})^{-1}$.

Where $p = 1$ and $\boldsymbol{\theta} = (\theta)$, the distribution of the m.l.e. $\hat{\theta}$ tends to $N[\theta, 1/\mathcal{I}(\theta)]$.

Proof (one parameter case; identically distributed Y_i)

Suppose that $y_1, \ldots, y_n$ are observations of independent identically distributed random variables $Y_1, \ldots, Y_n$, whose joint p.d.f. is therefore $f_{\mathbf{Y}}(\mathbf{y}; \theta) = \prod_{i=1}^{n} f_{\mathbf{Y}}(y_i; \theta)$. We can write the score as

$$u(\theta) = \frac{\partial}{\partial \theta} \log f_{\mathbf{Y}}(\mathbf{y}; \theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f_{\mathbf{Y}}(y_i; \theta)$$

so $U(\theta)$ can be expressed as the sum of n i.i.d. random variables. Therefore, asymptotically, as $n \to \infty$, by the central limit theorem , $U(\theta)$ is normally distributed. Furthermore, for the unknown true θ we know that $E[U(\theta)] = 0$ and $Var[U(\theta)] = \mathcal{I}(\theta)$, so $U(\theta)$ is asymptotically $N[0, \mathcal{I}(\theta)]$.

Now, a Taylor series expansion of $U(\hat{\theta})$ around the true θ gives

$$U(\hat{\theta}) = U(\theta) + (\hat{\theta} - \theta)U'(\theta) + \ldots$$

Now, $U(\hat{\theta}) = 0$, and if we approximate $U'(\theta) \equiv H(\theta)$ by $E[H(\theta)] \equiv -\mathcal{I}(\theta)$, and also ignore higher order terms,[1] we obtain

$$\hat{\theta} = \theta + \frac{1}{\mathcal{I}(\theta)} U(\theta)$$

As $U(\theta)$ is asymptotically $N[0, \mathcal{I}(\theta)]$, $\hat{\theta}$ is asymptotically $N[\theta, \mathcal{I}(\theta)^{-1}]$.

For 'large enough n', we can treat the asymptotic distribution of the m.l.e. as an approximation. The fact that $E(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{\theta}$ means that the maximum likelihood estimator is approximately *unbiased* (correct on average) for large samples. Furthermore, its variability, as measured by its variance $\mathcal{I}(\boldsymbol{\theta})^{-1}$ is the smallest possible amongst unbiased estimators, so the maximum

[1] This requires that $\hat{\theta}$ is close to θ in large samples, which is true but we do not prove it here.

likelihood has good precision. Therefore the m.l.e. is a desirable estimator in large samples (and therefore presumably also reasonable in small samples).

The usefulness of an estimate is always enhanced if some kind of measure of its precision can also be provided. Usually, this will be a *standard error*, an estimate of the standard deviation of the associated estimator. For the maximum likelihood estimator $\hat{\theta}$, a standard error is given by

$$s.e.(\hat{\theta}) = \frac{1}{\mathcal{I}(\hat{\theta})^{\frac{1}{2}}},$$

and for a vector parameter $\boldsymbol{\theta}$

$$s.e.(\hat{\theta}_i) = [\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}]_{ii}^{\frac{1}{2}} \qquad i = 1, \ldots, p.$$

An alternative summary of the information provided by the observed data about the location of a parameter θ and the associated precision is an *interval estimate* or *confidence interval*.

The asymptotic distribution of the maximum likelihood estimator can be used to provide approximate large sample confidence intervals. Asymptotically, $\hat{\theta}_i$ has a $N(\theta_i, [\mathcal{I}(\boldsymbol{\theta})^{-1}]_{ii})$ distribution and we can find h such that

$$P\left(h \leq \frac{\hat{\theta}_i - \theta_i}{[\mathcal{I}(\boldsymbol{\theta})^{-1}]_{ii}^{\frac{1}{2}}} \leq h\right) = \alpha.$$

Therefore

$$P\left(\hat{\theta}_i - h[\mathcal{I}(\boldsymbol{\theta})^{-1}]_{ii}^{\frac{1}{2}} \leq \theta_i \leq \hat{\theta}_i + h[\mathcal{I}(\boldsymbol{\theta})^{-1}]_{ii}^{\frac{1}{2}}\right) = \alpha.$$

The endpoints of this interval cannot be evaluated because they also depend on the unknown parameter vector $\boldsymbol{\theta}$. However, if we replace $\mathcal{I}(\boldsymbol{\theta})$ by its m.l.e. $\mathcal{I}(\hat{\boldsymbol{\theta}})$ we obtain the approximate large sample $100\alpha\%$ confidence interval

$$[\hat{\theta}_i - h[\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}]_{ii}^{\frac{1}{2}}, \hat{\theta}_i + h[\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}]_{ii}^{\frac{1}{2}}].$$

For $\alpha = 0.9, 0.95, 0.99$, $h = 1.64, 1.96, 2.58$.

