# Statistical Models

by

M.C.M. de Gunst

11/1/2013

# Contents


**1 The Analysis of Variance** **1**

1.1 Introduction . . . . . 1

1.2 Least Squares Estimation . . . . . 4

1.2.1 The One-factor Model . . . . . 5

1.2.2 The Two-factor Model . . . . . 7

1.2.3 Additive Models . . . . . 9

1.3 Hypothesis Testing . . . . . 10

1.3.1 Some Probability Distributions . . . . . 11

1.3.2 The Residual Sum of Squares . . . . . 13

1.3.3 A Test for One-factor Models . . . . . 14

1.3.4 Tests for Two-factor Models . . . . . 16

**2 Nonlinear Regression** **23**

2.1 Introduction . . . . . 23

2.2 Parameter Estimation . . . . . 25

2.3 Accuracy of Estimators; Confidence Regions . . . . . 28

2.3.1 Classical Asymptotic Results . . . . . 28

2.3.2 Using the Bootstrap . . . . . 30

2.3.3 Approximate Confidence Regions for the Expected Response . . . . 31

2.4 Assessment of Fit and Model Choice . . . . . 32

2.4.1 The Estimated Parameter Values . . . . . 34

2.4.2 Plots . . . . . 34

2.4.3 Model Comparison . . . . . 36

**3 Generalized Linear Models** **41**

3.1 Introduction . . . . . 41

3.1.1 Basic Concepts and Examples . . . . . 41

3.1.2 A Uniform Framework . . . . . 46

3.2 Parameter Estimation . . . . . 47

3.3 Inference in GLMs . . . . . 49

**4 Time Series** **56**
4.1 Introduction . . . . . 56
4.2 Stationarity, Trend, Seasonality . . . . . 58
4.2.1 Estimation and Elimination of Trend and Seasonal Components . . 62
4.2.2 Estimation and Elimination of Trend in Absence of Seasonality . . 63
4.2.3 Estimation and Elimination of both Trend and Seasonality . . . . . 66
4.2.4 Filtering . . . . . 70
4.3 Models for Stationary Time Series . . . . . 70
4.3.1 Moving Average Processes . . . . . 72
4.3.2 Autoregressive Processes . . . . . 72
4.3.3 ARMA Models . . . . . 74
4.3.4 ARIMA Models . . . . . 74
4.4 Estimation of Mean and (Partial) Autocorrelation Function . . . . . 76
4.5 Choice of Model and Determination of its Order . . . . . 77
4.6 Estimation of the Model Parameters . . . . . 78
4.7 Diagnostic Checking . . . . . 80
4.8 Final Remarks . . . . . 83

**5 References** **84**

# Chapter 1

# The Analysis of Variance

## 1.1 Introduction

The classical *analysis of variance* is a statistical method for investigating the dependence of the observations, the *response variables*, on certain explanatory variables, which are in this context referred to as *effects* or *factors*. One can, for instance, investigate in what way the blood pressure depends on socio-economic class, or study the dependence of the yield of corn on the kind of corn and the kind of manure. The observations are measured with a random error as opposed to the effects, which are assumed to be nonrandom and known exactly. In the theory of the analysis of variance the relation between the observations and the effects is via a *linear model*.

A linear model, which will be denoted by $\Omega$, is a statistical model where random observations $Y_1, \ldots, Y_n$ are described by a linear combination of $p+1$ unknown parameters $\beta_0, \ldots, \beta_p$ plus unobservable random errors $e_1, \ldots, e_n$,

$$\Omega : \begin{cases} Y_i = \beta_0 + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + e_i, \\ \mathrm{E}\, e_i = 0, \\ \mathrm{Cov}\,(e_i, e_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases}, \end{cases} \tag{1.1}$$

for $i,\ j = 1, \ldots, n$, and where the $\{x_{ij}\}$ are known constant coefficients. Note that the errors are uncorrelated and hence the observations $Y_1, \ldots, Y_n$ are uncorrelated as well. Sometimes it is practical to use the matrix notation for $\Omega$,

