Introduction to Intelligent Data Analysis

Ad Feelders

November 18, 2003

Contents

Introduction

The course *Introduction to Intelligent Data Analysis* (IDA) has been developed to fill the gap between an introductory bachelor course in probability and statistics, and the more advanced data analysis courses, such as statistical learning and data mining. IDA is a required course for the CI-track of the ACI master program.

Basically, IDA is an introductory statistical data analysis course, but we thought it would be a good idea marketingwise to have te word *intelligent* in the title. The central topic of the course is modelling relations between variables. We estimate (*learn* in AI terminology) these models from a set of observations

$$T = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}.$$

Here n is the number of observations in training sample T. More specifically we are looking at the relation between one or more explanatory (also called independent) variables $\mathbf{x} = x_1, x_2, \ldots, x_p$, and a response (also called dependent) variable y.

To illustrate, we consider an example from economics. Suppose we want to explain or predict the weekly expenditures on food (y) of households. Common sense suggests that the number of people in the household (x_1), and the weekly household income (x_2) may be relevant predictors for y:

$$y = f(x_1, x_2) + \varepsilon \tag{1}$$

Here f is an as yet unspecified function: our common sense does not reveal a functional form for the relation. Note that the relation between y and x_1, x_2 is not deterministic, but contains a random component ε. What this says is that even when we know the value of x_1 and x_2, the value of y is not uniquely determined. Suppose we were to gather data on all households with four people and weekly income of €500. We do not expect all these households to spend exactly the same amount on food! This variation is due

to the fact that we normally can not include all the variables that influence y. But we want to include the *important* influences in the model. Equation (1) is still not detailed enough, and to get a grip we typically (at least initially) assume that the relation between y and its predictors is linear:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{2}$$

To estimate equation (2) from data we can use a technique called linear regression. This is more or less the work horse of data analysis, and we will be spending quite some time on it. We will also see that the assumption of linearity is not so restrictive as it may seem at first sight. The reason to devote quite some time to the linear regression model is that we can learn a lot about general problems of data analysis by studying this relatively simple model. General issues such as the problem of overfitting and model specification can be analysed and understood more easily for linear regression than for complex models such as neural networks or other advanced data analysis techniques.

Linear regression is applicable when the dependent variable is numeric and can take on a lot of different values, such as in the food expenditure example. There are also a lot of problems where the dependent variable is a yes/no variable. Suppose, for example, that we want to predict whether an incoming e-mail message is spam or not. Anyone who has received spam can think of a number of good indicators, e.g. the fraction of capital letters in the message, or the number of occurrences of the word "free".

We could label a number of e-mail messages "by hand", i.e. we could study the messages and decide whether its spam or not, and use this data set to estimate a model

$$P(y = 1) = f(x_1, x_2) \tag{3}$$

Here $y = 1$ denotes a spam-message, and $y = 0$ a non-spam message. What we are saying in equation (3) is that the probability that we are dealing with a spam message depends on the value of x_1 (fraction of capital letters) and x_2 (number of occurrences of "free"). Again, we have not yet specified the functional form of this relationship. A popular technique for this type of problem is *logistic regression*. A different technique that can by used for the same type of problem is called *discriminant analysis*. Both approaches will be discussed in this course.

In summary then, the three main modelling techniques we discuss are: linear regression, logistic regression and discriminant analysis. For all these

techniques we study the important issues of: model specification, interpretation, estimation, testing, prediction, and model selection.

We conclude this introduction with an overview of the following chapters. The first three chapters (Probability, Sampling and sampling distributions, and Statistical Inference) do not belong to the required literature of this course, but contain a short review of material that is assumed known in the course. The chapters are provided as a service to the student. Chapter four deals with linear regression, the model that will be treated in the most detail. In chapter five we discuss logistic regression, and in chapter six statistical discriminant analysis. In chapter seven, we discuss a number of computer intensive techniques such as cross-validation and bootstrapping.

Most of the statistical procedures we study follow the frequentist approach to statistical inference. In chapter eight we look at a different school of thought called Bayesian statistics.

Chapter 1

Probability

The most important tool in statistical inference is probability theory. This chapter provides a short review of the important concepts.

1.1 Random Experiments

A *random experiment* is an experiment that satisfies the following conditions

1. all possible distinct outcomes are known in advance,
2. in any particular trial, the outcome is not known in advance, and
3. the experiment can be repeated under identical conditions.

The *outcome space* Ω of an experiment is the set of all possible outcomes of the experiment.

Example 1 *Tossing a coin is a random experiment with outcome space* $\Omega = \{H, T\}$

Example 2 *Rolling a die is a random experiment with outcome space* $\Omega = \{1,2,3,4,5,6\}$

Something that might or might not happen, depending on the outcome of the experiment, is called an *event*. Examples of events are "coin lands heads" or "die shows an odd number". An event A is represented by a subset of the outcome space. For the above examples we have $A = \{\text{H}\}$ and $A = \{1,3,5\}$ respectively. Elements of the outcome space are called elementary events.

1.2 Classical definition of probability

If all outcomes in Ω are equally likely, the probability of A is the number of outcomes in A, which we denote by $N(A)$ divided by the total number of outcomes N

$$P(A) = \frac{N(A)}{N}$$

If all outcomes are equally likely, the probability of {H} in the coin tossing experiment is $\frac{1}{2}$, and the probability of {5,6} in the die rolling experiment is $\frac{1}{3}$. The assumption of equally likely outcomes limits the application of the concept of probability: what if the coin or die is not 'fair'? Nevertheless there are random experiments where this definition of probability is applicable, most importantly in the experiment of random selection of a unit from a population. This special and important kind of experiment is discussed in the section 2.

1.3 Frequency definition of probability

Recall that a random experiment may be repeated under identical conditions. When the number of trials of an experiment is increased indefinitely, the relative frequency of the occurrence of an event approaches a constant number. We denote the number of trials by n, and the number of times A occurs by $n(A)$. The frequency definition of probability states that

$$P(A) = \lim_{n\to\infty} \frac{n(A)}{n}$$

The law of large numbers states that this limit does indeed exist. For a small number of trials, the relative frequencies may show strong fluctuation as the number of trials varies. The fluctuations tend to decrease as the number of trials increases.

Figure 1.1 shows the relative frequencies of heads in a sequence of 1000 coin tosses as the sequence progresses. In the beginning there is quite some fluctuation, but as the sequence progresses, the relative frequency of heads settles around 0.5.

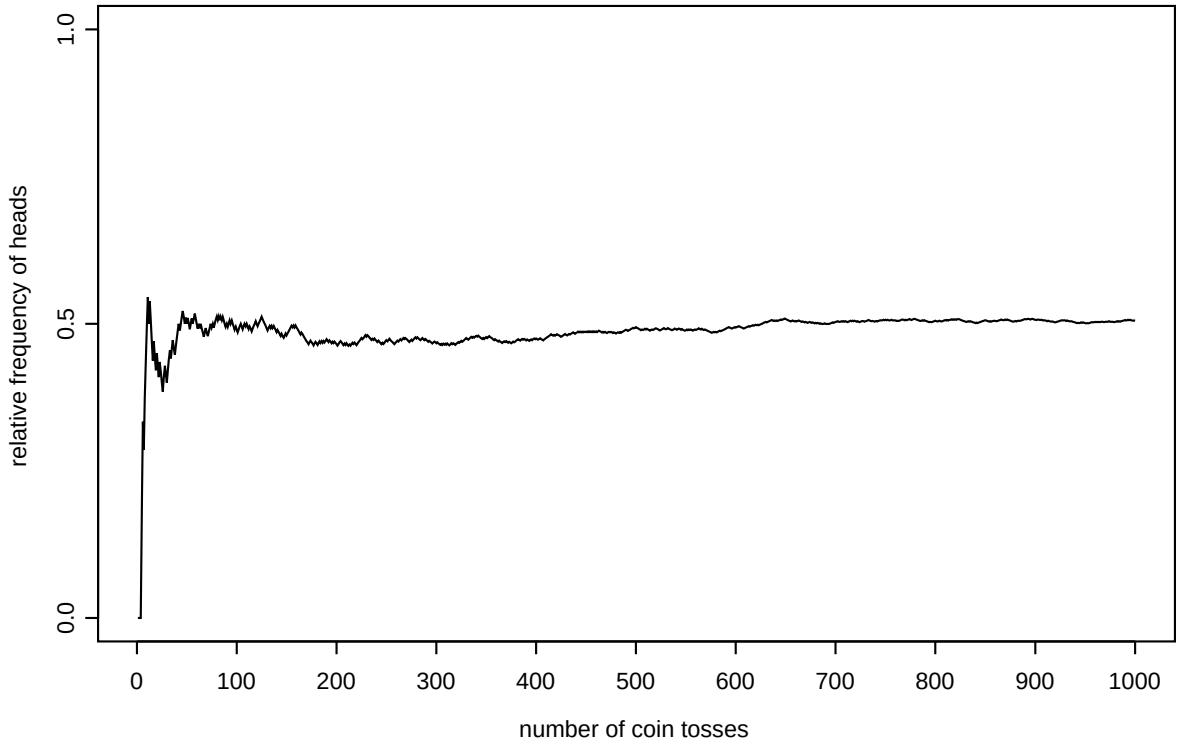

Figure 1.1: Relative frequency of heads in a sequence of 1000 coin tosses

1.4 Subjective definition of probability

Because of the demand of repetition under identical circumstances, the frequency definition of probability is not applicable to every event. According to the subjective definition, the probability of an event is a measure of the *degree of belief* that the event will occur (or has occured). Degree of belief depends on the person who has the belief, so my probability for event A may be different from yours.

Consider the statement: "There is extra-terrestrial life". The degree of belief in this statement could be expressed by a number between 0 and 1. According to the subjectivist definition we may interpret this number as the probability that there is extra-terrestrial life.

The subjective view allows the expression of all uncertainty through prob-

ability. This view has important implications for statistical inference (see section 8).

1.5 Probability axioms

Probability is defined as a function from subsets of Ω to the real line $\mathbb{R}$, that satisfies the following axioms

1. Non-negativity: $P(A) \geq 0$
2. Additivity: If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$
3. $P(\Omega) = 1$

The classical, frequency and subjective definitions of probability all satisfy these axioms. Therefore every property that may be deduced from these axioms holds for all three interpretations of probability.

1.6 Conditional probability and independence

The probability that event A occurs may be influenced by information concerning the occurrence of event B. The probability of event A, given that B will occur or has occurred, is called the *conditional probability* of A given B, denoted by $P(A \mid B)$. It follows from the axioms of probability that

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

for $P(B) > 0$. Intuitively we can appreciate this equality by considering that B effectively becomes the new outcome space. The events A and B are called *independent* if the occurrence of one event does not influence the probability of occurrence of the other event, i.e.

$$P(A \mid B) = P(A) \text{ , and consequently } P(B \mid A) = P(B)$$

Since independence of two events is always mutual, it is more concisely expressed by the product rule

$$P(A \cap B) = P(A)\, P(B)$$

1.7 Random variables

A random variable X is a *function* from the outcome space Ω to the real line

$$X : \Omega \to \mathbb{R}$$

Example 3 *Consider the random experiment of tossing a coin twice, and observing the faces turning up. The outcome space is*

$$\Omega = \{(H,T),(T,H),(H,H),(T,T)\}$$

The number of heads turning up is a random variable defined as follows

$$X((H,T)) = X((T,H)) = 1\,,\; X((H,H)) = 2\,,\; X((T,T)) = 0$$

1.8 Probability distribution

A probability function p assigns to each possible realisation x of a discrete random variable X the probability $p(x)$, i.e. $P(X = x)$. From the axioms of probability it follows that $p(x) \geq 0$, and $\sum_x p(x) = 1$.

Example 4 *The number of heads turning up in two tosses of a fair coin is a random variable with the following probability function:* $p(1) = 1/2$, $p(0) = 1/4$, $p(2) = 1/4$.

Since for continuous random variables, $P(X = x) = 0$, the concept of a probability function is useless. The probability distribution is now specified by representing probabilities as areas under a curve. The function $f : \mathbb{R} \to \mathbb{R}^+$ is called the probability density of X if for each pair $a \leq b$,

$$P(a < X \leq b) = \int_a^b f\,(x)\,dx$$

It follows from the probability axioms that $f\,(x) \geq 0$ and $\int_{-\infty}^{\infty} f\,(x)\,dx = 1$.

Example 5 *Consider the random variable X with the following density function*

$$f\,(x) = \begin{cases} \frac{1}{2} \text{ for } 0 \leq x \leq 2 \\ 0 \text{ otherwise} \end{cases}$$

It follows that

$$P(1/2 < X \leq 5/4) = \int_{1/2}^{5/4} 1/2dx = 1/2x|_{1/2}^{5/4} = 3/4$$

The *distribution function* is defined for both discrete and continuous random variables as the function F which gives for each $x \in \mathbb{R}$ the probability of an outcome of X at most equal to x:

$$F(x) = P(X \leq x), \quad \text{for } x \in \mathbb{R}$$

1.9 Expectation

For a discrete random variable, the *expected value* or mean is defined as

$$\mathrm{E}(X) = \sum_x x\, p(x) \text{ , and } \mathrm{E}[h(X)] = \sum_x h(x)\, p(x)$$

for arbitrary function $h : \mathbb{R} \to \mathbb{R}$.

Example 6 *Consider once more the coin tossing experiment of example 4 and corresponding probability distribution. The expected value or mean of X is*

$$\mathrm{E}(X) = 1/2 \cdot 1 + 1/4 \cdot 2 + 1/4 \cdot 0 = 1$$

The definition of expectation for a continuous random variable is analogous, with summation replaced by integration.

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx \text{ , and } \mathrm{E}[h(X)] = \int_{-\infty}^{\infty} h(x)\ f(x)\, dx$$

Example 7 *(Continuation of example 5) The mean or expected value of the random variable with probability density given in example 5 is*

$$\mathrm{E}(X) = \int_0^2 \frac{1}{2}\, dx = \frac{1}{2}x\Big|_0^2 = \frac{1}{2} \cdot 2 - \frac{1}{2} \cdot 0 = 1$$

The expected value $\mathrm{E}(X)$ of a random variable is usually denoted by μ. The variance σ^2 of a random variable is a measure of spread around the mean obtained by averaging the squared differences $(x - \mu)^2$, i.e.

$$\sigma^2 = \mathrm{V}(X) = \mathrm{E}(X - \mu)^2$$

The standard deviation $\sigma = \sqrt{\sigma^2}$ has the advantage that it has the same dimension as X.

x	4	6	8	10	12
$p(x \mid C)$	1/9	2/9	1/3	2/9	1/9

Table 1.1: Conditional probability function $p(x \mid C)$

1.10 Conditional probability distributions and expectation

For a discrete random variable X we define a conditional probability function as follows

$$p(x \mid C) = P(X = x \mid C) = \frac{P(\{X = x\} \cap C)}{P(C)}$$

Example 8 *Two fair dice are rolled, and the numbers on the top face are noted. We define the random variable X as the sum of the numbers showing. For example $X((3,2)) = 5$. Consider now the event C : both dice show an even number. We have $P(C) = \frac{1}{4}$ and $P(\{X = 6\} \cap C) = \frac{1}{18}$ since*

$$\begin{aligned} C &= \{(2,2),(2,4),(2,6),(4,2),(4,4),(4,6),(6,2),(6,4),(6,6)\} \\ \{X &= 6\} \cap C = \{(2,4),(4,2)\} \end{aligned}$$

The probability of $\{X = 6\}$ given C therefore is

$$P(X = 6 \mid C) = \frac{P(\{X = 6\} \cap C)}{P(C)} = \frac{1/18}{1/4} = \frac{2}{9}$$

The conditional probability function of X is shown in table 1.1. The conditional expectation of X given C is: $\mathrm{E}(X \mid C) = \sum_x x\, p(x \mid C) = 8$.

For continuous random variable X, the conditional density $f(x \mid C)$ of X given C is

$$f(x \mid C) = \begin{cases} f(x)/P(C) \text{ for } x \in C \\ 0 \text{ otherwise} \end{cases}$$

1.11 Joint probability distributions and independence

The joint probability distribution of a pair of discrete random variables (X, Y) is uniquely determined by their joint probability function $p : \mathbb{R}^2 \to \mathbb{R}$

$$p(x, y) = P((X, Y) = (x, y)) = P(X = x, Y = y)$$

From the axioms of probability it follows that $p(x, y) \geq 0$ and $\sum_x \sum_y p(x, y) = 1$.

The *marginal* probability function $p_X(x)$ is easily derived from the joint distribution

$$p_X(x) = p(X = x) = \sum_y P(X = x, Y = y) = \sum_y p(x, y)$$

The conditional probability function of X given $Y = y$

$$p(x \,|\, y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p(x, y)}{p_Y(y)}$$

Definitions for continuous random variables are analogous with summation replaced by integration. The function $f : \mathbb{R}^2 \to \mathbb{R}$ is the probability density of the pair of random variables (X, Y) if for all $a \leq b$ and $c \leq d$

$$P(a < X \leq b, c < Y \leq d) = \int_a^b \int_c^d f(x, y)\, dx\, dy$$

From the probability axioms it follows that

1. $f(x, y) \geq 0$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$

The marginal distribution of X is obtained from the joint distribution

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

and the conditional density of X given $\{Y = y\}$ is

$$f(x \,|\, y) = \frac{f(x, y)}{f_Y(y)}$$

According to the product rule discussed in section 1.6, the events $\{X = x\}$ and $\{Y = y\}$ are independent iff

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

We now generalize the concept of independence to pairs of random variables. Discrete random variables X and Y are independent iff

$$p(x, y) = p_X(x)p_Y(y) \text{ for all } (x, y),$$

and as a consequence $p(x \,|\, y) = p_X(x)$, and $p(y \,|\, x) = p_Y(y)$. Definitions are completely analogous for continuous random variables, with probability functions replaced by probability densities.

1.12 The law of total probability

In some cases the (unconditional) probability of an event may not be calculated directly, but can be determined as a weighted average of various conditional probabilities.

Let $B_1, B_2, \ldots, B_s$ be a partition of Ω, that is $B_i \cap B_j = \emptyset$ for all $i \neq j$ and $\bigcup_{i=1}^{s} B_i = \Omega$. It follows from the axioms of probability that

$$P(A) = \sum_{i=1}^{s} P(A|B_i)P(B_i)$$

Example 9 *Consider a box containing three white balls and one red ball. First we draw a ball at random, i.e. all balls are equally likely to be drawn from the box. Then a second ball is drawn at random (the first ball has not been replaced in the box). What is the probability that the second draw yields a red ball? This is most easily calculated by averaging conditional probabilities.*

$$P(R_2) = P(R_2|W_1)P(W_1) + P(R_2|R_1)P(R_1) = 1/3 \cdot 3/4 + 0 \cdot 1/4 = 1/4,$$

where R_i stands for "a red ball is drawn on i-th draw" and W_i for "a white ball is drawn on i-th draw".

	T^+	T^-
D	0.95	0.05
$\bar{D}$	0.02	0.98

Table 1.2: Performance of diagnostic test

1.13 Bayes' rule

Bayes' rule shows how probabilities change in the light of evidence. It is a very important tool in Bayesian statistical inference (see section 8). Let $B_1, B_2, \ldots, B_s$ again be a partition of Ω. Bayes' rule follows from the axioms of probability

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$$

Example 10 *Consider a physician's diagnostic test for the presence or absence of some rare disease D, that only occurs in 0.1% of the population, i.e. $P(D) = .001$. It follows that $P(\bar{D}) = .999$, where $\bar{D}$ indicates that a person does not have the disease. The probability of an event before the evaluation of evidence through Bayes' rule is often called the prior probability. The prior probability that someone picked at random from the population has the disease is therefore $P(D) = .001$.*

Furthermore we denote a positive test result by T^+, and a negative test result by T^-. The performance of the test is summarized in table 1.2.

What is the probability that a patient has the disease, if the test result is positive? First, notice that $D, \bar{D}$ is a partition of the outcome space. We apply Bayes' rule to obtain

$$P(D|T^+) = \frac{P(T^+|D)P(D)}{P(T^+|D)P(D) + P(T^+|\bar{D})P(\bar{D})} = \frac{.95 \cdot .001}{.95 \cdot .001 + .02 \cdot .999} = .045.$$

Only 4.5% of the people with a positive test result actually have the disease. On the other hand, the posterior probability (i.e. the probability after evaluation of evidence) is 45 times as high as the prior probability.

1.14 Some named discrete distributions

A random experiment that only distinguishes between two possible outcomes is called a *Bernoulli* experiment. The outcomes are usually referred to as *success* and *failure* respectively. We define a random variable X that denotes the number of successes in a Bernoulli experiment; X consequently has possible values 0 and 1. The probability distribution of X is completely determined by the probability of success, which we denote by π, and is: $p(X = 0) = 1-\pi$ and $p(X = 1) = \pi$. It easily follows that $\mathrm{E}(X) = \mu = \pi$ and $\sigma^2 = \pi(1 - \pi)$.

A number of *independent*, *identical* repetitions of a Bernoulli experiment is called a *binomial* experiment. We denote the number of successes in a binomial experiment by Y which has possible values $0, 1, \ldots, n$ (where n is the number of repetitions). Any particular sequence with y successes has probability

$$\pi^y(1 - \pi)^{n-y}$$

since the trials are independent. The number of distinct ways y successes may occur in a sequence of n is

$$\left(\begin{array}{c} n \\ y \end{array} \right) = \frac{n!}{y!(n - y)!}$$

so the probability distribution of Y is

$$p(y) = \left(\begin{array}{c} n \\ y \end{array} \right) \pi^y(1 - \pi)^{n-y} \quad \text{for } y = 0, 1, \ldots, n.$$

We indicate that Y has binomial distribution with parameters n and π by writing $Y \sim B(n, \pi)$ ($\sim$ should be read "has distribution"). We can derive easily that $\mathrm{E}(Y) = \mu = n\pi$ and $\sigma^2 = n\pi(1 - \pi)$.

The multinomial distribution is a generalization of the binomial distribution to random experiments with $m \geq 2$ possible outcomes or categories. Let y_i denote the number of results in category i, and let π_i denote the probability of a result in the i^{th} category on each trial (with $\sum_{i=1}^{m} \pi_i = 1$). The joint probability distribution of $Y_1, Y_2, \ldots, Y_m$ for a sequence of n trials is

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_m = y_m) = \frac{n!}{y_1!\, y_2! \ldots y_m!} \pi_1^{y_1} \pi_2^{y_2} \ldots \pi_m^{y_m}$$

The product of powers of the π_i represents the probability of any particular sequence with y_i results in category i for each $1 \leq i \leq m$, and the ratio of

factorials indicates the number distinct sequences with y_i results in category i for each $1 \leq i \leq m$.

A random variable Y has Poisson distribution with parameter μ if it has probability function

$$p(y) = \frac{\mu^y}{y!} e^{-\mu} \text{ for } y = 0, 1, 2, \ldots$$

where the single parameter μ is a positive real number. One can easily show that $\mathrm{E}(Y) = \mathrm{V}(Y) = \mu$. We write $Y \sim \mathrm{Po}(\mu)$. Use of the Poisson distribution as an approximation to the binomial distribution is discussed in chapter 2.

1.15 Some named continuous distributions

Continuous distributions of type

$$f(y) = \begin{cases} \frac{1}{\beta - \alpha} \text{ for } \alpha \leq y \leq \beta \\ 0 \text{ otherwise} \end{cases}$$

are called uniform distributions, denoted $U(\alpha, \beta)$. Mean and variance are respectively

$$\mu = \frac{\alpha + \beta}{2}, \text{ and } \sigma^2 = \frac{(\beta - \alpha)^2}{12}$$

Continuous distributions of type

$$f(y) = \frac{e^{-(y-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}} \quad \text{for } y \in \mathbb{R}$$

with $\sigma > 0$ are called *normal* or *Gaussian* distributions. Mean μ and variance σ^2 are the two parameters of the normal distribution, which we denote by $\mathcal{N}(\mu, \sigma^2)$. The special case with $\mu = 0$ and $\sigma^2 = 1$, is called the standard-normal distribution. A random variable of this type is often denoted by Z, i.e. $Z \sim \mathcal{N}(0, 1)$. If the distribution of a random variable is determined by many small independent influences, it tends to be normally distributed. In the next section we discuss why the normal distribution is so important in statistical inference.

The binormal distribution is a generelization of the normal distribution to the joint distribution of pairs (X, Y) of random variables. Its parameters are

μ_x, μ_y, σ_x^2, σ_y^2, and correlation coefficient ρ, with σ_x^2, $\sigma_y^2 > 0$ and $-1 \leq \rho \leq 1$. We write

$$(X, Y) \sim \mathcal{N}^2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$$

The parameter ρ is a measure for the linear dependence between X and Y. Further generelization to the joint distribution of $m \geq 2$ random variables $Y_1, Y_2, \ldots, Y_m$ yields the multivariate normal distribution. For convenience we switch to matrix notation for the parameters

$$(Y_1, Y_2, \ldots, Y_m) \sim \mathcal{N}^m(\mu, \Sigma)$$

where $\mu = (\mu_1, \mu_2, \ldots, \mu_m)$ is the vector of means and Σ is an $m \times m$ covariance matrix. The diagonal elements of Σ contain the variances $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)$ and element (i, j) with $i \neq j$ contains the covariance between Y_i and Y_j.

A random variable T has *exponential* distribution with rate λ ($\lambda > 0$) if T has probability density

$$f(t) = \lambda\, e^{-\lambda t} \quad (t \geq 0)$$

We may think of T as a random time of some kind, such as a time to failure for artifacts, or survival times for organisms. With T we associate a survival function

$$P(T > s) = \int_s^\infty f(t)dt = e^{-\lambda s}$$

representing the probability of surviving past time s. Characteristic for the exponential distribution is that it is *memoryless*, i.e.

$$P(T > t + s \,|\, T > t) = P(T > s) \quad (t \geq 0, s \geq 0)$$

Given survival to time t, the chance of surviving a further time s is the same as surviving to time s in the first place. This is obviously not a good model for survival times of systems with aging such as humans. It is however a plausible model for time to failure of some artifacts that do not wear out gradually but stop functioning suddenly and unpredictably.

A random variable Y has a Beta distribution with parameters $l > 0$ and $k > 0$ if it has probability density

$$f(y) = \frac{y^{l-1}(1-y)^{k-1}}{\int_0^1 y^{l-1}(1-y)^{k-1}dy} \quad (0 \leq y \leq 1)$$

For the special case that $l = k = 1$ this reduces to a uniform distribution over the interval $[0, 1]$. The Beta distribution is particularly useful in Bayesian inference concerning unknown probabilities, which is discussed in chapter 8.

Chapter 2

Sampling and sampling distributions

In many cases we would like to learn something about a big population, without actually inspecting every unit in that population. In that case we would like to draw a *sample* that permits us to draw conclusions about a population of interest. We may for example draw a sample from the population of Dutch men of 18 years and older to learn something about the joint distribution of height and weight in this population.

Because we cannot draw conclusions about the population from a sample without error, it is important to know how large these errors may be, and how often incorrect conclusions may occur. An objective assessment of these errors is only possible for a *probability sample*. For a probability sample, the probability of inclusion in the sample is *known* and *positive* for each unit in the population. Drawing a probability sample of size n from a population consisting of N units, may be a quite complex random experiment. The experiment is simplified considerably by subdividing it into n experiments, consisting of drawing the n consecutive units. In a *simple random sample* the n consecutive units are drawn with equal probabilities from the units concerned. In random sampling *with replacement* the subexperiments (drawing of one unit) are all identical and independent: n times a unit is randomly selected from the entire population. We will see that this property simplifies the ensuing analysis considerably.

For units in the sample we observe one or more population variables. For probability samples, each draw is a random experiment. Every observation may therefore be viewed as a random variable. The observation of a popula-

Unit	1	2	3	4	5	6
$\mathcal{X}$	1	1	2	2	2	3

Table 2.1: A small population

x	1	2	3
$p_1(x) = p_2(x)$	1/3	1/2	1/6

Table 2.2: Probability distribution of X_1 and X_2

tion variable $\mathcal{X}$ from the unit drawn in the i^{th} trial, yields a random variable X_i. Observation of the complete sample yields n random variables $X_1, ..., X_n$. Likewise, if we observe for each unit the pair of population variables $(\mathcal{X}, \mathcal{Y})$, we obtain pairs of random variables (X_i, Y_i) with outcomes (x_i, y_i). Consider the population of size $N = 6$, displayed in table 2.1.

A random sample of size $n = 2$ is drawn *with replacement* from this population. For each unit drawn we observe the value of $\mathcal{X}$. This yields two random variables X_1 and X_2, with identical probability distribution as displayed in table 2.2. Furthermore X_1 and X_2 are independent, so their joint distribution equals the product of their individual distributions,i.e.

$$p(x_1, x_2) = \prod_{i=1}^{2} p_i(x_i) = [p(x)]^2$$

The distribution of the sample is displayed in the table 2.3.

Usually we are not really interested in the individual outcomes of the sample, but rather in some sample statistic. A statistic is a function of the sample observations $X_1, ..., X_n$, and therefore is itself also a random variable. Some important sample statistics are the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, and sample fraction $Fr = \frac{1}{n}\sum_{i=1}^{n} X_i$ (for 0-1 variable $\mathcal{X}$). In table 2.3 we listed the values of sample statistics $\bar{x}$ and s^2, for all possible samples of size 2.

The probability distribution of a sample statistic is called its *sampling distribution*. The sampling distribution of $\bar{X}$ and S^2 is calculated easily from table 2.3; they are displayed in tables 2.4 and 2.5 respectively.

Note that $\mathrm{E}(\bar{X}) = \frac{11}{6} = \mu$, where μ denotes the population mean, and $\mathrm{E}(S^2) = \frac{17}{36} = \sigma^2$, where σ^2 denotes the population variance.

(x_1, x_2)	$p(x_1, x_2)$	$\bar{x}$	s^2
(1,1)	1/9	1	0
(2,2)	1/4	2	0
(3,3)	1/36	3	0
(1,2)	1/6	1.5	0.5
(1,3)	1/18	2	2
(2,1)	1/6	1.5	0.5
(2,3)	1/12	2.5	0.5
(3,1)	1/18	2	2
(3,2)	1/12	2.5	0.5

Table 2.3: Probability distribution of sample of size $n = 2$ by sampling with replacement from the population in table 2.1

$\bar{x}$	$p(\bar{x})$
1	1/9
1.5	1/3
2	13/36
2.5	1/6
3	1/36

Table 2.4: Sampling distribution of $\bar{X}$

s^2	$p(s^2)$
0	14/36
0.5	1/2
2	1/9

Table 2.5: Sampling distribution of S^2

In the above example, we were able to determine the probability distribution of the sample, and sample statistics, by complete enumeration of all possible samples. This was feasible only because the sample size and the number of distinct values of $\mathcal{X}$ was very small. When the sample is of realistic size, and $\mathcal{X}$ takes on many distinct values, complete enumeration is not possible. Nevertheless, we would like to be able to infer something about the shape of the sampling distribution of a sample statistic, from knowledge of the distribution of X. We consider here two options to make such inferences.

1. The distribution of X has some standard form that allows the mathematical derivation of the exact sampling distribution.

2. We use a limiting distribution to approximate the sampling distribution of interest. The limiting distribution may be derived from some characteristics of the distribution of X.

The exact sampling distribution of a sample statistic is often hard to derive analytically, even if the population distribution of $\mathcal{X}$ is known. As an example we consider the sample statistic $\bar{X}$. The mean and variance of the sampling distribution of $\bar{X}$ are $\mathrm{E}(\bar{X}) = \mu$ and $\mathrm{V}(\bar{X}) = \sigma^2/n$, but its exact shape can only be derived in a few special cases. For example, if the distribution of $\mathcal{X}$ is $\mathcal{N}(\mu, \sigma^2)$ then the distribution of $\bar{X}$ is $\mathcal{N}(\mu, \sigma^2/n)$. Of more practical interest is the exact sampling distribution of the sample statistic Fr, i.e. the fraction of successes in the sample, with $\mathcal{X}$ a 0-1 population variable. The number of successes in the sample has distribution $Y \sim B(n, \pi)$ where n is the sample size and π the fraction of successes in the population. We have $\mu_y = n\pi$ and $\sigma_y^2 = n\pi(1-\pi)$. Since $Fr = Y/n$, it follows that $\mu_{fr} = \pi$ and $\sigma_{fr}^2 = \pi(1-\pi)/n$. Since $P(Fr = fr) = P(Y = nfr)$, its sampling distribution is immediately derived from the sampling distribution of Y.

Example 11 *Consider a sample of size 10 from a population with fraction of successes $\pi = 0.8$. What is the sampling distribution of Fr , the sample fraction of successes? The distribution is immediately derived from the distribution of the number of successes $Y \sim B(10, 0.8)$.*

In practice, we often have to rely on approximations of the sampling distribution based on so called *asymptotic* results. To understand the basic idea, we have to introduce some definitions concerning the convergence of sequences of random variables. For present purposes we distinguish between

convergence in probability (to a constant) and convergence in distribution (weak convergence) of a sequence of random variables. The limiting arguments below are all with respect to sample size n.

Definition 1 *A sequence $\{X_n\}$ of random variables converges in probability to a constant c if, for every positive number ε and η, there exists a positive integer $n_0 = n_0(\varepsilon, \eta)$ such that*

$$P(\mid X_n - c \mid > \varepsilon) < \eta,\ n \geq n_0$$

Example 12 *Consider the sequence of random variables $\{X_n\}$ with probability distributions $P(x_n = 0) = 1 - 1/n$ and $P(x_n = n) = 1/n$. Then $\{X_n\}$ converges in probability to 0.*

Definition 2 *A sequence $\{X_n\}$ of random variables converges in distribution to a random variable X with distribution function $F(X)$ if for every $\varepsilon > 0$, there exists an integer $n_0 = n_0(\varepsilon, x)$, such that at every point where $F(X)$ is continuous*

$$\mid F_n(x) - F(x) \mid < \varepsilon,\ n \geq n_0$$

where $F_n(x)$ denotes the distribution function of x_n.

This is in fact the same as pointwise convergence of a sequence of functions. If $n_0 = n_0(\varepsilon)$, i.e. does not depend on x, we speak of uniform convergence.

Example 13 *Consider a sequence of random variables $\{X_n\}$ with probability distributions $P(x_n = 1) = 1/2 + 1/(n+1)$ and $P(x_n = 2) = 1/2 - 1/(n+1)$, $n = 1, 2, \ldots$. As n increases without bound, the two probabilities converge to $1/2$, and $P(X = 1) = 1/2$, $P(X = 2) = 1/2$ is called the limiting distribution of $\{X_n\}$.*

Convergence in distribution is a particularly important concept in statistical inference, because the limiting distributions of sample statistics may be used as an approximation in case the exact sampling distribution cannot be (or is prohibitively cumbersome) to derive. A crucial result in this respect is the *central limit theorem* : If $(x_1, ..., x_n)$ is a random sample from any probability distribution with finite mean μ and finite variance σ^2, and $\bar{x} = 1/n \sum x_i$ then

$$\frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0, 1)$$

regardless of the form of the parent distribution. In this expression, $\xrightarrow{D}$ denotes convergence in distribution. This property explains the importance of the normal distribution in statistical inference. Note that this theorem doesn't say anything however about the rate of convergence to the normal distribution. In general, the more the population distribution resembles a normal distribution, the faster the convergence. For extremely skewed distributions $n = 100$ may be required for the normal approximation to be acceptable.

A well-known application of the central limit theorem is the approximation of the distribution of the sample proportion of successes Fr by a normal distribution. Since a success is coded as 1, and failure as 0, the fraction of successes is indeed a mean. This means the central limit theorem is applicable and as a rule of thumb $Fr \approx \mathcal{N}(\pi, \pi(1-\pi)/n)$ if $n\pi \geq 5$ and $n(1-\pi) \geq 5$. Even though the exact sampling distribution can be determined in this case, as n becomes large it becomes prohibitively time-consuming to actually calculate this distribution.

If π is close to 0 or 1, quite a large sample is required for the normal approximation to be acceptable. In that case we may use the following convergence property of the binomial distribution

$$\left(\begin{array}{c} n \\ y \end{array} \right) \pi^y (1-\pi)^{n-y} \xrightarrow{D} \frac{(n\pi)^y}{y!} e^{-n\pi}$$

In words, the binomial distribution with parameters n and π converges to a Poisson distribution with parameter $\mu = n\pi$ as n gets larger and larger. Moreover, it can be shown that this approximation is quite good for $\pi \leq 0.1$, regardless of the value of n. This explains the use of the Poisson rather than the normal approximation to the binomial distribution when π is close to 0 or 1.

Chapter 3

Statistical Inference

The relation between sample data and population may be used for reasoning in two directions: from known population to yet to be observed sample data (as discussed in chapter 2), and from observed data to (partially) unknown population. This last direction of reasoning is of inductive nature and is addressed in statistical inference. It is the form of reasoning most relevant to data analysis, since one typically has available one set of sample data from which one intends to draw conclusions about the unknown population.

3.1 Frequentist Inference

According to frequentists, inference procedures should be interpreted and evaluated in terms of their behavior in hypothetical repetitions under the same conditions. To quote David S. Moore, the frequentist consistently asks "What would happen if we did this many times?" [9]. To answer this question, the sampling distribution of a statistic is of crucial importance. The two basic types of frequentist inference are estimation and testing. In estimation one wants to come up with a plausible value or range of plausible values for an unknown population parameter. In testing one wants to decide whether a hypothesis concerning the value of an unknown population parameter should be accepted or rejected in the light of sample data.

3.1.1 Point Estimation

In point estimation one tries to provide an estimate for an unknown population parameter, denoted by θ, with *one number*: the point estimate. If G denotes the estimator of θ, then the estimation error is a random variable $G - \theta$, which should preferably be close to zero.

An important quality measure from a frequentist point of view is the bias of an estimator

$$\mathrm{B}_\theta = \mathrm{E}_\theta(G - \theta) = \mathrm{E}_\theta(G) - \theta,$$

where expectation is taken with respect to repeated samples from the population. If $\mathrm{E}_\theta(G) = \theta$, i.e. the expected value of the estimator is equal to the value of the population parameter, then the estimator G is called *unbiased*.

Example 14 *If π is the proportion of successes in some population and Fr is the proportion of successes in a random sample from this population, then $\mathrm{E}_\pi(Fr) = \pi$, so Fr is an unbiased estimator of π.*

Another important quality measure of an estimator is its variance

$$\mathrm{V}_\theta(G) = \mathrm{E}_\theta(G - \mathrm{E}_\theta(G))^2$$

which measures how much individual estimates g tend to differ from $\mathrm{E}_\theta(G)$, the average value of g over a large number of samples.

An overall quality measure that combines bias and variance is the *mean squared error*

$$\mathrm{M}_\theta(G) = \mathrm{E}_\theta(G - \theta)^2$$

where low values indicate a good estimator. After some algebraic manipulation, we can decompose mean squared error into

$$\mathrm{M}_\theta(G) = \mathrm{B}_\theta^2(G) + \mathrm{V}_\theta(G)$$

that is mean squared error equals squared bias plus variance. It follows that if an estimator is unbiased, then its mean squared error equals its variance.

Example 15 *For the unbiased estimator Fr of π we have $\mathrm{M}_\pi(Fr) = \mathrm{V}_\pi(Fr) = \pi(1 - \pi)/m$.*

The so-called "plug-in" principle provides a simple and intuitively plausible method of constructing estimators. The plug-in estimate of a parameter $\theta = t(F)$ is defined to be $\hat{\theta} = t(\hat{F})$. Here F denotes the population distribution function and $\hat{F}$ its estimate, based on the sample. For example, to estimate the population mean μ use its sample analogue $\bar{x} = 1/n \sum x_i$, and to estimate population variance σ^2 use its sample analogue $s^2 = 1/n \sum (x_i - \bar{x})^2$. Another well-known method for finding point estimates is the method of least squares. The least squares estimate of population mean μ is the number g for which the sum of squared errors $(x_i - g)^2$ is at a minimum. If we take the derivative of this sum with respect to g, we obtain

$$\frac{\partial}{\partial g} \sum_{i=1}^{n} (x_i - g)^2 = \sum_{i=1}^{n} (x_i - g)(-2) = -2n(\bar{x} - g)$$

When we equate this expression to zero, and solve for g we obtain $g = \bar{x}$. So $\bar{x}$ is the least squares estimate of μ. A third important method of estimation is *maximum likelihood estimation*, which is discussed in section 3.2.