♡ **Example** 1.9. If $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. Bernoulli(p) random variables then asymptotically $\hat{p} = \overline{Y}$ has a $N(p, p(1-p)/n)$ distribution, and a large sample 95% confidence interval for p is

$$\begin{aligned} & [\hat{p} - 1.96[\mathcal{I}(\hat{p})^{-1}]^{\frac{1}{2}}, \hat{p} + 1.96[\mathcal{I}(\hat{p})^{-1}]^{\frac{1}{2}}] \\ = \; & [\hat{p} - 1.96[\hat{p}(1-\hat{p})/n]^{\frac{1}{2}}, \hat{p} + 1.96[\hat{p}(1-\hat{p})/n]^{\frac{1}{2}}] \\ = \; & [\overline{y} - 1.96[\overline{y}(1-\overline{y})/n]^{\frac{1}{2}}, \overline{y} + 1.96[\overline{y}(1-\overline{y})/n]^{\frac{1}{2}}] \end{aligned}$$

1.8 Comparing statistical models

If we have a set of competing probability models which might have generated the observed data, we may want to determine which of the models is most appropriate. In practice, we proceed by comparing models pairwise. Suppose that we have two competing alternatives, $f_{\mathbf{Y}}^{(0)}$ (model H_0) and $f_{\mathbf{Y}}^{(1)}$ (model H_1) for $f_{\mathbf{Y}}$, the joint distribution of $Y_1, \ldots, Y_n$. The most common situation is where H_0 and H_1 both take the same parametric form, $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$ but with $\boldsymbol{\theta} \in \Theta^{(0)}$ for H_0 and $\boldsymbol{\theta} \in \Theta^{(1)}$ for H_1, where $\Theta^{(0)}$ and $\Theta^{(1)}$ are alternative sets of possible values for $\boldsymbol{\theta}$.

A hypothesis test provides a mechanism for comparing the two competing statistical models, H_0 and H_1. A hypothesis test does not treat the two hypotheses (models) symmetrically. One hypothesis, H_0, is accorded special status, and referred to as the *null hypothesis*. The null hypothesis is the reference model, and will be assumed to be appropriate unless the observed data strongly indicate that H_0 is inappropriate, and that H_1 (the *alternative* hypothesis) should be preferred.

Hence, the fact that a hypothesis test does not reject H_0 should not be taken as evidence that H_0 is true and H_1 is not, or that H_0 is better supported by the data than H_1, merely that the data does not provide significant evidence to reject H_0 in favour of H_1.

A hypothesis test is defined by its *critical region* or *rejection region*, which we shall denote by C. C is a subset of $\mathcal{R}^n$ and is the set of possible $\mathbf{y}$ which would lead to rejection of H_0 in favour of H_1, *i.e.*

If $\mathbf{y} \in C$	H_0 is rejected in favour of H_1
If $\mathbf{y} \notin C$	H_0 is not rejected

As $\mathbf{Y}$ is a random variable, there remains the possibility that a hypothesis test will produce in an erroneous result. We define

$$\begin{aligned} \alpha &= \max_{\boldsymbol{\theta} \in \Theta^{(0)}} P(\mathbf{Y} \in C; \boldsymbol{\theta}) \\ \omega(\boldsymbol{\theta}) &= P(\mathbf{Y} \in C; \boldsymbol{\theta}) \end{aligned}$$

We call α the *size* (or *significance level*) of the test; it is the maximum probability of erroneously rejecting H_0, over all possible distributions for $\mathbf{Y}$ implied by H_0. The function $\omega(\boldsymbol{\theta})$ is called the *power function*. It represents the probability of rejecting H_0 for a particular value of $\boldsymbol{\theta}$. Clearly we would like to find a test with where $\omega(\boldsymbol{\theta})$ is large for every $\boldsymbol{\theta} \in \Theta^{(1)} \setminus \Theta^{(0)}$, while at the same time avoiding erroneous rejection of H_0. In other words, a good test will

have small size, but large power.

The general hypothesis testing procedure is to fix α to be some small value (often 0.05), so that the probability of erroneous rejection of H_0 is limited. In doing this, we are giving H_0 precedence over H_1. Given our specified α, we try to choose a test, defined by its rejection region C, to make $\omega(\boldsymbol{\theta})$ as large as possible for $\boldsymbol{\theta} \in \Theta^{(1)} \setminus \Theta^{(0)}$.