$$\Omega : \begin{cases} Y = X\beta + e, \\ \mathrm{E}\, e = 0, \\ \mathrm{Cov}(e) = \sigma^2 I_{n\times n}, \end{cases} \tag{1.2}$$

with $Y = (Y_1, \ldots, Y_n)^T$ the vector of observations, $X$ the $n \times (p+1)$ matrix ($n > (p+1)$) with the $i$-th row $x_i^T = (1, x_{i1}, \ldots, x_{ip})$, $\beta = (\beta_0, \ldots, \beta_p)^T$ the vector of unknown parameters, $I_{n\times n}$ the $n \times n$ identity matrix, and $e = (e_1, \ldots, e_n)^T$ the vector of random errors. The first column of $X$ is called the 0-th column. So the $(k+1)$-th column is referred to as the $k$-th column. The matrix $X$ is known as the *design* or *incidence matrix*.

If $rank(X) = p + 1$, that is $X$ has full rank, we deal with regression analysis with $\beta$ the vector of regression parameters. If $rank(X) < p + 1$ and all entries of the matrix $X$ are 0 or 1, referring to the absence or presence of the effects, we can use the methods of the analysis of variance. A mixture of regression and variance analysis is called the *analysis of covariance* (Scheffé (1959)). In this chapter we discuss the basic principles of the analysis of variance.

**Example 1.1** Blood pressure:
In order to investigate the relationship between blood pressure and socio-economic class, the systolic blood pressure (in mm Hg) was measured for 36 individuals participating in a health study in Los Angeles. All individuals were classified according to their age and socio-economic class. The results are shown in Table 1.1.

**Table 1.1**

| | socio-economic class | | |
|---|---|---|---|
| age class | low | middle | high |
| 30-45 | 116 108 160 116 | 136 124 112 118 | 128 104 132 112 |
| 46-59 | 108 110 134 122 | 138 124 160 157 | 120 136 174 166 |
| 60-75 | 192 148 138 136 | 156 110 188 158 | 214 146 138 148 |

The investigators had the idea that individuals with a better paid job in general have a higher blood pressure. Because age might affect the blood pressure too, the age factor was also introduced. Since older people generally have better paid jobs, not taking age into account may yield erroneous conclusions.

First we have to translate the problem into a mathematical model. For this we will use a more convenient description than the one in (1.1). Let the parameter $\mu$ represent a general mean, and let the parameters $\alpha_1, \alpha_2$ and $\alpha_3$ denote the unknown parameters for the age effect of the three age classes, and $\beta_1, \beta_2$ and $\beta_3$ those for the socio-economic effect. Hence the unknown parameter vector $\beta$ from the general theory is now given by $\beta = (\mu, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3)^T$. If Table 1.1 is read row by

row the model $\Omega$ can in this case be described by

$$(1.3)\ \Omega: \begin{cases} Y_i = \mu + x_{i1}\alpha_1 + x_{i2}\alpha_2 + x_{i3}\alpha_3 + x_{i4}\beta_1 + x_{i5}\beta_2 + x_{i6}\beta_3 + e_i, \\ \mathrm{E}\, e_i = 0, \\ \mathrm{Cov}\ (e_i, e_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases}, \end{cases}$$

for i, j= 1,...,36. The constant coefficients $\{x_{ij}\}$ are zero or one, referring to the absence or presence of the effects of the different age and socio-economic classes. The sixth individual, for instance, with observed blood pressure 124 mm Hg, has coefficients $x_{61} = x_{65} = 1$ and $x_{62} = x_{63} = x_{64} = x_{66} = 0$. In matrix notation this model is given by

$$(1.4)\qquad \Omega: \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ \vdots \\ Y_{19} \\ Y_{20} \\ Y_{21} \\ \vdots \\ Y_{34} \\ Y_{35} \\ Y_{36} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ \vdots \\ e_{19} \\ e_{20} \\ e_{21} \\ \vdots \\ e_{34} \\ e_{35} \\ e_{36} \end{pmatrix}$$

$$\begin{aligned} \mathrm{E}\, e &= 0, \\ \mathrm{Cov}(e) &= \sigma^2 I_{n\times n}, \end{aligned}$$

where $e = (e_1, \ldots, e_{36})^T$ is the vector of errors.