3.1.2 Interval Estimation

An interval estimator for population parameter θ is an interval of type (G_L, G_U). Two important quality measures for interval estimates are:

$$\mathrm{E}_\theta(G_U - G_L),$$

i.e. the expected width of the interval, and

$$P_\theta(G_L < \theta < G_U),$$

i.e. the probability that the interval contains the true parameter value. Clearly there is a trade-off between these quality measures. If we require a high probability that the interval contains the true parameter value, the interval itself has to become wider. It is customary to choose a confidence level $(1 - \alpha)$ and use an interval estimator such that

$$P_\theta(G_L < \theta < G_U) \geq 1 - \alpha$$

for all possible values of θ. A realisation (g_L, g_U) of such an interval estimator is called a $100(1 - \alpha)\%$ *confidence interval*.

The form of reasoning used in confidence intervals is most clearly reflected in the estimation of the mean of a normal population with variance σ^2 known, i.e. $X \sim \mathcal{N}(\mu, \sigma^2)$. The distribution of the sample mean for random samples of size n from this population is known to be $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. First $\bar{X}$ is standardized to obtain

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

which allows us to use a table for the standardnormal distribution $Z \sim \mathcal{N}(0, 1)$ to find the relevant probabilities. The probability that $\bar{X}$ is more than one standard error (standard deviation of the sampling distribution) larger than unknown μ is

$$P(\bar{X} > \mu + \frac{\sigma}{\sqrt{n}}) = P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > 1) = P(Z > 1) = 0.1587$$

But we can *reverse* this reasoning by observing that

$$P(\bar{X} - \frac{\sigma}{\sqrt{n}} < \mu) = 1 - 0.1587 = 0.8413$$

because $\bar{X} - \frac{\sigma}{\sqrt{n}} < \mu$ holds unless $\bar{X} > \mu + \frac{\sigma}{\sqrt{n}}$. Therefore, the probability that the interval $(\bar{X} - \sigma/\sqrt{n}, \infty)$ will contain the true value of μ equals 0.8413. This is called a left-sided confidence interval because it only states a lower bound for μ. In general a $100(1 - \alpha)\%$ left-sided confidence interval for μ reads $(\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty)$, where $P(Z > z_\alpha) = \alpha$. Likewise, we may construct a right-sided confidence interval $(-\infty, \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}})$ and a two-sided confidence interval

$$(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}).$$

If the distribution of X is unknown, i.e. $X \sim \mu, \sigma^2$, then for sufficiently large n we may invoke the central limit theorem and use $\bar{X} \approx \mathcal{N}(\mu, \sigma^2/n)$, and proceed as above.

In most practical estimation problems we don't know the value of σ^2, and we have to estimate it from the data as well. A rather obvious estimator is the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})$$

Now we may use

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

where t_{n-1} denotes the t-distribution with $n-1$ degrees of freedom. This distribution has a higher variance than the standardnormal distribution, leading to somewhat wider confidence intervals. This is the price we pay for the fact that we don't know the value of σ^2, but have to estimate it from the data. On the other hand we have $t_\nu \approx \mathcal{N}(0,1)$ for $\nu \geq 100$, so if n is large enough we may use the standardnormal distribution for all practical purposes.

3.1.3 Hypothesis Testing

A test is a statistical procedure to make a choice between two hypotheses concerning the value of a population parameter θ. One of these, called the *null hypothesis* and denoted by H_0, gets the "benefit of the doubt". The two possible conclusions are to reject or not to reject H_0. H_0 is only rejected if the sample data contains strong evidence that it is not true. The null hypothesis is rejected iff realisation g of test statistic G is in the *critical region* denoted by C. In doing so we can make two kinds of errors

Type I error: Reject H_0 when it is true.

Type II error: Accept H_0 when it is false.

Type I errors are considered to be more serious than Type II errors. Test statistic G is usually a point estimator for θ, e.g. if we test a hypothesis concerning the value of population mean μ, then $\bar{X}$ is an obvious choice of test statistic. As an example we look at hypothesis test

$$H_0 : \theta \geq \theta_0 \ , \ H_a : \theta < \theta_0$$

The highest value of G that leads to the rejection of H_0 is called the critical value c_u, it is the upper bound of the so-called critical region $C = (-\infty, c_u]$. All values of G to the left of c_u lead to the rejection of H_0, so this is called a left one-sided test. An overall quality measure for a test is its power β

$$\beta(\theta) = P_\theta(\text{Reject } H_0) = P_\theta(G \in C)$$

Because we would like a low probability of Type I and Type II errors, we like to have $\beta(\theta)$ small for $\theta \in H_0$ and $\beta(\theta)$ large for $\theta \in H_a$. It is common practice in hypothesis testing to restrict the probability of a Type I error to a maximum called the *significance level* α of the test, i.e.

$$\max_{\theta \in H_0} \beta(\theta) \leq \alpha$$

Since the maximum is reached for $\theta = \theta_0$ this reduces to the restriction $\beta(\theta_0) \leq \alpha$. If possible the test is performed in such a way that $\beta(\theta_0) = \alpha$ (This may not be possible for discrete sampling distributions). Common levels for α are 0.1, 0.05 and 0.01. If in a specific application of the test, the conclusion is that H_0 should be rejected, then the result is called *significant*.

Consider a left one-sided test on population mean μ with $X \sim \mathcal{N}(\mu, \sigma^2)$ and the value of σ^2 known. That is

$$H_0 : \mu \geq \mu_0 \ , \ H_a : \mu < \mu_0$$

We determine the sampling distribution of the test statistic $\bar{X}$ under the assumption that the $\mu = \mu_0$, i.e. $\bar{X} \sim \mathcal{N}(\mu_0, \sigma^2/n)$. Now

$$\alpha = P_{\mu_0}(\bar{X} \leq c_u) = P(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq \frac{c_u - \mu_0}{\sigma/\sqrt{n}}) = P(Z \leq \frac{c_u - \mu_0}{\sigma/\sqrt{n}})$$

and since $P(Z \leq -z_\alpha) = \alpha$, we obtain

$$\frac{c_u - \mu_0}{\sigma/\sqrt{n}} = -z_\alpha, \text{ and therefore } c_u = \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$$

Example 16 *Consider a random sample of size $n = 25$ from a normal population with known $\sigma = 5.4$ and unknown mean μ. The observed sample mean is $\bar{x} = 128$. We want to test the hypothesis*

$$H_0 : \mu \geq 130, \textit{ against } \ H_a : \mu < 130$$

i.e. $\mu_0 = 130$. The significance level of the test is set to $\alpha = 0.05$. We compute the critical value

$$c_u = \mu_0 - z_{0.05} \frac{\sigma}{\sqrt{n}} = 130 - 1.645 \frac{5.4}{\sqrt{25}} = 128.22$$

where $z_{0.05} = 1.645$ was determined using a statistical package (many books on statistics contain tables that can be used to determine the value of z_α). So the critical region is $(-\infty, 128.22]$ and since $\bar{x} = 128$ is in the critical region, we reject H_0.

Similarly, if

$$H_0 : \theta \leq \theta_0 \ , \ H_a : \theta > \theta_0$$

the critical region is $[c_l, \infty)$, and for a two-sided test

$$H_0 : \theta = \theta_0 \ , \ H_a : \theta \neq \theta_0$$

it has the form $(-\infty, c_u] \cup [c_l, \infty)$.

As with the construction of a confidence interval for the mean, for a hypothesis test concerning the mean we may invoke the central limit theorem if $X \sim \mu, \sigma^2$ and n is large. Furthermore, if σ^2 is unknown, we have to estimate it from the data and use a t_{n-1} distribution rather than the standardnormal distribution to determine the critical region.

Sometimes one doesn't want to specify the significance level α of the test in advance. In that case it us customary to report so-called p-values, indicating the *observed significance*.

Example 17 *Consider the test of example 16. The p-value of the observed outcome* $\bar{x} = 128$ *is*

$$P_{\mu_0}(\bar{X} \leq 128) = P(Z \leq \frac{128 - \mu_0}{\sigma/\sqrt{n}}) = P(Z \leq -1.852) = 0.0322$$

Since the p-value is 0.0322*, we would reject* H_0 *at* $\alpha = 0.05$*, but we would accept* H_0 *at* $\alpha = 0.01$*.*

3.2 Likelihood

The deductive nature of probability theory versus the inductive nature of statistical inference is perhaps most clearly reflected in the "dual" concepts of (joint) probability distribution and likelihood.

Given a particular probability model and corresponding parameter values, we may calculate the probability of observing different samples. Consider the experiment of 10 coin flips with probability of heads $\pi = 0.6$. The probability distribution of random variable "number of times heads comes up" is now the following function of the data

$$P(y) = \left(\begin{array}{c} 10 \\ y \end{array} \right) 0.6^y \, 0.4^{10-y}$$

We may for example compute that the probability of observing $y = 7$ is

$$\left(\begin{array}{c} 10 \\ 7 \end{array} \right) 0.6^7 0.4^3 \approx 0.215$$

y	π								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	.349	.107	.028	.006	.001				
1	.387	.269	.121	.04	.01	.002			
2	.194	.302	.234	.121	.044	.01	.002		
3	.057	.201	.267	.215	.117	.043	.009	.001	
4	.011	.088	.2	.251	.205	.111	.036	.005	
5	.002	.027	.103	.201	.246	.201	.103	.027	.002
6		.005	.036	.111	.205	.251	.2	.088	.011
7		.001	.009	.043	.117	.215	.267	.201	.057
8			.002	.01	.044	.121	.234	.302	.194
9				.002	.01	.04	.121	.269	.387
10					.001	.006	.028	.107	.349
	1	1	1	1	1	1	1	1	1

Table 3.1: Probability distributions (columns) and likelihood functions (rows) for $Y \sim B(10, \pi)$

In statistical inference however, we typically have one data set and want to say something about the relative likelihood of different values of some population parameter. Say we observed 7 heads in a sequence of ten coin flips. The likelihood is now a function of the unknown parameter π

$$L(\pi \,|\, y = 7) = \left(\begin{array}{c} 10 \\ 7 \end{array} \right) \pi^7 (1 - \pi)^3$$

where the constant term is actually arbitrary, since we are not interested in absolute values of the likelihood, but rather in ratios of likelihoods for different values of π.

In table 3.1, each column specifies the probability distribution of Y for a different value of π. Each column sums to 1, since it represents a probability distribution. Each row, on the other hand, specifies a likelihood function, or rather: it specifies the value of the likelihood function for 9 values of π. So for example, in the third row we can read off the probability of observing 2 successes in a sequence of 10 coin flips for different values of π.

In general, if $\mathbf{y} = (y_1, \ldots, y_n)$ are independent observations from a probability density $f(y \,|\, \theta)$, where θ is the parameter vector we wish to estimate,

then

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

The *likelihood function* then measures the relative likelihood that different θ have given rise to the observed $\mathbf{y}$. We can thus try to find that particular $\hat{\theta}$ which maximizes L, i.e. that $\hat{\theta}$ such that the observed $\mathbf{y}$ are more likely to have come from $f(y \,|\, \hat{\theta})$ than from $f(y \,|\, \theta)$ for any other value of θ.

For many parameter estimation problems one can tackle this maximization by differentiating L with respect to the components of θ and equating the derivatives to zero to give the *normal equations*

$$\frac{\partial L}{\partial \theta_j} = 0$$

These are then solved for the θ_j and the second order derivatives are examined to verify that it is indeed a maximum which has been achieved and not some other stationary point.

Maximizing the likelihood function L is equivalent to maximizing the (natural) log of L, which is computationally easier. Taking the natural log, we obtain the log-likelihood function

$$l(\theta \,|\, \mathbf{y}) = \ln(L(\theta \,|\, \mathbf{y})) = \ln(\prod_{i=1}^{n} f(y_i \,|\, \theta)) = \sum_{i=1}^{n} \ln f(y_i \,|\, \theta)$$

since $\ln ab = \ln a + \ln b$.

Example 18 *In a coin flipping experiment we define the random variable Y with $y = 1$ if heads comes up, and $y = 0$ when tails comes up. Then we have the following probability distribution for one coin flip*

$$f(y) = \pi^y (1 - \pi)^{1-y}$$

For a sequence of n coin flips, we obtain the joint probability distribution

$$f(\mathbf{y}) = f(y_1, y_2, ..., y_n) = \prod_{i=1}^{n} \pi^{y_i} (1 - \pi)^{1-y_i}$$

which defines the likelihood when viewed as a function of π. The log-likelihood consequently becomes

$$l(\pi \,|\, \mathbf{y}) = \sum_{i=1}^{n} y_i \ln(\pi) + (1 - y_i) \ln(1 - \pi)$$

In a sequence of 10 coin flips with seven times heads coming up, we obtain

$$l(\pi) = \ln(\pi^7(1-\pi)^3) = 7\ln\pi + 3\ln(1-\pi)$$

To determine the maximum we take the derivative and equate to zero

$$\frac{dl}{d\pi} = \frac{7}{\pi} - \frac{3}{1-\pi} = 0$$

which yields maximum likelihood estimate $\hat{\pi} = 0.7$.

The reader may notice that the maximum likelihood estimate in this case is simply the fraction of heads coming up in the sample, and we could have spared ourselves the trouble of maximizing the likelihood function to obtain the required estimate. Matters become more interesting (and complicated) however, when we make π a function of data *and* parameters. Suppose that for each y_i in our sample, we observe a corresponding measure x_i which we assume is a continuous variable. We could write $\pi_i = g(x_i)$, where g is some function. In so-called Probit analysis [6] we assume

$$\pi_i = \Phi(\alpha + \beta x_i)$$

where Φ denotes the standard normal distribution function. The parameters of the model are now α and β, and we can write the log-likelihood function as

$$l(\alpha, \beta) = \sum_{i=1}^{n} y_i \ln(\Phi(\alpha + \beta x_i)) + (1 - y_i)\ln(1 - \Phi(\alpha + \beta x_i))$$

This is the expression of the log-likelihood for the Probit model. By maximizing with respect to α and β, we obtain maximum likelihood estimates for these parameters.

Example 19 *Consider a random sample* $\mathbf{y} = (y_1, ..., y_n)$ *from a normal distribution with unknown mean* μ *and variance* σ^2. *Then we have likelihood*

$$L((\mu, \sigma^2)' \,|\, \mathbf{y}) = \prod_{i=1}^{n} \frac{e^{-(y_i-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}} = \frac{1}{\sigma^n(2\pi)^{m/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \mu}{\sigma}\right)^2\right]$$

The natural log of this expression is

$$l = \ln(L) = -n\ln\sigma - \left(\frac{m}{2}\right)\ln 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

To determine the maximum likelihood estimates of μ and σ, we take the partial derivative of l with respect to these parameters, and equate them to zero

$$\frac{\partial l}{\partial \mu} = \frac{n}{\sigma^2}(\bar{y} - \mu) = 0$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4}(s^2 + (\bar{y} - \mu)^2) = 0$$

Solving these equations for μ and σ, we obtain maximum likelihood estimates $\hat{\mu} = \bar{y}$ and $\hat{\sigma}^2 = s^2$, where $s^2 = 1/n \sum (y_i - \hat{\mu})^2$.

Another important aspect of the log-likelihood function is its shape in the region near the maximum. If it is rather flat, one could say that the likelihood contains little information in the sense that there are many values of θ with log-likelihood near that of $\hat{\theta}$. If, on the other hand, it is rather steep, one could say that the log-likelihood contains much information about $\hat{\theta}$. The log-likelihood of any other value of θ is approximately given by the Taylor expansion

$$l(\theta) = l(\hat{\theta}) + (\theta - \hat{\theta})\frac{dl}{d\theta} + \frac{1}{2}(\theta - \hat{\theta})^2 \frac{d^2 l}{d\theta^2} + ...$$

where the differential coefficients are evaluated at $\theta = \hat{\theta}$. At this point, $\frac{dl}{d\theta}$ is zero, so approximately

$$l(\theta) = l(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 \frac{d^2 l}{d\theta^2}.$$

Minus the second derivative of the log-likelihood function is known as the (Fisher) *information*. When evaluated at $\hat{\theta}$ (the maximum likelihood estimate of θ) it is called the *observed information*.

Some authors take the view that all statistical inference should be based on the likelihood function rather than the sampling distribution used in frequentist inference (see [3, 12]). In this sense likelihood inference differs from frequentist inference.

Example 20 *Figure 3.1 displays the likelihood function for π corresponding to 7 successes in a series of 10 coin flips. The horizontal line indicates the range of values of π for which the ratio of $L(\pi)$ to the maximum $L(0.7)$ is greater than 1/8. The 1/8 likelihood interval is approximately* $(0.38, 0.92)$.

Such an interval is similar in spirit to a confidence interval in the sense that it intends to provide a range of "plausible values" for π based on the sample data. A confidence interval for π is based however on the sampling distribution of some sample statistic (the sample proportion of successes is the most obvious choice) whereas a likelihood interval is based, as the name suggests, on the likelihood function.

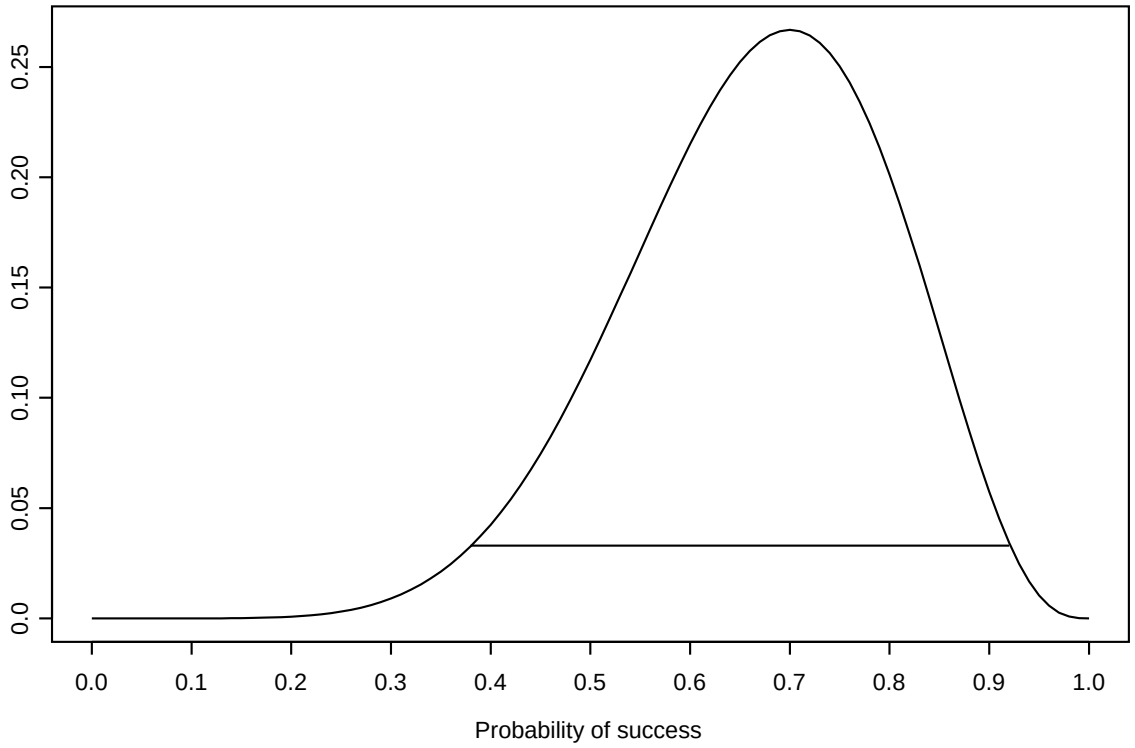

Figure 3.1: Likelihood function $L(\pi \mid y = 7) = 120\pi^7(1-\pi)^3$.

On the other hand, maximum likelihood estimation may be used and evaluated from a frequentist perspective. This motivates the study of the sampling distribution of maximum likelihood estimates. If we know the true value of $\theta = \theta^*$, we can determine the *expected* log-likelihood, i.e. the mean value of the log-likelihood conditional on $\theta = \theta^*$ (still expressed as a function of θ). The expected log-likelihood has a maximum at $\theta = \theta^*$. Minus

the second derivative of the expected log-likelihood evaluated at $\theta = \theta^*$, is called the *expected information*. Assuming parameter vector θ with several components the expected information matrix is defined as

$$I(\theta) = -\left\{ E\left(\frac{\partial^2 l}{\partial \theta_j \partial \theta_k} \right)_{\theta^*} \right\}$$

In large samples, the maximum likelihood estimate $\hat{\theta}$ is approximately normally distributed with mean θ^*, and covariance matrix $I(\theta)^{-1}$. Unfortunately, we cannot in practice determine $I(\theta)$, since θ^* is unknown. It is therefore set equal to $\hat{\theta}$ so that $I(\theta)$ can be calculated. An alternative estimate for the covariance matrix is the observed information matrix

$$-\left(\frac{\partial^2 l}{\partial \theta_j \partial \theta_k} \right)_{\hat{\theta}}$$

which is easier to compute since it does not involve an expectation. For the exponential family of distributions these two estimates are equivalent.

Example 21 *Consider a sequence of n coin tosses, with heads coming up y times. We are interested in the probability of heads π. We have seen that*

$$l(\pi) = y \ln(\pi) + (n - y) \ln(1 - \pi)$$

Setting the first derivative to zero and solving for π yields $\hat{\pi} = y/n$. The information is

$$-\frac{d^2 l}{d\pi^2} = \frac{y}{\pi^2} + \frac{(n - y)}{(1 - \pi)^2}$$

Evaluating this expression at $\hat{\pi} = y/n$ we obtain the observed information

$$\frac{n}{\hat{\pi}(1 - \hat{\pi})}.$$

In large samples, $\hat{\pi}$ is approximately normally distributed with mean π^* and variance $\pi^*(1 - \pi^*)/n$, i.e. the reciprocal of the expected information. The estimated variance of $\hat{\pi}$ is equal to the reciprocal of the observed information, i.e. $\hat{\pi}(1 - \hat{\pi})/n$.

Chapter 4

Linear Regression

4.1 Fitting a straight line to data

The relation between two variables can be used to predict the value of one when the value of the other is known. The basic idea is straightforward: draw a curve through the points of the scatterplot to represent the relationship and then use this curve for prediction. The simplest curve is a straight line, and that is the case we shall consider initially. Of course it would be foolish to draw a straight line when the pattern of the relationship is curved. The data must show a roughly linear trend like that in figure 4.1.

Least squares fitting looks at the *vertical* deviations of the points in a scatterplot from any straight line. Any line that is a good candidate for describing the data will pass above some points and below others, rather than miss the cloud of points entirely. So some of the deviations will be positive and some will be negative. But we need a total that ignores the signs of the deviations, otherwise positive and negative deviations could cancel eachother. The squares of the deviations are all positive. The least squares line is the line that makes the sum of the squared deviations as small as possible. Hence the name least squares.

In writing the equation for a line, x stands as usual for the explanatory variable and y for the response variable. The equation has the form

$$y = b_0 + b_1 x \tag{4.1}$$

The number b_1 is the *slope* of the line, the amount by which y changes when x increases by one unit. The number b_0 is the *intercept*, the value of y when $x = 0$.

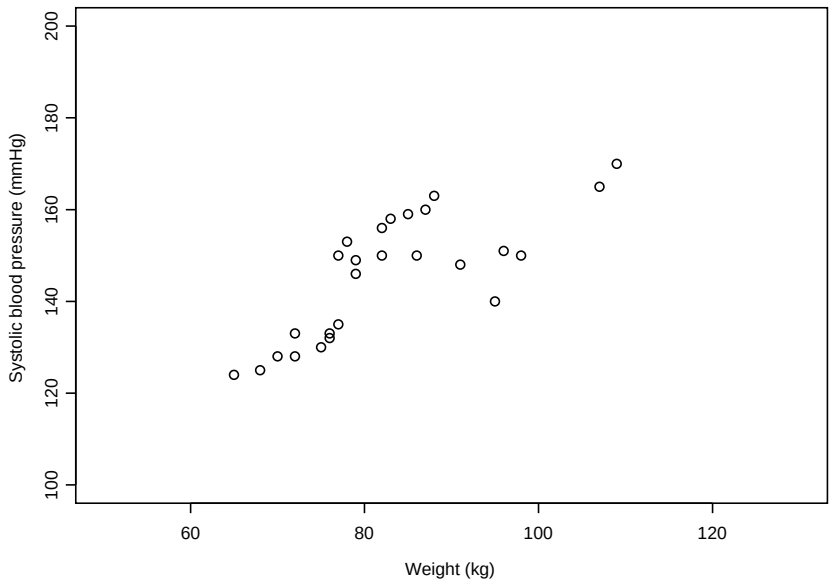

Figure 4.1: Scatterplot of weight against systolic bloodpressure

We are given a number of observations $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. We have to find the values of b_0 and b_1 such that the sum of squared deviations

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 \tag{4.2}$$

is minimized. The value $b_0 + b_1 x_i$ is the fitted value of y at x_i, and is denoted by $\hat{y}_i$. The difference between the actual value and the fitted value is called the error and is denoted by e_i, that is $e_i = y_i - \hat{y}_i$.

Note that b_0 and b_1 are the variables in expression (4.2), whereas x_i and y_i are fixed numbers once the data are observed.

We give a simple example to show the calculations. Suppose we have observations as shown below

i	x	y	$\hat{y} = b_0 + b_1 x$	$e = y - \hat{y}$
1	0	1	b_0	$1 - b_0$
2	1	3	$b_0 + b_1$	$3 - b_0 - b_1$
3	2	4	$b_0 + 2b_1$	$4 - b_0 - 2b_1$
4	3	3	$b_0 + 3b_1$	$3 - b_0 - 3b_1$
5	4	5	$b_0 + 4b_1$	$5 - b_0 - 4b_1$

Then

$$S(b_0, b_1) = (1-b_0)^2+(3-b_0-b_1)^2+(4-b_0-2b_1)^2+(3-b_0-3b_1)^2+(5-b_0-4b_1)^2$$

To find the minimum we take the partial derivatives with respect to b_0 and b_1 of this expression, and equate them to zero. We start with the partial derivative with respect to the intercept

$$\begin{aligned}\frac{\partial S}{\partial b_0} &= [2(1-b_0)(-1)] + [2(3-b_0-b_1)(-1)] + [2(4-b_0-2b_1)(-1)] + \\ &\quad [2(3-b_0-3b_1)(-1)] + [2(5-b_0-4b_1)(-1)] \\ &= -32 + 10b_0 + 20b_1\end{aligned}$$

The partial derivative with respect to the slope is

$$\begin{aligned}\frac{\partial S}{\partial b_1} &= 0 + [2(3-b_0-b_1)(-1)] + [2(4-b_0-2b_1)(-2)] + \\ &\quad [2(3-b_0-3b_1)(-3)] + [2(5-b_0-4b_1)(-4)] \\ &= -80 + 20b_0 + 60b_1\end{aligned}$$

A nice feature of taking the squared error is that the derivatives are linear: we have 2 linear equations in 2 unknowns:

$$\begin{aligned}10b_0 + 20b_1 &= 32 \\ 20b_0 + 60b_1 &= 80\end{aligned}$$

which give $b_0 = 8/5$ and $b_1 = 4/5$. So the least squares fitted line is

$$\hat{y} = \frac{8}{5} + \frac{4}{5}x$$

We now derive the general expressions for the partial derivative of the sum of squared errors with respect to intercept b_0 and slope b_1.

We start with the intercept:

$$\frac{\partial S}{\partial b_0} = \sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i)(-1) = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0 \qquad (4.3)$$

or equivalently

$$\sum_{i=1}^{n} e_i = 0 \qquad (4.4)$$

Note that it follows from this condition that the sum of the error terms $e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$ is zero.

The partial derivative with respect to the slope is:

$$\frac{\partial S}{\partial b_1} = \sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i)(-x_i) = -2\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0 \qquad (4.5)$$

from which it follows that

$$\sum_{i=1}^{n} x_i e_i = 0 \qquad (4.6)$$

Expanding (4.3) and (4.5) and collection terms yields what are commonly called the *normal equations*

$$\sum_{i=1}^{n} y_i = nb_0 + \sum_{i=1}^{n} x_i b_1 \qquad (4.7)$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i b_0 + \sum_{i=1}^{n} x_i^2 b_1 \qquad (4.8)$$

To solve for b_0 we divide both sides of (4.7) by n to obtain

$$\bar{y} = b_0 + b_1 \bar{x}$$

from which we can conclude that the least squares regression line passes through the point of means $(\bar{y}, \bar{x})$. Now isolate b_0 on the left hand side

$$b_0 = \bar{y} - b_1 \bar{x}. \qquad (4.9)$$

To solve for b_1, multiply the equation (4.7) by $\sum x_i$, and multiply equation (4.8) by n.

$$\begin{aligned} \sum x_i \sum y_i &= n \sum x_i b_0 + \left(\sum x_i\right)^2 b_1 \\ n \sum x_i y_i &= n \sum x_i b_0 + n \sum x_i^2 b_1 \end{aligned}$$

Subtracting the first equation from the second yields

$$\begin{aligned} n \sum x_i y_i - \sum x_i \sum y_i &= n \sum x_i^2 b_1 - \left(\sum x_i\right)^2 b_1 \\ &= b_1 \left(n \sum x_i^2 - \left(\sum x_i\right)^2 \right) \end{aligned}$$

Solving for b_1 gives

$$b_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2} \tag{4.10}$$

Now we have to verify that we have, indeed, found a minimum of the sum of squared errors. We have found a local minimum if the matrix of second derivatives (sometimes called the Hessian matrix) is positive definite at (b_0, b_1). This is the multivariable equivalent of the condition that the second derivative of a single-variable function must be positive to identify a minimum.

The matrix of second derivatives with respect to b_0 and b_1 is

$$\begin{bmatrix} \partial^2 S/\partial b_0^2 & \partial^2 S/\partial b_0 b_1 \\ \partial^2 S/\partial b_1 b_0 & \partial^2 S/\partial b_1^2 \end{bmatrix} = \begin{bmatrix} 2n & 2\sum x_i \\ 2\sum x_i & 2\sum x_i^2 \end{bmatrix}$$

To show that this matrix is positive definite, it suffices to show that all its principal minors are positive. We start by verifying that the determinant is positive. The determinant is

$$4n \sum x_i^2 - 4 \left(\sum x_i\right)^2 \tag{4.11}$$

But $\sum x_i = n\bar{x}$, so we can write (4.11) as

$$4n \left(\sum x_i^2 - n\bar{x}^2 \right) = 4n \left(\sum (x_i - \bar{x})^2 \right) \tag{4.12}$$

which is positive. Since $\partial^2 S/\partial b_0^2 = 2n$ is also positive, we know that b_0 and b_1 are the minimizers of the sum of squared errors.

Let's use these general formulas to compute the slope and intercept of our simple example. In the table below we compute the necessary quantities:

i	x	y	xy	x^2
1	0	1	0	0
2	1	3	3	1
3	2	4	8	4
4	3	3	9	9
5	4	5	20	16
$\sum$	10	16	40	30

From the quantities in this table we compute

$$b_1 = \frac{5 \cdot 40 - 10 \cdot 16}{5 \cdot 30 - 10^2} = \frac{4}{5} \qquad b_0 = \frac{16}{5} - \frac{4}{5}\frac{10}{5} = \frac{8}{5}$$

Expression (4.10) for the slope is convenient for computational purposes, but a more insightful expression for b_1 is

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \tag{4.13}$$

From this expression we can see immediately that the least squares solution is undefined when x has variance zero (i.e. only one value of x occurs).

We show below how to derive expression (4.13) from expression (4.10). Divide both the numerator and denominator of (4.10) by n:

$$b_1 = \frac{\sum x_i y_i - 1/n \sum x_i \sum y_i}{\sum x_i^2 - 1/n (\sum x_i)^2} \tag{4.14}$$

After rewriting the numerator becomes

$$\begin{aligned} \sum x_i y_i - 1/n \sum x_i \sum y_i &= \sum x_i y_i - \bar{x} \sum y_i \\ &= \sum x_i y_i - \sum \bar{x} y_i \\ &= \sum (x_i - \bar{x}) y_i \\ &= \sum (x_i - \bar{x})(y_i - \bar{y}) \end{aligned}$$

In the last step we make use of the following fact:

$$\begin{aligned} \sum (x_i - \bar{x})(y_i - \bar{y}) &= \sum x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y} \\ &= \sum (x_i - \bar{x}) y_i - \bar{y}(x_i - \bar{x}) \\ &= \sum (x_i - \bar{x}) y_i - \bar{y} \sum (x_i - \bar{x}) \\ &= \sum (x_i - \bar{x}) y_i \end{aligned}$$

since $\sum(x_i - \bar{x}) = 0$. Take good notice of the idea of this proof, it will prove very useful in other proofs as well.

Rewriting the denominator we get

$$\begin{aligned} \sum x_i^2 - 1/n(\sum x_i)^2 &= \sum x_i^2 - \sum x_i\bar{x} \\ &= \sum x_i^2 - x_i\bar{x} \\ &= \sum (x_i - \bar{x})x_i \\ &= \sum (x_i - \bar{x})^2 \end{aligned} \tag{4.15}$$

You should be able to justify the last step!

4.2 The coefficient of determination

We want to use x_i to explain as much of the variation in y_i as possible: we introduce the *explanatory* variable x_i in hope that its variation will *explain* the variation in y_i.

To develop a measure of the variation in y_i that is explained by the model, we begin by separating y_i into its explained and unexplained components:

$$y_i = \hat{y}_i + e_i \tag{4.16}$$

where $\hat{y}_i = b_0 + b_1 x_i$ and $e_i = y_i - \hat{y}_i$.

In figure 4.2 the "point of means" $(\bar{x}, \bar{y})$ is shown, with the least squares fitted line passing through it. This is a characteristic of the least squares fitted line whenever the regression model includes an intercept term. Subtract the sample mean from both sides in equation (4.16) to obtain

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + e_i \tag{4.17}$$

In figure 4.2, the difference between y_i and its mean value $\bar{y}$ consists of a part $(\hat{y}_i - \bar{y})$ that is "explained" by the fitted line, and a part e_i that is unexplained.

The breakdown in equation (4.17) leads to a useful decomposition of the total variability in y, within an entire sample, into explained and unexplained parts. There are many ways to measure the "total variation" in a variable. One convenient way is to square the differences between y_i and its mean

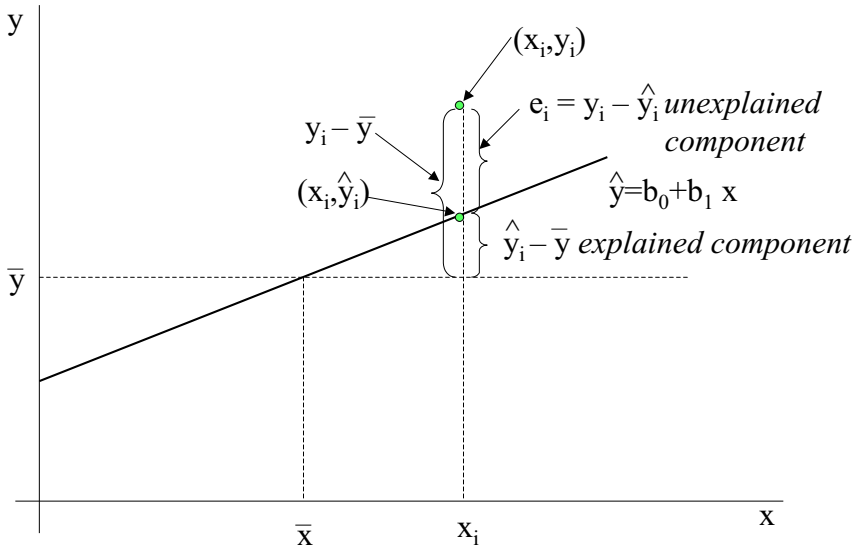

Figure 4.2: Decomposition of deviation from mean into explained and unexplained parts

value $\bar{y}$ and sum over the entire sample. If we square and sum both sides of equation (4.17), we obtain

$$\begin{aligned}\sum(y_i-\bar{y})^2 &= \sum[(\hat{y}_i-\bar{y})+e_i]^2 \\ &= \sum(\hat{y}_i-\bar{y})^2+\sum e_i^2+2\sum(\hat{y}_i-\bar{y})e_i \\ &= \sum(\hat{y}_i-\bar{y})^2+\sum e_i^2 \qquad (4.18)\end{aligned}$$

because the cross-product term $\sum(\hat{y}_i-\bar{y})e_i=0$ and drops out.

To see this substitute $b_0+b_1x_i$ for $\hat{y}_i$ in $\sum(\hat{y}_i-\bar{y})e_i$ to get

$$\sum(b_0+b_1x_i-\bar{y})e_i=\sum(\bar{y}-b_1\bar{x}+b_1x_i-\bar{y})e_i=\sum b_1(x_i-\bar{x})e_i \qquad (4.19)$$

since $b_0=\bar{y}-b_1\bar{x}$. Now we make use of the fact that $\sum e_i=0$ and $\sum e_ix_i=0$. This follows immediately from the first order conditions $\partial S/\partial b_0=0$ and $\partial S/\partial b_1=0$. Rewrite (4.19) to get

$$\sum b_1(x_i-\bar{x})e_i \;=\; b_1\left(\sum e_ix_i-\bar{x}e_i\right)$$

$$= \quad b_1 \left(\sum e_i x_i - \bar{x} \sum e_i \right) = 0 \tag{4.20}$$

Equation (4.18) is a decomposition of the "total sample variation" in y into explained and unexplained components. Specifically the sums of squares are:

1. $\sum(y_i - \bar{y})^2$ = total sum of squares = STT : a measure of the *total variation* in y around its sample mean.
2. $\sum(\hat{y}_i - \bar{y})^2$ = explained sum of squares = SSR: that part of total variation in y about its sample mean that is explained by the fitted line.
3. $\sum e_i^2$ = error sum of squares = SSE: that part in total variation in y about its sample mean that is not explained by the fitted line.

Thus equation (4.18) becomes

$$\text{SST} = \text{SSR} + \text{SSE} \tag{4.21}$$

One widespread use of this decomposition is to define a measure of the *proportion of variation* in y explained by x:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{4.22}$$

The measure R^2 is called the *coefficient of determination*. The closer R^2 is to 1, the better the job we have done in explaining the variation in y_i with $\hat{y}_i = b_0 + b_1 x_i$; and the greater is the predictive ability of our model over all the sample observations. If $R^2 = 1$, then all the sample data fall exactly on the fitted least squares line, so SSE=0, and the model fits the data "perfectly". If the sample data for y and x are uncorrelated and show no linear association then the least squares fitted line is horizontal and identical to $\bar{y}$, so that SSR=0 and $R^2 = 0$. When $0 < R^2 < 1$, it is interpreted as the percentage of variation in y about its mean that is explained by the fitted model.

Let's make the calculations for our simple example. The table below contains all the necessary numbers to compute R^2.

i	x	y	$\hat{y}$	e	e^2	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$
1	0	1	8/5	−3/5	9/25	121/25	64/25
2	1	3	12/5	3/5	9/25	1/25	16/25
3	2	4	16/5	4/5	16/25	16/25	0
4	3	3	20/5	−1	25/25	1/25	16/25
5	4	5	24/5	1/5	1/25	81/25	64/25
$\sum$	10	16	16	0	60/25	220/25	160/25

First of all, we verify that the sum of the errors is zero, as it should be. Secondly we see that

$$\begin{array}{ccccc} 220/25 & = & 60/25 & + & 160/25 \\ (SST) & & (SSE) & & (SSR) \end{array}$$

Finally, we compute R^2 as

$$R^2 = \frac{SSR}{SST} = \frac{160}{220} \approx 0.73$$

So about 73% of the variation in y is explained by the variation in x.

4.2.1 Example in Splus: Relation between weight and blood pressure

This example is taken from [8]. The weight (kg) and systolic blood pressure (mmHg) of 26 randomly selected males in the age group 25-30 are shown in table 4.1. Assume we are interested in modeling blood pressure as a function of weight. The scatterplot in figure 4.3 suggests that a straight line might give a reasonable fit.

Subject	Weight	Systolic BP	Subject	Weight	Systolic BP
1	75	130	14	78	153
2	76	133	15	72	128
3	82	150	16	76	132
4	70	128	17	79	149
5	96	151	18	83	158
6	79	146	19	98	150
7	86	150	20	88	163
8	95	140	21	82	156
9	91	148	22	65	124
10	68	125	23	109	170
11	72	133	24	107	165
12	77	135	25	87	160
13	77	150	26	85	159

Table 4.1: Data for bloodpressure example

We can use the `lm` function in Splus to fit a line by least squares.