Suppose that H_0 and H_1 both take the same parametric form, $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta^{(0)}$ for H_0 and $\boldsymbol{\theta} \in \Theta^{(1)}$ for H_1, where $\Theta^{(0)}$ and $\Theta^{(1)}$ are alternative sets of possible values for $\boldsymbol{\theta}$. A *generalised likelihood ratio test* of H_0 against H_1 has a critical region of the form

$$C = \left\{ \mathbf{y} : \frac{\max_{\boldsymbol{\theta} \in \Theta^{(1)}} f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \Theta^{(0)}} f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})} > k \right\}$$

where k is determined by α, the size of the test, so

$$\max_{\boldsymbol{\theta} \in \Theta^{(0)}} P(\mathbf{y} \in C; \boldsymbol{\theta}) = \alpha.$$

Therefore, we will only reject H_0 if H_1 offers a distribution for $Y_1, \ldots, Y_n$ which makes the observed data are more much more probable than any distribution under H_0. This is intuitively appealing and tends to produce good tests (large power) across a wide range of examples.

♡ **Example** 1.10. $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. Bernoulli(p) random variables. Suppose that we require a size α test of the hypothesis H_0: $p = p_0$ against the general alternative H_1: 'p is unrestricted' where α and p_0 are specified.

Here $\boldsymbol{\theta} = (p)$, $\Theta^{(0)} = \{p_0\}$ and $\Theta^{(1)} = (0, 1)$ and the generalised likelihood ratio test rejects H_0 when

$$\begin{aligned} & \frac{\max_{p \in (0,1)} f_{\mathbf{Y}}(\mathbf{y}; p)}{\max_{p = p_0} f_{\mathbf{Y}}(\mathbf{y}; p)} > k \\ \Rightarrow \quad & \frac{\overline{y}^{\Sigma y_i} (1 - \overline{y})^{n - \Sigma y_i}}{p_0^{\Sigma y_i} (1 - p_0)^{n - \Sigma y_i}} > k \\ \Rightarrow \quad & \left(\frac{\overline{y}}{p_0} \right)^{n\overline{y}} \left(\frac{1 - \overline{y}}{1 - p_0} \right)^{n(1 - \overline{y})} > k. \end{aligned}$$

Now the left hand side of (1.1), is minimised as a function of $\overline{y}$ at $\overline{y} = p_0$ and increases as $\overline{y}$ moves away from p_0 in either direction. Therefore, the rejection region (1.1) is equivalent to

$$C = \{ \mathbf{y} : \overline{y} > k' \text{ or } \overline{y} < k'' \}$$

where k' and k'' are chosen so that

$$P(\mathbf{y} \in C; p_0) = \alpha.$$

Therefore, we can use the binomial(n, p_0) distribution to find a precise rejection region for a test of specified size α.

Alternatively, if n is large, we can use the asymptotic distribution of $\overline{Y}$, $\mathrm{N}(p_0, p_0[1-p_0]/n)$

1.8.1 The log-likelihood ratio statistic

A *generalised likelihood ratio test* of H_0 against H_1 has a critical region of the form

$$C = \left\{ \mathbf{y} : \frac{\max_{\boldsymbol{\theta} \in \Theta^{(1)}} f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \Theta^{(0)}} f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})} > k \right\}$$

where k is determined by α, the size of the test, so

$$\max_{\boldsymbol{\theta} \in \Theta^{(0)}} P(\mathbf{y} \in C; \boldsymbol{\theta}) = \alpha.$$

Therefore, in order to determine k, we need to know the distribution of the likelihood ratio, or an equivalent statistic, under H_0. In general, this will not be available to us. However, we can make use of an important asymptotic result.

First we notice that, as log is a strictly increasing function, the rejection region is equivalent to

$$C = \left\{ \mathbf{y} : 2 \log \left(\frac{\max_{\boldsymbol{\theta} \in \Theta^{(1)}} f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \Theta^{(0)}} f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})} \right) > k' \right\}$$

where

$$\max_{\boldsymbol{\theta} \in \Theta^{(0)}} P(\mathbf{y} \in C; \boldsymbol{\theta}) = \alpha.$$

Now, provided that H_0 is *nested within* H_1, in other words $\Theta^{(0)} \subset \Theta^{(1)}$ ($\Theta^{(0)}$ is a subspace of $\Theta^{(1)}$) then under H_0: $\boldsymbol{\theta} \in \Theta^{(0)}$, asymptotically as $n \to \infty$

$$L_{01} \equiv 2 \log \left(\frac{\max_{\boldsymbol{\theta} \in \Theta^{(1)}} f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \Theta^{(0)}} f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})} \right)$$

has a chi-squared distribution with degrees of freedom equal to the difference in the dimensions of $\Theta^{(1)}$ and $\Theta^{(0)}$.