We see that in this model it is assumed that the expected blood pressure is the sum of a general mean, an effect of age and an effect of socio economic class; not taken into account are possible interaction effects of age and socio economic class. We also remark that the vector $\beta$ is not uniquely determined by $\mathrm{E}\, Y$ in this model: an increase of $\mu$ with a fixed amount $c$ together with a decrease of the $\alpha_i$ and $\beta_j$ each with an amount $c/2$ will yield the same $\mathrm{E}\, Y$. This is due to the fact that $X$ is not of full rank.

In the next section it will be explained how the unknown parameter vector $\beta$ can be estimated. It will be illustrated for one and two-factor models. Another important issue

is to determine which factors should be included in a model and which can be left out. This is usually investigated by means of testing hypotheses concerning the parameters, which will be treated in Section 1.3.

## 1.2 Least Squares Estimation

Consider the model $\Omega$ of (1.1) again. In this model the parameter vector $\beta$ is determined by $\mathrm{E}\,Y = X\beta$. In principle the vector $\beta = (\beta_0, \ldots, \beta_p)^T$ can be estimated in the usual way by the least squares estimator $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T$, that is the value which minimizes

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n} (Y_i - \mathrm{E}\,Y_i)^2 \\ &= \sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} x_{ij}\beta_j)^2 \\ &= (Y - X\beta)^T (Y - X\beta). \end{aligned} \tag{1.5}$$

Differentiating $S$ with respect to $\beta$ and setting this derivative equal to zero, we find that a least squares estimator $\hat{\beta}$ is a solution of the so-called *normal equations*,

$$\sum_{i=1}^{n} x_{ij}(Y_i - \sum_{j=0}^{p} x_{ij}\beta_j) = 0, \qquad j = 0, \ldots, p, \tag{1.6}$$

or in terms of matrices

$$X^T X\beta = X^T Y. \tag{1.7}$$

Unfortunately, in most analysis of variance models $X$ is not of full rank, and hence $X^T X$ does not have an inverse. This means, as was illustrated in Example 1.1, that in the case of the analysis of variance the parameter vector $\beta$ is in general not uniquely determined by $\mathrm{E}\,Y = X\beta$ only. To overcome this problem in practice some side conditions for the parameters which make $\beta$ uniquely determined, are usually assumed.

In some way, (1.7) does provide a solution for $\hat{\beta}$. It can be proved that

$$\hat{\beta} = (X^T X)^- X^T Y \tag{1.8}$$

with $(X^T X)^-$ a so-called *generalized inverse* [1] matrix of $X^T X$. When $X$ is not of full rank, $X^T X$ has in general several generalized inverse matrices, and we see that $\hat{\beta}$ is not

[1] A generalized inverse of a (not necessarily square) matrix $A$, denoted by $A^-$, is a matrix with the property that $A^- A A^- = A^-$ and $A A^- A = A$. We notice that the ordinary inverse matrix is a generalized inverse matrix, but the reverse is not necessarily true.

uniquely determined indeed. We note that in a linear regression model the generalized inverse $(X^TX)^-$ is equal to the real inverse $(X^TX)^{-1}$ of the matrix $X^TX$, because in this case $X$ has maximal rank. Therefore, as opposed to the situation in analysis of variance models, in linear regression models $\hat{\beta}$ is uniquely determined and unbiased.

As mentioned above, to solve the non-uniqueness problem some additional conditions on the parameters are usually assumed so that a unique choice among all possible least squares estimators can be made. To illustrate what kind of additional conditions are assumed for the parameters we consider some frequently used models where one or two factors are involved [2].