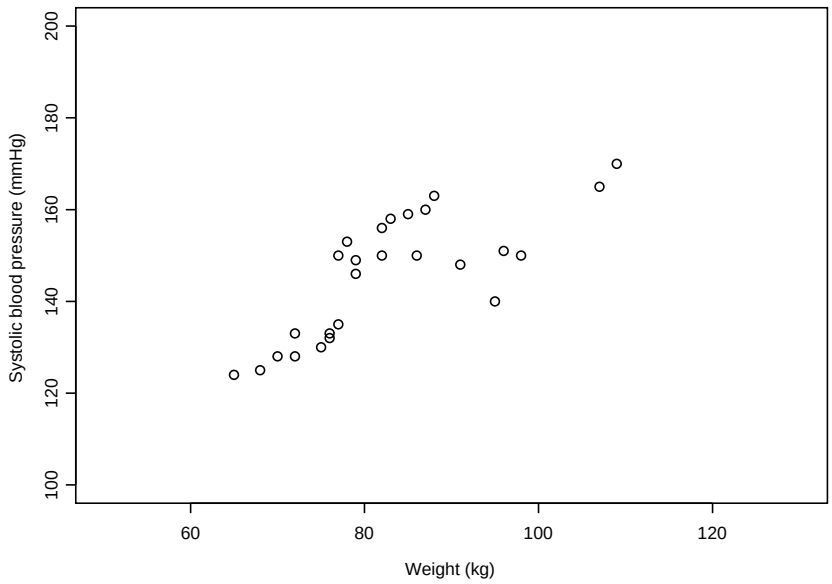

Figure 4.3: Scatterplot of weight against systolic bloodpressure

```
> bloodkg.fit <- lm(sys.bp ~ weight, data = bloodpressure.kg)

> summary(bloodkg.fit)

Call: lm(formula = sys.bp ~ weight, data = bloodpressure.kg)
Residuals:
    Min     1Q Median     3Q  Max
 -16.84 -6.66 -2.778 9.022 12.6

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 69.3578 12.9491     5.3562  0.0000
     weight  0.9209  0.1550     5.9410  0.0000

Residual standard error: 8.714 on 24 degrees of freedom
Multiple R-Squared: 0.5952
```

```
F-statistic: 35.3 on 1 and 24 degrees of freedom,
             the p-value is 3.94e-006

Correlation of Coefficients:
       (Intercept)
weight -0.9913
```

The variable `bloodkg.fit` gets the result of a call to `lm` assigned to it. Applying `summary` to `bloodkg.fit` gives quite some information, some of which we will discuss later in the course. We can read of the least squares estimates under the heading `Coefficients` in the `Value` column, so the least squares line is

$$\text{BLOODPRESSURE} = 69.3578 + 0.9209 \times \text{WEIGHT}$$

The fraction of variation in bloodpressure explained by variation in weight, R^2 is 0.5952. This is fairly high for a model with only one explanatory variable.

4.3 The Simple Linear Regression Model

So far, we have simply considered the problem of fitting a straight line to a given dataset by the method of least squares. Now we bring the problem within the realm of statistical inference. The data we have is in fact a sample from a larger population, and we want to draw conclusions about the population from the sample.

We assume that each unit in the population is described by two variables denoted by $\mathcal{X}$ and $\mathcal{Y}$ (we use calligraphic letters to denote population variables). The pair $(x_i, y_i), i = 1, \ldots, N$ denotes the values of element i for $\mathcal{X}$ and $\mathcal{Y}$.

Now for every value x of $\mathcal{X}$ different $\mathcal{Y}$ values can occur. We assume however that the *mean* of these $\mathcal{Y}$ values is always equal to $\beta_0 + \beta_1 x$:

$$\mu_{y.x} = \beta_0 + \beta_1 x \tag{4.23}$$

where $\mu_{y.x}$ denotes the mean of the $\mathcal{Y}$ values at x. Equation (4.23) is called the population regression line. The assumption is illustrated in figure 4.4. In

this population only four different $\mathcal{X}$ values occur, the means of the $\mathcal{Y}$ values are all on the line $\beta_0 + \beta_1 x$.

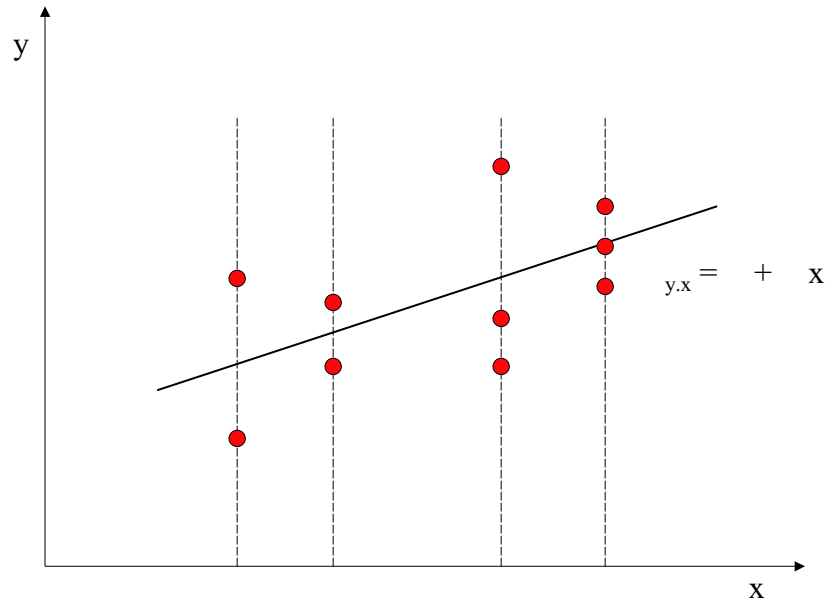

Figure 4.4: The central assumption of linear regression: $\mu_{y.x} = \beta_0 + \beta_1 x$.

If all pairs $(x_1, y_1), \ldots, (x_N, y_N)$ were known, then it would be easy to determine the value of β_0 and β_1. The problem of inductive statistics is to draw conclusions concerning β_0 and β_1 on the basis of a sample of observations. In doing so, it is of great importance to be able to assess the quality of those conclusions. Therefore, we have to make a few additional assumptions.

Depending on the way the observations are obtained, we distinguish between two cases. The most usual assumption is that the observation of the pair $(\mathcal{X}, \mathcal{Y})$ for one unit leads to a *pair* of random variables (X_i, Y_i). This would for example be the case if we sample the units at random from the population and observe their $(\mathcal{X}, \mathcal{Y})$ values.

Another possibility is that we select values x_i of $\mathcal{X}$ beforehand; in this case we speak of a *deterministic* explanatory variable. You can think of the x's as values the experimenter has chosen and set in a laboratory experiment. From the elements with value x_i one unit is selected at random, and for this

unit we observe the $\mathcal{Y}$ value. This case is relatively easy because we have to deal with only one random variable (Y_i). Therefore we will start with the deterministic case; we discuss random explanatory variables in section 4.5.

In the deterministic case we have observations

$$(x_i, Y_i), \qquad i = 1, 2, \ldots, n$$

The values $y_1, y_2, \ldots, y_n$ are observed values of independent random variables $Y_1, Y_2, \ldots, Y_n$. In terms of these observations, our basic assumption becomes

$$E(Y_i) = \beta_0 + \beta_1 x_i, \qquad i = 1, \ldots, n \tag{4.24}$$

where β_0 and β_1 are the true yet unknown intercept and slope respectively. Furthermore, we assume that for each value of x, the values of Y are distributed about their mean value, following probability distributions that all have the same (unknown) variance

$$\text{var}(Y_i) = \sigma^2 \tag{4.25}$$

Note that this assumption is not satisfied in the example of figure 4.4.

The model can also be expressed in this way. Assume that

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n \tag{4.26}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent random variables with

$$E(\varepsilon_i) = 0 \text{ and } \text{var}(\varepsilon_i) = \sigma^2$$

The $\varepsilon_1, \ldots, \varepsilon_n$ are called random errors (disturbances). Since Y_i depends only on ε_i, and the ε_i are independent, it follows that the Y_is are independent. It is also easily verified that

$$\begin{aligned} E(Y_i) &= E(\beta_0 + \beta_1 x_i + \varepsilon_i) \\ &= \beta_0 + \beta_1 x_i + E(\varepsilon_i) \\ &= \beta_0 + \beta_1 x_i \end{aligned}$$

Finally,

$$\begin{aligned} \text{var}(Y_i) &= \text{var}(\beta_0 + \beta_1 x_i + \varepsilon_i) \\ &= \text{var}(\varepsilon_i) = \sigma^2 \end{aligned}$$

For ease of reference we summarize the assumptions of the linear regression model

SLR1: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

SLR2: $E(\varepsilon_i) = 0 \Leftrightarrow E(Y_i) = \beta_0 + \beta_1 x_i$

SLR3: $\text{var}(\varepsilon_i) = \sigma^2 = \text{var}(Y_i)$

SLR4: $\varepsilon_i, \varepsilon_j$ independent $\Leftrightarrow Y_i, Y_j$ independent $(i \neq j)$.

SLR5: x_i is not random and must take at least two different values

4.3.1 Properties of Least Squares Estimators

How good are the least-squares estimators from a frequentist perspective? The frequentist consistently asks: *what happens if we do this a lot of times*? What happens if we sample $Y_1, \ldots, Y_n$ very many times and for each of those samples compute the least squares estimates b_0 and b_1 of β_0 and β_1?

We consider a number of questions

1. Are the least squares estimators unbiased?
2. What is their variance?
3. What more can we say about their sampling distribution?

Unbiasedness of the least squares estimators

Recall that we call an estimator G of parameter θ *unbiased* if

$$E_\theta(G) = \theta$$

where the expectation is taken with repect to repeated samples of some fixed size from the population. It is considered a desirable property that the estimator is right on average.

To show that b_1 is an unbiased estimator of β_1, we start with the following expression for b_1:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{4.27}$$

Since

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum(x_i - \bar{x})y_i$$

we can write b_1 as a weighted sum of the y values as follows

$$b_1 = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} = \sum \left[\frac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \right] y_i = \sum w_i y_i$$

Taking expectations, we get

$$\begin{aligned} E(b_1) &= E\left(\sum w_i y_i\right) \\ &= \sum w_i E(y_i) \\ &= \sum w_i(\beta_0 + \beta_1 x_i) \qquad \text{since } E(y_i) = \beta_0 + \beta_1 x_i \\ &= \beta_0 \sum w_i + \beta_1 \sum w_i x_i \\ &= \beta_1 \end{aligned} \tag{4.28}$$

In the last step we use the fact that $\sum w_i = 0$ (since $\sum(x_i - \bar{x}) = 0$), and $\sum w_i x_i = 1$.

Since we have shown that

$$E(b_1) = \beta_1,$$

so we may conclude that the least squares estimator b_1 is an unbiased estimator of the slope β_1 of the regression line.

Likewise, one can show that b_0 is an unbiased estimator of β_0. First we prove that

$$E(\bar{y}) = \beta_0 + \beta_1 \bar{x}$$

Since $\bar{y} = 1/n \sum y_i$, we get

$$\begin{aligned} E(\bar{y}) &= E(1/n \sum y_i) = 1/n E\left(\sum y_i\right) \\ &= 1/n \sum \beta_0 + \beta_1 x_i \\ &= 1/n(n\beta_0 + \beta_1 \sum x_i) \\ &= \beta_0 + \beta_1 1/n \sum x_i = \beta_0 + \beta_1 \bar{x} \end{aligned}$$

Now since

$$b_0 = \bar{y} - b_1 \bar{x}$$

we get

$$\begin{aligned} E(b_0) &= E(\bar{y} - b_1 \bar{x}) = E(\bar{y}) - E(b_1 \bar{x}) \\ &= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0 \end{aligned}$$

Variance and covariance of the least squares estimators

Unbiasedness is a nice property for an estimator, but on its own it doesn't say that much. If the estimator has a very high variance, any particular estimate is typically far from the true value. Therefore we are also interested in the variance of the least-squares estimators. First we present the formulas for variance and covariance of the least squares estimators, and try to make sense of them.

$$\begin{aligned} \text{var}(b_0) &= \sigma^2 \left(\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \right) \\ \text{var}(b_1) &= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \\ \text{cov}(b_0, b_1) &= \sigma^2 \left(\frac{-\bar{x}}{\sum (x_i - \bar{x})^2} \right) \end{aligned} \tag{4.29}$$

We consider the factors that affect the variances and covariance in (4.29).

1. The variance of the random error term σ^2 appears in all three expressions. The larger the variance term σ^2, the greater the uncertainty there is in the statistical model, and the larger the variances and covariance of the least squares estimators.

2. The sum of squares of the values of x about their sample mean, $\sum (x_i - \bar{x})^2$ appears in each of the variances and the covariance. The larger the sum of squares, the smaller the variances of the least squares estimators and hence the more precisely we can estimate the unknown parameters. The intuition behind this is demonstrated in figure 4.5. In the upper panel the x values are widely spread out along the x-axis. In the lower panel the data are bunched together. The data in the upper panel give more information on where the least squares line must fall, because they are more spread out along the x axis.

3. "More data is better than less data". The larger the sample size n, the smaller the variances and covariance of the least squares estimators. The sample size n appears in each of the variances and covariance, because each of the sums consists of n terms.

4. The term $\sum x_i^2$ appears in $\text{var}(b_0)$. The larger this term is, the larger the variance of the least squares estimator b_0. Recall that the intercept parameter β_0 is the expected value of y at $x = 0$. The farther our data are from $x = 0$, the more difficult it is to interpret β_0, and the more difficult it is to accurately estimate β_0. The term $\sum x_i^2$ measures the distance of the data from the origin $x = 0$.

5. The sample mean of the x-values appears in $\text{cov}(b_0, b_1)$. The least squares line must pass through the point of the means $(\bar{x}, \bar{y})$. Given a fitted line through the data, imagine the effect of increasing the estimated slope b_1. Since the line must pass through the point of the means, the effect must be to lower the point where the line hits the vertical axis, implying a reduced intercept estimate b_0. Thus, when the sample mean is positive, there is a negative covariance between the least squares estimators of the slope and intercept.

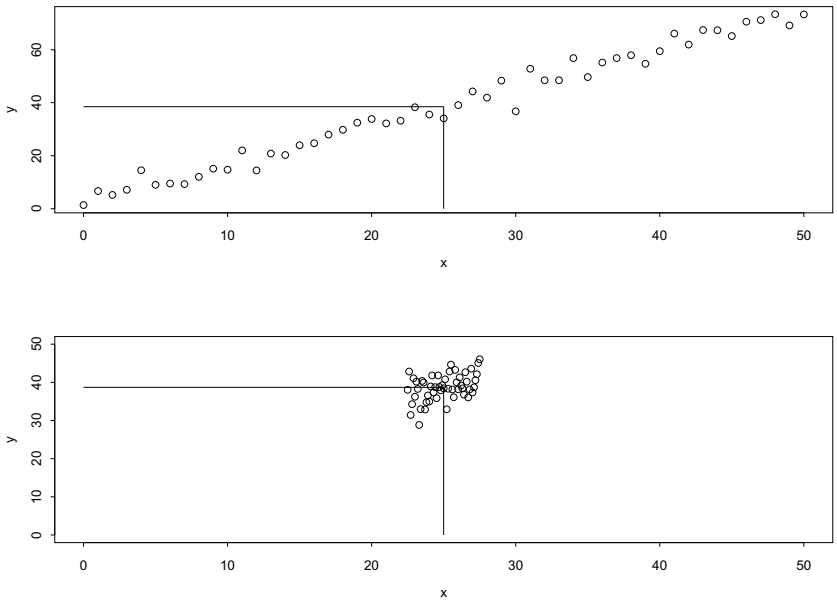

Figure 4.5: The influence of variation in the explanatory variable on the precision of estimation

As an example we show the derivation of the variance of b_1. We start from expression

$$b_1 = \sum w_i y_i,$$

with

$$w_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$$

Taking the variance, we get

$$\begin{aligned}
\text{var}(b_1) &= \text{var}\left(\sum w_i y_i\right) \\
&= \sum w_i^2 \text{var}(y_i) \quad \text{since var}(cX) = c^2 \text{ var}(X) \text{ and } y_i, y_j \text{ independent} \\
&= \sigma^2 \sum w_i^2 \quad \text{since var}(y_i) = \sigma^2 \\
&= \frac{\sigma^2}{\sum (x_i - \bar{x})^2}
\end{aligned}$$

Since

$$\sum w_i^2 = \frac{1}{\sum (x_i - \bar{x})^2}$$

The variance of b_0, and covariance of b_0 and b_1 can be derived in a similar manner.

The Gauss-Markov Theorem

So now we know the expected value and the variance of the least-squares estimators. We have shown that they are unbiased, and have derived expressions for the variance. The question remains whether there perhaps are unbiased estimators of β_0 and β_1 that have a smaller variance than the least squares estimators and would therefore be preferrable. For the class of *linear* estimators, the answer is given by the Gauss-Markov theorem

> Under the assumptions SLR1-SLR5 of the linear regression model, the estimators b_0 and b_1 have the smallest variance of all linear and unbiased estimators of β_0 and β_1. They are the Best Linear Unbiased Estimators (BLUE) of β_0 and β_1.

An estimator is called *linear* when it can be written as a linear combination of the y_i. We have shown for example that $b_1 = \sum w_i y_i$ so b_1 is a linear estimator of β_1. Likewise, b_0 is a linear estimator of β_0. Some further remarks about the significance of this theorem:

1. Why this pre-occupation with linear estimators? The reasons are primarily computational and analytic convenience.

2. Note that the theorem does not require normality of the error term ε_i.

3. When ε_i is normally distributed, then the least squares estimators are the best of *all* unbiased estimators (linear or non-linear).

4. When we drop the normality assumption, there can be in some cases non-linear (robust) estimators that are better than least squares.

As an illustration we prove the Gauss-Markov theorem for the estimator b_1 of β_1. Let $\hat{\beta}_1 = \sum k_i y_i$ be any other linear estimator of β_1. To make the comparison we write $k_i = w_i + c_i$, that is choose $c_i = k_i - w_i$. We substitute y_i into this new estimator and simplify using the properties $\sum w_i = 0$ and $\sum w_i x_i = 1$.

$$\begin{aligned}
\hat{\beta}_1 &= \sum k_i y_i = \sum (w_i + c_i) y_i = \sum (w_i + c_i)(\beta_0 + \beta_1 x_i + \varepsilon_i) \\
&= \sum (w_i + c_i)\beta_0 + \sum (w_i + c_i)\beta_1 x_i + \sum (w_i + c_i)\varepsilon_i \\
&= \beta_0 \sum w_i + \beta_0 \sum c_i + \beta_1 \sum w_i x_i + \beta_1 \sum c_i x_i + \sum (w_i + c_i)\varepsilon_i \\
&= \beta_0 \sum c_i + \beta_1 + \beta_1 \sum c_i x_i + \sum (w_i + c_i)\varepsilon_i \qquad (4.30)
\end{aligned}$$

Take the expectation of (4.30) and use the assumption that $E(\varepsilon_i) = 0$ (SLR2)

$$\begin{aligned}
E(\hat{\beta}_1) &= \beta_0 \sum c_i + \beta_1 + \beta_1 \sum c_i x_i + \sum (w_i + c_i) E(\varepsilon_i) \\
&= \beta_0 \sum c_i + \beta_1 + \beta_1 \sum c_i x_i \qquad (4.31)
\end{aligned}$$

In order for $\hat{\beta}_1$ to be unbiased, it must be true that $E(\hat{\beta}_1) = \beta_1$ for all values of β_0 and β_1. Using (4.31) we see that this implies that

$$\sum c_i = 0 \text{ and } \sum c_i x_i = 0 \qquad (4.32)$$

We use these constraints to simplify expression (4.30)

$$\hat{\beta}_1 = \beta_1 + \sum (w_i + c_i)\varepsilon_i \qquad (4.33)$$

Using the properties of variance we can now write the variance of $\hat{\beta}_1$ as follows

$$\begin{aligned}
\text{var}(\hat{\beta}_1) &= \text{var}(\beta_1 + \sum(w_i + c_i)\varepsilon_i) = \sum(w_i + c_i)^2 \text{var}(\varepsilon_i) \\
&= \sigma^2 \sum(w_i + c_i)^2 = \sigma^2 \sum w_i^2 + \sigma^2 \sum c_i^2 \quad (\text{since } \textstyle\sum c_i w_i = 0) \\
&= \text{var}(b_1) + \sigma^2 \sum c_i^2 \\
&\geq \text{var}(b_1) \text{ since } \sum c_i^2 \geq 0
\end{aligned} \tag{4.34}$$

Probability distributions of the least squares estimators

If we make one additional assumption

SLR6: $\varepsilon_i \sim N(0, \sigma^2) \Leftrightarrow Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$,

then it follows that

$$\begin{aligned}
b_0 &\sim N\left(\beta_0, \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}\right) \\
b_1 &\sim N\left(\beta_1, \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right)
\end{aligned} \tag{4.35}$$

This is true because the least squares estimators are linear estimators, and weighted sums of normal random variables are normally distributed themselves. If the errors are not normally distributed, but assumptions SLR1-SLR5 hold, then if the sample size n is *sufficiently large*, by the central limit theorem, the least squares estimators have a distribution that is well approximated by the normal distributions shown in (4.35).

Unfortunately, the variance of the error term σ^2 that appears in the formulas for the variance of b_0 and b_1 is typically unknown. This means we have to estimate it from the data. Recall that

$$\text{var}(\varepsilon_i) = \sigma^2 = E[\varepsilon_i - E(\varepsilon_i)]^2 = E(\varepsilon_i^2) \tag{4.36}$$

since $E(\varepsilon_i) = 0$ by assumption SLR2. Since the expectation is an average value, we consider estimating σ^2 as the average of the squared errors

$$\hat{\sigma}^2 = \frac{\sum \varepsilon_i^2}{n} \tag{4.37}$$

The random errors

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

are however unobservable, since we don't know the values of β_0 and β_1. It seems reasonable to replace the random errors in equation 4.37 by the least squares residuals

$$e_i = y_i - b_0 - b_1 x_i$$

to obtain

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n} \tag{4.38}$$

If we compute the expected value of the numerator, we get

$$E\left(\sum e_i^2\right) = (n-2)\sigma^2,$$

which implies that an *unbiased* estimator of σ^2 is

$$s^2 = \frac{\sum e_i^2}{n-2} \tag{4.39}$$

The number that is subtracted from the sample size, is the number of regression parameters (β_0, β_1) in the model that we have to estimate before we can estimate the variance of the error term. To estimate β_0 and β_1 we use up two *degrees of freedom*, which then leaves $n-2$ degrees of freedom to estimate the error variance.

We can see the reason most clearly in case we have only $n = 2$ observed data points. In that case the least squares line will always provide a perfect fit. For *any* two points, a line can always be drawn that goes through them exactly. Thus, although b_0 and b_1 would be easily determined in that case, there would be no "information left over" to tell us anything about σ^2, the variance of the observations about the regression line. Only to the extent that n exceeds 2 can we get information about σ^2. That is $n-2$ degrees of freedom remain when we use s^2 to estimate σ^2.

Now that we have an unbiased estimator of the error variance, we can estimate the variance of the least squares estimators b_0 and b_1, as well as the covariance between them. Replace the unknown error variance σ^2 in (4.29) by its estimator to obtain

$$\widehat{\text{var}}(b_0) \;=\; s^2\left(\frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\right)$$

$$\begin{aligned} \widehat{\text{var}}(b_1) &= \frac{s^2}{\sum(x_i - \bar{x})^2} \\ \widehat{\text{cov}}(b_0, b_1) &= s^2 \left(\frac{-\bar{x}}{\sum(x_i - \bar{x})^2} \right) \end{aligned} \tag{4.40}$$

Furthermore, we define the standard errors $\text{se}(b_0)$ and $\text{se}(b_1)$ of b_0 and b_1 respectively, to be the square root of the estimated variances.

4.3.2 Interval Estimation and Hypothesis Testing

If the assumptions SLR1-SLR6 are correct, then the least squares estimators b_0 and b_1 are normally distributed random variables with means and variances as follows:

$$\begin{aligned} b_0 &\sim N\left(\beta_0, \frac{\sigma^2 \sum x_i^2}{n \sum(x_i - \bar{x})^2} \right) \\ b_1 &\sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \right) \end{aligned} \tag{4.41}$$

We can create a standard normal random variable based on the normal distribution of the least squares estimator. A standardized random variable is obtained from b_1 by subtracting its mean and dividing by its standard deviation

$$Z = \frac{b_1 - \beta_1}{\text{sd}(b_1)} \sim N(0, 1) \tag{4.42}$$

where $\text{sd}(b_1) = \sqrt{\text{var}(b_1)}$ is called the standard deviation of b_1. That is, the standardized random variable Z is normally distributed with mean 0 and variance 1.

When we replace the unknown parameter σ^2 with its unbiased estimator s^2, then

$$t = \frac{b_1 - \beta_1}{\text{se}(b_1)} \sim t_{(n-2)} \tag{4.43}$$

where $\text{se}(b_1) = \sqrt{\widehat{\text{var}}(b_1)}$.

The shape of the t-distribution is completely determined by the degrees of freedom parameter, m, and the distribution is symbolized by $t_{(m)}$. The t-distribution is symmetric with mean $E[t_{(m)}] = 0$ and variance $\text{var}[t_{(m)}] = m/(m-2)$ As the degrees of freedom parameter $m \to \infty$, the $t_{(m)}$ distribution approaches the standard normal $N(0, 1)$. Result (4.43) is used to construct

confidence intervals for β_1 and to perform hypothesis tests concerning the value of β_1.

Interval Estimation

We have seen that if assumptions SLR1-SLR6 hold, then

$$t = \frac{b_1 - \beta_1}{\text{se}(b_1)} \sim t_{(n-2)} \tag{4.44}$$

and similarly

$$t = \frac{b_0 - \beta_0}{\text{se}(b_0)} \sim t_{(n-2)} \tag{4.45}$$

The random variable t in (4.44) and (4.45) will be the basis for interval estimation and hypothesis testing in the simple linear regression model.

Using a computer or a statistical table, we can find critical values $t_{(m);\alpha/2}$ from a $t_{(m)}$ distribution such that

$$P(t \geq t_{(m);\alpha/2}) = P(t \leq -t_{(m);\alpha/2}) = \frac{\alpha}{2}$$

where α is a probability value often taken to be $\alpha = 0.01$ or $\alpha = 0.05$.

The critical values $t_{(m);\alpha/2}$ and $-t_{(m);\alpha/2}$ are depicted in figure 4.6. Each of the shaded tail areas contains $\alpha/2$ of the probability, so that $1 - \alpha$ of the probability is contained in the center portion. Therefore, we can make the probability statement

$$P(-t_{(m);\alpha/2} \leq t \leq t_{(m);\alpha/2}) = 1 - \alpha \tag{4.46}$$

Now, we put all these pieces together to create a procedure for interval estimation. Substitute t from (4.44) in (4.46) to obtain

$$P\left[-t_{(n-2);\alpha/2} \leq \frac{b_1 - \beta_1}{\text{se}(b_1)} \leq t_{(n-2);\alpha/2})\right] = 1 - \alpha \tag{4.47}$$

Simplify the expression to obtain

$$P[b_1 - t_{(n-2);\alpha/2}\ \text{se}(b_1) \leq \beta_1 \leq b_1 + t_{(n-2);\alpha/2}\ \text{se}(b_1)] = 1 - \alpha \tag{4.48}$$

In the interval endpoints $b_1 - t_{(n-2);\alpha/2}\ \text{se}(b_1)$ and $b_1 + t_{(n-2);\alpha/2}\ \text{se}(b_1)$, both b_1 and $\text{se}(b_1)$ are random variables, since their values are not known

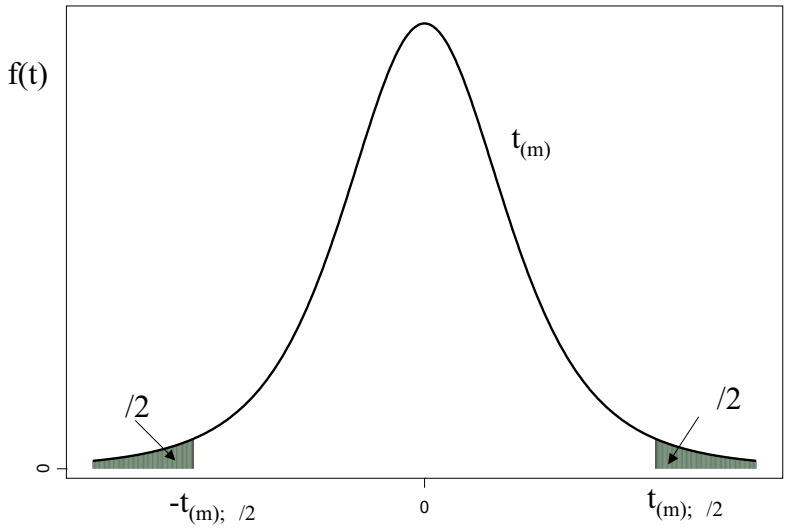

Figure 4.6: Critical values from a t-distribution

until a sample of data is drawn. The random endpoints of the interval define an *interval estimator* of β_1. The probability statement in (4.48) says that the interval

$$b_1 \pm t_{(n-2);\alpha/2}\ \mathrm{se}(b_1),$$

with random endpoints, has probability $1-\alpha$ of containing the true but unknown parameter β_1. This interval estimation procudure and its properties are established based on model assumptions SLR1-SLR6 and may be applied to any sample of data we might obtain.

When b_1 and $\mathrm{se}(b_1)$ in (4.48) are *estimated values*(numbers), based on a sample of data, then $b_1 \pm t_{(n-2);\alpha/2}\ \mathrm{se}(b_1)$ is called a $(1-\alpha)\times 100\%$ confidence interval for β_1. The interpretation of interval estimators and interval estimates requires a great deal of care. The properties of the *random interval estimator* are based on the notion of repeated sampling. If we were to select *many* random samples of size n, compute the least squares estimate b_1, and its standard error $\mathrm{se}(b_1)$ for each sample, then $(1-\alpha)\times 100\%$ of all the intervals constructed would contain the true parameter β_1. This we know before any data are actually collected.

Any one interval estimate, based on a sample of data, may or may not

contain the true parameter β_1, and since β_1 is unknown, we will *never* know if it does or not. Our confidence is in *the procedure used to construct the interval estimate*; it is not in any one interval estimate calculated from a sample of data.

To illustrate, we construct a 95% confidence interval for β_1 in the blood pressure example. Here's the output that Splus provides when we apply linear regression to the bloodpressure data:

```
> summary(bloodkg.fit)

Call: lm(formula = sys.bp ~ weight, data = bloodpressure.kg)
Residuals:
    Min     1Q Median     3Q  Max
 -16.84 -6.66 -2.778 9.022 12.6

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) 69.3578 12.9491     5.3562  0.0000
     weight  0.9209  0.1550     5.9410  0.0000

Residual standard error: 8.714 on 24 degrees of freedom
Multiple R-Squared: 0.5952
F-statistic: 35.3 on 1 and 24 degrees of freedom,
             the p-value is 3.94e-006

Correlation of Coefficients:
       (Intercept)
weight -0.9913
```

Since weight (kg) is the explanatory variable in this model and blood pressure (mmHg) the response variable, the slope β_1 measures the expected change in blood pressure when the weight increases by 1 kg. The point estimate b_1 of the slope is 0.9209 (see the `Value` column in the `Coefficient` table). Next to the point estimate we find the standard error 0.1550. Since we have 26 observations in our data set, we have to find the critical value $t_{(24);0.025}$. This critical value is computed in S-plus as follows

```
> qt(0.975,df=24)
[1] 2.063899
```

So $t_{(24);0.025} \approx 2.064$. The 95% confidence interval therefore becomes

$$0.9209 \pm 2.064 \times 0.1550 = (0.60, 1.24)$$

Let's compare this interval with the one we would get if we *knew for a fact* that $\text{sd}(b_1) = 0.1550$. In that case we could use (4.42) instead of (4.43) to create the confidence interval, i.e. we can use the critical values of the standardnormal distribution rather than the $t_{(n-2)}$ distribution. The 95% confidence interval then becomes

$$0.9209 \pm z_{\alpha/2} \times 0.1550 = 0.9209 \pm 1.96 \times 0.1550 = (0.62, 1.22)$$

The t-distribution has fatter tails than the standardnormal, so the critical values $t_{(m);\alpha/2}$ are larger than the corresponding critical values $z_{\alpha/2}$. This results in a wider interval at the same confidence level. This makes sense: because the standard deviation of b_1 has to be estimated, uncertainty is added, which results in less precise conclusions.

Hypothesis testing

Before we concern ourselves with statistical hypothesis testing, let's look at the simpler deterministic case. Suppose we want to consider the hypothesis: *two objects of different weights will fall at the same speed*. To test this hypothesis, we drop two canonballs, one large and one small, from the tower of Pisa. The outcome of the test is that the canon balls indeed strike at nearly the same instant, which supports our hypothesis.

The general structure of the hypothesis testing method is as follows

1. According to the hypothesis an observable quantity x should have the value x_0.

2. If we observe a value of x different from x_0, we must reject the hypothesis.

3. The observation that $x = x_0$ serves to support the hypothesis.

In a *statistical* hypothesis test we reason in a similar fashion. We state some hypothesis concerning the value of a population parameter (β_1 or β_0 in the linear regression model) and use some observable quantity (the point estimates b_1 and b_0 and their standard errors) to make a decision. The observable quantity used to make the decision is called a test statistic. The most important difference with the deterministic case is that a statistical hypothesis will usually assign a positive probability to all possible values of the test statistic. This means that whatever value of the test statistic we observe, we are never able to conclude with absolute certainty that the maintained hypothesis is false. What we sometimes *can* say is that the observed value of the test statistic is highly unlikely if the maintained hypothesis were true. In that case we would reject the maintained hypothesis.

The components of a statistical hypothesis test are:

1. A *null* hypothesis, H_0.
2. An *alternative hypothesis*, H_a.
3. A test *statistic*.
4. A *rejection* region.

The null hypothesis specifies a value for a population parameter, and is stated

$$H_0 : \beta_1 = c$$

where c is a constant, and is an important value in the context of a specific regression model. A null hypothesis is the belief we will maintain until we are convinced by the sample evidence that it is not true, in which case we *reject* the null hypothesis.

For the null hypothesis $H_0 : \beta_1 = c$, three possible alternatives are

- $H_a : \beta_1 \neq c$. Rejecting the null hypothesis that $\beta_1 = c$ implies the conclusion that β_1 takes some other value greater than or less than c.
- $H_a : \beta_1 > c$. Rejecting the null hypothesis that $\beta_1 = c$ leads to the conclusion that it is greater than c. Using this alternative *completely* discounts the possibility that $\beta_1 < c$. It implies that these values are logically unacceptable alternatives to the null hypothesis.
- $H_a : \beta_1 < c$. Following the previous discussion, use this alternative when there is no chance that $\beta_1 > c$.

The sample information about the null hypothesis is embodied in the sample value of a *test statistic*. Based on the value of a test statistic, which itself is a random variable, we decide either to reject the null hypothesis or not to reject it. A test statistic has a very special characteristic: its probability distribution must be *completely known when the null hypothesis is true*, and it must have some *other* distribution if the null hypothesis is not true.

The *rejection region* is the range of values of the test statistic that leads to rejection of the null hypothesis. It is possible to construct a rejection region only if we have a test statistic whose distribution is known when the null hypothesis is true. In practice, the rejection region is a set of test statistic values that, *when the null hypothesis is true*, are unlikely and have *low probability* of occurring. If a sample value of the test statistic is obtained that falls in a region of low probability, then it is unlikely that the test statistic has the assumed distribution, and thus it is unlikely that the null hypothesis is true.

If the null hypothesis $H_0 : \beta_1 = c$ is true, then it follows from (4.43) that the test statistic

$$t = \frac{b_1 - c}{\text{se}(b_1)} \sim t_{(n-2)}$$

Thus, if the hypothesis is true, then the distribution of t is that shown in figure 4.6. If the alternative hypothesis $H_a : \beta_1 \neq c$ is true, then values of the test statistic will tend to be unusually "large" or unusually "small". The terms *large* and *small* are determined by choosing a probability α, called the level of significance of the test, which provides a meaning for "an unlikely event". The rejection region is determined by finding critical values $t_{(n-2);\alpha/2}$ such that

$$P(t \geq t_{(n-2);\alpha/2}) = P(t \leq -t_{(n-2);\alpha/2}) = \frac{\alpha}{2}$$

Thus the rejection region consists of the two "tails" of the t-distribution.

When the null hypothesis is true, the probability of obtaining a sample value of the test statistic that falls in either tail area is "small" and, combined, is equal to α. Sample values of the test statistic that are in the tail areas are incompatible with the null hypothesis and are evidence against the null hypothesis being true. When testing the null hypothesis $H_0 : \beta_1 = c$ against the alternative $H_a : \beta_1 \neq c$ we are led to the following rule: if the value of the test statistic falls in either tail of the t-distribution, then we reject the null hypothesis and accept the alternative. If the value of the test statistic falls between the critical values $-t_{\alpha/2}$ and $t_{\alpha/2}$, then *we do not reject* the null

hypothesis. The test decision rules are summarized in figure 4.7.

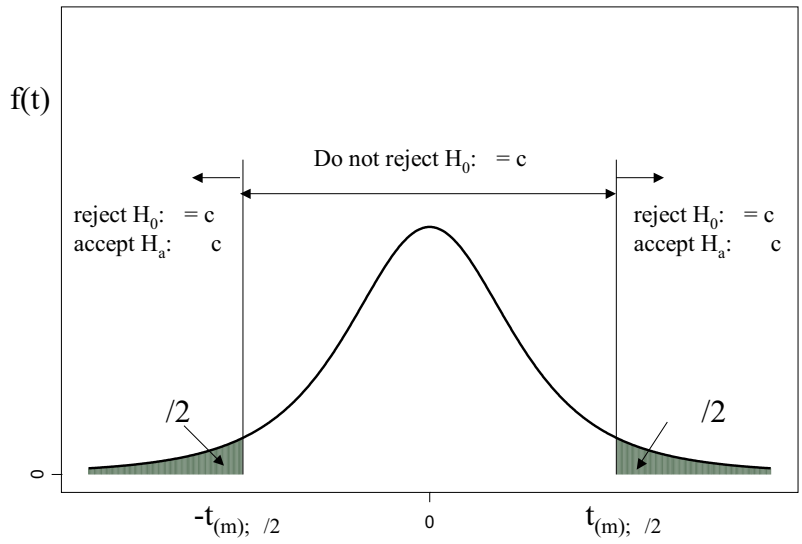

Figure 4.7: Rejection region for a test of $H_0 : \beta_1 = c$ against $H_a : \beta_1 \neq c$.

Let's use our blood pressure data to consider some examples of statistical tests. Suppose you claim that when weight increases with one kilogram, the systolic blood pressure is expected to rise with 1 mmHg. Your roommate doubts this and says it must be some different value. You decide to collect relevant data, and test the null hypothesis $H_0 : \beta_1 = 1$ against $H_a : \beta_1 \neq 1$. You decide to select $\alpha = 0.05$. The critical value $t_{(24);0.024}$ is 2.064, so we reject the null hypothesis when $|t| \geq 2.064$.

The observed t-value is

$$t = \frac{b_1 - c}{\text{se}(b_1)} = \frac{0.9209 - 1}{0.1550} = -0.51,$$

which means we do not reject H_0 since the observed t-value does not fall in the reject region. As usual you can claim you were right.

Now suppose your other roommate claims that weight has no effect at all on blood pressure. Now you decide to test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$. The observed t-value is

$$t = \frac{b_1 - c}{\text{se}(b_1)} = \frac{0.9209}{0.1550} = 5.94,$$

and since this is bigger than 2.064, we reject H_0.

Rather than computing critical values for specific values of α, we can also do the following. Compute the probability under H_0 of observing a t-value *at least as far from zero* as the t value we actually observed, i.e. compute $P(|t_{(n-2)}| > |t|)$. This probability is called the observed significance or p-value.

For the first test, we compute $P(|t_{(24)}| > 0.51)$ as follows in S-plus

```
> 2*pt(-0.51,df=24)
[1] 0.6147099
```

This means that if H_0 is true, the probability of observing a t value at least as far from zero as the one we actually observed is 0.615, which is pretty high. Since the p-value is larger than α, we accept H_0. For the second test we compute $P(|t_{(24)}| > 5.94)$ as follows

```
> 2*pt(-5.94,df=24)
[1] 3.949709e-006
```

Since the p-value is smaller than α, we reject H_0.

4.3.3 Estimation of expected value and prediction

Drawing conclusions about population parameters β_0 and β_1, is not the only possible objective of regression analysis. On many occasions we want to use the regression line for the purpose of prediction of future observations. We consider two closely related problems

1. Point and interval *estimation* of the expected value $E(y_0)$ of y_0 at some point x_0.
2. Point and interval *prediction* of y_0 for a future observation with x-value x_0.

According to the assumptions the y values are drawn independently, so y_0 is independent of the earlier observations and has distribution

$$y_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$$

We start with the first problem.