Proof First we note that in the case where $\boldsymbol{\theta}$ is one-dimensional and $\boldsymbol{\theta} = (\theta)$, a Taylor series expansion of $\log f_{\mathbf{Y}}(\mathbf{y};\theta)$ around around the m.l.e. $\hat{\theta}$ gives

$$\log f_{\mathbf{Y}}(\mathbf{y};\theta) = \log f_{\mathbf{Y}}(\mathbf{y};\hat{\theta}) + (\theta - \hat{\theta})U(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 U'(\hat{\theta}) + \ \ldots$$

Now, $U(\hat{\theta}) = 0$, and if we approximate $U'(\hat{\theta}) \equiv H(\hat{\theta})$ by $E[H(\theta)] \equiv -\mathcal{I}(\theta)$, and also ignore higher order terms, we obtain

$$2[\log f_{\mathbf{Y}}(\mathbf{y};\hat{\theta}) - \log f_{\mathbf{Y}}(\mathbf{y};\theta)] = (\theta - \hat{\theta})^2 \mathcal{I}(\theta)$$

As $\hat{\theta}$ is asymptotically $N[\theta, \mathcal{I}(\theta)^{-1}]$, $(\theta - \hat{\theta})^2\mathcal{I}(\theta)$ is asymptotically χ^2_1, and hence so is $2[\log f_{\mathbf{Y}}(\mathbf{y};\hat{\theta}) - \log f_{\mathbf{Y}}(\mathbf{y};\theta)]$.

Similarly it can be shown that when $\boldsymbol{\theta} \in \Theta$, a multidimensional space, $2[\log f_{\mathbf{Y}}(\mathbf{y};\hat{\boldsymbol{\theta}}) - \log f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})]$ is asymptotically χ^2_p, where p is the dimension of Θ.

Now, suppose that H_0 is true and $\boldsymbol{\theta} \in \Theta^{(0)}$ and therefore $\boldsymbol{\theta} \in \Theta^{(1)}$. Furthermore, suppose that $\log f_{\mathbf{Y}}(\mathbf{y};\hat{\boldsymbol{\theta}})$ is maximised in $\Theta^{(0)}$ by $\hat{\boldsymbol{\theta}}^{(0)}$ and is maximised in $\Theta^{(1)}$ by $\hat{\boldsymbol{\theta}}^{(1)}$. Then

$$\begin{aligned} L_{01} &\equiv 2\log\left(\frac{\max_{\boldsymbol{\theta}\in\Theta^{(1)}} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}\in\Theta^{(0)}} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})}\right) \\ &= 2\log f_{\mathbf{Y}}(\mathbf{y};\hat{\boldsymbol{\theta}}^{(1)}) - 2\log f_{\mathbf{Y}}(\mathbf{y};\hat{\boldsymbol{\theta}}^{(0)}) \\ &= 2[\log f_{\mathbf{Y}}(\mathbf{y};\hat{\boldsymbol{\theta}}^{(1)}) - \log f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})] - 2[\log f_{\mathbf{Y}}(\mathbf{y};\hat{\boldsymbol{\theta}}^{(0)}) - \log f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})] \\ &= L_1 - L_0 \end{aligned}$$

Therefore $L_1 = L_{01} + L_0$ and we know that, under H_0, L_1 has a $\chi^2_{d_1}$ distribution and L_0 has a $\chi^2_{d_0}$ distribution. Furthermore, it is possible to show (although we will not do so here) that under H_0, L_{01} and L_0 are independent. Therefore, from the properties of the chi-squared distribution, it follows that under H_0, the log likelihood ratio statistic L_{01} has a $\chi^2_{d_1-d_0}$ distribution.

♡ **Example** 1.11. $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. Bernoulli(p) random variables. Suppose that we require a size α test of the hypothesis H_0: $p = p_0$ against the general alternative H_1: 'p is unrestricted' where α and p_0 are specified.

Here $\boldsymbol{\theta} = (p)$, $\Theta^{(0)} = \{p_0\}$ and $\Theta^{(1)} = (0,1)$ and the log likelihood ratio statistic is

$$L_{01} = 2n\overline{y}\log\left(\frac{\overline{y}}{p_0}\right) + 2n(1-\overline{y})\log\left(\frac{1-\overline{y}}{1-p_0}\right).$$

As $d_1 = 1$ and $d_0 = 0$, under H_0, the log-likelihood ratio statistic has an asymptotic χ^2_1 distribution. For a log likelihood ratio test, we only reject H_0 in favour of H_1 when the test statistic is too large (observed data are much more probable under model H_1 than under model H_0), so in this case we reject H_0, when the observed value of the test statistic above is 'too large' to have come from a χ^2_1 distribution. What we mean by 'too large' depends on the significance level α of the test. For example, if $\alpha = 0.05$, a common choice, then we should reject H_0 if the test statistic is greater than the 3.84, the 95% point of the χ^2_1 distribution.