### 1.2.1 The One-factor Model

Suppose that there is only one factor of interest that can influence the outcome of the response vector. Let there be $I$ possible values or *levels* for this factor and let there be $J_i$ observations for level $i$, so that, if the total number of observations equals $n$, $\sum_{i=1}^{I} J_i = n$. This situation is sometimes called the *k-sample problem*, even though there are in fact $I$ samples. Instead of using only one index to distinguish our observations, as in model (1.1), it is more convenient here to use two indices, one for the level, and one for the number of the observation within the sample for that level. We remark that for a given level $i$, all responses $Y_{ij}$ have the same expectation. We write $\mathrm{E}\, Y_{ij} = \eta_i$. The general one-factor model is thus given by

$$\Omega : \begin{cases} Y_{ij} = \eta_i + e_{ij}, \\ \mathrm{E}\, e_{ij} = 0, \\ \mathrm{Cov}\ (e_{ij}, e_{kl}) = \begin{cases} \sigma^2, & (i,j) = (k,l) \\ 0, & (i,j) \neq (k,l) \end{cases}, \end{cases} \tag{1.9}$$

for $j = 1, \ldots, J_i,\ i = 1, \ldots, I$.

Parametrizing $\eta_i$ such that (1.9) is a linear model, we obtain $\eta_i = \mu + \alpha_i$, and the model becomes

$$\Omega : \begin{cases} Y_{ij} = \mu + \alpha_i + e_{ij}, \\ \mathrm{E}\, e_{ij} = 0, \\ \mathrm{Cov}\ (e_{ij}, e_{kl}) = \begin{cases} \sigma^2, & (i,j) = (k,l) \\ 0, & (i,j) \neq (k,l) \end{cases}, \end{cases} \tag{1.10}$$

for $j = 1, \ldots, J_i,\ i = 1, \ldots, I$. Here $\mu$ is an unknown general mean, and $\alpha_i$ is an unknown effect due to the factor having level $i$—also called the *main effect* of the factor at the level

[2] The theory for models with $3, 4, \ldots$ factors is similar, but the formulas are more complex.

$i$. We note that the independent variables itself are not in the model anymore: since for the response variable $Y_{ij}$ it holds that $x_{ij}$ equals 1 and all other $x_{kl}$ equal zero, the only contribution of the factor is via $x_{ij}$, namely $\alpha_i x_{ij} = \alpha_i$.

The unknown parameter vector for (1.10) is $\beta = (\mu, \alpha_1, \ldots, \alpha_I)^T$. However, it can easily be seen that, as in Example 1.1, $\beta$ is not uniquely determined by (1.10). One way to solve this problem is to assume that

$$\mu = 0, \tag{1.11}$$

so that

$$\alpha_i = \eta_i \qquad i = 1, \ldots, I. \tag{1.12}$$

This results in the design matrix $X$ having full rank, $rank(X) = I$. The least squares estimators for the $\alpha_i$ can hence be derived from (1.8) where the inverse of $X^T X$ is used instead of the generalized inverse, or, more directly, from (1.5). Differentiating $S(\beta) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \mathrm{E}\, Y_{ij})^2 = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \alpha_i))^2$ with respect to the $\alpha_i$ yields the normal equations

$$\frac{\partial}{\partial \alpha_i} S(\beta) \;=\; -2 \sum_{j=1}^{J_i} (Y_{ij} - \alpha_i) = 0, \qquad i = 1, \ldots, I, \tag{1.13}$$

and hence the estimator

$$\hat{\alpha}_i = \frac{1}{J_i} \sum_{j=1}^{J_i} Y_{ij} = Y_{i\cdot}\ , \qquad i = 1, \ldots, I, \tag{1.14}$$

which is what one would expect from (1.12). Here a dot as a subscript means that the mean is taken over the corresponding index.