Estimation of $E(y_0)$

Obviously, $\hat{y}_0 = b_0 + b_1 x_0$ is an unbiased estimator of $E(y_0) = \beta_0 + \beta_1 x_0$, since

$$E(\hat{y}_0) = E(b_0 + b_1 x_0) = \beta_0 + \beta_1 x_0$$

To be able to make confidence intervals we have to know the variance of the point estimator $\hat{y}_0$.

$$\begin{aligned}
\text{var}(\hat{y}_0) &= \text{var}(b_0 + b_1 x_0) \\
&= \text{var}(b_0) + x_0^2 \text{var}(b_1) + 2x_0 \text{cov}(b_0, b_1) \\
&= \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2} + \frac{\sigma^2 x_0^2}{\sum (x_i - \bar{x})^2} - \frac{2\sigma^2 x_0 \bar{x}}{\sum (x_i - \bar{x})^2} \\
&= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \left[\frac{1}{n} \sum x_i^2 + x_0^2 - 2x_0 \bar{x} \right] \\
&= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \left[\frac{1}{n} \sum x_i^2 - \bar{x}^2 + \bar{x}^2 + x_0^2 - 2x_0 \bar{x} \right] \\
&= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \left[\frac{1}{n} \left\{ \sum x_i^2 - \frac{1}{n} \left(\sum x_i \right)^2 \right\} + (x_0 - \bar{x})^2 \right] \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]
\end{aligned}$$

Here we use the rule that

$$\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y) + 2ab \text{cov}(X, Y)$$

In the last step we use the fact that

$$\sum x_i^2 - \frac{1}{n} \left(\sum x_i \right)^2 = \sum (x_i - \bar{x})^2$$

as was shown in (4.15).

The formula for $\text{var}(\hat{y}_0)$ implies that the farther x_0 is from the sample mean, the less reliable the estimation of the mean of y_0. This result is reasonable, since we would not expect to be able to estimate the mean of y_0 very accurately for an x about which we have little sample information. Graphically this point is illustrated in figure 4.8. This figure shows two lines

with different slopes, both passing through the point of means $(\bar{x}, \bar{y})$. For values of x near the mean the change in slope results in a small change in $\hat{y}$. The further x is from its mean, the larger the change in $\hat{y}$ due to the change in the slope of the line.

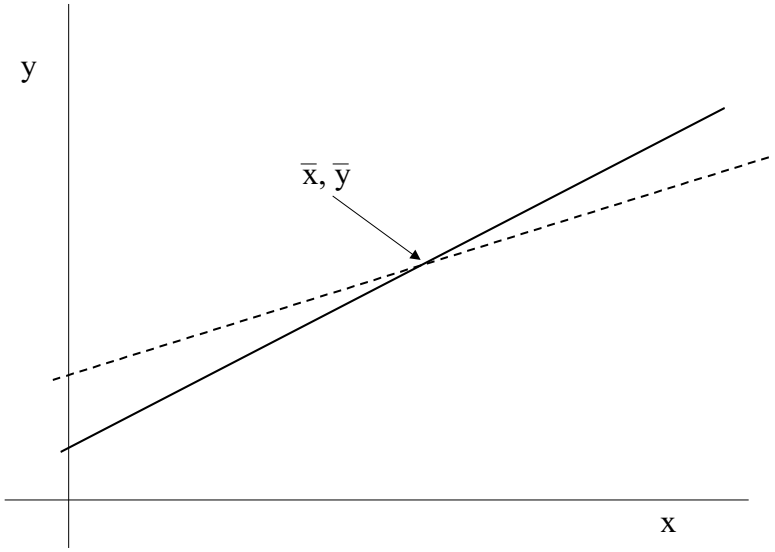

Figure 4.8: The further we get from the mean, the larger the change in $\hat{y}$ due to a change in the slope of the line.

Since $\hat{y}_0$ is a linear combination of b_0 and b_1, it is normally distributed as well, so we have

$$\hat{y}_0 \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]\right)$$

As usual, we have to estimate σ^2 by s^2. Standardization then gives

$$\frac{\hat{y}_0 - (\beta_0 + \beta_1 x_0)}{\text{se}(\hat{y}_0)} \sim t_{(n-2)}$$

where

$$\text{se}(\hat{y}_0) = \sqrt{\widehat{\text{var}}(\hat{y}_0)}$$

and

$$\widehat{\text{var}}(\hat{y}_0) = s^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

Analogous to the way we constructed confidence intervals for β_0 and β_1, we can now conclude that

$$\hat{y}_0 \pm t_{(n-2);\alpha/2} \text{ se}(\hat{y}_0)$$

is a $(1 - \alpha) \times 100\%$ confidence interval for $E(y_0) = \beta_0 + \beta_1 x_0$.

For an example, we return to our by now familiar blood pressure example. Let's create a 95% confidence interval for the mean blood pressure of people who weigh 90kg. The point estimate of mean blood pressure at 90kg is

$$\hat{y}_0 = 69.3578 + 0.9209 \times 90 = 152.24$$

The estimated error variance s^2 is 75.9277 (S-plus gives its square root as *residual standard error*), so we compute

$$\begin{aligned} \widehat{\text{var}}(\hat{y}_0) &= s^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] \\ &= 75.93 \left[\frac{1}{26} + \frac{(90 - 82.81)^2}{3160.038} \right] = 4.16 \end{aligned}$$

So

$$\text{se}(\hat{y}_0) = \sqrt{4.16} = 2.04$$

and

$$\hat{y}_0 \pm t_{(24);0.025} \text{ se}(\hat{y}_0) = 152.24 \pm 2.064 \times 2.04 = (148.03, 156.45)$$

is a 95% confidence interval for mean blood pressure at weight=90kg.

The relationship between point and interval estimates for different values of x_0 is illustrated in figure 4.9. A point estimate is always given by the fitted least squares line, $\hat{y}_0 = b_0 + b_1 x_0$. The confidence intervals take the form of two bands around the least squares line. Since the estimation variance increases the farther x_0 is from the sample mean, the confidence bands increase in width as $|x_0 - \bar{x}|$ increases.

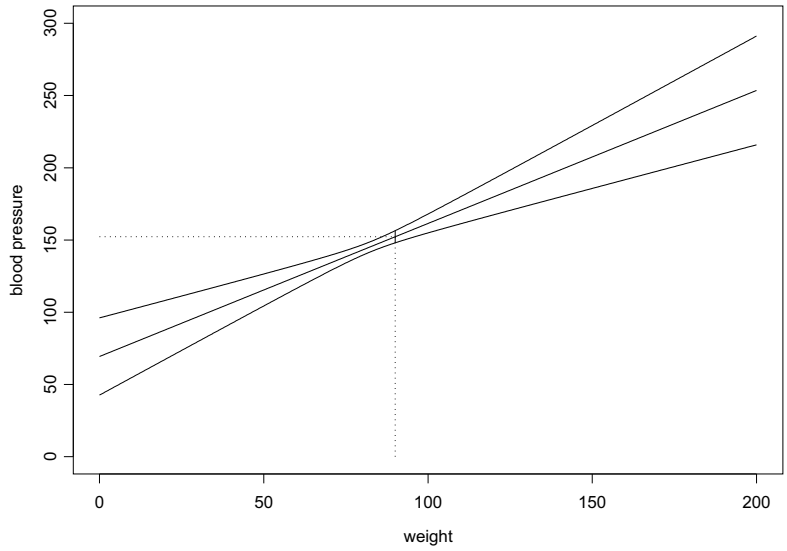

Figure 4.9: 95% confidence intervals for the blood pressure example. The interval we computed is indicated by the solid vertical line at weight=90kg

Prediction of y_0

The least squares predictor $\hat{y}_0$ of y at x_0 is

$$\hat{y}_0 = b_0 + b_1 x_0 \tag{4.49}$$

This prediction is given by the point on the least squares fitted line at $x = x_0$. To evaluate the sampling properties of this predictor we examine the prediction error $\hat{y}_0 - y_0$. Since

$$E(\hat{y}_0) = \beta_0 + \beta_1 x_0 = E(y_0)$$

it follows that

$$E(\hat{y}_0 - y_0) = 0$$

which means that $\hat{y}_0 = b_0 + b_1 x_0$ is an unbiased predictor of y_0. Now y_0 is independent of $y_1, \ldots, y_n$ and therefore also of b_0, b_1, and $\hat{y}_0$ since they are computed from $y_1, \ldots, y_n$. This means we can write

$$\text{var}(\hat{y}_0 - y_0) = \text{var}(\hat{y}_0) + \text{var}(y_0)$$

Looking at this formula we can see there are two sources of uncertainty in the prediction of y_0. The first is the uncertainty concerning the true mean of y_0 at x_0, and the second is the uncertainty due to the spread of y_0 around its mean.

Since by assumption $\text{var}(y_0) = \sigma^2$ and $\text{var}(\hat{y}_0)$ was derived in the previous section, we can write

$$\text{var}(\hat{y}_0 - y_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$$

From normality of $\hat{y}_0$ and y_0 it follows that

$$\hat{y}_0 - y_0 \sim N\left(0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]\right)$$

Estimation of σ^2 by s^2 and standardization gives

$$\frac{\hat{y}_0 - y_0}{\text{se}(\hat{y}_0 - y_0)} \sim t_{(n-2)}$$

so

$$\hat{y}_0 \pm t_{(n-2);\alpha/2}\ \text{se}(\hat{y}_0 - y_0)$$

is a $(1 - \alpha) \times 100\%$ prediction interval for y_0.

As an example we create a 95% prediction interval for someone who weighs 90kg, i.e. $x_0 = 90$. The point prediction is

$$\hat{y}_0 = b_0 + b_1 x_0 = 69.3578 + 0.9209(90) = 152.24$$

This means that we predict that a person who weighs 90kg, will have a systolic bloodpressure of 152.24 mmHg. The estimated variance of the prediction error is

$$\begin{aligned}\widehat{\text{var}}(\hat{y}_0 - y_0) &= s^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] \\ &= 75.9277 \left[1 + \frac{1}{26} + \frac{(90 - 82.81)^2}{3160.038}\right] = 79.02075\end{aligned}$$

The standard error of the prediction is then

$$\text{se}(\hat{y}_0 - y_0) = \sqrt{\widehat{\text{var}}(\hat{y}_0 - y_0)} = \sqrt{79.02075} = 8.8894$$

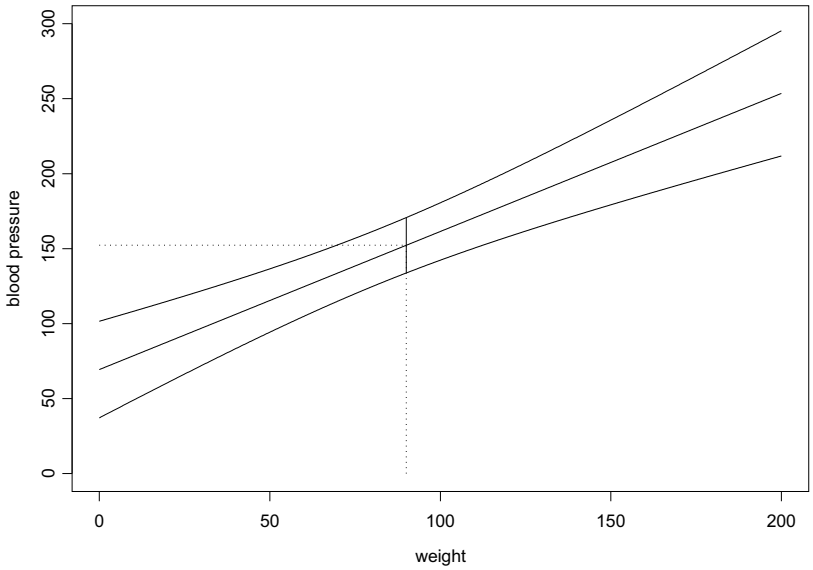

Figure 4.10: 95% prediction intervals for bloodpressure example.The interval we computed is indicated by the solid vertical line at weight=90kg

We select $\alpha = 0.05$. Since $t_{(24);\alpha/2} = 2.064$, the 95% prediction interval for y_0 is

$$\hat{y}_0 \pm t_{(n-2);\alpha/2}\ \mathrm{se}(\hat{y}_0 - y_0) = 152.24 \pm 2.064(8.8894) = (133.8924, 170.5876)$$

This interval is indicated by the solid vertical line in figure 4.10. Note that, as expected, the 95% prediction interval is wider than the 95% confidence interval for the mean. As we already indicated this is due to the additional uncertainty caused by the spread of y_0 around its mean.

4.4 Maximum likelihood estimation of the simple linear regression model

We can also apply the method of maximum likelihood to find estimates of the unknown parameters of the linear regression model. This means we aim to find those values of $(\beta_0, \beta_1, \sigma^2)$ that maximize the probability of the sample

we have actually observed. In order to do so, we have to make assumptions about the probability distribution of ε_i (otherwise we can not compute the probability of the sample). The usual assumption is that $\varepsilon_i \sim N(0, \sigma^2)$, and as a consequence $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.

Recall that the density function $N(\mu, \sigma^2)$ is given by

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}(y-\mu)^2/\sigma^2} \quad \text{for } y \in \mathbb{R}$$

Plugging in $\beta_0 + \beta_1 x_i$ for μ_i we get

$$f(y_i) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}(y_i-\beta_0-\beta_1 x_i)^2/\sigma^2}$$

The likelihood function establishes the probability of observing all the n observations in our sample, i.e. $\mathcal{L}(\beta_0, \beta_1, \sigma^2) = f(y_1, y_2, \ldots, y_n)$. Assuming the observations are independent, we get

$$\begin{aligned} \mathcal{L}(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}(y-\beta_0-\beta_1 x_i)^2/\sigma^2} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2} \end{aligned}$$

Taking logs yields

$$\ln \mathcal{L}(\beta_0, \beta_1, \sigma^2) = n \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \qquad (4.50)$$

It can be seen from equation 4.50 that regardless of the value of σ^2, the values of β_0 and β_1 for which the likelihood function is a maximum will be the values for which the following sum of squares is a minimum

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

In other words, the MLE's of the regression coefficients β_0 and β_1 are the same as the least squares estimators b_0 and b_1.

The MLE of σ^2 can be found by first replacing β_0 and β_1 in equation 4.50 by their MLE's $\hat{\beta}_0$ and $\hat{\beta}_1$ and then maximizing the resulting expression with respect to σ^2. Taking the derivative of

$$n \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

with respect to σ and equating to zero, we get

$$-\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sigma^3} = 0$$

Solving this expression for σ^2 yields

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{\sum e_i^2}{n}$$

We have already seen that this is a biased estimator of σ^2 (see equation (4.38)). As n increases however, the bias gets smaller and smaller. Hence the maximum likelihood estimator of σ^2 is *asymptotically* unbiased.

We have shown that the least squares estimates and maximum likelihood estimates of β_0 and β_1 coincide when the error has a normal distribution. To show that this is by no means always the case, we see what happens when the errors have some other distribution. Assume that ε_i has an exponential distribution with mean $\mu_i = \beta x_i$ (for ease of exposition we assume the intercept is zero). That is

$$f_i(y) = \frac{1}{\mu_i} e^{-y/\mu_i} \qquad\qquad y \geq 0$$

For observations $y_1, \ldots, y_n$, the log-likelihood is

$$\begin{aligned}
\mathcal{L}(\beta) &= \sum_{i=1}^{n} \log \frac{1}{\beta x_i} e^{-y_i/\beta x_i} \\
&= \sum_{i=1}^{n} \log \frac{1}{\beta x_i} - \frac{y_i}{\beta x_i} \\
&= -\sum_{i=1}^{n} \log \beta x_i + \frac{y_i}{\beta x_i}
\end{aligned}$$

To obtain the maximum we compute the derivative of the log-likelihood with respect to β

$$\frac{\partial \mathcal{L}}{\partial \beta} = -\sum_{i=1}^{n} \frac{1}{\beta} - \frac{y_i}{\beta^2 x_i}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \left(\frac{y_i}{\beta^2 x_i} - \frac{1}{\beta} \right) \\
&= \sum_{i=1}^{n} \frac{y_i}{\beta^2 x_i} - \frac{n}{\beta}
\end{aligned}
$$

and equate to zero:

$$\sum_{i=1}^{n} \frac{y_i}{\beta^2 x_i} - \frac{n}{\beta} = 0$$

$$\frac{1}{\beta^2} \sum_{i=1}^{n} \frac{y_i}{x_i} - \frac{n}{\beta} = 0$$

$$\frac{1}{\beta} \sum_{i=1}^{n} \frac{y_i}{x_i} = n$$

$$nb = \sum_{i=1}^{n} \frac{y_i}{x_i}$$

So

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i}$$

This estimator is different from the least squares estimator wich was shown to be

$$b = \frac{\sum x_i y_i}{\sum x_i^2}$$

It is straightforward to show that $\hat{\beta}$ is an unbiased estimator:

$$
\begin{aligned}
E(\hat{\beta}) &= E\left(\frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i} \right) \\
&= \frac{1}{n} E\left(\sum_{i=1}^{n} \frac{y_i}{x_i} \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{E(y_i)}{x_i} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\beta x_i}{x_i}
\end{aligned}
$$

$$= \frac{1}{n} n\beta = \beta$$

Now for the variance of $\hat{\beta}$. Since y_i has exponential distribution with mean $\mu_i = \beta x_i$, the variance of y_i is

$$V(y_i) = \mu_i^2 = (\beta x_i)^2 = \beta^2 x_i^2$$

Now

$$\begin{aligned} V(\hat{\beta}) &= V\left(\frac{1}{n}\sum \frac{y_i}{x_i}\right) \\ &= \frac{1}{n^2} V\left(\sum \frac{y_i}{x_i}\right) \\ &= \frac{1}{n^2}\sum \frac{1}{x_i^2} V(y_i) \\ &= \frac{1}{n^2}\sum \frac{\beta^2 x_i^2}{x_i^2} = \frac{\beta^2}{n} \end{aligned}$$

Finally, we compute the variance of the least-squares estimator:

$$\begin{aligned} V(b) &= V\left(\sum \frac{x_i}{\sum x_j^2} y_i\right) \\ &= \sum \left(\frac{x_i}{\sum x_j^2}\right)^2 V(y_i) \\ &= \sum \frac{x_i^2}{\left(\sum x_j^2\right)^2} \beta^2 x_i^2 \\ &= \beta^2 \frac{\sum x_i^4}{\left(\sum x_j^2\right)^2} \\ &= \frac{\beta^2}{n} \frac{\sum x_i^4}{\frac{1}{n}\left(\sum x_i^2\right)^2} \end{aligned}$$

So the relative efficiency of b as compared to $\hat{\beta}$ is $\frac{1}{n}\left(\sum x_i^2\right)^2 / \sum x_i^4$ which is always ≤ 1. This means that the maximum likelihood estimator is always at least as good as the least squares estimator. The more spread out the x_i are, the smaller the efficiency of least squares. Since $V(y_i) = \beta^2 x_i^2$, the more diverse the values of x, the more unequal the variances.

4.5 When X is random

So far we have assumed that the explanatory variable x was fixed in repeated samples, and y was the only random variable we had to deal with. This assumption is realistic when we think of an experimenter that selects the x values and measures the outcome y. In actual data analysis practice this is hardly ever the case. Typically we draw a random sample of units from a population and both x and y are observed for the units selected into the sample. Hence they are *both* random variables. To what extent does this influence the results we have obtained so far?

One cop out is to condition the entire analysis on the x values we have actually observed. This means the results we obtained so far go through with the added assumption that the x values are kept fixed at the values observed in the sample (rather than actually being preselected by the experimenter).

In fact this cop out is not necessary, since most of the results derived under deterministic x values, still go through when x is random. It is, for example, straightforward to show that the least-squares estimators of β_0 and β_1 are still unbiased when x is random. First we obtained the desired result conditioned on arbitrary x. This is the same as the deterministic case and so according to (4.28) we have

$$E(b_1 \mid x) = \beta_1$$

By the law of iterated expectations, we then have

$$E(b_1) = E_x(E(b_1 \mid x)) = E_x(\beta_1) = \beta_1$$

So the unbiasedness of the least squares estimators does not depend on the probability distribution of x. It only depends on the assumption that $E(\varepsilon \mid x) = 0$. Likewise, we can show (but we won't) that the Gauss-Markov theorem still holds when x is random, and that the maximum likelihood estimators and least squares estimators still coincide.

4.6 Diagnosis/Analysis of residuals

The procedures we discussed (estimation, testing, prediction) are only correct if the assumptions SLR1-SLR6 are satisfied. Therefore it is important to check whether these assumptions are correct. In regression analysis we use

the collected observations to verify whether the modelling assumptions are likely to be correct. We will look at the following requirements of the linear regression model:

1. $E(y) = \beta_0 + \beta_1 x$: Linearity in x
2. $\text{var}(\varepsilon_i) = \sigma^2$: Homoskedasticity
3. $\varepsilon_i, \varepsilon_j$ independent $(i \neq j)$
4. $\varepsilon_i \sim N(0, \sigma^2)$: Normality

These requirements all concern the disturbances ε_i. Because the least squares residuals e_i can be viewed as outcomes of the random variables ε_i, it makes sense to use the residuals to check the modelling assumptions.

4.6.1 Linearity

Linearity in x means that the disturbances in the linear model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

should satisfy $E(\varepsilon_i) = 0$, $i = 1, \ldots, n$.

This implies that the residuals should fluctuate randomly around 0, irrespective of the value x_i of the explanatory variable. By plotting e_i against x_i, we can verify whether this is the case.

The following example is taken from [10]. According to economists the need to borrow money decreases as there is more money circulating in the economy. This will then in turn cause the interest rate to decrease.

Take as explanatory variable the liquidity quote x, which is defined as the liquidity mass expressed as a percentage of national income. The independent variable y is the interest rate on government loans. We have observations for the years 1980-1989 (time series).

Using a linear specification, we obtain the following estimates

$$\hat{y} = 21.13 - 0.3165x$$

with $R^2 = 0.803$. The theory is supported by the data, since $b_1 < 0$. In figure 4.11 a scatterplot of (x_i, e_i) is displayed.

Note that for high and low values of x we have positive residuals, whereas in the middle we have negative residuals. This is indicative for a curvilinear

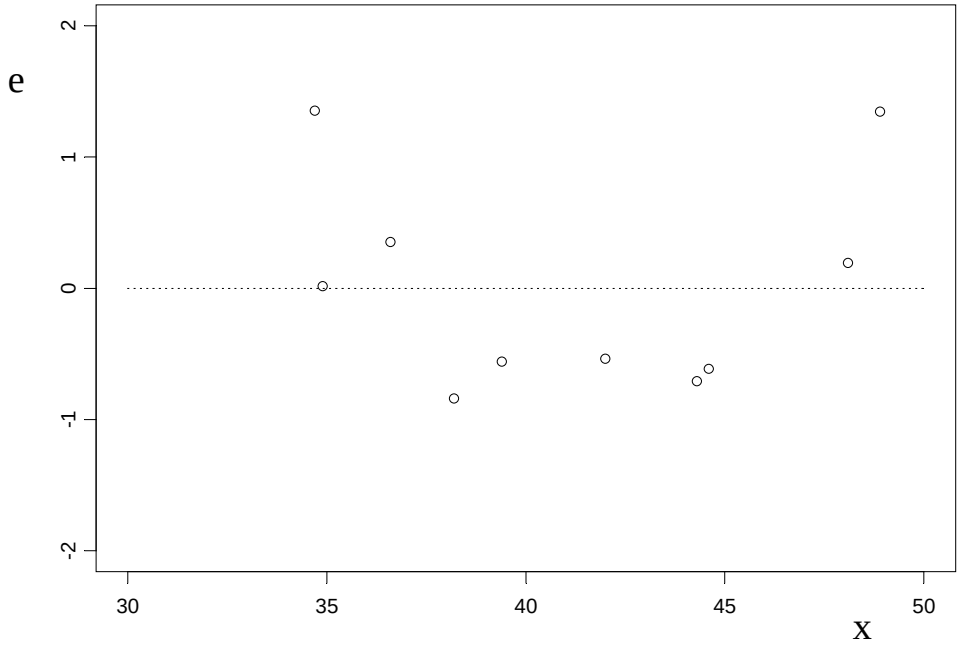

Figure 4.11: Plot of x_i against e_i for interest rate data.

relationship. Therefore we also include x^2 as an explanatory variable. When we estimate this specification, we het

$$\hat{y} = 80.22 - 3.1965x + 0.0364x^2$$

with $R^2 = 0.9516$. Note that the explained variance increases considerably (from 0.803 to 0.9516). Because the second model has an extra explanatory variable (x^2) included, we cannot immediately conclude from the increase in R^2 that the model is better. We return to this issue later (see 4.10).

Note furthermore that the model

$$y = \beta_0 + \beta_1 x_i + \beta_2 x^2 + \varepsilon$$

is not linear in the variables. We can however still apply linear regression analysis, as long as the model is linear in the *parameters*. By linear in the parameters we mean that the parameters are not multiplied together, divided, squared, etc. The variables, however, can be tranformed in any convenient way, as long as the resulting model satisfies assumptions SLR1-SLR5 of the simple linear regression model.

4.6.2 Homoskedasticity

This means that all disturbance terms have equal variance, i.e.

$$\text{var}(\varepsilon_i) = \sigma^2 \qquad\qquad i = 1, \ldots, n$$

The spread around the regression line is equally big everywhere. In practice this assumption is regularly violated.

In the food expenditure example (taken from [7]), e_i clearly increases with x_i (see figure 4.12). The data is given in table 4.2.

Household	Food Expenditure	Weekly Income	Household	Food Expenditure	Weekly Income
1	52.25	258.3	21	98.14	719.8
2	58.32	343.1	22	123.94	720.0
3	81.79	425.0	23	126.31	722.3
4	119.90	267.5	24	146.47	722.3
5	125.80	482.9	25	115.98	734.4
6	100.46	487.7	26	207.23	742.5
7	121.51	496.5	27	119.80	747.7
8	100.08	519.4	28	151.33	763.3
9	127.75	543.3	29	169.51	810.2
10	104.94	548.7	30	108.03	818.5
11	107.48	564.6	31	168.90	825.6
12	98.48	588.3	32	227.11	833.3
13	181.21	591.3	33	84.94	834.0
14	122.23	607.3	34	98.70	918.1
15	129.57	611.2	35	141.06	918.1
16	92.84	631.0	36	215.40	929.6
17	117.92	659.6	37	112.89	951.7
18	82.13	664.0	38	166.25	1014.0
19	182.28	704.2	39	115.43	1141.3
20	139.13	704.8	40	269.03	1154.6

Table 4.2: Food expenditure data

This means we have to let go of the homoskedasticity assumption

$$\text{var}(y_i) = \text{var}(\varepsilon_i) = \sigma^2$$

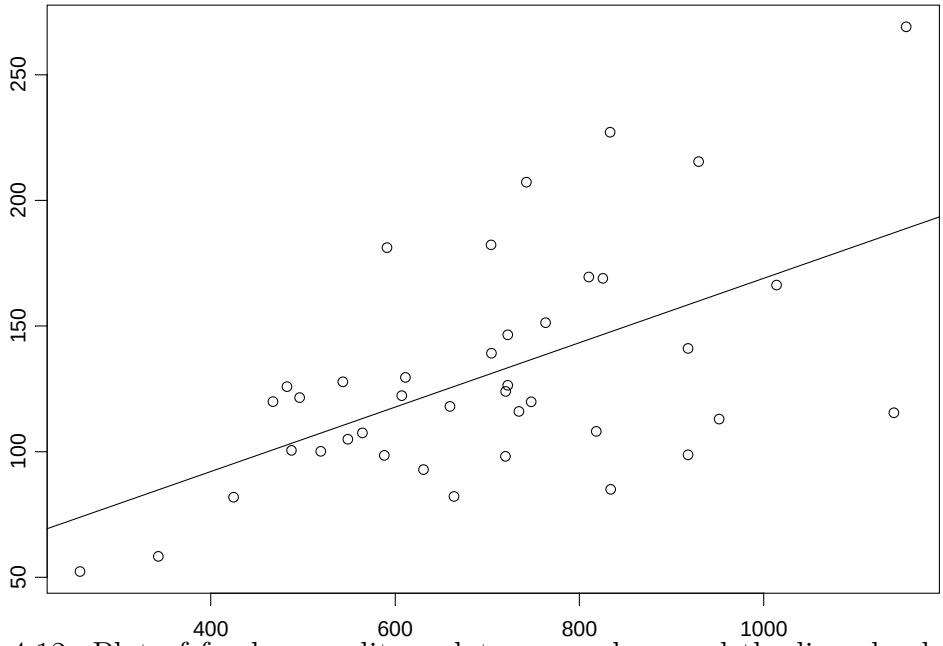

Figure 4.12: Plot of food expenditure data: spread around the line clearly increases with x (income).

This means the least squares estimators are no longer best; more specifically

- The least squares estimators are still unbiased, but they are no longer BLUE.
- The standard errors computed with the least squares estimators are incorrect. This means that confidence intervals and tests based on these standard errors can be misleading.

The most general assumption with respect to the disturbance term

$$\text{var}(y_i) = \text{var}(\varepsilon_i) = \sigma_i^2$$

is not useful because this would mean we have to estimate n different variances (and β_0 and β_1) from n observations. We can however use the assumption of proportional variance

$$\text{var}(\varepsilon_i) = \sigma_i^2 = \sigma^2 x_i$$

How do we fit this model to the data? The idea is to transform this model in order to get a model with homoskedastic disturbance. This works as follows. We start out with the model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Divide left and right by $\sqrt{x_i}$

$$\frac{y_i}{\sqrt{x_i}} = \beta_0 \left(\frac{1}{\sqrt{x_i}}\right) + \beta_1 \left(\frac{x_i}{\sqrt{x_i}}\right) + \left(\frac{\varepsilon_i}{\sqrt{x_i}}\right)$$

Define the following transformed variables:

$$y_i^* = \frac{y_i}{\sqrt{x_i}}, \quad x_{i1}^* = \frac{1}{\sqrt{x_i}}, \quad x_{i2}^* = \frac{x_i}{\sqrt{x_i}}, \quad \varepsilon_i^* = \frac{\varepsilon_i}{\sqrt{x_i}}$$

Then we can write

$$y_i^* = \beta_0 x_{i1}^* + \beta_1 x_{i2}^* + \varepsilon_i^*$$

The point of this whole exercise is that this model is homoskedastic, since

$$\text{var}(\varepsilon_i^*) = \text{var}\left(\frac{\varepsilon_i}{\sqrt{x_i}}\right) = \left(\frac{1}{\sqrt{x_i}}\right)^2 \text{var}(\varepsilon_i) = \frac{1}{x_i}\sigma^2 x_i = \sigma^2$$

So the procedure is simply to

1. Compute the transformed variables.

2. Apply ordinary least-squares to the transformed model.

This whole procedure can be interpreted as a *weighted* least squares method. The ordinary least squares method finds those values of b_0 and b_1 that minimize the sum of squared errors $\sum e_i^2$. In this case we minimize the transformed errors

$$\sum e_i^{*2} = \sum \left(\frac{e_i}{\sqrt{x_i}}\right)^2 = \sum \frac{e_i^2}{x_i}$$

So the squared error is weighted by $1/x_i$. When x_i is small (large), the data contains more (less) information about the regression function and the observations have a bigger (smaller) weight.

In the Splus session below, we first fit a regression model with ordinary least squares using the `lm` function. Note that the standard errors provided in the output are not reliable, because the errors are clearly heteroskedastic. To fit the model presented in this section (i.e. $\text{var}(\varepsilon_i) = \sigma_i^2 = \sigma^2 x_i$), we use the `gls` (for generalized least squares) function. By specifying `weights=varFixed( income)` in the call, we select the proportional variance model. Note that we get different estimates for the slope and intercept of the regression line.

```
> food.ols <- lm(y~x,data=fooddata)
> summary(food.ols)

Call: lm(formula = foodexp ~ income, data = fooddata)
Residuals:
    Min     1Q Median    3Q   Max
 -71.75 -19.67 -5.969 17.75 80.14

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) 40.7676 22.1387     1.8415  0.0734
   income    0.1283  0.0305     4.2008  0.0002

Residual standard error: 37.81 on 38 degrees of freedom
Multiple R-Squared: 0.3171
F-statistic: 17.65 on 1 and 38 degrees of freedom, the p-value is 0.000155

Correlation of Coefficients:
       (Intercept)
income -0.9629


> food.gls <- gls(foodexp ~ income,data=fooddata,weights=varFixed(~income))
> summary(food.gls)
Generalized least squares fit by REML
  Model: foodexp ~ income
  Data: fooddata
       AIC      BIC    logLik
```

```
  401.6017 406.5144 -197.8008

Variance function:
 Structure: fixed weights
 Formula:  ~ income

Coefficients:
               Value Std.Error  t-value p-value
(Intercept) 31.92438  17.98608 1.774949  0.0839
     income  0.14096   0.02700 5.221574  <.0001

 Correlation:
       (Intr)
income -0.955

Standardized residuals:
       Min         Q1        Med        Q3      Max
 -1.703248 -0.5866877 -0.1512335 0.6116881 2.016665

Residual standard error: 1.344599
Degrees of freedom: 40 total; 38 residual
```

4.6.3 Independence of the error terms

The assumption that the error terms are independent ($\text{cov}(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$) for example excludes the possibility that ε_i is influenced by ε_{i-1}. In case we have *time series* data, this assumption is usually not satisfied. The "disturbing influences" that are operative at time $i-1$, usually still exert their influence at time i. If that is the case, we speak of *autocorrelation*: the sequence $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ is correlated with itself because each component depends on its predecessors. The same will be true for the residuals $e_1, e_2, \ldots, e_n$. By plotting e_i against i or e_{i-1} we can check whether this is the case. The following example is taken from [7].

It makes sense that when the price of a particular crop is high, farmers tend to plant more of that crop than when the price is low. Let A denote the area planted and P the output price. We assume a log-log (constant

elasticity) functional form

$$\ln(A) = \beta_0 + \beta_1 \ln(P) + \varepsilon$$

Now β_1 is the percentage change in A arising from a one percent increase in P. This is what economists call the *elasticity* of A with respect to P. Suppose we have 34 annual observations on area and price. We fit a linear model is Splus:

```
> sugar.lm <- lm(log(area) ~ log(price),data=sugar)
> summary(sugar.lm)

Call: lm(formula = log(area) ~ log(price), data = sugar)
Residuals:
     Min      1Q   Median     3Q    Max
 -0.6508 -0.1882 -0.03096 0.2491 0.6049

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  6.1113  0.1686    36.2540  0.0000
 log(price)  0.9706  0.1106     8.7733  0.0000

Residual standard error: 0.3088 on 32 degrees of freedom
Multiple R-Squared: 0.7063
F-statistic: 76.97 on 1 and 32 degrees of freedom, the p-value is 5.031e-010

Correlation of Coefficients:
           (Intercept)
log(price) 0.9494
```

The plot of e_{i-1} against e_i suggests there is positive autocorrelation. Although the least squares estimators are still unbiased, they are no longer BLUE. There are more efficient estimators that exploit the correlation structure in the data. Furthermore, the standard errors reported by least squares are not reliable: with positive autocorrelation, they tend to underestimate the standard deviation of the estimators. This means, for example, that confidence intervals tend to be narrower than they should be.

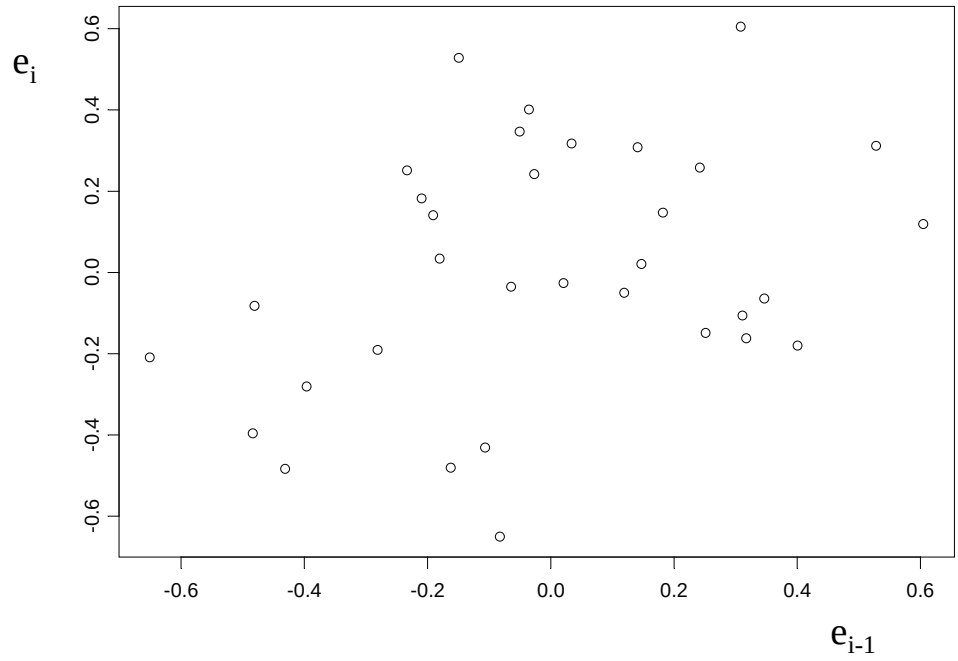

Figure 4.13: Plot of e_{i-1} against e_i for sugar cane data.

4.6.4 Normality of the error term

When we want to give a confidence interval, or test a hypothesis, for the slope β_1 and intercept β_0 of the regression line, we invoke the assumption that the error term ε is normally distributed. For large samples, the central limit theorem kicks in, and we don't have to worry about normality of the error term too much. Usually we are a few observations shy of infinity however.

One of the ways to compare the distribution of the residuals with a normal distribution is to use a so called Q-Q plot (short for quantile-quantile plot). For a sample x the quantile function is the inverse of the empirical distribution function

$$\text{quantile}(p) = \min\{z \mid \text{proportion } p \text{ of the data} \leq z\}$$

To check whether the residuals are approximately normally distributed, we plot the quantiles of the residuals against the quantiles of the standard normal distribution. When the residuals are normally distributed, the points

should lie approximately on a straight line.

We make a normal probability plot for the food data residuals in Splus.

```
> qqnorm(food.lm$residuals)
> qqline(food.lm$residuals)
```

This yields the graph displayed in figure 4.14. The greater spread of the extreme quantiles for the residuals is indicative for a distribution with longer tails than the normal distribution.

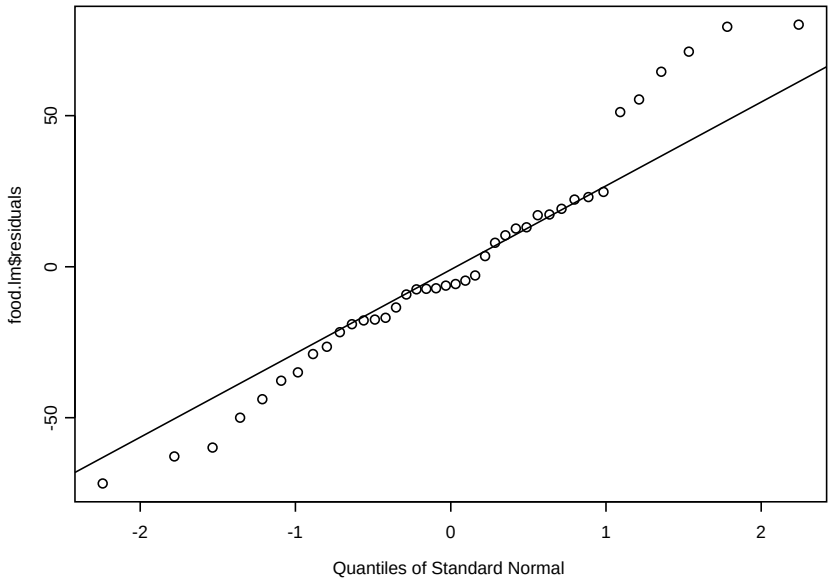

Figure 4.14: Normal probability plot of residuals of food expenditure model

4.7 Linear regression in matrix terms

When we go from one to several explanatory variables, it will prove useful to use matrix notation to state the linear regression model and the least squares solution. We start with vector/matrix notation for simple linear regression

through the origin. This also enables us to interpret the least squares solution geometrically.