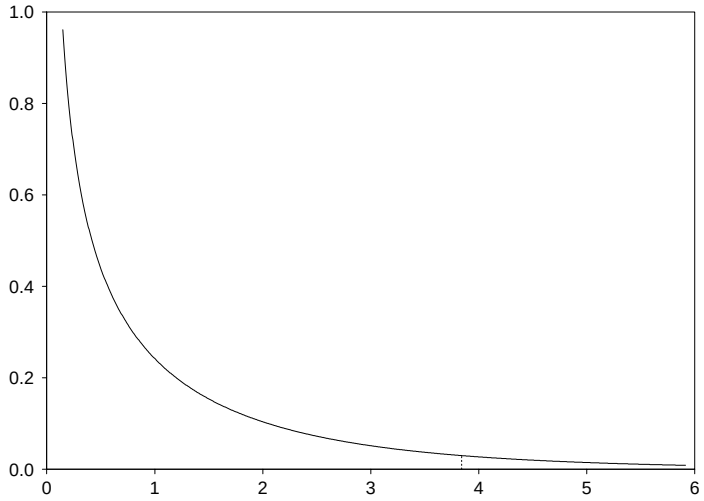

Figure 1.5: The χ^2_1 distribution

1.9 Linear Models (a brief revision)

1.9.1 Introduction

In practical applications, we often distinguish between a *response* variable and a group of *explanatory* variables. The aim is to determine the pattern of dependence of the response variable on the explanatory variables. We denote the n observations of the response variable by $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$. In a statistical model, these are assumed to be observations of *random variables* $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$. Associated with each y_i is a vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ of values of p explanatory variables.

Linear models are those for which the relationship between the response and explanatory variables is of the form

$$\begin{aligned} E(Y_i) &= \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \\ &= \sum_{j=1}^{p} x_{ij}\beta_j \\ &= \mathbf{x}_i^T \boldsymbol{\beta} \\ &= [\mathbf{X}\boldsymbol{\beta}]_i, \qquad i = 1, \ldots, n \end{aligned} \tag{1}$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of fixed but unknown parameters describing the dependence of Y_i on $\mathbf{x}_i$. The four ways of describing the linear model in (1) are equivalent, but the most economical is the matrix form

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}. \tag{2}$$

The $n \times p$ matrix $\mathbf{X}$ consists of known (observed) constants and is called the *design matrix*. The ith row of $\mathbf{X}$ is $\mathbf{x}_i^T$, the explanatory data corresponding to the ith observation of the response. The jth column of $\mathbf{X}$ contains the n observations of the jth explanatory variable.

♡ **Example** 1.12. The null model

$$E(Y_i) = \beta_1 \qquad i = 1, \ldots, n$$

$$\mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \qquad \boldsymbol{\beta} = (\beta_1).$$

One (dummy) explanatory variable. In practice, this variable is present in all models.

♡ **Example** 1.13. Simple linear regression

$$E(Y_i) = \beta_1 + \beta_2 x_i \qquad i = 1, \ldots, n$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

Two explanatory variables; the dummy variable and one 'real' variable.

♡ **Example** 1.14. **Polynomial regression**

$$E(Y_i) = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \ldots + \beta_p x_i^{p-1} \qquad i = 1, \ldots, n$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{p-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{p-1} \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

p explanatory variables; the dummy variable and one 'real' variable, transformed to $p-1$ variables.

♡ **Example** 1.15. **Multiple regression**

$$E(Y_i) = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \ldots + \beta_p x_{i\,p-1} \qquad i = 1, \ldots, n$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1\,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2\,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n\,p-1} \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

p explanatory variables; the dummy variable and $p-1$ 'real' variables.

Strictly, the only requirement for a model to be linear is that the relationship between the response variables, $\mathbf{Y}$, and any explanatory variables can be written in the form (2). No further specification of the joint distribution of $Y_1, \ldots, Y_n$ is required. However, the linear model is more useful for statistical analysis if we can make three further assumptions:

1. $Y_1, \ldots, Y_n$ are independent random variables.
2. $Y_1, \ldots, Y_n$ are normally distributed.
3. $Var(Y_1) = Var(Y_2) = \cdots = Var(Y_n)$ ($Y_1, \ldots, Y_n$ are homoscedastic).

We denote this common variance by σ^2

With these assumptions the linear model completely specifies the distribution of $\mathbf{Y}$, in that $Y_1, \ldots, Y_n$ are independent and

$$Y_i \sim N\left(\mathbf{x}_i^T \boldsymbol{\beta}\ ,\ \sigma^2\right) \qquad i = 1, \ldots, n.$$

Another way of writing this is

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} \; + \; \epsilon_i \qquad i = 1, \ldots, n$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathrm{N}(0, \sigma^2)$ random variables.