Alternatively, to solve the non-uniqueness problem one could assume that all levels contribute equally to the general mean by postulating

$$\mu = \frac{1}{I} \sum_{i=1}^{I} \eta_i = \eta_{\cdot}\ , \tag{1.15}$$

which is equivalent to postulating

$$\sum_{i=1}^{I} \alpha_i = 0. \tag{1.16}$$

It makes the $\alpha_i$ in this case represent the additional effect contributing to the expectation at level $i$, $\eta_i$, due to the the factor being at level $i$, that is

$$\alpha_i = \eta_i - \mu = \eta_i - \eta_{\cdot}\ , \qquad i = 1, \ldots, I. \tag{1.17}$$

The sum of squares $S(\beta)$ is now given by

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \mathrm{E}\,Y_{ij})^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \mu - \alpha_i)^2, \end{aligned}$$

and the normal equations are given by

$$\begin{aligned} \frac{\partial}{\partial\mu}S(\beta) &= -2\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \mu - \alpha_i) = 0, \\ \frac{\partial}{\partial\alpha_i}S(\beta) &= -2\sum_{j=1}^{J_i}(Y_{ij} - \mu - \alpha_i) = 0, \qquad i = 1, \ldots, I. \end{aligned} \tag{1.18}$$

Solving these equations with use of the side condition $\sum_{i=1}^{I}\alpha_i = 0$, we find that the least squares estimators for $\mu$ and the $\alpha_i$ are in this case given by

$$\begin{aligned} \hat{\mu} &= \frac{1}{I}\sum_{i=1}^{I}\frac{1}{J_i}\sum_{j=1}^{J_i}Y_{ij} = Y_{\cdot\cdot}\ , \\ \hat{\alpha}_i &= \frac{1}{J_i}\sum_{j=1}^{J_i}Y_{ij} - \frac{1}{I}\sum_{i=1}^{I}\frac{1}{J_i}\sum_{j=1}^{J_i}Y_{ij} = Y_{i\cdot} - Y_{\cdot\cdot}\ . \end{aligned} \tag{1.19}$$

We remark that also $\sum_{i=1}^{I}\hat{\alpha}_i = 0$. Furthermore, we see from (1.12) and (1.14), (1.15) and (1.19), and (1.17) and (1.19) that the estimators satisfy similar relationships as the parameters they are supposed to estimate except that expectations $\mathrm{E}\,Y_i$ are replaced by means $Y_{i\cdot}$. Obviously the estimators are unbiased.

In practice almost always the first solution is used. The second solution we included, because its structure is similar to the solution that is most often used for the two- (see below) and more-factor models. It therefore gives insight in how such models could be constructed.

### 1.2.2 The Two-factor Model

Suppose now that two factors, $A$ and $B$ say, vary in an experiment, like age and socio-economic class in Example 1.1. If there are $I$ disjunct levels of factor $A$ and $J$ disjunct levels of factor $B$, we have $I \times J$ combinations (in the example: $I = J = 3$). Suppose also that every combination $(i, j)$ is observed $K_{ij}$ times, $K_{ij} \geq 1$. In this situation it is convenient to index the observations by $i$, $j$ and $k$. Of course, the expectation of an observation $Y_{ijk}$ now depends on its levels $i$ and $j$ of the two factors, and not on $k$.

Analogously to the notation for one-factor models we write $\mathrm{E}\, Y_{ijk} = \eta_{ij}$, for all $i$ and $j$. The general two-factor model thus becomes

$$\Omega : \begin{cases} Y_{ijk} = \eta_{ij} + e_{ijk}, \\ \mathrm{E}\, e_{ijk} = 0, \\ \mathrm{Cov}\,(e_{ijk}, e_{lmr}) = \begin{cases} \sigma^2, & (i,j,k) = (l,m,r) \\ 0, & (i,j,k) \neq (l,m,r) \end{cases} , \end{cases} \tag{1.20}$$

for $k = 1, \ldots, K_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$.