4.7.1 Geometrical interpretation of least squares: regression through the origin

In the previous sections we derived the least squares estimators

$$b_0 = \bar{y} - b_1\bar{x} \qquad\qquad b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{4.51}$$

using calculus. Recall also that the least squares estimator for the slope in the regression through the origin model was

$$b = \frac{\sum x_i y_i}{\sum x_i^2} \tag{4.52}$$

Another way to derive the least squares estimator uses a geometrical argument. To view the reasoning most clearly, we consider simple linear regression through the origin. The fitted values are then given by

$$\hat{y}_i = bx_i$$

We now consider how to obtain the least squares solution for this problem. We consider an example and then derive the general solution. To be able to draw a picture, we assume that we only have two observations

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \text{ and } Y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$$

We will also use the vector of fitted values $\hat{Y}$ and the error vector e:

$$e = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \text{ and } \hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix}$$

We have drawn the vectors X and Y in figure 4.15. Now $\hat{Y}$ has to be some multiple of X, so $\hat{Y}$ is somewhere on the line in the direction of X. The least squares criterion states that we choose the point on the line through X such that the length $\sqrt{e \cdot e}$ of the vector $e = Y - \hat{Y}$ is as small as possible (Note that the length of e is just the square root of the sum of squared errors). In

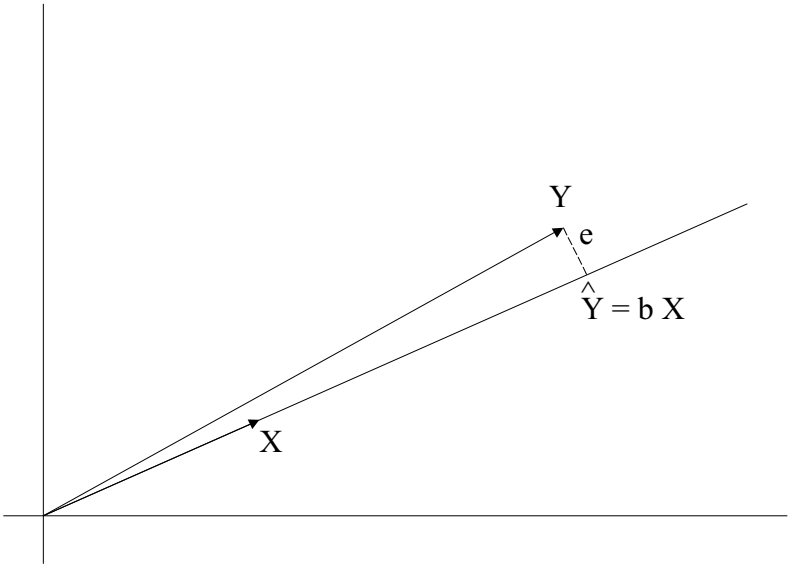

Figure 4.15: $\hat{Y}$ is the orthogonal projection of Y onto X

other words, we choose the point on the line through X that is *closest* to Y. This is achieved by taking the error vector orthogonal to X, as is shown in figure 4.15. The point $\hat{Y}$ is called the *orthogonal projection* of Y on X.

From this observation we can derive the value of b as follows. Since e must be perpendicular to X, we have $X \cdot e = 0$. So

$$X \cdot e = X \cdot (Y - bX) = X \cdot Y - bX \cdot X = 0$$

Therefore

$$b = \frac{X \cdot Y}{X \cdot X}$$

which is of course the same as the result

$$b = \frac{\sum x_i y_i}{\sum x_i^2}$$

that we got before using calculus and summation rather than vector notation.

For consistency with later notation we also give the matrix notation of this result. In matrix notation, the dot product is written as a row times a

column vector. To make a row vector of X, we take its transpose X^T. We then get

$$b = \frac{X^T Y}{X^T X} \qquad \text{or} \qquad b = (X^T X)^{-1} X^T Y$$

If we apply the solution to the numerical example, we get

$$X^T Y = [2 \;\; 1] \begin{bmatrix} 5 \\ 3 \end{bmatrix} = 13 \qquad \text{and} \qquad X^T X = [2 \;\; 1] \begin{bmatrix} 2 \\ 1 \end{bmatrix} = 5$$

which yields

$$b = \frac{X^T Y}{X^T X} = \frac{13}{5} = 2\frac{3}{5}$$

4.7.2 Simple linear regression model in matrix terms

Next we look at simple linear regression in matrix terms. We start with the least squares solution.

Least Squares solution

We can write the observed y values as

$$y_i = b_0 + b_1 x_i + e_i \tag{4.53}$$

which is short for

$$\begin{aligned} y_1 &= b_0 + b_1 x_1 + e_1 \\ y_2 &= b_0 + b_1 x_2 + e_2 \\ &\vdots \\ y_n &= b_0 + b_1 x_n + e_n \end{aligned}$$

In matrix notation we can write this system of equations much more compact as follows Let

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \quad b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

Then we can write

$$Y = Xb + e \tag{4.54}$$

since

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

$$= \begin{bmatrix} b_0 + b_1 x_1 \\ b_0 + b_1 x_2 \\ \vdots \\ b_0 + b_1 x_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} b_0 + b_1 x_1 + e_1 \\ b_0 + b_1 x_2 + e_2 \\ \vdots \\ b_0 + b_1 x_n + e_n \end{bmatrix}$$

The column of 1s in the X matrix may be viewed as consisting of the dummy variable $x_0 \equiv 1$ in the alternative model

$$y_i = b_0 x_0 + b_1 x_i + e_i \qquad \text{where } x_0 \equiv 1 \tag{4.55}$$

The fitted value $\hat{Y}$ is a linear combination of the columns of X, i.e.

$$\hat{Y} = Xb$$

Typically, the observed values Y are not in the column space of X, but we want to find the value of $\hat{Y}$ that is closest to Y. For this to be the case, the error vector

$$e = Y - Xb$$

must be orthogonal to *all columns* of X.

In other words,

$$X^T e = X^T(Y - Xb) = X^T Y - X^T X b = 0,$$

from which it follows that

$$X^T X b = X^T Y \tag{4.56}$$

Equation (4.56) states that

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum x_i Y_i \end{bmatrix}$$

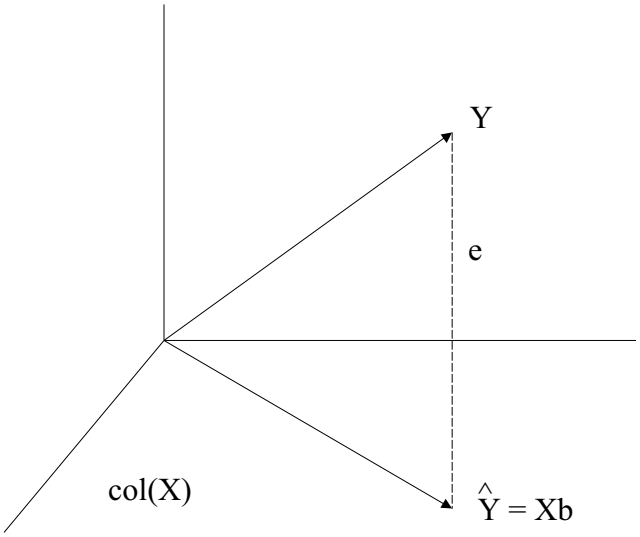

Figure 4.16: $\hat{Y}$ is the orthogonal projection of Y onto col(X)

or

$$\begin{bmatrix} nb_0 + \sum b_1 x_i \\ b_0 \sum x_i + b_1 \sum x_i^2 \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum x_i Y_i \end{bmatrix}$$

which are precisely the normal equations we derived using calculus.

To obtain the estimated regression coefficients from the normal equations (4.56) by matrix methods, we premultiply both sides by the inverse of $X^T X$ (assuming it exists):

$$(X^T X)^{-1} X^T X b = (X^T X)^{-1} X^T Y$$

We then find, since $(X^T X)^{-1} X^T X = I$ and $Ib = b$:

$$b = (X^T X)^{-1} X^T Y \tag{4.57}$$

Simple linear regression model

In terms of population parameters, the observed y values can be written

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{4.58}$$

which is short for

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \varepsilon_n \end{aligned}$$

In matrix notation we can write this system of equations much more compact as follows Let

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Then we can write

$$Y = X\beta + \varepsilon \tag{4.59}$$

with Y and X as defined before.

The condition $E(\varepsilon_i) = 0$ in matrix terms is

$$E(\varepsilon) = 0 \tag{4.60}$$

since (4.60) states that

$$\begin{bmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \\ \vdots \\ E(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

The condition that the error terms have constant variance σ^2 and that all covariances $\text{cov}(\varepsilon_i, \varepsilon_j)$ for $i \neq j$ are zero (since the ε_i are independent) is expressed in matrix terms through the variance-covariance matrix of the error terms

$$\text{var}(\varepsilon) = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix} \tag{4.61}$$

Since this is a scalar matrix, we can express it in the following simple fashion

$$\text{var}(\varepsilon) = \sigma^2 I \tag{4.62}$$

Numeric example

Suppose we have observations $T = \{(0,1),(1,1),(2,2),(3,2)\}$. The relevant data matrices are

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \quad Y = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \end{bmatrix} \quad X^T X = \begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix} \quad X^T Y = \begin{bmatrix} 6 \\ 11 \end{bmatrix}$$

Now, since

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

we get

$$(X^T X)^{-1} X^T Y = \frac{1}{20} \begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix} \begin{bmatrix} 6 \\ 11 \end{bmatrix} = \frac{1}{20} \begin{bmatrix} 18 \\ 8 \end{bmatrix} = \begin{bmatrix} 9/10 \\ 4/10 \end{bmatrix}$$

4.8 Multiple Linear Regression

In general, a model with a single explanatory variable is not very realistic. The extension of linear regression to more than one explanatory variable is straightforward.

In terms of population parameters, the observations can be written

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \tag{4.63}$$

which is again short for

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{1,1} + \beta_2 x_{1,2} + \ldots + \beta_{p-1} x_{1,p-1} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{2,1} + \beta_2 x_{2,2} + \ldots + \beta_{p-1} x_{2,p-1} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \ldots + \beta_{p-1} x_{n,p-1} + \varepsilon_n \end{aligned}$$

Here the number of explanatory variables is $p-1$, and the number of parameters p, assuming there is an intercept coefficient. In matrix notation we can

write this system of equations much more compact as follows Let

$$X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p-1} \\ \vdots & & & & \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p-1} \end{bmatrix} Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

Then, like with simple linear regression, we can write

$$Y = X\beta + \varepsilon \tag{4.64}$$

The least squares estimator of β still is

$$b = (X^T X)^{-1} X^T Y \tag{4.65}$$

4.8.1 Inferences about regression parameters

We start by proving that the least squares estimators of β are unbiased. This turns out to be quite straightforward in matrix notation. We have to prove that

$$E(b) = \beta \tag{4.66}$$

Proof

$$\begin{aligned} E(b) &= E((X^T X)^{-1} X^T Y) \\ &= (X^T X)^{-1} X^T E(Y) \\ &= (X^T X)^{-1} X^T X\beta \\ &= \beta \end{aligned}$$

In this proof we used the fact that $(X^T X)^{-1}$ is a matrix of constants, and therefore its expected value is simply the matrix itself. Secondly, we used $E(Y) = X\beta$ which follows directly from the assumption that $E(\varepsilon) = 0$ and $Y = X\beta + \varepsilon$.

The variance-covariance matrix $\text{var}(b)$

$$\text{var}(b) = \begin{bmatrix} \text{var}(b_0) & \text{cov}(b_0, b_1) & \dots & \text{cov}(b_0, b_{p-1}) \\ \text{cov}(b_1, b_0) & \text{var}(b_1) & \dots & \text{cov}(b_1, b_{p-1}) \\ \vdots & \vdots & & \vdots \\ \text{cov}(b_{p-1}, b_0) & \text{cov}(b_{p-1}, b_1) & \dots & \text{var}(b_{p-1}) \end{bmatrix}$$

is given by

$$\text{var}(b) = \sigma^2(X^TX)^{-1}$$

The proof is given below. Write

$$b = (X^TX)^{-1}X^TY = CY$$

where C is a constant matrix

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

Now since $\text{var}(CY) = C\ \text{var}(Y)\ C^T$ (This is the matrix equivalent of the rule $\text{var}(cy) = c^2\text{var}(y)$) we get

$$\text{var}(b) = C\ \text{var}(Y)\ C^T$$

Now $\text{var}(Y) = \sigma^2 I$. Furthermore, since $(X^TX)^{-1}$ is symmetric, we have

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

We find therefore

$$\begin{aligned}
\text{var}(b) &= (X^TX)^{-1}X^T\sigma^2 IX(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}I \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}$$

The estimated variance-covariance matrix $\widehat{\text{var}}(b)$

$$\widehat{\text{var}}(b) = \begin{bmatrix}
\widehat{\text{var}}(b_0) & \widehat{\text{cov}}(b_0, b_1) & \dots & \widehat{\text{cov}}(b_0, b_{p-1}) \\
\widehat{\text{cov}}(b_1, b_0) & \widehat{\text{var}}(b_1) & \dots & \widehat{\text{cov}}(b_1, b_{p-1}) \\
\vdots & \vdots & & \vdots \\
\widehat{\text{cov}}(b_{p-1}, b_0) & \widehat{\text{cov}}(b_{p-1}, b_1) & \dots & \widehat{\text{var}}(b_{p-1})
\end{bmatrix}$$

is given by

$$\widehat{\text{var}}(b) = s^2(X^TX)^{-1}$$

From $\widehat{\text{var}}(b)$ one can obtain $\widehat{\text{var}}(b_0)$, $\widehat{\text{var}}(b_1)$, or whatever other variance or covariance is needed.

For the normal error regression model we have

$$\frac{b_k - \beta_k}{\text{se}(b_k)} \sim t_{(n-p)} \qquad\qquad k = 0, 1, \ldots, p-1$$

Hence

$$b_k \pm t_{(n-p);\alpha/2}\, \text{se}(b_k)$$

is a $(1-\alpha) \times 100\%$ confidence interval for β_k.

Tests for β_k are set up in the usual fashion. To test

$$H_0 : \beta_k = 0 \qquad\qquad H_a : \beta_k \neq 0$$

we use the test statistic

$$t = \frac{b_k}{\text{se}(b_k)}$$

and the decision rule: if $|t| > t_{(n-p);\alpha/2}$ reject H_0, otherwise accept H_0. To test

$$H_0 : \beta_k = 0 \qquad\qquad H_a : \beta_k > 0$$

we use the decision rule: if $t > t_{(n-p);\alpha}$ reject H_0, otherwise accept H_0.

4.8.2 Coefficient of multiple determination

Recall that the definition of R^2 is

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

In the context of multiple regression, R^2 is called the coefficient of *multiple* determination. It measures how much of the variation in y around its mean is explained by the variation in $x_1, x_2, \ldots, x_{p-1}$ together.

To see how the sums of squares are expressed in matrix notation, we begin with the total sum of squares

$$\text{SST} = \sum (y_i - \bar{y})^2 = \sum (y_i - \bar{y}) y_i = \sum y_i^2 - \frac{1}{n} \left(\sum y_i \right)^2$$

We know that

$$Y^T Y = \sum y_i^2$$

The subtraction term $1/n(\sum y_i)^2$ in matrix form uses an $n \times n$ matrix of 1's which we call J

$$\frac{1}{n}\left(\sum y_i\right)^2 = \left(\frac{1}{n}\right) Y^T J Y$$

For instance, if $n = 2$ we have

$$\left(\frac{1}{2}\right) \begin{bmatrix} y_1 & y_2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{(y_1 + y_2)(y_1 + y_2)}{2}$$

Hence it follows that

$$\text{SST} = Y^T Y - \left(\frac{1}{n}\right) Y^T J Y$$

Furthermore, we have

$$\text{SSE} = e^T e = (Y - Xb)^T (Y - Xb)$$

Expanding, we get

$$\begin{aligned} (Y - Xb)^T (Y - Xb) &= (Y^T - (Xb)^T)(Y - Xb) \\ &= (Y^T - b^T X^T)(Y - Xb) \\ &= Y^T Y - Y^T Xb - b^T X^T Y + b^T X^T Xb \end{aligned}$$

Now $Y^T Xb$ is a scalar and hence equal to its transpose $b^T X^T Y$, so we get

$$Y^T Y - Y^T Xb - b^T X^T Y + b^T X^T Xb = Y^T Y - 2b^T X^T Y + b^T X^T Xb$$

Now to simplify the expression, replace the rightmost b by $(X^T X)^{-1} X^T Y$ to get

$$\begin{aligned} Y^T Y - 2b^T X^T Y + b^T X^T Xb &= Y^T Y - 2b^T X^T Y + b^T X^T X (X^T X)^{-1} X^T Y \\ &= Y^T Y - 2b^T X^T Y + b^T I X^T Y \\ &= Y^T Y - b^T X^T Y \end{aligned}$$

Finally, since SSR = SST - SSE, we get

$$\text{SSR} = (Y^T Y - \left(\frac{1}{n}\right) Y^T J Y) - (Y^T Y - b^T X^T Y) = b^T X^T Y - \left(\frac{1}{n}\right) Y^T J Y$$

4.8.3 Multicollinearity

Most data that are used for estimating relationships are nonexperimental: the data are simply collected for administrative or other purposes. In *controlled experiments* the right-hand-side variables in the statistical model can be assigned values in such a way that their individual effects can be identified and estimated with precision. When we are dealing with *observational data*, many of the variables may move together in systematic ways. Such variables are said to be *collinear*, and the problem is labeled *collinearity* or *multicollinearity* when several variables are involved. In this case there is no guarantee that the data will be "rich in information" nor that it will be possible to isolate the relationship or parameters of interest.

To see this point intuitively, consider the following example. Suppose you are on the cycling team. Before some of your cycling meets your grandmother prepares you a terrific pasta dinner and gives you one of her famous pep talks. When you get this special treatment you invariably cycle well. Now, would you ever be able to figure out whether it is the pasta dinner or the pep talk that produces this wonderful result? The answer is no. The two things (pasta dinner/pep talk) always happen together. We wouldn't be able to disentangle their separate effects unless we sometimes had the pep talk with no dinner or the dinner with no pep talk. In regression terms: we couldn't figure out the *ceteris paribus* effects of the two variables since they are perfectly correlated. Figuring out how an explanatory variable affects the dependent variable requires that there is some *independent* variation in that explanatory variable.

Consider the multiple regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \qquad i = 1, 2, \ldots, n$$

and suppose that x_2 is a multiple of x_1, for example

$$x_{i2} = \tfrac{1}{2} x_{i1}, \qquad i = 1, 2, \ldots, n$$

In the matrix X, the third column is half the second column. We can write this model as

$$Y_i = \beta_0 + (\beta_1 + \frac{1}{2}\beta_2) x_{i1} + \varepsilon_i$$

Under the usual assumptions, the regression coefficient $\beta_1 + \frac{1}{2}\beta_2$ is uniquely determined, but this is not the case for the individual constants β_1 and β_2.

y	x_1	x_2	$\hat{y}$
10	2	1	10.6
22	4	2	20.2
28	6	3	29.8
40	8	4	39.4
100	20	10	

Table 4.3: Example of exact colinearity: $x_2 = \frac{1}{2}x_1$

The least-squares estimate of $\beta_1 + \frac{1}{2}\beta_2$ is

$$b_{1,2} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{4(596) - 20(100)}{4(120) - 20^2} = \frac{384}{80} = 4.8$$

The least-squares estimate of the intercept is

$$b_0 = \bar{y} - b_{1,2}\bar{x} = 25 - 4.8(5) = 1$$

So the least-squares regression line is

$$\hat{y} = 1 + 4.8x_1$$

Now, if we regress y on x_1 and x_2, any pair $(\hat{\beta}_1, \hat{\beta}_2)$ such that $\hat{\beta}_1 + \frac{1}{2}\hat{\beta}_2 = 4.8$ will give the same solution. So $(4, 1.6), (3, 3.6), (100, -190.4)$ all yield the same fit. The *separate* influence of x_1 and x_2 on y can not be determined. It is therefore pointless to try to estimate β_1 and β_2 separately.

Let's consider the more general situation of a *linear dependence* between the values of the explanatory variables x_i, that is for some constants $c_0, c_1, \ldots, c_{p-1}$ (not all equal to zero) we have

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + \ldots + c_{p-1} x_{ip-1} = 0, \qquad i = 1, 2, \ldots, n \tag{4.67}$$

In this case we speak of *exact collinearity*. The columns of matrix X are linearly dependent: each column can be written as a linear combination of the others.

Since $n \geq p$ dependence between the columns means that X is not of full rank p. From linear algebra we know the rule

$$\text{rank}(X^T X) = \text{rank}(X)$$

so X^TX is not of full rank p either. This means matrix X^TX is not invertible (the inverse of a matrix exists only if it is of full rank). So in case of exact collinearity the inverse of X^TX does not exist: the regression coefficients β_k are not uniquely determined, and can not be estimated separately. In case of simple linear regression exact collinearity means that the explanatory variable x is constant. In that case the denominator of the equation for b_1 is 0, which does indeed make it useless.

Exact collinearity rarely occurs in practice and if it occurs it usually is because the analyst has made a mistake in constructing the data matrix. However, a mild form of the problem also occurs when there is a high correlation between the explanatory variables. As an example (taken from [11]), table 4.4 contains the data for a study of the relation of amount of body fat to several possible predictor variables, based on a sample of 20 healthy females 25-34 years of age. The possible predictor variables are triceps skinfold thickness, thigh circumference and midarm circumference. The amount of bodyfat in table 4.4 for each of the 20 persons was obtained by a cumbersome and expensive procedure requiring the immersion of the person in water. It would therefore be very helpful if a regression model with some or all of these predictor variables could provide reliable predictions of the amount of body fat, since the measurements needed for the predictor variables are easy to obtain.

Figure 4.17 contains the scatterplot matrix of the dependent and predictor variables. Table 4.5 contains the correlation matrix.

It is evident from the scatterplot matrix that the predictor variables *triceps skinfold thickness* and *thigh circumference* are highly correlated; the correlation matrix shows that their coefficient of simple correlation is 0.924. On the other hand *midarm circumference* is not so highly correlated to *triceps skinfold thickness* and *thigh circumference* individually; the correlation matrix shows that the correlation coefficients are 0.458 and 0.085 respectively. But *midarm circumference* is highly correlated with *triceps skinfold thickness* and *thigh circumference* together; the coefficient of multiple determination when *midarm circumference* is regressed on *triceps skinfold thickness* and *thigh circumference* is 0.998.

```
> summary(bodyfat.fit)

Call: lm(formula = body.fat ~ ., data = bodyfat)
```

subject	triceps	thigh	midarm	body.fat
1	19.5	43.1	29.1	11.9
2	24.7	49.8	28.2	22.8
3	30.7	51.9	37.0	18.7
4	29.8	54.3	31.1	20.1
5	19.1	42.2	30.9	12.9
6	25.6	53.9	23.7	21.7
7	31.4	58.5	27.6	27.1
8	27.9	52.1	30.6	25.4
9	22.1	49.9	23.2	21.3
10	25.5	53.5	24.8	19.3
11	31.1	56.6	30.0	25.4
12	30.4	56.7	28.3	27.2
13	18.7	46.5	23.0	11.7
14	19.7	44.2	28.6	17.8
15	14.6	42.7	21.3	12.8
16	29.5	54.4	30.1	23.9
17	27.7	55.3	25.7	22.6
18	30.2	58.6	24.6	25.4
19	22.7	48.2	27.1	14.8
20	25.2	51.0	27.5	21.1

Table 4.4: Data for bodyfat example

```
Residuals:
    Min     1Q Median    3Q   Max
 -3.726 -1.611 0.3923 1.466 4.128

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)  117.0847   99.7824     1.1734    0.2578
    triceps    4.3341    3.0155     1.4373    0.1699
      thigh   -2.8568    2.5820    -1.1064    0.2849
     midarm   -2.1861    1.5955    -1.3701    0.1896

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-Squared: 0.8014
```

	triceps	thigh	midarm	body.fat
triceps	1.000	0.924	0.458	0.843
thigh	0.924	1.000	0.085	0.878
midarm	0.458	0.085	1.000	0.142
body.fat	0.843	0.878	0.142	1.000

Table 4.5: Correlation matrix for bodyfat example

Variables in model	b_1	b_2
x_1	0.8572	$-$
x_2	$-$	0.8565
x_1, x_2	0.2224	0.6594
x_1, x_2, x_3	4.3341	-2.8568

Table 4.6: Value of coefficients b_1, b_2 for different models

Note from table 4.6 that the regression coefficient for x_1, triceps skinfold thickness, varies markedly depending on which other variables are included in the model. The regression coefficient b_2 even changes sign when x_3 is added to the model that includes x_1 and x_2. When predictor variables are correlated, the regression coefficient of any one variable depends on which other predictor variables are included in the model. Thus, a regression coefficient does not reflect any inherent effect of the particular predictor variable on the response variable but only a marginal or partial effect, given whatever other correlated predictor variables are included in the model.

The consequences of collinear relationships among explanatory variables may be summarized as follows:

1. Whenever there are one or more exact linear relationships among the explanatory variables, then the condition of exact collinearity, or exact multicollinearity exists. In this case the least squares estimator is not defined.

2. When nearly exact linear dependencies exist among the explanatory variables, some of the variances, standard errors and covariances of the

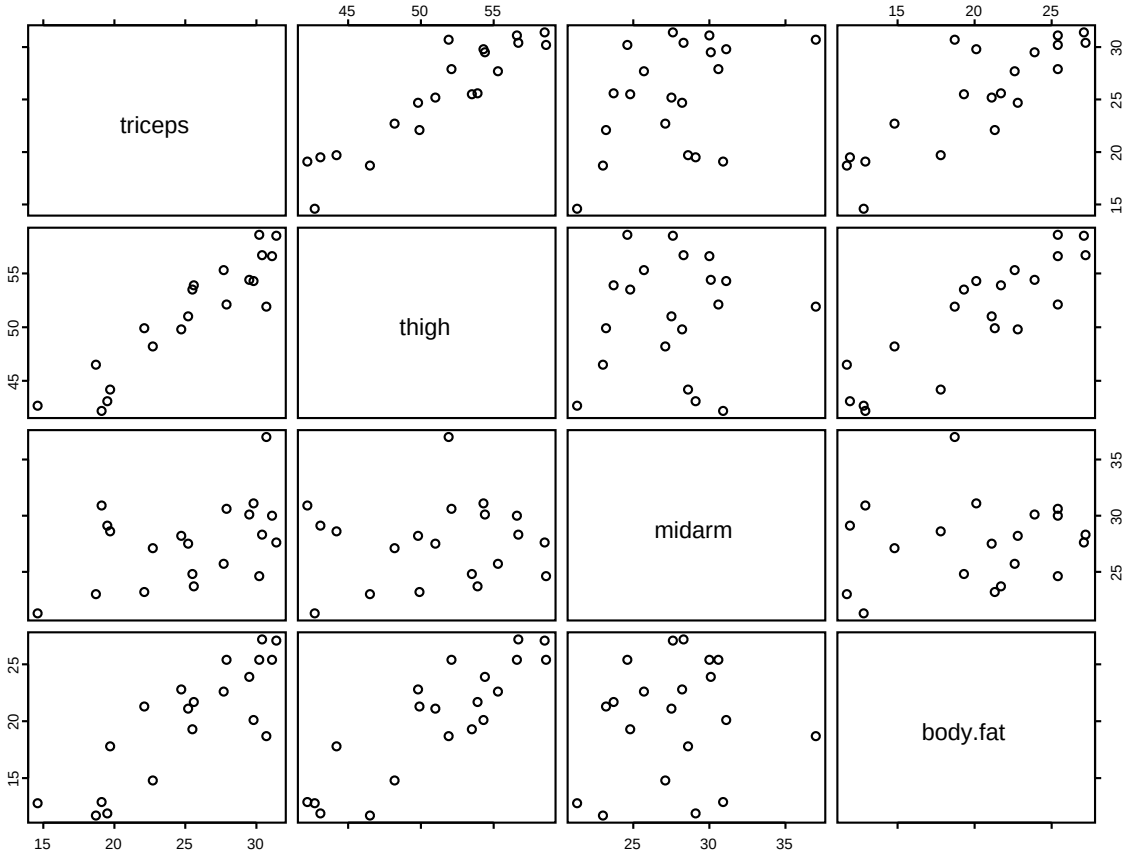

Figure 4.17: Pairwise scatterplots for bodyfat example

least squuares estimators may be large. Large standard errors for the least squares estimators imply: high sampling variability, coefficient estimates that are unstable to small changes in the sample or model specification, interval estimates that are wide, and relatively imprecise information provided by the sample data about the unknown parameters.

3. When estimator standard errors are large, it is likely that the usual t-tests will lead to the conclusion that parameter estimates are not significantly different from zero. This outcome occurs despite possibly high R^2 indicating "significant" explanatory power of the model as a

whole. The problem is that collinear variables do not provide enough information to estimate their separate effects, even though theoretical considerations may indicate their importance in the relationship.

4. Despite the difficulties in isolating the effects of individual variables from such a sample, accurate *predictions* may still be possible if the nature of the collinear relationship remains the same within the new (future) sample observations.

4.8.4 Omitted variable bias

Suppose we inadvertently omit a variable from the regression model. Suppose the true model is

$$y = \beta_0 + \beta_1 x + \beta_2 h + \varepsilon \tag{4.68}$$

but we estimate the model

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{4.69}$$

omitting h from the model.

Then we use the estimator

$$\begin{aligned} b_1^* &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} \\ &= \frac{\sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i + \beta_2 h_i + \varepsilon_i)}{\sum(x_i - \bar{x})^2} \\ &= \frac{\beta_0 \sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} + \frac{\beta_1 \sum(x_i - \bar{x})x_i}{\sum(x_i - \bar{x})^2} + \frac{\beta_2 \sum(x_i - \bar{x})h_i}{\sum(x_i - \bar{x})^2} + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2} \\ &= \beta_1 + \beta_2 \sum w_i h_i + \sum w_i \varepsilon_i \end{aligned}$$

where

$$w_i = \frac{x_i - \bar{x}}{\sum(x_j - \bar{x})^2}$$

So,

$$E(b_1^*) = \beta_1 + \beta_2 \sum w_i h_i \neq \beta_1$$

We can write

$$\sum w_i h_i = \frac{(x_i - \bar{x})h_i}{\sum(x_i - \bar{x})^2}$$

$$
\begin{aligned}
&= \frac{(x_i - \bar{x})(h_i - \bar{h})}{\sum(x_i - \bar{x})^2} \\
&= \frac{(x_i - \bar{x})(h_i - \bar{h})/(n-1)}{\sum(x_i - \bar{x})^2/(n-1)} \\
&= \frac{\widehat{\text{cov}}(x_i, h_i)}{\widehat{\text{var}}(x_i)}
\end{aligned}
$$

so

$$E(b_1^*) = \beta_1 + \beta_2 \frac{\widehat{\text{cov}}(x_i, h_i)}{\widehat{\text{var}}(x_i)}$$

The sign of β_2 and the sign of the covariance between x_i and h_i tells us the direction of the bias. If the sample covariance, or sample correlation, between x_i and the omitted variable h_i is zero, then the least squares estimator in the misspecified model is still unbiased.

4.9 Binary explanatory variables

Binary variables allow us to construct models in which some or all of the regression parameters, including the intercept, change for some observations in the sample. To illustrate the different uses of binary variables, we consider an example from real estate economics: the prediction of the value of a house. We assume the price of a house is explained by its characteristics, such as its size, location, number of bedrooms, age, etc.

We begin with a simple model where the price P of the house only depends on its size S

$$P_i = \beta_0 + \beta_1 S_i + \varepsilon_i$$

In this model, β_1 is the value of an additional square meter of living area, and β_0 is the value of the land alone. How can we take into account the effect of a property being in a desirable neighborhood such as one near a university? Binary variables are used to account for such qualitative factors. We usually code a binary variable as 0 or 1, to indicate the presence or absence of a characteristic. That is a binary variable B is

$$B = \begin{cases} 1 & \text{if characteristic is present} \\ 0 & \text{if characteristic is absent} \end{cases}$$

For the house price model, we can define a binary variable to account for a desirable neighborhood as

$$B_i = \begin{cases} 1 & \text{if house is in a desirable neighborhood} \\ 0 & \text{if house is not in a desirable neighborhood} \end{cases}$$

If we add this variable and corresponding parameter δ to the model we obtain

$$P_i = \beta_0 + \delta B_i + \beta_1 S_i + \varepsilon_i$$

The effect of inclusion of a binary variable B_i into the regression model is best seen by examining the regression function $E(P_i)$, in the two locations.

If the model is correctly specified then $E(\varepsilon_i) = 0$ and

$$E(P_i) = \begin{cases} (\beta_0 + \delta) + \beta_1 S_i & \text{when } B_i = 1 \\ \beta_0 + \beta_1 S_i & \text{when } B_i = 0 \end{cases}$$

The two situations are depicted in figure 4.18, assuming $\delta > 0$.

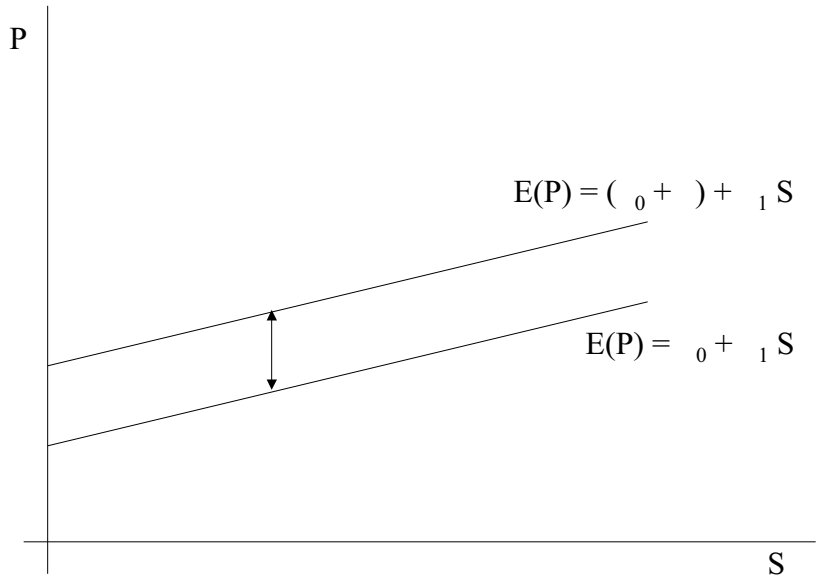

Figure 4.18: An intercept binary variable

In the context of the house price model, the interpretation of δ is that of a location premium: the difference in house price due to being located in a desirable neighborhood.

It is also reasonable to assume that the effect of location on house price causes a change in the slope of the regression equation, instead of the intercept. We can allow for a change in the slope by including in the model an additional explanatory variable that is equal to the product of a binary variable and a numeric variable. In the house price model, the slope of the relationship is the value of an additional square foot of living area. If we assume this is one value for homes in a desirable neighborhood and another value for homes in other neighborhoods, then the correct specification is

$$P_i = \beta_0 + \beta_1 S_i + \gamma(S_i B_i) + \varepsilon_i$$

Examining the regression function for the two different locations best illustrates the effect of the inclusion of the interaction variable into the model

$$E(P_i) = \begin{cases} \beta_0 + (\beta_1 + \gamma)S_i & \text{when } B_i = 1 \\ \beta_0 + \beta_1 S_i & \text{when } B_i = 0 \end{cases}$$

In the desirable neighborhood per square meter of a home is $(\beta_1 + \gamma)$, in other locations it is $\beta 1$.

4.10 Model Selection for Linear Regression

In many cases we want to use the linear regression model to predict for new observations the value of Y, when the value of X is known. If we have many potential explanatory variables, we have to consider many different model specifications. In this section we address the problem how to select the model with the best predictive performance from the space of potential regression models.

4.10.1 Prediction and the danger of overfitting

We first show why the naive approach doesn't work. The naive approach would be to select the model with the lowest error (highest R^2) on the sample we use to fit the model. We consider a simple regression example to illustrate

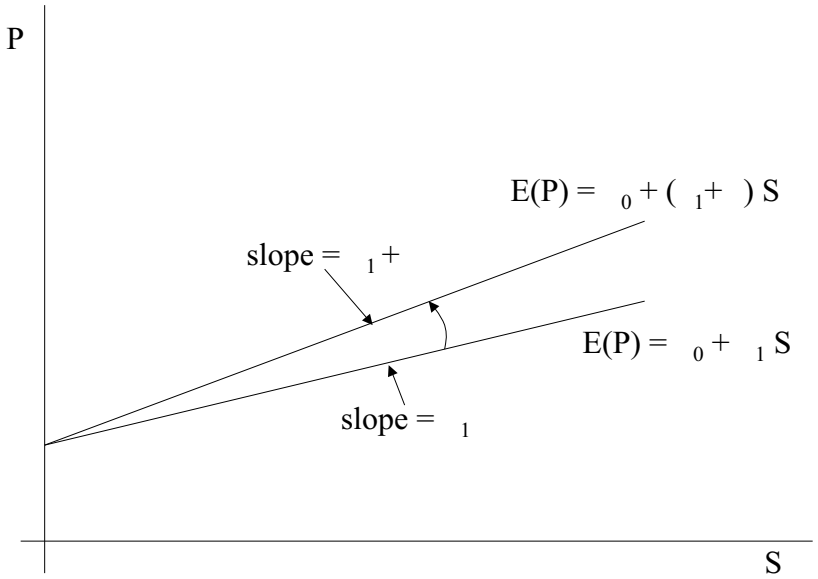

Figure 4.19: A slope binary variable

why this doesn't work. Suppose that $Y_i \sim N(\mu = 2.0 + 0.5x_i, \sigma^2 = 1)$, i.e. the true relation between $\mathrm{E}(Y)$ and x is

$$\mathrm{E}(Y) = 2.0 + 0.5x.$$

We have a sample T of ten (x, y) observations, which is displayed in the scatterplot of Fig. 4.20(a). Note that x is not a random variable but its values are chosen by us to be $1, 2, \ldots, 10$. Although $\mathrm{E}(Y)$ is a linear function of x, the observations do not lie on a straight line due to the inherent variability of Y. We pretend we don't know the relation between x and y, but only have T at our disposal, as would be the case in most data analysis settings. We consider three classes of models to describe the relation between x and y

Linear Model: $\mathrm{E}(Y) = f_1(x) = \beta_0 + \beta_1 x$

Quadratic Model: $\mathrm{E}(Y) = f_2(x) = \beta_0 + \beta_1 x + \beta_2 x^2$

Cubic Model: $\mathrm{E}(Y) = f_3(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

Note that (2) encompasses (1) in the sense that if $\beta_2 = 0$, (2) reduces to the linear function (1). Likewise, (3) encompasses (2), and consequently also (1). The β_j are the parameters of the model, whose estimates are chosen in such a way that the sum of squared *vertical* distances from the points (x_i, y_i) to the fitted equation is minimized. For example, for the simple linear regression model we choose the estimates b_0 and b_1 of β_0 and β_1 such that

$$\sum_{i=1}^{n} [y_i - (b_0 + b_1\, x_i)]^2$$

is minimal. The expression $b_0 + b_1\, x_i$ denotes the predicted value for y_i, so one effectively minimizes the sum of squared differences between predicted values and realisations of y. The estimates b_j of the β_j thus obtained are called the *least squares* estimates.

The equations obtained by least squares estimation for the respective models are displayed in Fig. 4.20 (b) to (d). Without performing the actual calculations, one can easily see that the linear model gives the worst fit, even though the true (population) relation is linear. The quadratic model gives a somewhat better fit, and the cubic model gives the best fit of the three. In general, the more parameters the model has, the better it is able to adjust to the data in T. Does this mean that the cubic model gives better predictions than the linear model? It does on T, but how about on data that were not used to fit the equation? We drew a second random sample, denoted by T', and looked at the fit of the equations to T' (see Fig. 4.21). The fit of the cubic model is clearly worse than that of the linear model. The reason is that the cubic model has adjusted itself to the random variations in T, leading on average to bad predictive performance on new samples. This phenomenon is called *overfitting*.

In the next section we discuss the decomposition of prediction error into its components to gain a further understanding of the phenomenon illustrated by the above example.

4.10.2 Decomposition of prediction error in regression

Once we have obtained estimates b_j by estimation from some training set T, we may use the resulting function to make predictions of y when we know the corresponding value of x. Henceforth we denote this prediction by $\hat{f}(x)$. The difference between prediction $\hat{f}(x)$ and realisation y is called prediction error.