A linear model can now be expressed in matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ has a multivariate normal distribution with mean vector $\mathbf{0}$ and variance covariance matrix $\sigma^2\mathbf{I}$, (because all $Var(\epsilon_i) = \sigma^2$ and $\epsilon_1, \ldots, \epsilon_n$ are independent implies all $Cov(\epsilon_i, \epsilon_j) = 0$). It follows from (3) that the distribution of $\mathbf{Y}$ is multivariate normal with mean vector $\mathbf{X}\boldsymbol{\beta}$ and variance covariance matrix $\sigma^2\mathbf{I}$, *i.e.* $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

1.9.2 Least squares estimation

The regression coefficients $\beta_1, \ldots, \beta_p$ describe the pattern by which the response depends on the explanatory variables. We use the observed data $y_1, \ldots, y_n$ to *estimate* this pattern of dependence.

In least squares estimation, roughly speaking, we choose $\hat{\boldsymbol{\beta}}$, the estimates of $\boldsymbol{\beta}$ to make the fitted values $\hat{E}(\mathbf{Y}) = \mathbf{X}\hat{\boldsymbol{\beta}}$ *as close as possible* to the observed values $\mathbf{y}$, *i.e.* $\hat{\boldsymbol{\beta}}$ minimises the sum of squares

$$\begin{aligned} \sum_{i=1}^{n} [y_i - E(Y_i)]^2 &= \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \end{aligned}$$

as a function of $\beta_1, \ldots, \beta_p$. The sum of squares may also be written as $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.

Differentiating w.r.t. $\beta_k, \; k = 1, \ldots, p$, and setting equal to 0 gives

$$\begin{aligned} -2\sum_{i=1}^{n} x_{ik}\left(y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j\right) &= 0 \qquad k = 1, \ldots, p \\ \Rightarrow \qquad \sum_{i=1}^{n} x_{ik} y_i &= \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ik}x_{ij}\hat{\beta}_j \qquad k = 1, \ldots, p \end{aligned}$$

$$\begin{aligned} \Rightarrow \quad [\mathbf{X}^T\mathbf{y}]_k &= [\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}]_k \\ \Rightarrow \quad \mathbf{X}^T\mathbf{y} &= \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

The least squares estimates $\hat{\boldsymbol{\beta}}$ are the solutions to this set of p simultaneous linear equations, which are known as the *normal equations*. If $\mathbf{X}^T\mathbf{X}$ is invertible (as it usually is) then the least squares estimates are given by

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

The corresponding fitted values are

$$\begin{aligned} & \hat{E}(\mathbf{Y}) = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ \Rightarrow \quad & \hat{E}(Y_i) = \mathbf{x}_i^T\hat{\boldsymbol{\beta}} \qquad i = 1, \ldots, n. \end{aligned}$$

Recalling that $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, we notice that the least squares estimates for $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ are obtained as the difference between the observed and fitted values

$$\begin{aligned} & \hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ \Rightarrow \quad & \hat{\epsilon}_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}} \qquad i = 1, \ldots, n. \end{aligned}$$

$\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ describe the variability in the observed responses $y_1, \ldots, y_n$ which has not been explained by the linear model. It is residual variability, and we call $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$, the residuals. We call

$$\begin{aligned} D &= \sum_{i=1}^{n} \hat{\epsilon}_i^2 \\ &= \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}} \right)^2 \end{aligned}$$

the *residual sum of squares* or *deviance* for the linear model. It is the actual minimum value attained in the least squares estimation.

1.9.3 Properties of the least squares estimator

1. $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$

This follows by recalling that $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is a linear function of $\mathbf{y}$. Therefore, from the properties of expectation and variance of a vector random variable we have

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E[\mathbf{Y}] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned}$$

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= Var[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Var[\mathbf{Y}][(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{I}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \end{aligned}$$

2. Assuming that $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathrm{N}(0, \sigma^2)$ the least squares estimate $\hat{\boldsymbol{\beta}}$ is also the maximum likelihood estimate.

   This is obvious when one considers the likelihood for a linear model

   $$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2\right). \tag{4}$$

3. $\hat{\boldsymbol{\beta}}$ is multivariate normal (with mean and variance given above).

   As $\mathbf{Y}$ is normally distributed, and $\hat{\boldsymbol{\beta}}$ is a linear function of $\mathbf{Y}$, then $\hat{\boldsymbol{\beta}}$ must also be normally distributed.

1.9.4 Estimation of σ^2

In addition to the linear coefficients $\beta_1, \ldots, \beta_p$ estimated using least squares, a linear model usually involves the unknown *residual variance* σ^2, representing the variability of observations about their mean.

We can estimate σ^2 using maximum likelihood. Maximising (4) with respect to $\boldsymbol{\beta}$ and σ^2 gives

$$\hat{\sigma}^2 = \frac{D}{n} = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2.$$

It is possible to prove (although we shall not do so here) that if the model is correct,

$$\frac{D}{\sigma^2} \sim \chi^2_{n-p}$$

which implies that

$$E(\hat{\sigma}^2) = \frac{n-p}{n}\sigma^2,$$

so the maximum likelihood estimate is biased for σ^2 (although still asymptotically unbiased as $\frac{n-p}{n} \to 1$ as $n \to \infty$). We usually prefer to use the unbiased estimate of σ^2

$$\tilde{\sigma}^2 = \frac{D}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

The denominator $n-p$, the number of observations minus the number of linear coefficients in the model is called the *degrees of freedom* of the model. Therefore, we estimate the residual variance by the deviance divided by the degrees of freedom.

1.9.5 Comparing linear models

If we have a set of competing linear models which might explain the dependence of the response on the explanatory variables, we will want to determine which of the models is most appropriate. Recall that we have three main requirements of a statistical model; plausibility, parsimony and goodness of fit, of which parsimony and goodness of fit are statistical issues.

The goodness of fit of a linear model to the observed data is encapsulated by its deviance or residual sum of squares. Models which fit the data well have a low deviance, whereas those which fit the data poorly have a high deviance. However, the calibration of 'low' and 'high' is unclear, and depends on the scale of measurement of the response. We can calibrate the deviance by dividing it by the *natural variation in the data*, $\sum_{i=1}^{n}(y_i - \overline{y})^2$. Then

$$R^2 = 1 - \frac{D}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

is the proportion of natural variation in the data which has been accounted for, or explained, by the linear model.

R^2 is still not an entirely satisfactory measure by which to compare models (although it is an easily interpretable summary of the goodness-of-fit of a selected model) because adding terms to a model increases R^2 whether or not the increased complexity is justified.

The C_p statistic combines the deviance (a measure of goodness of fit) and the number of linear parameters of the model (a measure of complexity) into an overall measure for a model. The C_p statistic calculated by S-Plus is

$$C_p = D + 2p\sigma^2$$

which clearly penalises a model for lack of fit (through D) and complexity (through p, the number of linear parameters). Here σ^2 is replaced by the most reliable estimate of σ^2 available, in practice the estimate based on the most complex model under consideration.

Although C_p provides a mechanism for comparing any pair of models, it is essentially an *ad hoc* summary measure. In practice, we use it as a guide to direct a model comparison strategy based on hypothesis tests.

As described previously, we proceed by comparing models pairwise using a generalised likelihood ratio test. For linear models this kind of comparison is restricted to situations where one of the models, H_0, is *nested* in the other, H_1. For linear models, this usually means that the explanatory variables present in H_0 are a subset of those present in H_1. In this case model H_0 is a special case of model H_1, where certain coefficients are set equal to zero. We let $\boldsymbol{\theta}$ represent the collection of linear parameters for model H_1, together with the residual variance σ^2, and let $\Theta^{(1)}$ be the unrestricted parameter space for $\boldsymbol{\theta}$. Then $\Theta^{(0)}$ is the parameter space corresponding to model H_0, *i.e.* with the appropriate coefficients constrained to zero.

We will assume that model H_1 contains p linear parameters and model H_0 a subset of $q < p$ of these. Without loss of generality, we can think of H_1 as the model

$$E(Y_i) = \sum_{j=1}^{p} x_{ij}\beta_j \qquad i = 1, \ldots, n$$

and H_0 being the same model with

$$\beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0.$$

Now, a *generalised likelihood ratio test* of H_0 against H_1 has a critical region of the form

$$C = \left\{ \mathbf{y} : \frac{\max_{(\boldsymbol{\beta},\sigma^2)\in\Theta^{(1)}} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\beta},\sigma^2)}{\max_{(\boldsymbol{\beta},\sigma^2)\in\Theta^{(0)}} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\beta},\sigma^2)} > k \right\}$$

where k is determined by α, the size of the test, so

$$\max_{\boldsymbol{\theta}\in\Theta^{(0)}} P(\mathbf{y} \in C;\boldsymbol{\beta},\sigma^2) = \alpha.$$

For a linear model,

$$f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\beta},\sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2\right).$$

This is maximised with respect to $(\boldsymbol{\beta},\sigma^2)$ at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and $\sigma^2 = \hat{\sigma}^2 = D/n$. Therefore

$$\begin{aligned}\max_{\boldsymbol{\beta},\sigma^2} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\beta},\sigma^2) &= (2\pi D/n)^{-\frac{n}{2}} \exp\left(-\frac{n}{2D}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2\right) \\ &= (2\pi D/n)^{-\frac{n}{2}} \exp\left(-\frac{n}{2}\right)\end{aligned}$$