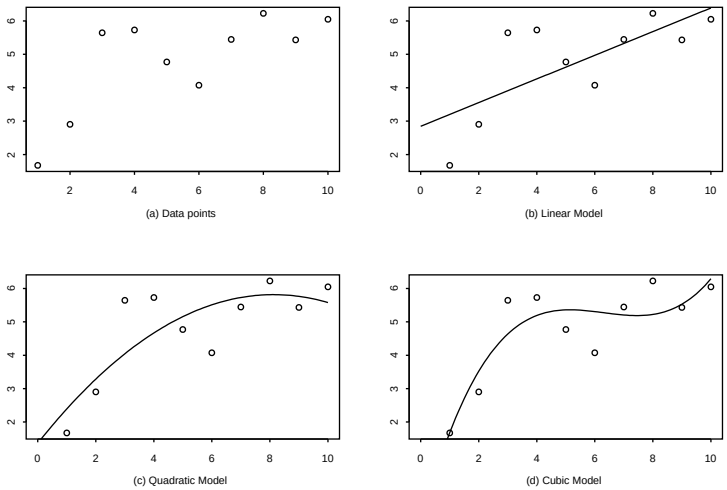

Figure 4.20: Equations fitted by least squares to the data in T

It should preferrably take values close to zero. A natural quality measure of $\hat{f}$ as a predictor of Y is the mean squared error. For fixed T and x

$$\mathrm{E}\,(Y - \hat{f}(x \,|\, T))^2$$

where the expectation is taken with respect to $p(Y \,|\, x)$, the probability distribution of Y at x. We may decompose this overall error into a *reducible* part, and an *irreducible* part that is due to the variability of Y at x, as follows

$$\mathrm{E}\,(Y - \hat{f}(x \,|\, T))^2 = [f(x) - \hat{f}(x \,|\, T)]^2 + \mathrm{E}(y - f(x))^2$$

where $f(x) \equiv \mathrm{E}[Y \,|\, x]$. The last term in this expression is the mean square error of the best possible (in the mean squared error sense) prediction $\mathrm{E}[Y \,|\, x]$. Since we can't do much about it, we focus our attention on the other source of error $[f(x) - \hat{f}(x \,|\, T)]^2$. This tells us something about the quality of the estimate $\hat{f}(x \,|\, T)$ for a particular realisation of T. To say something about the quality of the estimator $\hat{f}$, we take its expectation over all possible training samples (of fixed size) and decompose it into its bias and variance components as follows:

$$\mathrm{E}_T(f(x) - \hat{f}(x \,|\, T))^2 = (f(x) - \mathrm{E}_T \hat{f}(x \,|\, T))^2 + \mathrm{E}_T(\hat{f}(x \,|\, T) - \mathrm{E}_T \hat{f}(x \,|\, T))^2$$

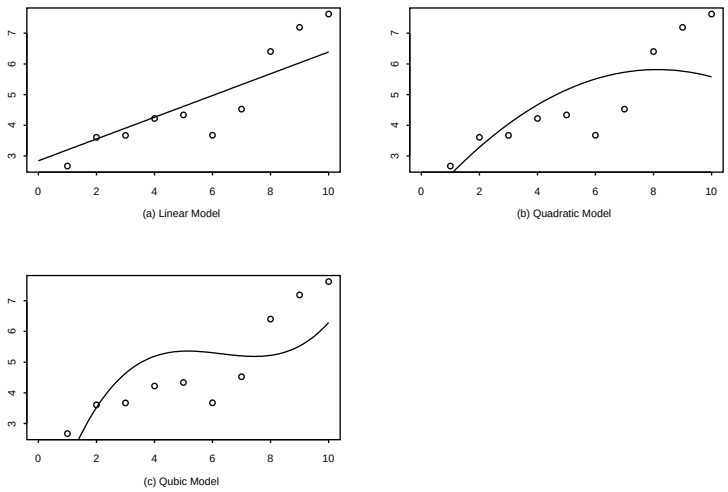

Figure 4.21: Fit of equations to new sample T'

The first component is the squared bias, where bias is the difference between the best prediction $f(x)$ and its average estimate over all possible samples of fixed size. The second component, variance, is the expected squared difference between an estimate obtained for a single training sample and the average estimate obtained over all possible samples.

We illustrate these concepts by a simple simulation study using the models introduced in the previous section. The expectations in the above decomposition are taken over *all possible* training samples, but this is a little bit too much to compute. Instead we use the computer to draw a large number of random samples to obtain an estimate of the desired quantities. In the simulation we sampled 1000 times from

$$Y_i \sim N(\mu = 2 + 0.5x_i, \sigma^2 = 1)$$

with $x_i = 1, 2, \ldots, 10$. In other words we generated 1000 random samples, $T_1, T_2, \ldots, T_{1000}$ each consisting of 10 (x, y) pairs. For each sample, the least squares parameter estimates for the three models were computed. Using the estimated models we computed the predicted values $\hat{f}(x)$. From the 1000 predicted values we computed the mean to estimate the expected value $\mathrm{E}\hat{f}(x)$ and variance to estimate $\mathrm{V}(\hat{f}(x))$. The results of this simulation study are summarized in Table 4.7. Consider the fourth row of this table for the

x	$f(x)$	$\mathrm{E}(\hat{f}_1)$	$\mathrm{E}(\hat{f}_2)$	$\mathrm{E}(\hat{f}_3)$	$\mathrm{V}(\hat{f}_1)$	$\mathrm{V}(\hat{f}_2)$	$\mathrm{V}(\hat{f}_3)$
1	2.50	2.48	2.48	2.49	0.34	0.61	0.84
2	3.00	2.99	2.98	2.98	0.25	0.27	0.29
3	3.50	3.49	3.49	3.48	0.18	0.18	0.33
4	4.00	3.99	4.00	3.99	0.13	0.20	0.32
5	4.50	4.50	4.50	4.50	0.10	0.23	0.25
6	5.00	5.00	5.00	5.01	0.10	0.22	0.23
7	5.50	5.50	5.51	5.52	0.13	0.19	0.28
8	6.00	6.01	6.01	6.02	0.17	0.18	0.31
9	6.50	6.51	6.51	6.51	0.24	0.28	0.30
10	7.00	7.01	7.01	7.00	0.33	0.62	0.80

Table 4.7: Expected value and variance of $\hat{f}_j$

moment. It contains the simulation results of the predictions of the different models for $x = 4$. The expected value is $f(4) = \mathrm{E}(Y|x = 4)$ is $2 + 0.5 \cdot 4 = 4$. From the first three columns we conclude that all models have no or negligable bias; in fact we can prove they are unbiased since all three models encompass the correct model. But now look at the last three columns of table 4.7. We see that the linear model has lowest variance, the cubic model has highest variance, and the quadratic model is somewhere inbetween. This is also illustrated by the histograms displayed in Fig. 4.22. We clearly see the larger spread of the cubic model compared to the linear model. Although all three models yield unbiased estimates, the linear model tends to have a lower prediction error because its variance is smaller than that of the quadratic and cubic model.

The so-called bias/variance dillema lies in the fact that there is a trade-off between the bias and variance components of error. Incorrect models lead to high bias, but highly flexible models suffer from high variance.

For a fixed bias, the variance tends to decrease when the training sample gets larger and larger. Consequently, for very large training samples, bias tends to be the most important source of prediction error. This phenomenon is illustrated by a second simulation. We generated the training sets by drawing from

$$Y_i \sim N(\mu = 2 + 0.5x_i + 0.02x_i^2, \sigma^2 = 1).$$

The true model is quadratic, so the linear model is biased whereas the

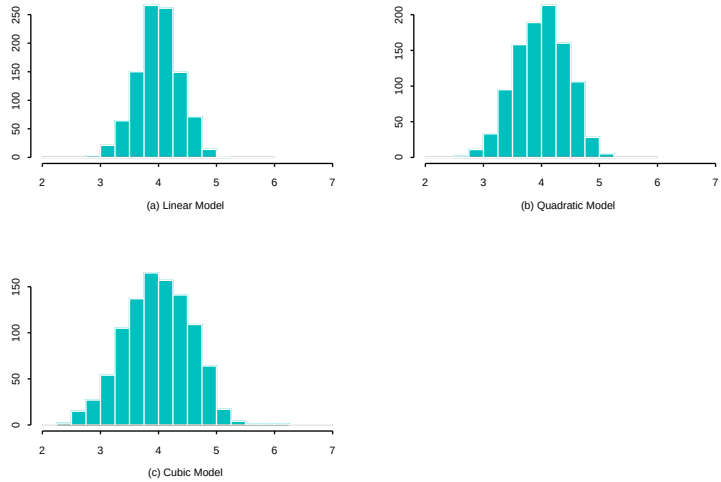

Figure 4.22: Histograms of $\hat{f}_j(4)$ based on 1000 samples

	Squared bias			Variance			Mean square error		
n	10	100	1000	10	100	1000	10	100	1000
Linear (f_1)	.021	.022	.022	.197	.022	.002	.218	.043	.024
Quadratic (f_2)	.000	.000	.000	.299	.037	.004	.299	.037	.004
Cubic (f_3)	.001	.000	.000	.401	.054	.006	.401	.054	.006

Table 4.8: Bias, variance and mean squared estimation error for different sample sizes

quadratic and cubic model are unbiased. We generated 1000 training samples of size 10, 100 and 1000 respectively. The first three columns of Table 4.8 summarize the estimated squared bias for the different models and sample sizes.

The results confirm that the linear model is biased, and furthermore indicate that the bias component of error does not decrease with sample size. Now consider the variance estimates shown in the middle three columns of Table 4.8. Looking at the rows, we observe that variance does decrease with the size of the sample. Taking these two phenomena together results in the summary of mean square error given in the last three columns of Table 4.8. The linear model outperforms the other models for small sample size, de-

spite its bias. Because variance is a substantial component of overall error the linear model profits from its smaller variance. As the sample size gets larger, variance becomes only a small part of total error, and the linear model becomes worse due to its bias.

4.10.3 Model Selection

The discussion in the previous section has shown that there are two reasons why we obtain a better fit of the data when we move from simple to more complex models

1. The more complex models generally have a smaller bias; their average tends to be closer to the population regression curve than for simple models.

2. Overfitting: the higher the number of adjustable parameters, the more prone the model is to fit to noise in the data.

We want to favour more complex models if the SSE goes down because of factor (1), but not if its decline is largely due to (2). If only we could correct the SSE value for overfitting, then the *corrected* SSE value would be a good indication of what we are interested in, the mean squared error of the fitted model.

Many suggestions have been made for estimating the size of the overfitting factor. We discuss here a result due to Akaike, called the Akaike Information Criterion (AIC). For linear regression models it looks like this:

$$\text{AIC} = \text{SSE} + 2\sigma^2 p$$

where SSE is the sum of squared errors of the fitted model, p is the number of adjustable parameters of the fitted model, and σ^2 is the variance of the error term.

Let's see if the different components of this expression are intuitively plausible. The SSE indicates the goodness-of-fit of the fitted model and clearly for models of the same complexity low values of SSE are desireable. The second term corrects for the average degree of overfitting for the family. Since overfitting has the effect of reducing SSE, any correction should be positive. That this correction is proportional to p, the number of adjustable parameters, reflects the intuition that overfitting will increase as the models become more flexible, allowing them to fit to the noise in the data.

That the expected degree of overfitting is also proportional to σ^2 is plausible as well. The bigger the error deviations from the population regression curve, the greater the potential for misleading fluctuations in the data. Also note that if there is no error ($\sigma^2 = 0$), then AIC reduces to SSE. If there is no error then simplicity of the model (as measured by p) is no longer relevant, and any model that fits the data perfectly scores equally well.

Let's look at some consequences of using AIC to select a model. It is clear that a simple model is preferrable if it fits the data about as well as a more complex model. AIC describes how much of an improvement in goodness-of-fit the move to a more complex model must provide for it to make sense to prefer the more complex model. The improvement must be large enough to overcome the penalty for complexity (represented by p).

Another feature of AIC is that the relative weight we give to simplicity declines as the number of data points increases. As the number of data points increases, the SSE becomes the dominant component of AIC, since SSE is the squared error summed over all data points. On the other hand, with small amounts of data, simplicity plays a more determining role. This is consistent with our earlier observation that with large amounts of data, the bias component of error starts to dominate, whereas the variance component gets smaller and smaller as the amount of data increases.

A problem with the practical application of AIC is that σ^2 is of course unknown, and has to be estimated. We return to this issue when we discuss the possibilities of model selection in Splus.

Model selection in Splus

The problem of model selection has now been reduced to finding the model with the lowest AIC score. We now look at the problem of how to search the space of possible models. If the pool of potential explanatory variables is small, one can use exhaustive search (all possible subsets regression) but since the number of possible subsets is 2^k for k potential explanatory variables, this strategy has its limitations. Therefore, often hill-climbing algorithms are used. We give an example to illustrate how model selection can be performed in Splus.

We use the bodyfat example for illustration. We start the search with the model that includes `triceps`, `thigh` and `midarm` as explanatory variables. With the command `drop1` we can inspect the AIC value of models that can be obtained from the current model by removing a single variable. In

computing the AIC value, σ^2 is estimated by the s^2 of the current model. We apply `drop1` to the bodyfat model:

```
> drop1(bodyfat.fit)
Single term deletions

Model:
body.fat ~ triceps + thigh + midarm
        Df Sum of Sq      RSS       Cp
 <none>               98.4049 147.6073
triceps  1  12.70489 111.1098 148.0116
  thigh  1   7.52928 105.9342 142.8360
 midarm  1  11.54590 109.9508 146.8526
```

The column labeled `Cp` contains the AIC values. It turns out that we can reduce the AIC value the most by dropping `thigh` from the model. We drop it from the model, and obtain the new model:

```
> bodyfat.fit2 <- lm(body.fat ~ triceps + midarm, data=bodyfat)
> summary(bodyfat.fit2)

Call: lm(formula = body.fat ~ triceps + midarm, data = bodyfat)
Residuals:
    Min     1Q Median    3Q   Max
 -3.879 -1.963 0.3811 1.269 3.894

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept)  6.7916  4.4883     1.5132  0.1486
    triceps  1.0006  0.1282     7.8029  0.0000
     midarm -0.4314  0.1766    -2.4428  0.0258

Residual standard error: 2.496 on 17 degrees of freedom
Multiple R-Squared: 0.7862
```

Now we consider dropping yet another term:

```
> drop1(bodyfat.fit2)
```

```
Single term deletions

Model:
body.fat ~ triceps + midarm
        Df Sum of Sq      RSS       Cp
 <none>              105.9342 143.3227
triceps  1  379.4037 485.3379 510.2636
 midarm  1   37.1855 143.1197 168.0454
```

Notice that the AIC value of the current model is different from its value in the previous table. This is because a different estimate s^2 of σ^2 is used. We see that there is no single term deletion that reduces the AIC value, so our search stops here.

We can also start with the model that only contains the intercept term, and start adding variables:

```
> bodyfat.fit0 <- lm(body.fat ~ 1, data=bodyfat)
> summary(bodyfat.fit0)

Call: lm(formula = body.fat ~ 1, data = bodyfat)
Residuals:
    Min     1Q Median   3Q   Max
 -8.495 -3.145  1.005 4.08 7.005

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 20.1950  1.1418    17.6873  0.0000

Residual standard error: 5.106 on 19 degrees of freedom
Multiple R-Squared: 5.86e-031
```

The `1` in the formula

```
 body.fat ~ 1
```

denotes the intercept term. Now we look at the effect of adding a single term:

```
> add1(bodyfat.fit0, ~ triceps + thigh + midarm)
Single term additions
```

```
Model:
body.fat ~ 1
        Df Sum of Sq      RSS       Cp
 <none>              495.3895 547.5358
triceps  1  352.2698 143.1197 247.4122
  thigh  1  381.9658 113.4237 217.7162
 midarm  1   10.0516 485.3379 589.6304
```

In the second argument of `add1` we specify the terms that can be added. Adding `thigh` gives the most reduction of AIC so the new model becomes:

```
> bodyfat.fit3 <- lm(body.fat ~ thigh,data=bodyfat)
> summary(bodyfat.fit3)

Call: lm(formula = body.fat ~ thigh, data = bodyfat)
Residuals:
    Min     1Q Median    3Q   Max
 -4.495 -1.567 0.1241 1.336 4.408

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept) -23.6345   5.6574    -4.1776   0.0006
      thigh   0.8565   0.1100     7.7857   0.0000

Residual standard error: 2.51 on 18 degrees of freedom
Multiple R-Squared: 0.771
```

We consider adding yet another term:

```
> add1(bodyfat.fit3, ~ triceps + thigh + midarm)
Single term additions

Model:
body.fat ~ thigh
        Df Sum of Sq      RSS       Cp
 <none>              113.4237 138.6289
triceps  1  3.472892 109.9508 147.7587
 midarm  1  2.313901 111.1098 148.9177
```

There is no single term addition that would reduce the AIC value, so we stop here.

We can also do a fully atomatic search using the `step` function. Again we start with the full model.

```
> bodyfat.step1 <- step(bodyfat.fit,trace=F)

> bodyfat.step1$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
body.fat ~ triceps + thigh + midarm

Final Model:
body.fat ~ triceps + midarm


     Step Df Deviance Resid. Df Resid. Dev      AIC
1                             16    98.4049 147.6073
2 - thigh  1 7.529278        17   105.9342 142.8360
```

Let's try starting with the empty model again.

```
> bodyfat.step2 <- step(bodyfat.fit0, scope = ~ triceps+thigh+midarm)
...
> bodyfat.step2$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
body.fat ~ 1

Final Model:
body.fat ~ thigh


     Step Df  Deviance Resid. Df Resid. Dev      AIC
```

```
1                             19   495.3895 547.5358
2 + thigh -1 -381.9658        18   113.4237 217.7162
```

In both cases we end up with the same result as with the manual search.

In starting from the empty model, the estimate of σ^2 used by step is way too big, which might discourage the addition of variables. Therefore we retry with what we consider to be a more reasonable estimate of σ^2, namely the estimate based on the full model. This estimate is passed as the parameter `scale` to the step function:

```
> bodyfat.step3 <- step(bodyfat.fit0, scope = ~ triceps+thigh+midarm,
                        scale=2.48^2)

> bodyfat.step3$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
body.fat ~ 1

Final Model:
body.fat ~ thigh


     Step Df  Deviance Resid. Df Resid. Dev      AIC
1                             19   495.3895 507.6903
2 + thigh -1 -381.9658        18   113.4237 138.0253
```

In this case the result is the same, but we can see that the AIC values are considerably lower because of the smaller scaling value used.

4.11 Monte Carlo simulation

According to the frequentist viewpoint, statistical procedures (estimation, testing, prediction) should be evaluated on the basis of their properties when they are repeated many times. We have seen that sometimes we can use theoretical analysis to determine these properties. For example, under the assumptions SLR1-SLR6, we were able to derive the distributions of the

least squares estimators in the linear regression model. When we drop SLR6 (normality of the error term), we can invoke the Central Limit Theorem to approximate the distribution of the least squares estimators for large sample sizes. To determine at what sample size this approximation becomes reasonable is much harder however.

Another possibility is to use the computer to actually perform these many repetitions and to study the behaviour of an estimator or test procedure in that way. This technique is often called Monte Carlo simulation because we use the computer to simulate random draws from some population. Let's look at a simple example.

Suppose we draw a sample of size n from a distribution which is uniform on $[0, u]$, where u is unknown and has to be estimated from the data. One way to reason is as follows. Since the sample points $x_1, x_2, \ldots, x_n$ are drawn from a uniform distribution over the range 0 to u, their average

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

should be nearly $u/2$. So we estimate u as $2\bar{x}$.

We could also apply the maximum likelihood principle: find that value of u for which the observed sample has higher probability than for any other possible value of u. Obviously, we should estimate u by the sample maximum in that case. Even more obviously, this estimator is downward biased.

Which of these estimators do you think is better (let's say in terms of mean square error)? This is far less obvious. We can use computer simulation to find out. We just mimic the repeated sampling from $U(0, u)$ on the computer. For each sample we compute the two proposed estimators, and compare their mean and variance. Here's a small S program to do that

```
function(m, n, u)
{
        # m : number of samples
        # n : sample size
        # u : max of uniform distribution
call <- match.call()
uhat.1 <- uhat.2 <- vector(length = m)
for(i in 1:m) {
# draw sample of size n from uniform distribution U(0,u)
```

```
x <- runif(n, min = 0, max = u)
# compute first estimator
uhat.1[i] <- 2 * mean(x)
# compute second estimator
uhat.2[i] <- max(x)
}
m1 <- mean(uhat.1)
m2 <- mean(uhat.2)
v1 <- var(uhat.1)
v2 <- var(uhat.2)
list(call = call, m1 = m1, m2 = m2, v1 = v1, v2 = v2,
             uhat.1 = uhat.1, uhat.2 = uhat.2)
}
```

The program is self-explanatory as usual. The last expression of a function definition is returned as a result of a call. In this case it is a list with components enumerated between parentheses. We can call the function, which we named `mcunif` as follows:

```
> mc.1 <- mcunif(50,20,5)
> mc.1[2:5]
$m1:
[1] 4.950275

$m2:
[1] 4.75027

$v1:
[1] 0.4080421

$v2:
[1] 0.04212668
```

We assign the result of the call to `mcunif` to the variable ("object") `mc.1`. Then we select list components 2 to 5 for printing. We can see that the mean `m1` of the first estimator (based on drawing 50 samples of size 20) is 4.95. This

0.05 below the true value of u, but it is easy to show that it is in fact an unbiased estimator. The mean of the second estimator is 4.75 which is 0.25 below the true value. This confirms our intuition that this estimator must be downward biased. The most interesting result however is that the computed variance of the first estimator is about ten times as high as the computed variance of the second estimator.

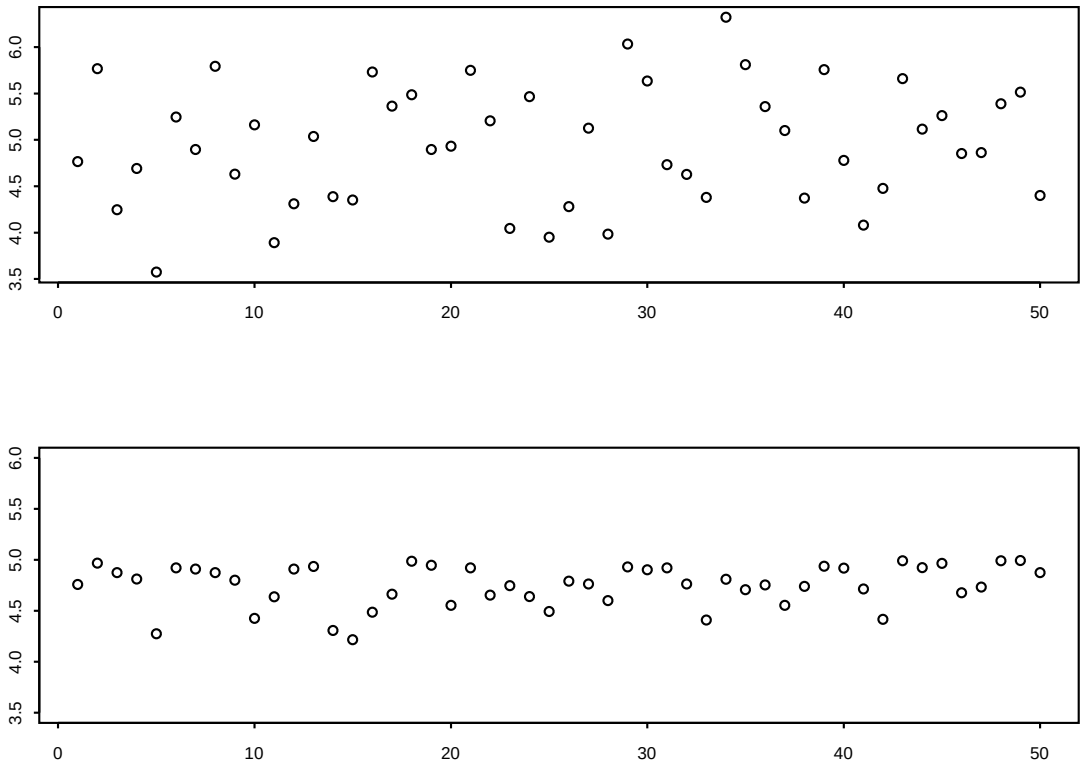

Figure 4.23: Plot of 50 estimates of u using first estimator (top panel) and second estimator (bottom panel)

Combining squared bias + variance into mean squared error, we get a mean squared error of $0.05^2 + 0.41 = 0.4125$ for the first estimator and $0.25^2 + 0.042 = 0.1045$ for the second estimator. So we should prefer the second estimator despite its bias.

Let's look at a second example from linear regression. We have stated

that OLS is BLUE under the standard assumptions of the linear regression model. In case of heteroskedastic error this is no longer the case: we can get an estimator with smaller variance by downweighting the unreliable observations, i.e. some form of weighted least squares. The simplest model was to assume that

$$\text{var}(\varepsilon_i) = \sigma_i^2 = \sigma^2 x_i$$

We did however not prove that the weighted least squares estimator indeed has lower variance. We perform a small simulation study to illustrate that this is indeed the case. We generate the data according to the simple heteroskedastic model given above.

```
function(m, n, beta0, beta1, sigma.square)
{       # m: number of samples
        # n: sample size
        # beta0: intercept of population regression line
        # beta1: slope of population regression line
        # sigma.square: common factor of variance
b1.ls <- vector(length = m)
b1.gls <- vector(length = m)
x <- 1:n
        # draw m samples and compute slopes
for(i in 1:m) {
                # generate heteroskedastic sample of size n
y <- beta0 + beta1 * x + rnorm(n, sd = sqrt(sigma.square * x))
                # ordinary least squares fit
ls.fit <- lsfit(x, y)
                # weighted least squares fit
gls.fit <- gls(y ~ x, data = data.frame(y, x), weights = varFixed( ~ x))
                # extract slope
b1.ls[i] <- ls.fit$coef[2]
b1.gls[i] <- gls.fit$coef[2]
}
list(b1.ls = b1.ls, b1.gls = b1.gls)
}
```

Note that x is kept fixed in repeated samples: it is outside the sampling loop. We called this function `mclinregr`. Here's a small example simulation:

```
> mclin.1 <- mclinregr(100,20,1,3,9)
> var(mclin.1$b1.gls)
[1] 0.1064367
> var(mclin.1$b1.ls)
[1] 0.1562418
> mean(mclin.1$b1.gls)
[1] 3.047714
> mean(mclin.1$b1.ls)
[1] 3.027393
```

On the basis of this simulation we would estimate the bias of ordinary least squares to be about 0.03 (for this sample size and parameter values) and of weighted least squares 0.05 (of course we know they are both unbiased). The variance of OLS is estimated at about 0.16, and of WLS at 0.11. Combining the two into mean square error, we get $0.03^2 + 0.16 = 0.161$ for OLS and $0.05^2 + 0.11 = 0.113$. Hence we should prefer WLS on the basis of this simulation.

Proof of bias-variance decomposition

In this section we prove the decomposition of mean square estimation error into its bias and variance components. To save space, we write f for $f(x)$, $\hat{f}$ for $\hat{f}(x|T)$ and drop the subscript T from the expectations.

The mean square error of $\hat{f}$ as an estimator of f is defined as

$$M(\hat{f}) = E(\hat{f} - f)^2$$

The bias of $\hat{f}$ as an estimator of f is defined as

$$B(\hat{f}) = E(\hat{f} - f)$$

The variance of $\hat{f}$ is defined as

$$V(\hat{f}) = E(\hat{f} - E(\hat{f}))^2$$

We prove that

$$M(\hat{f}) = B^2(\hat{f}) + V(\hat{f}),$$

i.e., mean square error equals squared bias + variance.

Proof: we can write $\hat{f} - f$ as

$$\hat{f} - f = (\hat{f} - E(\hat{f})) + (E(\hat{f}) - f)$$

Square on the left and on the right

$$(\hat{f} - f)^2 = (\hat{f} - E(\hat{f}))^2 + (E(\hat{f}) - f)^2 + 2(\hat{f} - E(\hat{f}))(E(\hat{f}) - f)$$

We take expectations left and right. Since $E(\hat{f} - E(\hat{f})) = E(\hat{f}) - E(\hat{f}) = 0$ and $E(\hat{f}) - f$ is a constant, the cross term then drops out. So we get

$$E(\hat{f} - f)^2 = E(\hat{f} - E(\hat{f}))^2 + (E(\hat{f}) - f)^2 = V(\hat{f}) + B^2(\hat{f})$$

Proof that $E(Y|X)$ minimizes mean square prediction error

The problem here is to predict the value of Y by a function of X, call it $f(X)$. We pick a point $X = x$.

One measure of the goodness of the predictor $f(x)$ of Y at x is its *mean square error*

$$\mathrm{M}(f(x)) = E(Y - f(x))^2$$

It is a measure of, on average, how far off the prediction is. We show that $f(x) = E(Y|x)$ minimizes the mean square error.

$$\begin{aligned} E(Y - f(x))^2 &= E(Y^2 - 2f(x)Y + f(x)^2) \\ &= E(Y^2|x) - 2f(x)E(Y|x) + f(x)^2 \end{aligned}$$

So

$$\frac{d}{df} = -2E(Y|x) + 2f(x),$$

which is zero when $f(x) = E(Y|x)$. Since this is true for any value of X, $f(X) = E(Y|X)$ minimizes the mean square error.

4.12 Exercises

1. Prove that $\sum(x_i - \bar{x})x_i = \sum(x_i - \bar{x})^2$.

2. Fitting a line through the origin. Suppose we know that the fitted line must go through the origin, i.e. when x is zero, y must be zero as well. Use least squares to find a general expression for the slope of the fitted line.

3. Suppose we know for a fact that the slope of the fitted line must be zero. What kind of relationship is there between x and y? What value would you now predict for y if you want to minimize the sum of squared errors?

4. (Refresher) Use the properties of expectations and variances to show that if X is a random variable with expectation $E(X) = \mu$, and variance $V(X) = \sigma^2$, then
$$Z = \frac{X - \mu}{\sigma}$$
has expectation $E(Z) = 0$ and variance $V(Z) = 1$.

5. (Regression through the origin). We have observations $(x_i, y_i), i = 1, \ldots, n$ where the y_i are the observed values of random variables $Y_1, \ldots, Y_n$. The x_i are fixed by the experimenter. Assume that
$$E(Y_i) = \beta x_i$$
Show that the least squares estimator of β it is unbiased.

6. Show that $\sum c_i w_i = 0$ in the proof of the Gauss-Markov theorem for b_1 (Hint: use the constraints $\sum c_i = 0$ and $\sum c_i x_i = 0$).

7. (Regression through the origin continued) Suppose the usual assumptions of the linear regression model apply, but the true value of the intercept is zero. We have already derived the least squares estimator for the slope of this model. Compare the variance of this estimator to that of the slope estimator computed with an unnecessary intercept term.

8. Suppose you are estimating a linear regression model. If you multiply all the x_i values by 10, but not the y_i values, what happens to the parameter values β_0 and β_1? What happens to the least squares estimates b_0 and b_1? What happens to the variance of the error term?

9. We have the following 6 observations on x and y and want to fit a linear regression model.

i	1	2	3	4	5	6
x_i	4	1	2	3	3	4
y_i	16	5	10	15	13	22

   a) Compute Y^TY, X^TX and X^TY.

   b) Find $(X^TX)^{-1}$.

   c) Find the vector of estimated regression coefficients, the vector of residuals (errors),SSR,SSE, and the estimated variance-covariance matrix of b. Give a point estimate of $E(y)$ for $x = 4$.

   d) From the estimated variance-covariance matrix in c) obtain the following: $\widehat{\text{cov}}(b_0, b_1)$, $\widehat{\text{var}}(b_0)$, and $\widehat{\text{var}}(b_1)$. Test the null hypothesis that $\beta_1 = 0$ against a two-sided alternative at $\alpha = 0.05$.

10. (Taken from [11]) A substance that is used in biological and medical research is shipped by airfreight to users in cartons of 1000 ampules. The data below, involving 10 shipments, were collected on the number of times the carton was transferred from one aircraft to another over the shipment route (x), and the number of ampules found to be broken on arrival (y).

i	1	2	3	4	5	6	7	8	9	10
x_i	1	0	2	0	3	1	0	1	2	0
y_i	16	9	17	12	22	13	8	15	19	11

Using these observations, we estimate the model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Assume that the usual assumptions of the linear regression model are appropriate.

a) Compute the least-squares estimates of β_0 and β_1.

b) Obtain a point estimate of the expected number of broken ampules when one transfer is made. Estimate the increase in the expected number of broken ampules when there are two transfers as compared to one transfer.

c) Compute a 95% confidence interval for β_1. Interpret this interval estimate.

d) Perform a test to decide whether or not there is a linear association between number of times a carton is transferred (x) and the number of broken ampules (y), at $\alpha = 0.05$. State the null hypothesis, the alternative hypothesis, the decision rule and the conclusion. What is the P-value of the test?

e) A consultant claims, based on previous experience, that the mean number of broken ampules should not exceed 9.0 when no transfers are made. Conduct an appropriate test to verify this claim at $\alpha = 0.025$. State the null hypothesis, the alternative hypothesis, the decision rule and the conclusion. What is the P-value of the test?

f) What percentage of the variation in y is explained by the variation in x ?

g) The next shipment will entail two transfers. Compute a 99% prediction interval for the number of broken ampules of this shipment. Interpret this prediction interval.

11. Analysis of the food expenditure data with Splus yields the following output:

```
> food.fit <- lm(foodexp ~ income,data=food)
> summary(food.fit)

Call: lm(formula = foodexp ~ income, data = food)
Residuals:
```

```
    Min     1Q Median    3Q   Max
 -71.75 -19.67 -5.969 17.75 80.14

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) 40.7676 22.1387     1.8415  0.0734
     income  0.1283  0.0305     4.2008  0.0002

Residual standard error: 37.81 on 38 degrees of freedom
Multiple R-Squared: 0.3171
F-statistic: 17.65 on 1 and 38 degrees of freedom,
             the p-value is 0.000155

Correlation of Coefficients:
       (Intercept)
income -0.9629
```

a) Write down the estimated regression function.

b) Construct a 95% confidence interval for β_0 and explain what it means.

c) Test the null hypothesis that β_0 is zero against the alternative that it is not, at the 5% level of significance without using the reported p-value. What is your conclusion?

d) Draw a sketch showing the p-value 0.0734 given in the Splus output, the critical value from the t-distribution used in (c) and how the p-value could have been used to answer (c).

e) Test the null hypothesis that β_0 is zero, against the alternative that it is positive, at the 5% level of significance. Draw a sketch of the rejection region. What is your conclusion? Repeat for β_1.

12. (Taken from [7]) Suppose you wish to estimate the slope of the regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. The problem you have is threefold:

    1) you are stranded on a desert island and have no computer or calculator, and

    2) you have only three observations and

3) you don't remember the formulas for the least squares estimators

Recalling that 2 points determine a line, you form an average of observations 2 and 3 as follows

$$y^* = \frac{y_2 + y_3}{2} \qquad\qquad x^* = \frac{x_2 + x_3}{2}$$

The slope of the line connecting the points is $b^* = (y^* - y_1)/(x^* - x_1)$.

a) Show that this estimator of β_1 is a linear estimator.

b) Show that this estimator of β_1 is unbiased.

c) Determine the variance of b^*.

d) Is b^* just as good as the least squares estimator? Explain.

e) Write a simulation program in Splus to doublecheck your theoretical results.

13. (Taken from [10]) Extensive studies have shown that the performance of employees depends on the temperature of the working environment according to the following model

$$Y = 230 - 2x + \varepsilon$$

Here x denotes the temperature in degrees Celcius and Y the performance of an employee (according to some measure); the relationship holds for $20 \leq x \leq 35$.

An employer suspects that in his company temperature has an even stronger negative influence on performance. He decides to make some observations, with the following outcomes:

i	1	2	3	4	5	6	7
x_i	31	25	27	23	32	22	29
y_i	80	105	120	105	70	120	100

Using these observations we estimate the model

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Assume that the usual assumptions of the linear regression model apply.

a) Compute the least squares estimates of β_0 and β_1.

b) What percentage of the variation in performance is explained by the variation in temperature ?

c) Use a test to check whether the suspicion of the employer is confirmed by the data. Use $\alpha = 0.05$. State the null hypothesis, alternative hypothesis, decision rule and conclusion.

d) To present the results on a conference in the US the temperature has to be expressed in degrees Fahrenheit. Give the regression equation that you will present at the conference (You can convert from degrees Celsius to degrees Fahrenheit by multiplying the number of degrees Celsius by 9/5 and adding 32 to the resulting figure).

Chapter 5

Logistic Regression

5.1 Introduction

In many regression problems, the response variable of interest has only two possible qualitative outcomes, and therefore can be represented by a binary indicator variable taking on values 0 and 1. For example:

1. In a model for credit scoring, the response variable may be defined to have two possible outcomes: the loan defaulted (1) or it didn't (0). Explanatory variables may be income of the applicant, his or her occupation and so on.

2. In building a SPAM filter, we would like to determine whether an e-mail message is SPAM (1) or not (0). Among the variables that discriminate SPAM from non-SPAM are word counts (e.g. relative frequency of the word "free"), number of CAPITAL LETTERS, and so on.

In section 5.2 we explain why linear regression does not work very well for this kind of problem. In section 5.3 we presented a model that is more appropriate for binary classification problems.

5.2 The linear probability model

Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad\qquad Y_i = 0, 1 \tag{5.1}$$

where the outcome Y_i is binary taking on the value of either 0 or 1. Since by assumption, $E(\varepsilon_i) = 0$ we have

$$E(Y_i) = \beta_0 + \beta_1 x_i \tag{5.2}$$

Now Y_i is a Bernoulli random variable with probability distribution

Y_i	Probability
1	$P(Y_i = 1) = \pi_i$
0	$P(Y_i = 0) = 1 - \pi_i$

Thus, π_i is the probability that $Y_i = 1$, and $1 - \pi_i$ is the probability that $Y_i = 0$. The expected value of Y_i then is

$$E(Y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i \tag{5.3}$$

Equating (5.2) and (5.3) we get

$$E(Y_i) = \beta_0 + \beta_1 x_i = \pi_i \tag{5.4}$$

The mean response $E(Y_i) = \beta_0 + \beta_1 x_i$ as given by the response function is therefore simply the probability that $Y_i = 1$ when the level of the predictor variable is x_i.

When the response variable is binary, assumption SLR3 (constant variance of the error term) of the linear regression model is false. To see this, we derive the variance of Y_i for the simple linear regression model in (5.1).

$$\begin{aligned} \text{var}(Y_i) &= E[(Y_i - E(Y_i))^2] \\ &= (1 - \pi_i)^2 \pi_i + (0 - \pi_i)^2 (1 - \pi_i) \\ &= \pi_i (1 - \pi_i) = E(Y_i)(1 - E(Y_i)) \\ &= (\beta_0 + \beta_1 x_i)(1 - \beta_0 - \beta_1 x_i) \end{aligned}$$

Note that $\text{var}(Y_i)$ depends on x_i, hence the error variances will be different for different levels of x and ordinary least squares will no longer be best.

A more serious problem is that the linear response function does not necessarily satisfy the constraint

$$0 \leq E(Y) = \pi \leq 1$$

For example, in the credit scoring example with a linear response function, some applicant with an extremely high income may be predicted to have a

negative probability of defaulting. If this probability is then set to zero (which seems reasonable), this would have the effect that all low (i.e. below a certain threshold) income applicants are assigned probability 1 of being defaulters, and all high income applicants get probability 0. As a consequence, there may be defaulters in the sample that have probability zero of being a defaulter according to the model. This problem is illustrated in figure 5.1.

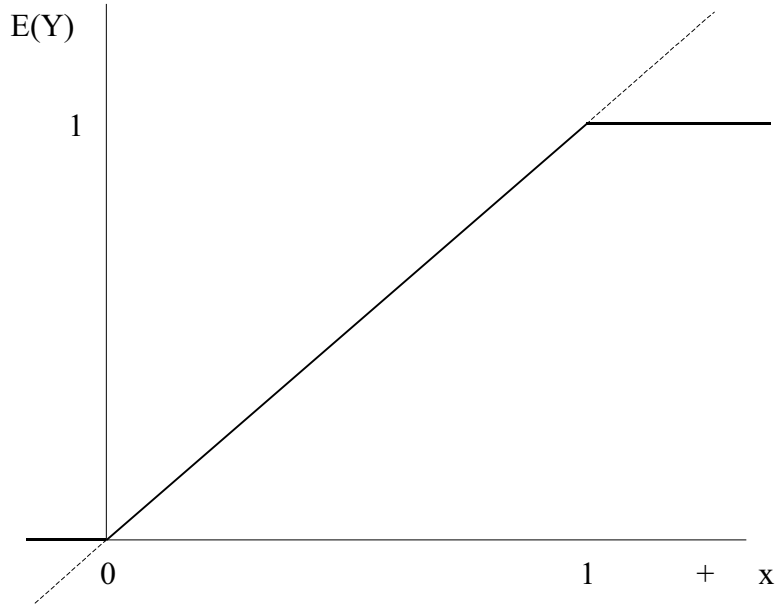

Figure 5.1: Linear response function

Such a model would often be considered unreasonable. Instead, a model where the probabilities 0 and 1 are reached asymptotically, as illustrated in figure 5.2, would usually be more appropriate.

5.3 Simple logistic regression

The response function plotted in figure 5.2 is called the *logistic response function* and is of the form

$$E(Y) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{5.5}$$

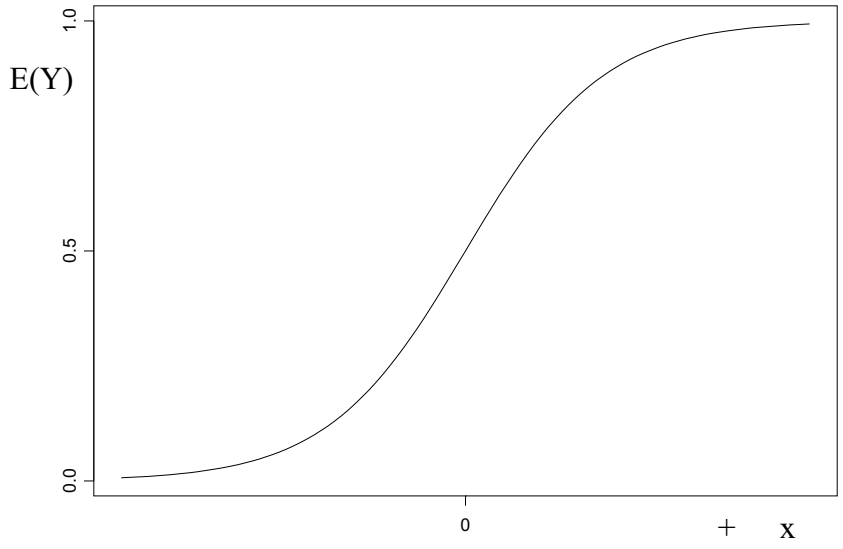

Figure 5.2: Logistic response function

or (divide numerator and denominator by $e^{\beta_0+\beta_1 x}$)

$$E(Y) = \frac{1}{1 + e^{-\beta_0-\beta_1 x}} \tag{5.6}$$

An interesting property of the logistic response function is that it can be linearized easily. Let us denote $E(Y)$ by π, since the mean response is a probability when the response variable is a 0,1 indicator variable. Then if we make the transformation (we substitute z for $\beta_0 + \beta_1 x$)

$$\begin{aligned}
\pi' &= \ln\left(\frac{\pi}{1-\pi}\right) \\
&= \ln\left(\frac{(1+e^{-z})^{-1}}{1-(1+e^{-z})^{-1}}\right) \\
&= \ln\left(\frac{1}{(1+e^{-z})-1}\right) \\
&= \ln\left(\frac{1}{e^{-z}}\right) \\
&= \ln(e^z) = z = \beta_0 + \beta_1 x
\end{aligned}$$

Where in the second step, we divided the numerator and the denominator by $(1 + e^{-z})^{-1}$. This transformation is called the *logit transformation* of the probability π. The ratio $\pi/(1 - \pi)$ is called the *odds*.

5.3.1 Maximum likelihood estimation of logistic regression model

We state the simple logistic regression model as follows: the Y_i are independent Bernoulli random variables with expected values $E(Y_i) = \pi_i$, where

$$E(Y_i) = \pi_i = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}} \tag{5.7}$$

Since each Y_i observation is a Bernoulli random variable, where

$$\begin{aligned} P(Y_i = 1) &= \pi_i \\ P(Y_i = 0) &= 1 - \pi_i \end{aligned}$$

we can represent its probability distribution as follows

$$p_i(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \qquad y_i = 0, 1; \;\; i = 1, \ldots, n \tag{5.8}$$

Note that $p_i(1) = \pi_i$, $p_i(0) = 1 - \pi_i$ as required.

Since the y_i observations are independent (e.g. random sampling), their joint probability is simply the product of their individual probabilities

$$p(y_1, \ldots, y_n) = \prod_{i=1}^{n} p_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \tag{5.9}$$

For convenience we work with the loglikelihood, since products become sums, and if we take the natural log, we can get rid of some powers of e.

$$\begin{aligned} \ln p(y_1, \ldots, y_n) &= \ln \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \\ &= \sum_{i=1}^{n} y_i \ln\left(\frac{\pi_i}{1 - \pi_i}\right) + \sum_{i=1}^{n} \ln(1 - \pi_i) \\ &= \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \ln(1 + e^{\beta_0+\beta_1 x_i}) \end{aligned}$$

In the last step we use the fact that

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_i$$

and

$$1 - \pi_i = (1 + e^{\beta_0+\beta_1 x_i})^{-1}$$

Hence the loglikelihood function is

$$\mathcal{L}(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \ln(1 + e^{\beta_0+\beta_1 x_i}) \tag{5.10}$$

The maximum likelihood estimates of β_0 and β_1 are those values $\hat{\beta}_0$ and $\hat{\beta}_1$ that maximize the log-likelihood in (5.10). Unfortunately, there are no formulas that give us the values of $\hat{\beta}_0$ and $\hat{\beta}_1$, as there are in least squares estimation of the linear regression model. Computer intensive numerical search procedures are required to find the maximum likelihood estimates. Once they are found, we substitute these values into the response function in (5.7) to obtain the *fitted response function*

$$\hat{\pi} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 x}} \tag{5.11}$$

5.3.2 Example

Suppose we want to model the probability of succesfully completing a programming assignment within a limited amount of time. The explanatory variable is "number of months programming experience". Below we give the data used to estimate the regression equation, as well as the fitted response for each observation. In the `Value` column of the `Coefficients` table we find $\hat{\beta}_0 = -3.0597$ and $\hat{\beta}_1 = 0.1615$. This means that

$$\hat{\pi}_i = \frac{e^{-3.0597+0.1615x_i}}{1 + e^{-3.0597+0.1615x_i}}$$

So for example, if someone has 14 months of programming experience, we would estimate his probability of successfully completing the task to be

$$\hat{\pi}_i = \frac{e^{-3.0597+0.1615(14)}}{1 + e^{-3.0597+0.1615(14)}} \approx 0.31$$

```
> summary(programming.fit)

Coefficients:
                 Value Std. Error   t value
(Intercept) -3.0596954  1.2589852 -2.430287
  month.exp  0.1614859  0.0649625  2.485833

   month.exp success   fitted     month.exp success   fitted
 1        14       0 0.310262  16        13       0 0.276802
 2        29       0 0.835263  17         9       0 0.167100
 3         6       0 0.109996  18        32       1 0.891664
 4        25       1 0.726602  19        24       0 0.693379
 5        18       1 0.461837  20        13       1 0.276802
 6         4       0 0.082130  21        19       0 0.502134
 7        18       0 0.461837  22         4       0 0.082130
 8        12       0 0.245666  23        28       1 0.811825
 9        22       1 0.620812  24        22       1 0.620812
10         6       0 0.109996  25         8       1 0.145815
11        30       1 0.856299
12        11       0 0.216980
13        30       1 0.856299
14         5       0 0.095154
15        20       1 0.542404
```

Now let's look at the interpretation of the coefficient $\hat{\beta}_1$. Unfortunately it is not as simple as in the linear regression model. There $\hat{\beta}_1$ indicated the expected change in y when x increased with one unit.

Let's see what happens to the fitted logit respons when x increases with one unit. Pick any value of x, say $x = x_j$. Then the fitted logit response is

$$\hat{\pi}'(x_j) = \hat{\beta}_0 + \hat{\beta}_1 x_j$$

The fitted logit response for $x = x_j + 1$ is

$$\hat{\pi}'(x_j + 1) = \hat{\beta}_0 + \hat{\beta}_1(x_j + 1)$$

The difference between the two is

$$\hat{\pi}'(x_j + 1) - \hat{\pi}'(x_j) = \hat{\beta}_1$$

If we write ln(odds1) for $\hat{\pi}'(x_j)$ and ln(odds2) for $\hat{\pi}'(x_j + 1)$ then

$$\ln(\text{odds2}) - \ln(\text{odds1}) = \ln\left(\frac{\text{odds2}}{\text{odds1}}\right) = \hat{\beta}_1$$

From which it follows that

$$\widehat{\text{OR}} = \frac{\text{odds2}}{\text{odds1}} = e^{\hat{\beta}_1}$$

Here $\widehat{\text{OR}}$ is short for the estimated odds ratio.

We continue our programming task example to illustrate the interpretation of $\hat{\beta}_1$. We found that $\hat{\beta}_1 = 0.1615$, so

$$\widehat{\text{OR}} = e^{0.1615} = 1.175$$

This means the odds increase with 17.5% with every extra month of experience. The estimated odds ratio for an increase with c months is simply $e^{c\hat{\beta}_1}$. So if we compare someone with 10 months of experience to someone with twenty-five months of experience ($c = 15$), then

$$\widehat{\text{OR}} = e^{15(0.1615)} = 11.3$$

so the odds for the programmer with 25 months of experience are about 11 times as high as for the programmer with only 10 months of experience.

5.4 Multiple Logistic Regression

Like in linear regression, we usually want to include more than one explanatory variable in a logistic regression model. Basically we just replace

$$\beta_0 + \beta_1 x$$

by

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1}$$

in all formulas. So we get, for example

$$E(y_i) = \pi_i = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1})}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1})}$$

It is more convenient to use vector notation. We define the vectors

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad X = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_{p-1} \end{bmatrix} \qquad X_i = \begin{bmatrix} 1 \\ x_{i,1} \\ \vdots \\ x_{i,p-1} \end{bmatrix}$$

so we get

$$\beta^T X = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1}$$

and

$$\beta^T X_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1}$$

The response function can now be written as

$$E(y) = \pi = \frac{\exp(\beta^T X)}{1 + \exp(\beta^T X)}$$

and the logit transformation of the response is

$$\pi' = \ln\left(\frac{\pi}{1-\pi}\right) = \beta^T X$$

The parameters are estimated by maximizing the log-likelihood

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} y_i(\beta^T X_i) - \sum_{i=1}^{n} \ln(1 + \exp(\beta^T X_i))$$

Again we have to resort to numeric methods to determine the maximum likelihood estimate $\hat{\beta}$ of β. Once we have computed $\hat{\beta}$, we can determine the fitted response function

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}^T X_i)}{1 + \exp(\hat{\beta}^T X_i)}$$

In order to be able to compute confidence intervals and to perform hypothesis tests concerning the parameters, we have to know something about the distribution of $\hat{\beta}$, the maximum likelood estimator of β. This is somewhat harder to determine than in the case of the least-squares estimators for linear regression. In fact the only results we have are so-called asymptotic results, i.e. they only apply if we have enough observations. Simply put, if we have

enough observations, then $\hat{\beta}$ is nearly unbiased, and also nearly normally distributed. We will not discuss the variance-covariance matrix of $\hat{\beta}$. If this matrix is given, we can use it together with the standard normal distribution to compute confidence intervals and perform hypothesis tests in the usual fashion.

5.5 Discrete choice models

We can arrive at the logistic regression and similar models via another path as well. We view the outcome $(y = 0, 1)$ as a discretization of an underlying regression. Consider for example the decision to make a large purchase. Micro-economic theory states that the consumer makes a cost-benefit calculation. Since benefit is obviously not observable, we model the difference between cost and benefit as an unobserved variable y^*, such that

$$y^* = \beta_0 + \beta_1 x + \varepsilon$$

We typically assume that ε has a logistic or standard normal distribution. We do not observe the net benefit of the purchase, only whether it is made or not. Therefore, our observation is

$$y = \begin{cases} 1 & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leq 0 \end{cases}$$

Now the probability that $y = 1$ is

$$\begin{aligned} P(y = 1) &= P(y^* > 0) \\ &= P(\beta_0 + \beta_1 x + \varepsilon > 0) \\ &= P(\varepsilon > -\beta_0 - \beta_1 x) \end{aligned}$$

If the distribution of ε is symmetric (e.g. normal or logistic), then

$$\begin{aligned} P(\varepsilon > -\beta_0 - \beta_1 x) &= P(\varepsilon < \beta_0 + \beta_1 x) \\ &= F(\beta_0 + \beta_1 x) \end{aligned}$$

Here F is the distribution function of ε.

5.5.1 The probit model

In the probit model we assume $\varepsilon \sim N(0,1)$. The assumption of unit variance is a harmless normalization. Suppose we assume that $\varepsilon \sim N(0,\sigma^2)$ Then

$$P(\varepsilon < \beta_0 + \beta_1 x) = P\left(\frac{\varepsilon}{\sigma} < \frac{\beta_0 + \beta_1 x}{\sigma}\right)$$

Clearly, $\varepsilon/\sigma \sim N(0,1)$, so we can divide β_0 and β_1 by σ and get exactly the same probabilities as in the other model. Since we only observe whether y is 0 or 1 (and not the value of y^*), these models are observationally equivalent. The assumption of zero for the threshold is likewise innocent if the model contains a constant term β_0.

So for the probit model

$$P(y=1) = \Phi(\beta_0 + \beta_1 x)$$

where $\Phi(\cdot)$ is the standard normal distribution function.

And for the logit model

$$P(y=1) = \Lambda(\beta_0 + \beta_1 x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

where $\Lambda(\cdot)$ indicates the cumulative logistic distribution function.

5.6 Model Selection for Logistic Regression

Like in linear regression we may be confronted with the problem of finding the model with the best predictive performance from a large set of potential models. As a measure of model fit we use the value of the log-likelihood function evaluated at $\hat{\beta}$ (the maximum likelihood estimate of β)

$$\mathcal{L}(\hat{\beta}) = \sum_{i=1}^{n} y_i \ln(\hat{\pi}_i) + (1-y_i)\ln(1-\hat{\pi}_i)$$

The larger this value, the better the fit of the model. This makes sense, because the likelihood gives the probability of the data given the model, and the higher this probability the better the fit. A measure that is comparable with SSE in the linear regression model is

$$\text{Deviance} = -2\mathcal{L}(\hat{\beta})$$

The model with the smallest deviance gives the best fit. Of course we have to beware of overfitting, so just picking the model with the smallest deviance would be a bad strategy. Again we have to balance the fit against the model complexity as measured by the number of parameters. The Akaike Information Criterion for logistic regression models is

$$\text{AIC}(\text{model}) = \text{Deviance}(\text{model}) + 2p$$

The lower the AIC value the better. The term $2p$ is a penalty for the complexity of the model: if we move from a simple to a more complex model then the reduction of deviance it achieves must be more than twice the number of additional parameters for the complex model to be preferred.

The strategies to search the space of possible models for the lowest AIC value are essentially the same as those discussed for the linear regression model.

5.7 Exercises

1. A health clinic in Utrecht sent flyers to inhabitants to encourage everyone, but especially older persons, to get a flu shot in time for protection against an expected flu epidemic. In a small pilot study, 50 inhabitants were randomly selected and asked whether they actually received a flu shot. In addition, data were collected on their age (x_1) and their health awareness. The latter data were combined into a health awareness index (x_2), for which higher values indicate greater awareness. An inhabitant who received a flu shot was coded $y = 1$, and and an inhabitant who did not receive a flu shot was coded $y = 0$.

   Initially, we estimate the model

   $$E(y) = P(y = 1) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}$$

   using maximum likelihood. This give the following results

```
Coefficients:
                  Value Std. Error
(Intercept) -21.5821259 6.33965854
```

```
        age   0.2217512 0.07359717
      index   0.2034849 0.06206469

Deviance: 32.41631 on 47 degrees of freedom
```

a) We compute $\exp(\hat{\beta}_1) = \exp(0.2217512) \approx 1.25$. Does this number have a simple interpretation? Explain.

Somebody claims that the influence of age on whether or not someone gets a flu shot depends on the health awareness of this person. Therefore we estimate the alternative model

$$E(y) = P(y = 1) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2))}$$

with maximum likelihood. This yields the following results

```
Coefficients:
                     Value Std. Error
   (Intercept)  26.75512936 23.2458565
           age  -0.88140146  0.5399075
         index  -0.82228216  0.4948594
     age:index   0.02365021  0.0117723

Deviance: 24.28312 on 46 degrees of freedom
```

Here the row `age:index` contains the results for the interaction term $x_1 x_2$.

b) We compute $\exp(\hat{\beta}_1) = \exp(-0.88140146) \approx 0.4$. Does this number have a simple interpretation? Explain.

c) Compute the AIC score for both models. Which model is preferred?

d) Test the claim that the influence of age on whether or not someone gets a flu shot depends on health awareness at $\alpha = 0.05$. (Assume that the sample is big enough for the asymptotic distribution of the maximum likelihood estimators to give a good approximation).

e) What is the estimated probability that inhabitants aged 55 with a health awareness index of 60 will receive a flu shot?

2. (Taken from [8]) The table below presents the test-firing results for 25 surface-to-air anti aircraft missiles at targets of varying speed. The result of each test is either a hit ($y = 1$) or a miss ($y = 0$). The explanatory variable x gives the speed of the target in knots.

Test	Target speed (x) in knots	y	Test	Target speed (x) in knots	y
1	400	0	14	330	1
2	220	1	15	280	1
3	490	0	16	210	1
4	210	1	17	300	1
5	500	0	18	470	1
6	270	0	19	230	0
7	200	1	20	430	0
8	470	0	21	460	0
9	480	0	22	220	1
10	310	1	23	250	1
11	240	1	24	200	1
12	490	0	25	390	0
13	420	0			

We estimate the model

$$E(y) = P(y = 1) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

using maximum likelihood. This gives the following results:

```
Coefficients:
                   Value  Std. Error
  (Intercept)  6.07086259 2.105830002
target.speed  -0.01770463 0.006065314

Model Deviance: 20.36366 on 23 degrees of freedom
```

a) We compute $\exp(50 \times \hat{\beta}_1) = \exp(-0.89) \approx 0.41$. Interpret this number.

b) Test whether an increase in target speed has a negative influence on the probability of a hit, at $\alpha = 0.05$. (Assume the sample size is sufficient for the asymptotic distribution of the maximum likelihood estimators to give a reasonable approximation). State the null hypothesis, alternative hypothesis, decision rule and conclusion.

Someone claims that a quadratic term in target speed should be included in the model. We estimate the alternative model

$$E(y) = P(y = 1) = \frac{\exp(\beta_0 + \beta_1 x + \beta_2 x^2)}{1 + \exp(\beta_0 + \beta_1 x + \beta_2 x^2)}$$

using maximum likelihood. This gives the following results:

```
Coefficients:
                          Value    Std. Error
      (Intercept)  6.192757e+000 9.01877932046
     target.speed -1.846563e-002 0.05505815155
   target.speed.2  1.100548e-006 0.00007910137

Model Deviance: 20.36346 on 22 degrees of freedom
```

Here the row `target.speed.2` contains the estimated coefficient and standard error for the quadratic term.

c) Compute the AIC score for the linear and quadratic model and indicate which one is preferred on the basis of this score.

Chapter 6

Statistical Discriminant Analysis

6.1 Introduction

In this chapter we are concerned with the interconnected problems of

a) Assigning/allocating an object to a class, on the basis of a number of variables that describe the object.

b) Estimating the probability that a particular object belongs to a specific class.

The problems are interconnected, since an allocation rule is usually based on the estimated probabilities.

In this kind of *classification* problem there is an output/dependent variable y that assumes values in an unordered discrete set. An important special case is when there are only two classes, in which case we usually assume $y \in \{0, 1\}$. The goal of a classification procedure is to predict the output value given a set of input/independent/explanatory variables $\mathbf{x} = \{x_1, \ldots, x_p\}$ measured on the same object.

At a particular point $\mathbf{x}$ the value of y is not uniquely determined. It can assume both its values with respective probabilities that depend on the location of the point $\mathbf{x}$ in the input space. We write

$$P(y = 1|\mathbf{x}) = 1 - P(y = 0|\mathbf{x}) = f(\mathbf{x}). \tag{6.1}$$

Here $f(\mathbf{x})$ is a single-valued deterministic function that at every point $\mathbf{x}$ specifies the probability that $y = 1$. We assume the goal of a classification procedure is to produce an estimate $\hat{f}(\mathbf{x})$ of $f(\mathbf{x})$ at every input point.

There are two basic approaches to producing such an estimate, sometimes called *function estimation* and *density estimation* respectively. We have already encountered an example of the function estimation approach: logistic regression. In the next sections we give a general description of the two approaches. For simplicity we assume there are only two classes.

6.2 Function estimation

In the usual function estimation setting, one assumes that the output variable y is related to a set of input variables $\mathbf{x}$ by

$$y = f(\mathbf{x}) + \varepsilon \tag{6.2}$$

where $f(\mathbf{x})$ (*target function*) is a single-valued deterministic function of p arguments and ε is a random variable distributed according to some probability distribution. By definition its average is $E(\varepsilon|\mathbf{x}) = 0$ for all $\mathbf{x}$ so that the target function is defined by

$$f(\mathbf{x}) = E(y|\mathbf{x}), \tag{6.3}$$

the expected value of y at $\mathbf{x}$.

The goal is to obtain an estimate

$$\hat{f}(\mathbf{x}|T) = \hat{E}(y|\mathbf{x}, T) \tag{6.4}$$

using some training set T. The classification problem can be cast in the function estimation setting by observing that (6.3) holds for y and $f(\mathbf{x})$ in (6.1) so that they can be related by (6.2) where ε has a binomial distribution with variance $\sigma^2 = f(\mathbf{x})(1 - f(\mathbf{x}))$. Thus, regular function estimation technology (6.4) can be applied to obtain the estimate $\hat{f}(\mathbf{x}|T)$. This paradigm is used by many popular classification methods, including logistic regression, neural networks, and classification trees. Notice that we only model the conditional distribution of y given $\mathbf{x}$, the probability distribution of $\mathbf{x}$ itself is not modeled. This means we either assume the $\mathbf{x}$ values are chosen by the experimenter, or that we condition our inferences on the observed $\mathbf{x}$ values.

6.3 Density estimation

An alternative paradigm for estimating $f(\mathbf{x})$ in the classification setting is based on density estimation. Here Bayes' theorem

$$f(\mathbf{x}) = \frac{\pi_1 p_1(\mathbf{x})}{\pi_0 p_0(\mathbf{x}) + \pi_1 p_1(\mathbf{x})} \tag{6.5}$$

is applied where $p_i(\mathbf{x}) = p(\mathbf{x}|y = i)$ are the class conditional probability density functions and $\pi_i = P(y = i)$ are the unconditional ("prior") probabilities of each class. The training data are partioned into subsets $T = \{T_0, T_1\}$ with the same class label. The data in each subset are separately used to estimate its respective probability density $\hat{p}_i(\mathbf{x}|T_i)$, and prior probabilities $\hat{\pi}_i$. These estimates are plugged into (6.5) to obtain an estimate $\hat{f}(\mathbf{x}|T)$. Examples of this approach are discriminant analysis, mixture modeling and bayesian networks. A general expression to compute the probability of group i at $\mathbf{x}$ is given by

$$P(y = i|\mathbf{x}) = \frac{p_i(\mathbf{x})\pi_i}{\sum_{j=1}^{g} p_j(\mathbf{x})\pi_j} \tag{6.6}$$

where g denotes the number of groups/classes.

6.4 Density estimation: example

We start with a simple example of how the density estimation approach works. Many companies send so called test mailings to potential customers, and record whether or not a person responds to such a mailing. Since attributes such as age, income etc. are also recorded, this allows us to analyse which groups of customers have a high probability of responding.

Suppose we send a mailing to 300 potential customers, and only record the age of the customer, and whether or not he or she responded. The results are given in table 6.1.

We have divided age into a number of categories. We estimate the distribution of age within the group of respondents simply by calculating the relative frequency of each age category within the group. So for example:

$$p(\text{age=36-50}|\text{respondent}) = \frac{20}{100} = 0.2$$

age	respondents	p(age)	non-respondents	p(age)
18-25	25	0.25	10	0.05
26-35	35	0.35	20	0.1
36-50	20	0.20	30	0.15
51-64	15	0.15	80	0.4
65+	5	0.05	60	0.3
Total	100	1	200	1

Table 6.1: Distribution of age within the two groups

We do the same for the non-respondents. Estimation of the group prior probabilities is also straightforward. There are 300 mailings in total, and 100 respondents, so the prior probability of the respondent group is 1/3, and for the non-respondent group 2/3. Now let's see how we would use Bayes rule to calculate the probability that someone in age category 18-25 belongs to the group of respondents.

$$\begin{aligned} P(\text{respondent}|\text{age=18-25}) &= \frac{P(\text{age=18-25}|\text{respondent})P(\text{respondent})}{P(\text{age=18-25})} \\ &= \frac{1/4 \times 1/3}{1/4 \times 1/3 + 1/20 \times 2/3} = 5/7 \end{aligned}$$

When asked for an outright assignment to one of the two groups, we would assign this person to the group of respondents because this is the group with the highest probability at age=18-25.

In order to estimate the probability distributions of age within each group, we simply constructed 5 age categories and estimated the probability of each category by computing its relative frequency. This means we have to estimate 4 probabilities per group (the fifth is determined by the other four since they must add to one). This approach is not easily extended to the case where we have many input variables. For example, if we have a second variable income, also divided into 5 categories, then we have to estimate the probability of each age-income combination which means estimating $5^2 - 1 = 24$ probabilities per group. In general, if we have p variables with k possible values each, we would have to estimate $k^p - 1$ probabilities per group. With $p = 10$ and $k = 5$ this means estimating $5^{10} - 1 = 9765624$ probabilities. We would have to have an enormous amount of data to do this reliably. For example, if we have a 1000 observations, almost all cells are empty, i.e. we have 0 observations

for almost all possible value combinations of the 10 variables. This problem is sometimes called the *curse of dimensionality*: in high dimensions almost all of the input space is empty.

One way to tackle this problem is to introduce additional assumptions that allow us to reduce the number of parameters to be estimated. A rather drastic approach is the so called naive Bayes assumption: assume that the x variables are independent within each group, i.e.

$$P(\mathbf{x} \mid y) = P(x_1 \mid y)P(x_2 \mid y) \ldots P(x_p \mid y)$$

This means that instead of $k^p - 1$ parameters, we only have to estimate $kp - 1$ parameters per group. So with $p = 10$ and $k = 5$, we only have to estimate 49 probabilities per group. Although the naive Bayes assumption is almost always demonstrably false, it may perform quite well as a classifier. This is because the performance of a classifier is usually evaluated by looking at the fraction of cases it assigns to the wrong class. So as long as the classifier's estimate of $P(y \mid \mathbf{x})$ is at the right side of 0.5 (for a two-class problem) its bias goes unpunished. In terms of the bias-variance decomposition of prediction error: its bias may be harmless whereas the variance component of prediction error will be relatively low for naive Bayes.

Another way to reduce the number of parameters to be estimated for each group (which may also be motivated by theoretical considerations) is to assume that the input variables can be modeled by a multivariate normal distribution. We explore the consequences of this assumption in the next section.

6.5 Density estimation: normal distribution

If within each group the variables that make up the input vector $\mathbf{x}$ have a multivariate normal distribution, then the form of $p_i(\mathbf{x})$ is known, that is

$$p_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp[-\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_i^{-1}(\mathbf{x} - \mu_i)] \tag{6.7}$$

In this case, estimating $p_i(\mathbf{x})$ comes down to estimating two parameters for each group, the group mean vector μ_i, and the group covariance matrix Σ_i. If there are p variables in $\mathbf{x}$, then there are p means in the mean vector and $p(p+1)/2$ elements in the covariance matrix, making a total of $(p^2+3p)/2$ parameters to be estimated for each group.

6.5.1 The multivariate normal distribution

We start with the bivariate normal distribution, i.e. suppose

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \mathbf{x} \sim N(\mu, \Sigma) \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

The bivariate normal distribution has two parameters, denoted by μ and Σ. The vector of means μ specifies the means of x_1 and x_2; it determines the location of the distribution.

Warning: In this section μ_1 denotes the mean or expected value of x_1, *not* the mean vector of group 1!

The covariance matrix Σ contains the variance of x_1 and x_2 on the main diagonal, and the covariance between x_1 and x_2 in the off-diagonal entries. Notice that $\sigma_{12} = \sigma_{21}$, i.e. the covariance matrix is symmetric. The covariance matrix determines the shape and orientation of the distribution. Instead of the covariance, it is also common to report the correlation coefficient

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

The advantage of the correlation coefficient is that it is a dimensionless number between -1 and 1, where $\rho_{12} = 1$ means there is a perfect positive linear relation between x_1 and x_2, and $\rho_{12} = 0$ means that x_1 and x_2 are not linearly related at all. To get a feeling for the influence of the parameters on the location and shape of the distribution, it is insightfull to look at so called contour plots. In such a plot, we connect the points of equal probability density. Figure 6.1 contains a contour plot of a bivariate normal density with $\mu_1 = \mu_2 = 0$, $\rho = 0$ and $\sigma_1^2 = \sigma_2^2 = 1$. Because there is no correlation and equal variance in both directions, the contours have the shape of a circle.

In figure 6.2 we see a contour plot of a distribution with $\mu_1 = 10$, $\mu_2 = 25$, $\sigma_1^2 = \sigma_2^2 = 1$ and $\rho_{12} = 0.7$. The contours now have the shape of an ellipse. Because $\rho_{12} > 0$, the principle axis of the ellipse has a positive slope. In figure 6.3 we see a contour plot of a distribution with $\mu_1 = 15$, $\mu_2 = 5$, $\sigma_1^2 = \sigma_2^2 = 1$ and $\rho_{12} = -0.6$. Because $\rho_{12} < 0$, the principle axis of the ellipse has a negative slope.

In general, if we have p variables, i.e. $\mathbf{x} = [x_1, \ldots, x_p]^T$ that follow a

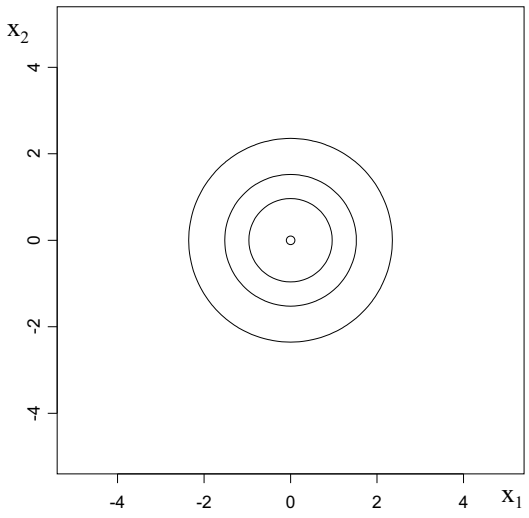

Figure 6.1: Contour plot with $\mu = [0\ \ 0]^T$, $\rho_{12} = 0$,$\sigma_1^2 = \sigma_2^2 = 1$

multivariate normal distribution, then the relevant parameters are

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \dots & \sigma_{2p} \\ \vdots & & & & \\ \sigma_{p1} & \sigma_{p2} & \sigma_{p3} & \dots & \sigma_p^2 \end{bmatrix}$$

6.5.2 Allocation rule for normal densities

When the outright assignment of an object $\mathbf{x}$ to one of the classes/groups is required, then the rule that gives the smallest overall error is to assign to group i if $P(y = i|\mathbf{x})$ is larger than $P(y = j|\mathbf{x})$ for all $j \neq i$. That is, assign $\mathbf{x}$ to the group with the highest probability at $\mathbf{x}$. Via Bayes formula this leads to the rule to assign to group i if

$$p_i(\mathbf{x})\pi_i > p_j(\mathbf{x})\pi_j \qquad \text{for all } j \neq i$$

Notice that we ignore the denominator of equation 6.6, since it is equal for all groups. It merely acts as a normalising constant.

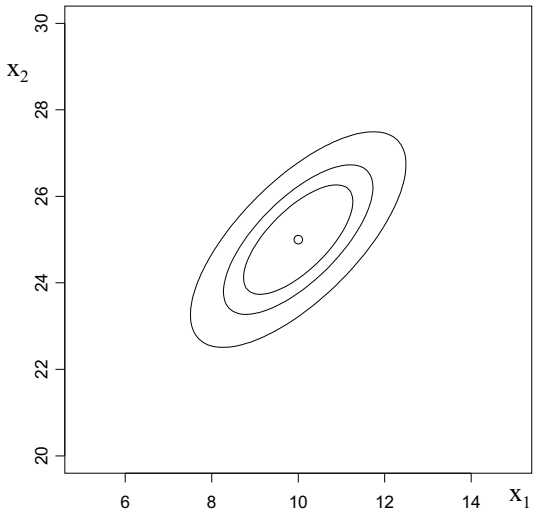

Figure 6.2: Contour plot with $\mu = [10\ \ 25]^T$, $\rho_{12} = 0.7$,$\sigma_1^2 = \sigma_2^2 = 1$

Application of the normal distribution leads to the following assignment rule. Assign to group i if

$$\frac{\pi_i}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp[-\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i)] > \\ \frac{\pi_j}{(2\pi)^{p/2}|\Sigma_j|^{1/2}} \exp[-\frac{1}{2}(\mathbf{x}-\mu_j)^T\Sigma_j^{-1}(\mathbf{x}-\mu_j)] \qquad \text{for all} j \neq i$$

Taking the natural logarithm of both sides of the inequality preserves the order, since all quantities are positive, and ln is strictly increasing on $(0, \infty)$. This gives the rule: assign to group i if

$$-1/2p\ln(2\pi) - 1/2\ln(|\Sigma_i|) - 1/2(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i) + \ln(\pi_i) > \\ -1/2p\ln(2\pi) - 1/2\ln(|\Sigma_j|) - 1/2(\mathbf{x}-\mu_j)^T\Sigma_j^{-1}(\mathbf{x}-\mu_j) + \ln(\pi_j) \qquad \text{for all } j \neq i$$

Cancelling all the terms that are common to both sides gives:

$$-\ln(|\Sigma_i|) - (\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i) + 2\ln(\pi_i) > \\ -\ln(|\Sigma_j|) - (\mathbf{x}-\mu_j)^T\Sigma_j^{-1}(\mathbf{x}-\mu_j) + 2\ln(\pi_j) \qquad \text{for all } j \neq i$$

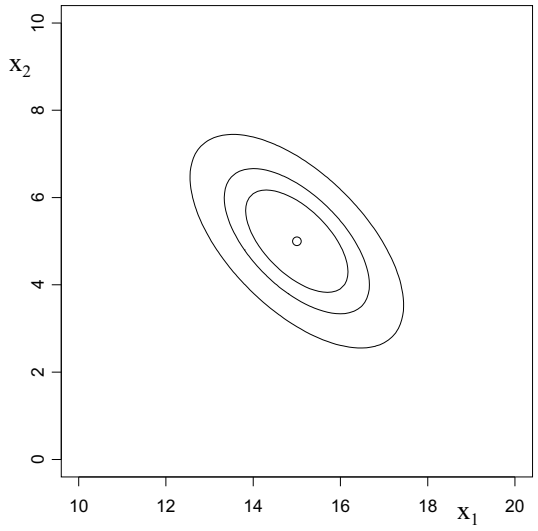

Figure 6.3: Contour plot with $\mu = [15 \;\; 5]^T$, $\rho_{12} = -0.6$,$\sigma_1^2 = \sigma_2^2 = 1$

Multiplication of both sides by -1 and reversal of the inequality gives:

$$\begin{aligned} &\ln(|\Sigma_i|) + (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) - 2\ln(\pi_i) < \\ &\ln(|\Sigma_j|) + (\mathbf{x} - \mu_j)^T \Sigma_j^{-1} (\mathbf{x} - \mu_j) - 2\ln(\pi_j) \qquad \text{for all } j \neq i \end{aligned}$$

The quantity

$$\ln(|\Sigma_i|) + (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) - 2\ln(\pi_i)$$

is often referred to as the *discriminant score* of $\mathbf{x}$ for group i, and

$$d_i(\mathbf{x}) = \ln(|\Sigma_i|) + (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i)$$

is called a *discriminant function*.

Summarizing, normal class distributions lead to the allocation rule: assign to group i if

$$d_i(\mathbf{x}) - 2\ln(\pi_i) < d_j(\mathbf{x}) - 2\ln(\pi_j) \qquad \text{for all } j \neq i$$

It is interesting to consider how this allocation rule divides the feature space. Considering a case with only two groups and two variables simplifies

the situation and loses nothing of the principles involved. The region of the feature space that belongs to group 1 is characterized by the values x_1 and x_2 such that

$$d_1(x_1, x_2) - 2\ln(\pi_1) < d_2(x_1, x_2) - 2\ln(\pi_2)$$

The dividing line between the region "belonging" to group 1 and the region "belonging" to group 2 is given by

$$d_1(x_1, x_2) - 2\ln(\pi_1) = d_2(x_1, x_2) - 2\ln(\pi_2)$$

It can be shown by matrix algebra or geometry that this dividing line has the form of a quadratic curve. This fact can also be seen from an inspection of a picture of the equal probability contours for the two groups (see figure 6.4). If the prior probabilities π_1 and π_2 are assumed to be equal, then the classification rule is equivalent to

$$p_1(\mathbf{x}) > p_2(\mathbf{x})$$

and the dividing line between the two regions is given by

$$p_1(\mathbf{x}) = p_2(\mathbf{x})$$

In other words, the dividing line between the two groups passes through the intersection of the equal probability contours of the two groups.

6.5.3 Equal Covariances

In many cases the correlations between the variables are the same within each group, and this property can be used to simplify the classifier yet further. When all groups have the same covariance matrix Σ, terms involving this constant can be cancelled from both sides of the inequality. Recall that we defined the *discriminant function*:

$$d_i(\mathbf{x}) = \ln(|\Sigma_i|) + (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i)$$

Since by assumption $\Sigma_i = \Sigma$, we can write

$$d_i(\mathbf{x}) = \ln(|\Sigma|) + (\mathbf{x} - \mu_i)^T \Sigma^{-1} (\mathbf{x} - \mu_i)$$

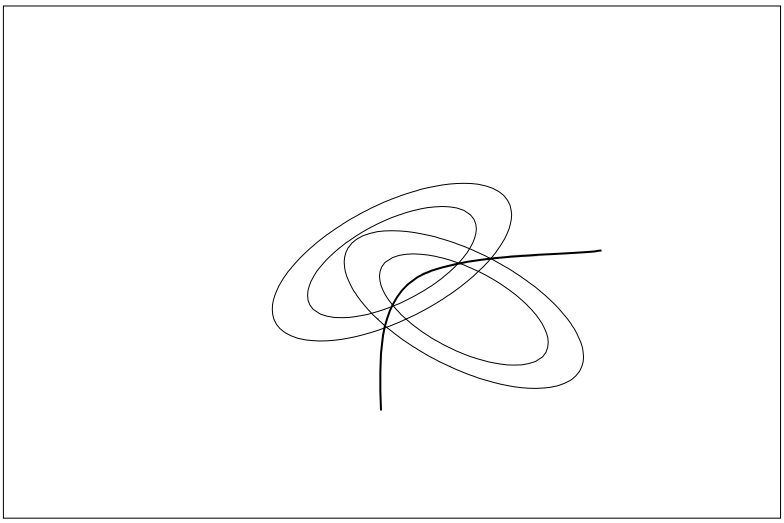

Figure 6.4: Intersection points of the contours are connected by a quadratic curve if the group covariance matrices are different

Since $\ln(|\Sigma|)$ is the same for each group, it can be dropped. Then we have

$$\begin{aligned} d_i(\mathbf{x}) &= (\mathbf{x}-\mu_i)^T\Sigma^{-1}(\mathbf{x}-\mu_i) \\ &= \mathbf{x}^T\Sigma^{-1}\mathbf{x} - 2\mu_i^T\Sigma^{-1}\mathbf{x} + \mu_i^T\Sigma^{-1}\mu_i \end{aligned}$$

Since the quadratic term $\mathbf{x}^T\Sigma^{-1}\mathbf{x}$ is now the same for each group, it can be dropped as well and we get (after division by 2):

$$d_i(\mathbf{x}) = -\mu_i^T\Sigma^{-1}\mathbf{x} + \frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i$$

Due to cancelling constants, the classification function has become negative again, and so for convenience it is usual to multiply both sides of the inequality by -1 and define a new function $f_i(\mathbf{x}) = -d_i(\mathbf{x})$. We then get the rule to assign to group i if

$$f_i(\mathbf{x}) + \ln(\pi_i) > f_j(\mathbf{x}) + \ln(\pi_j) \qquad \text{for all } j \neq i$$

where

$$f_i(\mathbf{x}) = \mu_i^T\Sigma^{-1}\mathbf{x} - \frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i$$

The second term in $f_i(\mathbf{x})$ doesn't involve $\mathbf{x}$ and so it can be written as a constant c_{0i} for each group. The matrix multiplication in the first term can be worked out in advance to give a single vector $\mathbf{c}_i$, that is

$$\mathbf{c}_i^T = \mu_i^T \Sigma^{-1}$$

and

$$f_i(\mathbf{x}) = \mathbf{c}_i^T \mathbf{x} + c_{0i}$$

Or, in summation notation

$$f_i(\mathbf{x}) = \sum_{k=1}^{p} c_{ki} x_k + c_{0i}$$

For example, in the case of two variables ($p = 2$), $f_i(\mathbf{x})$ becomes

$$f_i(\mathbf{x}) = c_{1i} x_1 + c_{2i} x_2 + c_{0i}$$

This looks pretty much like the all-familiar linear regression equation, except that there is a different $f_i(\mathbf{x})$ for each group, and this means that there are $p + 1$ parameters to be estimated for each group.