Let the deviances under models H_0 and H_1 be denoted D_0 and D_1 respectively. Then the critical region for the generalised likelihood ratio test is of the form

$$\begin{aligned}&\frac{(2\pi D_1/n)^{-\frac{n}{2}}}{(2\pi D_0/n)^{-\frac{n}{2}}} > k \\ \Rightarrow \quad &\left(\frac{D_0}{D_1}\right)^{\frac{n}{2}} > k \\ \Rightarrow \quad &\left(\frac{D_0}{D_1} - 1\right)\frac{n-p}{p-q} > k' \\ \Rightarrow \quad &\frac{(D_0 - D_1)/(p-q)}{D_1/(n-p)} > k'.\end{aligned}$$

We refer to the left hand side of this inequality as the F-statistic. We reject the simpler model H_0 in favour of the more complex model H_1 if F is 'too large'.

As we have required H_0 to be nested in H_1, F has a known distribution when H_0 is true. It is an F distribution with $p-q$ degrees of freedom in the numerator and $n-p$ degrees of freedom in the denominator. To see this, note that

$$\frac{D_0}{\sigma^2} = \frac{D_0 - D_1}{\sigma^2} + \frac{D_1}{\sigma^2}.$$

Furthermore, we know from §1.4.3 that, under H_0, D_1/σ^2 has a χ^2_{n-p} distribution and D_0/σ^2 has a χ^2_{n-q} distribution. It is possible to show (although we will not do so here) that under H_0, $(D_0 - D_1)/\sigma^2$ and D_0/σ^2 are independent. Therefore, from the properties of the chi-squared distribution, it follows that under H_0, $(D_0 - D_1)/\sigma^2$ has a χ^2_{p-q} distribution, and F has a $F_{p-q,\,n-p}$ distribution.

Therefore, the precise critical region can be evaluated given the size, α, of the test. We reject H_0 in favour of H_1 when

$$\frac{(D_0 - D_1)/(p-q)}{D_1/(n-p)} > k$$

where k is the $100(1-\alpha)\%$ point of the $F_{p-q,\,n-p}$ distribution.

1.9.6 Assessing a selected model

An easily interpreted overall measure of goodness of fit of a model is provided by the R^2 coefficient, discussed in §1.4.4. If predictions are required, then confidence intervals for the predictions will provide an appropriate measure for assessing the likely accuracy of the model. Remember that it is always risky to use a model to make predictions for values of the explanatory variables which are not in the range of those which were used to fit the model.

Confidence intervals and hypothesis tests for linear models may be unreliable if all the model assumptions are not justified. In particular, we have made three assumptions about the distribution of $Y_1, \ldots, Y_n$.

1. $Y_1, \ldots, Y_n$ are independent random variables.
2. $Y_1, \ldots, Y_n$ are normally distributed.
3. $Var(Y_1) = Var(Y_2) = \cdots = Var(Y_n)$.

The validity of these assumptions can be checked using residual plots.

1. In general, independence is difficult to validate, but where observations have been collected in serial order, serial correlation may be detected by a plot of the residuals $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ against the serial order $i = 1, \ldots, n$.
2. A simple check for non-normality is obtained by plotting the ordered residuals against the expected order statistics of a sample of size n from a standard normal distribution. The plot should look like a straight line. Beware of any obvious curves in the plot.
3. A simple check for non-constant variance is obtained by plotting the residuals $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ against the corresponding fitted values $\mathbf{x}_i^T\hat{\boldsymbol{\beta}}, \quad i = 1, \ldots, n$. The plot should look like a random scatter. Beware of 'funneling'.

Other residual plots may also be useful. For example, if a plot of the residuals against the values of an explanatory variable reveals a pattern, then this suggests that the explanatory variable, or perhaps some function of it, should be included in the model.

Another place where residual diagnostics are useful is in assessing *influence*. An observation is influential if deleting it would lead to estimates of model parameters being substantially changed. S-Plus calculates Cook's distance for each observation. Cook's distance is a measure of the change in $\hat{\boldsymbol{\beta}}$ when the observation is omitted from the dataset.

$$\text{Cook's distance} = \frac{(\hat{\boldsymbol{\beta}}' - \hat{\boldsymbol{\beta}})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}}' - \hat{\boldsymbol{\beta}})}{p\hat{\sigma}^2}$$

where $\hat{\boldsymbol{\beta}}'$ is the least squares estimate of $\boldsymbol{\beta}$ based on the observed data with the ith observation omitted.

A rule of thumb is that a Cook's distance of 1 or more indicates a potentially important change in $\hat{\boldsymbol{\beta}}$ and may be worthy of further investigation.