Once again, it is worth asking how the input space is divided by the classification rule. Considering the two group/two variable case for simplicity it is clear that the dividing line between the two areas of the input space (one "belonging" to group 1, and the other to group 2) is given by the values of x_1, x_2 satisfying (assuming $\pi_1 = \pi_2$):

$$f_1(x_1, x_2) = f_2(x_1, x_2)$$

Clearly, this is a straight line. This also fits in with what would be expected from an examination of the way the equal probability lines intersect (see figure 6.5). The linear form of $f_i(\mathbf{x})$ and the "straight line" division of the input space has resulted in $f_i(\mathbf{x})$ being known as a linear discriminant function.

The most often encountered classification problem involves assignment to one of two groups. In the two group case with equal covariance we get the following assigment rule. Assign to group 1 if

$$f_1(\mathbf{x}) + \ln(\pi_1) > f_2(\mathbf{x}) + \ln(\pi_2)$$

and otherwise assign to group 2. Taking $f_2(\mathbf{x})$ and $\ln(\pi_1)$ from both sides gives the rule: assign to group 1 if

$$f_1(\mathbf{x}) - f_2(\mathbf{x}) > \ln(\pi_2) - \ln(\pi_1)$$

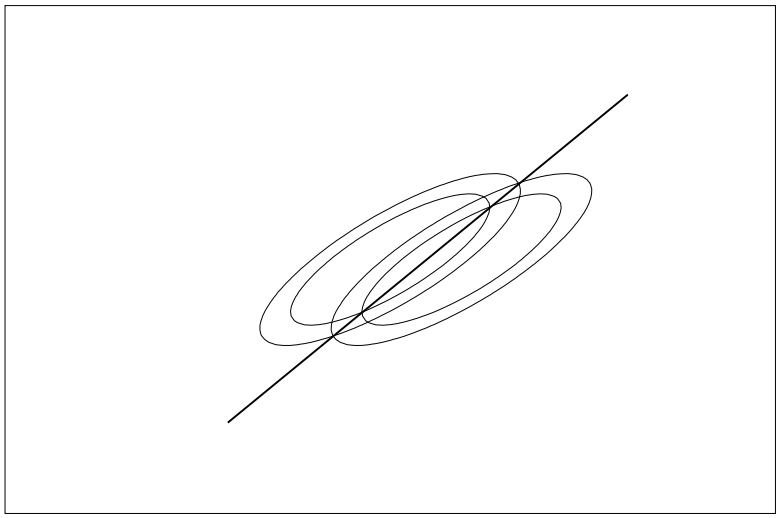

Figure 6.5: Intersection points of the contours are connected by a straight line when the group covariance matrices are equal

and to group 2 otherwise. The difference between the two linear discriminant functions can be written as a single new function:

$$w(\mathbf{x}) = f_1(\mathbf{x}) - f_2(\mathbf{x})$$

In other words, in the two group case only one function is required for classification. Values greater than $\ln(\pi_2/\pi_1)$ implying assignment to group 1 and values less than this implying assignment to group 2.

6.6 Plug-in estimates for normal densities

Estimation of the discriminant functions is pretty straightforward. We simply estimate the μ_i, Σ_i and π_i from the data and plug these estimates into the discriminant functions. We consider the heteroscedastic case (i.e. groups have different covariance matrices) and the homoscedastic case (covariance matrix the same for all groups).

6.6.1 Heteroscedastic Normal Model: Quadratic Discriminant Analysis

For ease of notation, we define

$$z_{ij} = \begin{cases} 1 & \text{if observation } j \text{ belongs to group } i \\ 0 & \text{otherwise} \end{cases}$$

The maximum likelihood estimates of μ_i and Σ_i are given by the sample mean $\bar{\mathbf{x}}_i$ and the sample covariance matrix $\hat{\Sigma}_i$, respectively, where

$$\bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n} z_{ij} \mathbf{x}_j$$

and

$$\hat{\Sigma}_i = \frac{1}{n_i} \sum_{j=1}^{n} z_{ij} (\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)^T$$

for $i = 1, \ldots, g$. Furthermore

$$n_i = \sum_{j=1}^{n} z_{ij}$$

denotes the number of observations from group i in the training data. The usual practice is to estimate Σ_i by the *unbiased* estimator

$$\mathbf{S}_i = \frac{n_i}{n_i - 1} \hat{\Sigma}_i \qquad\qquad (i = 1, \ldots, g)$$

6.6.2 Homoscedastic Normal Model: Linear Discriminant Analysis

The maximum likelihood estimate $\hat{\Sigma}$ of the common group-covariance matrix Σ is the pooled (within-group) sample covariance matrix

$$\begin{aligned} \hat{\Sigma} &= \sum_{i=1}^{g} (n_i/n) \hat{\Sigma}_i \\ &= 1/n \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} (\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)^T \end{aligned}$$

Again, it is customary to use the unbiased estimator

$$\mathbf{S} = \frac{n}{n-g}\hat{\Sigma}$$

instead of $\hat{\Sigma}$.

6.7 Linear Discriminant analysis vs. Logistic Regression

It is interesting to compare the first-order logistic regression model with the linear discriminant model, since both result in a straight line division of the input space into areas belonging to the different groups.

For ease of comparison we assume there are only two groups, labeled 0 and 1. We have seen that for logistic regression

$$\ln\left(\frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})}\right) = \beta^T\mathbf{x}$$

This leads to the allocation rule that we assign to group 1 if $\beta^T\mathbf{x} > 0$, and to group 0 otherwise.

The homoscedastic normal model also gives a linear boundary between the groups, i.e.

$$\ln\left(\frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})}\right) = \ln(p_1(\mathbf{x})\pi_1) - \ln(p_0(\mathbf{x})\pi_0) = \alpha^T\mathbf{x}$$

This follows easily from our derivation of the linear discriminant function.

Does this mean that logistic regression and linear discriminant analysis give exactly the same solution? No it does not, but usually they are close.

In logistic regression we assume that

$$\ln\left(\frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})}\right) = \beta^T\mathbf{x}$$

which is *exactly* true when

a) $\mathbf{x}$ is normally distributed in all groups, and the groups have the same covariance matrix.

b) $\mathbf{x}$ consists of binary variables that are independent within each group.

c) some other cases as well.

Linear discriminant analysis (LDA) assumes that

$$\mathbf{x} \sim N(\mu_i, \Sigma) \qquad \text{for } i = 0, 1$$

from which it follows that

$$\ln\left(\frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})}\right) = \alpha^T \mathbf{x}$$

where α is some function of $\mu_0, \mu_1, \Sigma, \pi_0, \pi_1$. This assumption is more *specific* than the assumption of logistic regression (LR). When the LDA assumption is exactly true, then it will do better than LR in the sense that it has a smaller asymptotic variance (it is more efficient). On the other hand, when the LR assumption is true but the LDA assumption is not (e.g. the case under b) above) then LR is consistent and LDA is not. However, these are all highly theoretical comparisons, and as it turns out, LDA has proven to be quite *robust* against violations of the normality assumption. This means that when the group distributions are not exactly normal, or even no where near normal, linear discriminant analysis may still give reasonable results. This is especially true if we are only interested in the *allocation* of $\mathbf{x}$ to a group, and not in the group probabilities at $\mathbf{x}$. For example, in the two goup case with $y \in \{0, 1\}$, we assign to group 1 if

$$\hat{P}(y=1|\mathbf{x}) > 0.5$$

and to group 0 otherwise. If $P(y=1|\mathbf{x}) = 0.8$ in reality, then any estimate $\hat{P}(y=1|\mathbf{x}) > 0.5$ will give the proper allocation.

6.8 Exercises

1. Suppose random variable x has a normal distribution with variance 4. If x is from group 1, its mean is 10; if it is from group 2 its mean is 14. Assume equal group prior probabilities, i.e. $\pi_1 = \pi_2$. We decide that we shall allocate (classify) x to group 1 if $x \leq c$ and to group 2 if $x > c$, for some c to be determined. Let A_1 denote the event that x is from

group 1, and A_2 that x is from group 2. Likewise, let B_1 be the event that x is classified to group 1, and B_2 that x is classified to group 2. Make a table showing the following: $P(B_1|A_2)$, $P(B_2|A_1)$, $P(A_1, B_2)$, $P(A_2, B_1)$ and P(misclassification).

c	$P(B_1\|A_2)$	$P(B_2\|A_1)$	$P(A_1, B_2)$	$P(A_2, B_1)$	P(misclassification)
10					
⋮					
14					

2. Suppose we have the following training sample:

$$X_1 = \begin{bmatrix} 2 & 12 \\ 4 & 10 \\ 3 & 8 \end{bmatrix} \qquad X_2 = \begin{bmatrix} 5 & 7 \\ 3 & 9 \\ 4 & 5 \end{bmatrix}$$

where X_1 contains three observations of $\mathbf{x} = [x_1 \quad x_2]^T$ for group 1, and X_2 contains three observations of $\mathbf{x} = [x_1 \quad x_2]^T$ for group 2. For example: the first observation from group 1 has values $x_1 = 2$ and $x_2 = 12$. We assume the covariance matrix is the same in group 1 and 2.

   a) Estimate the group means, covariance matrix, and group prior probabilities from this training sample.

   b) Estimate the linear discriminant functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ for groups 1 and 2 respectively.

   c) Give one linear classification function for this problem and construct a *confusion matrix* for this classification function by applying it to the training sample. What is the *in-sample* or *apparent* error rate of the classification function?

   d) Draw the border between the areas that (according to the classification function computed under c) belong to group 1 and 2 respectively, in a scatterplot of the data. Can you find a straight line that has a lower apparent error rate ?

   e) Based on the training sample, do you think the assumption of equal covariance matrices for both groups is justified ?

3. Suppose we have the following training sample:

$$X_1 = \begin{bmatrix} -2 & 5 \\ 0 & 3 \\ -1 & 1 \end{bmatrix} \qquad X_2 = \begin{bmatrix} 0 & 6 \\ 2 & 4 \\ 1 & 2 \end{bmatrix} \qquad X_3 = \begin{bmatrix} 1 & -2 \\ 0 & 0 \\ -1 & -4 \end{bmatrix}$$

where X_1 contains three observations of $\mathbf{x} = [x_1 \;\; x_2]^T$ for group 1, X_2 contains three observations of $\mathbf{x}$ for group 2, and X_3 contains three observations of $\mathbf{x}$ for group 3. For example: the first observation from group 1 has values $x_1 = -2$ en $x_2 = 5$. We assume the covariance matrix is the same in all three groups.

a) Estimate the group means, covariance matrix, and group prior probabilities from this training sample.

b) Estimate the linear discriminant functions $f_1(\mathbf{x})$, $f_2(\mathbf{x})$ and $f_3(\mathbf{x})$ for the three groups.

c) Classify the new observation $\mathbf{x}_0 = [-2 \quad -1]^T$ using the result obtained under b).

d) Use the training sample to compute $\mathbf{S}_1$ (the unbiased estimator of the covariance matrix of group 1) and $\mathbf{S}_2$. Does it seem the assumptions of linear discriminant analysis are met ? Explain.

Chapter 7

Resampling

7.1 Introduction

Resampling techniques are computationally expensive techniques that reuse the available sample to make statistical inferences. Because of their computational requirements these techniques were infeasible at the time that most of "classical" statistics was developed. With the availability of ever faster and cheaper computers, their popularity has grown very fast in the last decade. In this section we provide a brief introduction to some important resampling techniques.

7.2 Cross-Validation

Cross-Validation is a resampling technique that is often used for model selection and estimation of the prediction error of a classification- or regression function. We have seen already that squared error is a natural measure of prediction error for regression functions:

$$\mathrm{PE} = \mathrm{E}(y - \hat{f})^2$$

Estimating prediction error on the same data used for model estimation tends to give downward biased estimates, because the parameter estimates are "fine-tuned" to the peculiarities of the sample. For very flexible methods, e.g. neural networks or tree-based models, the error on the training sample can usually be made close to zero. The true error of such a model will usually be much higher however: the model has been "overfitted" to the training

sample. One way of dealing with this problem is to include a penalty term for model complexity (e.g. AIC, BIC) as we have seen in section 4.10.

An alternative is to divide the available data into a training sample and a test sample, and to estimate the prediction error on the test sample. If the available sample is rather small, this method is not preferred because the test sample may not be used for model estimation in this scenario. Cross-validation accomplishes that all data points are used for training as well as testing. The general K-fold cross-validation procedure works as follows

1. Split the data into K roughly equal-sized parts.
2. For the kth part, estimate the model on the other $K-1$ parts, and calculate its prediction error on the kth part of the data.
3. Do the above for $k = 1, 2, \ldots, K$ and combine the K estimates of prediction error.

If $K = n$, we have the so-called *leave-one-out* cross-validation: one observation is left out at a time, and $\hat{f}$ is computed on the remaining $n-1$ observations.

Now let $k(i)$ be the part containing observation i. Denote by $\hat{f}_i^{-k(i)}$ the value predicted for observation i by the model estimated from the data with the $k(i)$th part removed. The cross-validation estimate of mean squared error is now

$$\widehat{\mathrm{PE}}_{cv} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_i^{-k(i)})^2$$

We consider a simple application of model selection using cross-validation, involving the linear, quadratic and cubic model introduced in section 4.10.1. In a simulation study we draw 50 (x, y) observations from the probability distributions

$$X \sim U(0, 10) \quad \text{and} \quad Y \sim \mathcal{N}(\mu = 2 + 3x + 1.5x^2, \sigma_\varepsilon = 5),$$

i.e. $\mathrm{E}(Y)$ is a quadratic function of x. For the purposes of this example, we pretend we don't know the true relation between x and y, as would usually be the case in a practical data analysis setting. We consider a linear, quadratic and cubic model as the possible candidates to be selected as the model with lowest prediction error, and we use leave-one-out cross validation to compare the three candidates.

	in-sample	leave-one-out
linear	150.72	167.63
quadratic	16.98	19.89
cubic	16.66	20.66

Table 7.1: Mean square error of candidate models: in-sample and leave-one-out

The first column of Table 7.1 contains the "in-sample" estimate of the mean square error of all three models. Based on the in-sample comparison one would select the cubic model as the best model since it has the lowest prediction error. We already noted however that this estimate tends to be too optimistic, and the more flexible the model the more severe the optimism tends to be. In the second column the cross-validation estimates of prediction error are listed. As one would expect they are higher than their in-sample counterparts. Furthermore, we see that the quadratic model (the true model) has the lowest cross-validation prediction error of the three. The lower in-sample prediction error of the cubic was apparently due to a modest amount of overfitting.

7.3 Bootstrapping

In chapter 2 we saw that in some special cases we can derive mathematically the exact distribution of a sample statistic, and in some other cases we can rely on limiting distributions as an approximation to the sampling distribution for a finite sample. For many statistics that may be of interest to the analyst, such exact or limiting distributions cannot be derived analytically. In yet other cases the asymptotic approximation may not provide a good fit for a finite sample. In such cases an alternative approximation to the sampling distribution of a statistic $t(\mathbf{x})$ may be obtained using just the data at hand, by a technique called *bootstrapping* [4, 2]. To explain the basic idea of the *non-parametric* bootstrap, we first introduce the *empirical distribution function*

$$\hat{F}(z) = \frac{1}{n}\sum_{i=1}^{n} I\left(x_i \leq z\right) \quad -\infty < z < \infty$$

where I denotes the indicator function and $\mathbf{x} = (x_1, x_2, ..., x_n)$ is a random sample from population distribution function F. We now approximate the sampling distribution of $t(\mathbf{x})$ by repeated sampling from $\hat{F}$. This is achieved by drawing samples $\mathbf{x}^{(r)}$ of size n by sampling independently *with replacement* from $(x_1, x_2, ..., x_n)$. If all observations are distinct, there are $\begin{pmatrix} 2n-1 \\ n \end{pmatrix}$ distinct samples in

$$\mathcal{B} = \{\mathbf{x}^{(r)}, r = 1, ..., \begin{pmatrix} 2n-1 \\ n \end{pmatrix}\}$$

with respective multinomial probabilities (see section 1.14)

$$P(\mathbf{x}^{(r)}) = \frac{m!}{j_1^{(r)}!\, j_2^{(r)}! \ldots j_n^{(r)}!}(\frac{1}{n})^n$$

where $j_i^{(r)}$ is the number of copies of x_i in $\mathbf{x}^{(r)}$. The bootstrap distribution of $t(\mathbf{x})$ is derived by calculating the realisation $t(\mathbf{x}^{(r)})$ for each of the resamples and assigning each one probability $P(\mathbf{x}^{(r)})$. As $n \to \infty$, the empirical distribution $\hat{F}$ converges to the underlying distribution F, so it is intuitively plausible that the bootstrap distribution is an asymptotically valid approximation to the sampling distribution of a statistic.

We can in principle compute all $\begin{pmatrix} 2n-1 \\ n \end{pmatrix}$ values of the statistic to obtain its "ideal" bootstrap distribution, but this is computationally infeasible even for moderate n. For $n = 15$ there are already 77558760 distinct samples. The usual alternative is to use Monte-Carlo simulation, by drawing a number B of samples and using them to approximate the bootstrap distribution.

If a *parametric* form is adopted for the underlying distribution, where θ denotes the vector of unknown parameters, then the parametric bootstrap uses an estimate $\hat{\theta}$ formed from $\mathbf{x}$ in place of θ. If we write F_θ to signify its dependence on θ, then bootstrap samples are generated from $\hat{F} = F_{\hat{\theta}}$.

The non-parametric bootstrap makes it unneccesary to make parametric assumptions about the form of the underlying distribution. The parametric bootstrap may still provide more accurate answers than those provided by limiting distributions, and makes inference possible when no exact or limiting distributions can be derived for a sample statistic.

We present an elementary example to illustrate the parametric and non-parametric bootstrap. The population parameter of interest is the correlation

coefficient, denoted by ρ. We first discuss this parameter before we show how to use bootstrapping to make inferences about it.

The linear dependence between population variables $\mathcal{X}$ and $\mathcal{Y}$ is measured by the covariance

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

A term $(x_i - \mu_x)(y_i - \mu_y)$ from this sum is positive if both factors are positive or both are negative, i.e. if x_i and y_i are both above or both below their mean. Such a term is negative if x_i and y_i are on opposite sides of their mean. The dimension of σ_{xy} is the product of the dimensions of X and Y; division by both σ_x and σ_y yields a dimensionless number called the correlation coefficient, i.e.

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Evidently ρ has the same sign as σ_{xy}, and always lies between -1 and $+1$. If $\rho = 0$ there is no linear dependence: the two variables are uncorrelated. The linear dependence increases as $|\rho|$ gets closer to 1. If all pairs (x, y) are on a straight line with positive slope, then $\rho = 1$; if all pairs are on a straight line with negative slope then $\rho = -1$.

To make inferences about ρ we use the sample correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The sampling distribution of this statistic can't be mathematically derived in general, in fact there is no general expression for the expected value of r_{xy}. Therefore it makes sense to use the bootstrap to make inferences concerning ρ.

In our study, we draw 30 (x, y) pairs from a standard binormal distribution with $\rho = 0.7$, i.e.

$$(X, Y) \sim \mathcal{N}^2(\mu_x = 0, \mu_y = 0, \sigma_x^2 = 1, \sigma_y^2 = 1, \rho = 0.7)$$

Based on this dataset, bootstrapping proceeds as follows

Non-parametric: Draw samples of 30 (x, y) pairs (with replacement) from the data. For each bootstrap sample, compute r, to obtain an empirical sampling distribution.

Parametric: Make appropriate assumptions about the joint distribution of X and Y. In our study we assume

$$(X, Y) \sim \mathcal{N}^2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$$

which happens to be correct. In a practical data analysis situation we would evidently not know that, and it would usually be hard to ascertain that our assumptions are appropriate. We build an empirical sampling distribution by drawing samples of size 30 from

$$\mathcal{N}^2(\bar{x}, \bar{y}, s_x^2, s_y^2, r)$$

In both cases we draw 1000 samples to generate the empirical sampling distribution of r. To construct $100(1-\alpha)\%$ confidence intervals for ρ, we simply take the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of this distribution.

In order to determine whether the bootstrap provides reliable confidence intervals with the right coverage, we repeated the following procedure 100 times

1. Draw a sample of size 30 from the population.
2. Build a bootstrap distribution for r, and construct 90% confidence intervals for ρ. (both parametric and non-parametric)
3. Determine whether the true value of ρ is inside the confidence interval.

Like any conventional method for constructing confidence intervals, the bootstrap will sometimes miss the true value of the population parameter. This happens when the data is not representative for the population. For example, in 1 of the 100 samples the sample correlation coefficient was 0.36. This is highly unlikely to occur when sampling from a population with $\rho = 0.7$ but it will occur occasionally. In such a case the bootstrap distribution of r is bound to be way off as well. In Fig. 7.1 the non-parametric bootstrap distribution for this particular sample is displayed. The 90% confidence interval computed from this distribution is $(0.064, 0.610)$. Not surprisingly it does not contain the true value of ρ.

On average, one would expect a 90% confidence interval to miss the true value in 10% of the cases; that's why it's called a 90% confidence interval. Furthermore the narrower the confidence intervals, the more informative they are. Both the parametric and non-parametric bootstrap missed the true value

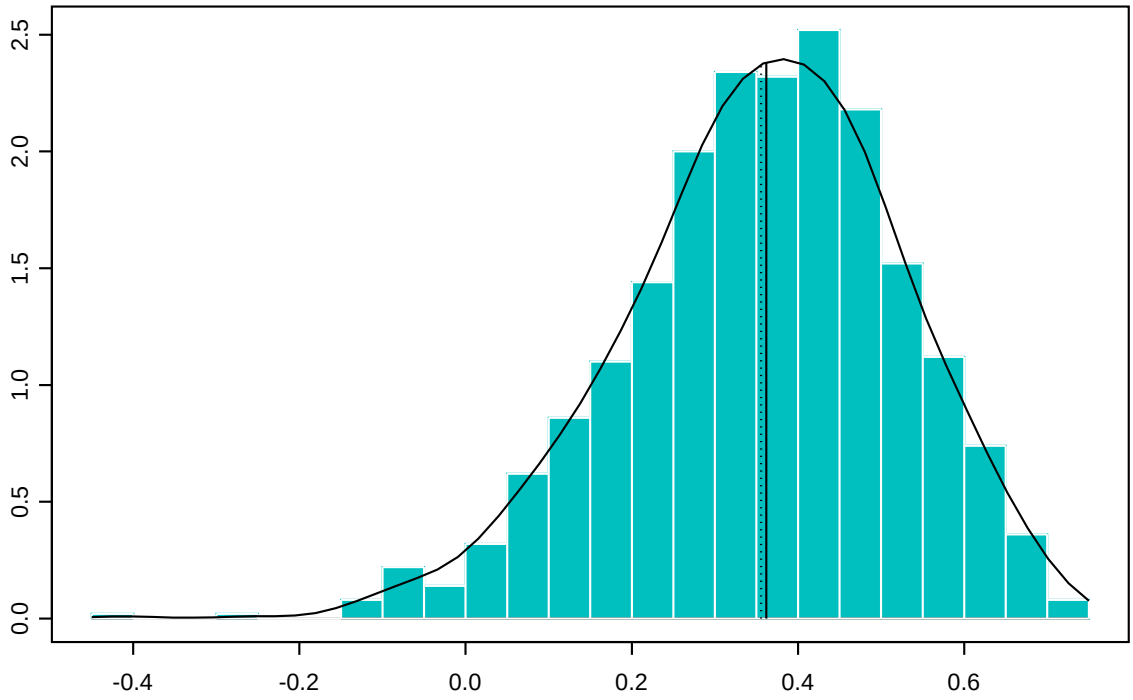

Figure 7.1: Bootstrap distribution for r. Observed value of r is 0.36.

of ρ 13 times out of 100, where one would expect 10 misses. Now we may test whether the bootstrap confidence intervals have the right coverage:

$$H_0 : \alpha = 0.1 \quad \text{against} \quad H_a : \alpha \neq 0.1$$

We observed 13 misses out of 100, so the observed value of our test statistic is $a = 0.13$. The distribution of $\hat{\alpha}$ under H_0 (the null-hypothesis) may be approximated by

$$\hat{\alpha} \approx_{H_0} \mathcal{N}(\mu = \alpha, \sigma^2 = \alpha(1-\alpha)/n)$$

which yields $\hat{\alpha} \approx \mathcal{N}(0.1, 0.0009)$. We may now compute the p-value of the observed value under the null-hypothesis as follows

$$P_{H_0}(\hat{\alpha} > a) = P_{H_0}(\hat{\alpha} > 0.13) = P(Z > \frac{0.13 - 0.1}{\sqrt{0.0009}}) = P(Z > 1) = 0.1587$$

where the value 0.1587 was looked-up in a table for the standardnormal distribution. Since we are performing a two-sided test this probability should be doubled, so we obtain a p-value of $2 \times 0.1587 = 0.3174$. This means we would not reject the null-hypothesis under any conventional significance level. The probability under the null-hypothesis of obtaining a result at least as far from $\alpha_0 = 0.1$ (to either side) as the one we observed is "pretty high".

The mean length of the confidence intervals is 0.31 for the non-parametric bootstrap, and 0.32 for the parametric bootstrap. Even though the assumptions of the parametric bootstrap were correct it did not give shorter confidence intervals on average.

Chapter 8

Bayesian Statistics

In this section we briefly consider the principal idea of Bayesian inference. In [1, 5], Bayesian inference is discussed in greater detail.

Consider the following three claims

1. An English lady claims that she can taste whether the milk or tea has been poured first into the cup.

2. A music expert claims that he can distinguish a page from a Haydn score from a page of a Mozart score.

3. A drunk friend of yours claims that he can predict whether a fair coin will land heads or tails.

Suppose that we set up suitable experiments to test their claims, and in all three cases the persons make 9 correct predictions out of 10. What would be your opinion concerning the three claims after observing the outcome of the experiments? Probably you would still think your friend cannot predict the whether the coin lands heads or tails, but was simply extremely lucky in this particular experiment. The outcome of the experiment is not the only factor that influences your opinion: it also depends on how plausible you think the claim is a priori. This idea is at the heart of Bayesian statistical inference: our posterior beliefs are not just determined by the sample data, but also by what we believed before seeing the data.

How would this combination of prior belief and sample data into posterior belief work technically? Consider again the coin tossing experiment. We stated that the probability of heads, denoted by π, is a fixed yet unknown

	Prior $P(M_i)$	Likelihood $P(y = 5 \mid M_i)$	Posterior $P(M_i \mid y = 5)$
M_1: $\pi = 0.8$	0.7	0.027	0.239
M_2: $\pi = 0.4$	0.3	0.201	0.761

Table 8.1: Prior and posterior probabilities of M_1 and M_2

quantity. From a relative frequency viewpoint, it makes no sense to talk about the probability distribution of π since it is not a random variable. In Bayesian inference one departs from this strict interpretation of probability. We may express prior, yet incomplete, knowledge concerning the value of π through the construction of a *prior distribution*. This prior distribution is then combined with sample data (using Bayes rule, see section 1.13) to obtain a posterior distribution. The posterior distribution expresses the new state of knowledge, in light of the sample data. We reproduce Bayes' rule using symbols that are more indicative for the way it is used in Bayesian inference:

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{\sum_j P(D|M_j)P(M_j)}$$

where the M_i specify different models for the data, i.e. hypotheses concerning the parameter value(s) of the probability distribution from which the data were drawn. Note that in doing so, we actually assume that this probability distribution is known up to a fixed number of parameter values.

Example 22 *Consider the somewhat artificial situation where two hypotheses concerning the probability of heads of a particular coin are entertained, namely M_1: $\pi = 0.8$ and M_2: $\pi = 0.4$ (see table 8.1). Prior knowledge concerning these models is expressed through a prior distribution as specified in the first column of table 8.1. Next we observe 5 times heads in a sequence of 10 coin flips, i.e. $y = 5$. The likelihood of this outcome under the different models is specified in the second column of table 8.1 (the reader can also find them in table 3.1). The posterior distribution is obtained via Bayes' rule, and is specified in the last column of table 8.1. Since the data are more likely to occur under M_2, the posterior distribution has clearly shifted towards this model.*

In general, the probability distribution of interest is indexed by a number of continuous valued parameters, which we denote by parameter vector θ.

Replacing probabilities by probability densities and summation by integration, we obtain the probability density version of Bayes' rule

$$f(\theta \,|\, \mathbf{y}) = \frac{f(\mathbf{y} \,|\, \theta)\ f(\theta)}{\int_\Omega f(\mathbf{y} \,|\, \theta)\ f(\theta)\ d\theta}$$

where $\mathbf{y}$ denotes the observed data and Ω denotes the parameter space, i.e. the space of possible values of θ.

Consider the case where we have no prior knowledge whatsoever concerning the probability of heads π. How should this be reflected in the prior distribution? One way of reasoning is to say that all values of π are considered equally likely, which can be expressed by a uniform distribution over $\Omega = [0, 1]$: the range of possible values of π. Let's consider the form of the posterior distribution in this special case.

$$f(\pi \,|\, \mathbf{y}) = \frac{f(\mathbf{y} \,|\, \pi) f(\pi)}{\int_0^1 f(\mathbf{y} \,|\, \pi) f(\pi) d\pi}$$

If we observe once again 7 times heads in a sequence of 10 coin flips, then $f(\mathbf{y} \,|\, \pi) = \pi^7(1-\pi)^3$. Since $f(\pi) = 1$, the denominator of the above fraction becomes

$$\int_0^1 \pi^7(1-\pi)^3 d\pi = \frac{1}{1320}$$

and so the posterior density becomes

$$f(\pi \,|\, \mathbf{y}) = 1320\, \pi^7(1-\pi)^3$$

It is reassuring to see that in case of prior ignorance the posterior distribution is proportional to the likelihood function of the observed sample. Note that the constant of proportionality merely acts to make the integral of the expression in the numerator equal to one, as we would expect of a probability density.

In general, the computationally most difficult part of obtaining the posterior distribution is the evaluation of the (multiple) integral in the denominator of the expression. For this reason, a particular class of priors, called *conjugate* priors, have received special attention in Bayesian statistics. Assume our prior knowledge concerning the value of π may be expressed by a Beta(4,6) distribution (see section 1.15), i.e.

$$f(\pi) = \frac{\pi^3(1-\pi)^5}{\int_0^1 \pi^3(1-\pi)^5 d\pi}$$

Since $\int_0^1 \pi^3(1-\pi)^5 d\pi = \frac{1}{504}$, we get $f(\pi) = 504\,\pi^3(1-\pi)^5$.

Multiplied with the likelihood this results in $504\,\pi^3(1-\pi)^5\pi^7(1-\pi)^3 = 504\pi^{10}(1-\pi)^8$, so the denominator becomes

$$\int_0^1 504\,\pi^{10}(1-\pi)^8 = \frac{28}{46189}$$

and the posterior density becomes

$$f(\pi \,|\, \mathbf{y}) = 831402\,\pi^{10}(1-\pi)^8$$

which is a Beta(11,9) distribution.

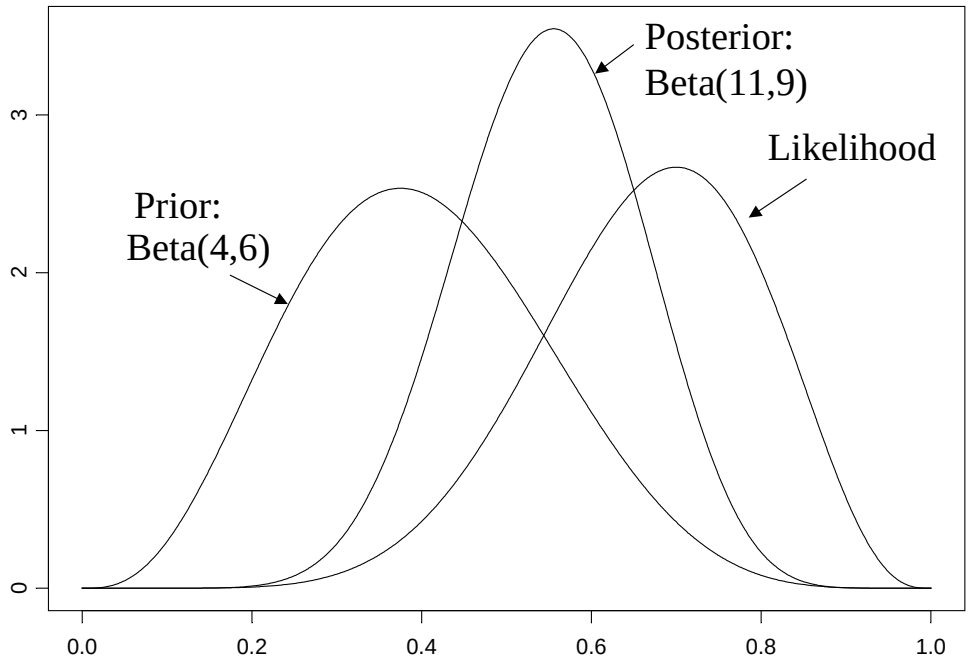

Figure 8.1: Prior, likelihood and posterior for coin tossing experiment.

In general, when we have a binomial sample of size n with r successes, and we combine that with a Beta(l, k) prior distribution, then the posterior distribution is Beta$(l + r, k + n - r)$. Loosely speaking, conjugate priors allow for simple rules to update the prior with sample data to arrive at the posterior distribution. Furthermore, the posterior distribution belongs to the

same family as the prior distribution. Since the uniform distribution over the interval $[0, 1]$ is the same as a Beta(1,1) distribution (see section 1.15), we could have used this simple update rule in the "prior ignorance" case as well: combining a Beta(1,1) prior with a binomial sample of size 10 with 7 successes yields a Beta(8,4) posterior distribution.

Once we have calculated the posterior distribution, we can extract all kinds of information from it. We may for example determine the mode of the posterior distribution which represents the value of π for which the posterior density is maximal. When asked to give a point estimate for π, it makes sense to report this value. When asked for a range of plausible values for π we may use the posterior distribution to determine a so-called $100(1-\alpha)\%$ probability interval, which is an interval $[g_l, g_u]$ such that $P(\pi < g_l) = \alpha/2$ and $P(\pi > g_u) = \alpha/2$ where the relevant probabilities are based on the posterior distribution for π.

The following example is taken from [1]. A study was designed to evaluate the effectiveness of a chemotherapeutic agent, called 6-mercaptopurine (6MP), for the treatment of acute leukemia. Patients were randomized into the therapy or placebo group by coin tosses. The first patient was assigned to the 6MP group if the coin landed heads and to the placebo group otherwise. The second patients was then assigned to the other group, and so on. For each pair of patiens the investigators recorded whether the 6MP patient or the placebo patient stayed in remission longer. There were 21 pairs of patients in the study, and 6MP was more effective on 18 of the 21 ($\approx 86\%$) pairs of patients. Let π denote the probability that a randomly selected patient will stay in remission longer if treated with 6MP than if not treated. We observe 18 out of 21 cases for which this is true, so the likelihood function is $\pi^{18}(1-\pi)^3$. Suppose there are two doctors, say A and B, with different prior probability distributions for π. Their prior distributions are displayed in figure 8.2 and figure 8.3 respectively.

Combination of these priors with the sample likelihood gives a Beta(19,4) posterior for A, and a Beta(20,4) posterior for B. Although the doctors had quite different opinions before the experiment, they almost agree after seeing the outcome of the experiment. For example, the posterior predictive probability that the next patient will benefit from treatment with 6MP is equal to the expected value of the posterior distribution, which is

$$\frac{l+r}{l+k+n} = \frac{1+18}{1+1+21} = \frac{19}{23} = 0.826$$

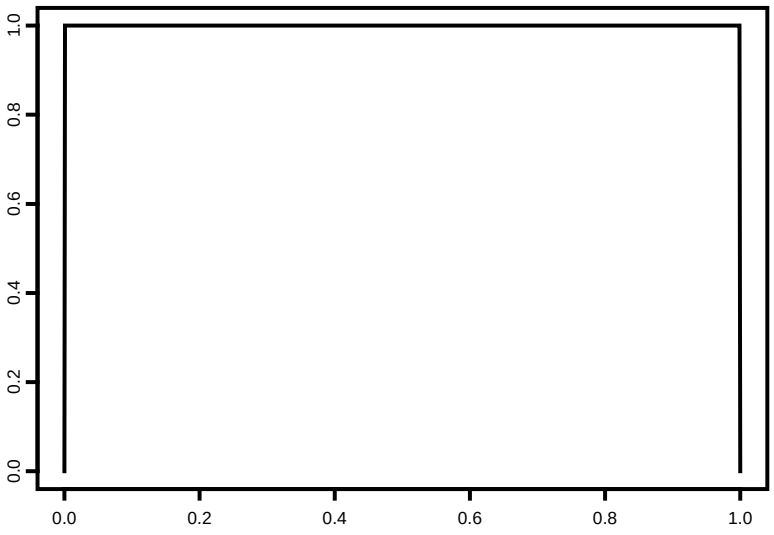

Figure 8.2: Prior distribution for π of dr. A: Beta(1,1).

for A, and 20/24=0.833 for B. Hence, after seeing the data they almost agree on this.

A 95% posterior probability interval for π is constructed as follows. The posterior distribution of doctor A is Beta(19,4). This is approximately equal to a normal distribution with mean

$$\pi^* = \frac{19}{23} = 0.826$$

and standard deviation

$$\sqrt{\frac{\pi^*(1-\pi^*)}{n^*+1}} = \sqrt{\frac{0.826(0.174)}{24}} \approx 0.0774$$

where $n^* = l + k + n$. So a 95% probability interval for π is given by

$$\pi^* \pm z_{0.05/2}\sqrt{\frac{\pi^*(1-\pi^*)}{n^*+1}} = 0.826 \pm 1.96 \times 0.0774 \approx [0.67, 0.98]$$

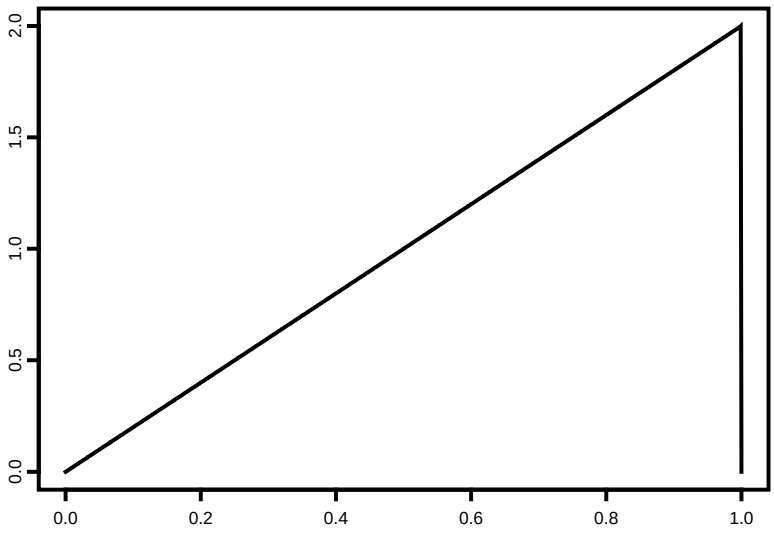

Figure 8.3: Prior distribution for π of dr. B: Beta(2,1).

Bibliography

[1] D.A. Berry. *Statistics: a Bayesian perspective*. Wadsworth, Belmont (CA), 1996.

[2] A.C. Davison and D.V. Hinkley. *Bootstrap methods and their application*. Cambridge University Press, Cambridge, 1997.

[3] A.W.F Edwards. *Likelihood*. The John Hopkins University Press, Baltimore, 1992.

[4] B. Efron and R.J Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

[5] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.

[6] W.H. Greene. *Econometric Analysis (second edition)*. Macmillan, New York, 1993.

[7] R.C. Hill, W.E. Griffiths, and G.G. Judge. *Undergraduate Econometrics (second edition)*. Wiley, New York, 2001.

[8] D.C. Montgomery, E.A. Peck, and G.G. Vining. *Introduction to linear regression analysis (third edition)*. Wiley, New York, 2001.

[9] D.S. Moore. Bayes for beginners: some reasons to hesitate. *The American Statistician*, 51(3):254–261, 1997.

[10] J.J.A. (in Dutch) Moors. *Statistiek in de economie (deel twee: steekproeftheorie en analyserende statistiek)*. Academic Service, Schoonhoven, 1991.

[11] J. Neter, M.H. Kutner, C.J. Nachtsheim, and W. Wasserman. *Applied linear statistical models (fourth edition)*. McGraw-Hill, Boston, 1996.

[12] R.M. Royall. *Statistical evidence: a likelihood paradigm*. Chapman & Hall, London, 1997.