To turn (1.20) into a linear model, we may, as for the one-factor model, use a parameter $\mu$ to denote a general mean, $\alpha_1, \ldots, \alpha_I$ to denote the main effects due to factor $A$ and $\beta_1, \ldots, \beta_J$ to denote the main effects due to factor $B$. However, for an observation having level $i$ of factor $A$ and level $j$ of factor $B$, there may also be a contribution to its expectation due to an *interaction effect* of the two factors. We denote this contribution by $\gamma_{ij}$. The linear two-factor model obtained from (1.20) then is

$$\Omega : \begin{cases} Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \\ \mathrm{E}\, e_{ijk} = 0, \\ \mathrm{Cov}\,(e_{ijk}, e_{lmr}) = \begin{cases} \sigma^2, & (i,j,k) = (l,m,r) \\ 0, & (i,j,k) \neq (l,m,r) \end{cases} , \end{cases} \tag{1.21}$$

for $k = 1, \ldots, K_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$. The vector of all unknown parameters is $\beta = (\mu, \alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_J, \gamma_{11}, \ldots, \gamma_{IJ})^T$.

Once more, without additional conditions on the parameters this model cannot be identified. The side conditions on the parameters are usually chosen as follows.

$$\begin{aligned} &\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0 \\ &\sum_{i=1}^{I} \gamma_{ij} = \sum_{j=1}^{J} \gamma_{ij} = 0, \end{aligned} \tag{1.22}$$

for all $i$ and $j$, which is similar to (1.16) in the second case of the one-factor model. It means that we have

$$\begin{aligned} &\mu = \eta_{..} \,, \\ &\alpha_i = \eta_{i.} - \eta_{..} \,, \qquad i = 1, \ldots, I, \\ &\beta_j = \eta_{.j} - \eta_{..} \,, \qquad j = 1, \ldots, J, \\ &\gamma_{ij} = \eta_{ij} - \eta_{i.} - \eta_{.j} + \eta_{..} \,, \qquad i = 1, \ldots, I,\ j = 1, \ldots, J. \end{aligned} \tag{1.23}$$

To simplify the notation we assume from now on that $K_{ij} = K \geq 1$, $i = 1, \ldots, I, j = 1, \ldots, J$. For unequal $K_{ij}$ the procedures are the same, but the formulas quite a bit more complicated. In practice this does not cause problems, because nowadays all analyses of variance are performed on a computer with standard statistical packages.

In order to find the normal equations that determine the least squares estimator $\hat{\beta}$ we again differentiate $S(\beta)$, the sum of squares of the observations minus their expectation, with respect to all parameters, and solve these under the side conditions (1.22). For the two-factor model (1.21) $S(\beta)$ is given by

$$S(\beta) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (Y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2.$$

A little bit of computation yields

$$\begin{aligned} \hat{\mu} &= Y_{\cdots}\,, \\ \hat{\alpha}_i &= Y_{i\cdot\cdot} - Y_{\cdots}\,, \qquad i = 1, \ldots, I, \\ \hat{\beta}_j &= Y_{\cdot j\cdot} - Y_{\cdots}\,, \qquad j = 1, \ldots, J, \\ \hat{\gamma}_{ij} &= Y_{ij\cdot} - Y_{i\cdot\cdot} - Y_{\cdot j\cdot} + Y_{\cdots}\,, \qquad i = 1, \ldots, I, j = 1, \ldots, J. \end{aligned} \tag{1.24}$$

We notice that $\sum_{i=1}^{I} \hat{\alpha}_i = \sum_{j=1}^{J} \hat{\beta}_j = \sum_{i=1}^{I} \hat{\gamma}_{ij} = \sum_{j=1}^{J} \hat{\gamma}_{ij} = 0$ for every $i$ and $j$. Once again, we see from (1.23) and (1.24) that the estimators satisfy similar relationships as the parameters they are supposed to estimate except that expectations are replaced by means, and we see that the estimators are unbiased.

### 1.2.3 Additive Models

A model with two or more factors is called *additive* if there are no interactions. In model (1.21) we only have to consider one type of interactions: model (1.21) is called *additive* if $\gamma_{ij} = 0$ for all $i$ and $j$. In a model with three main effects $A, B$ and $C$ we have to consider four types of interactions: those between $A$ and $B$, $A$ and $C$, $B$ and $C$, and those between $A, B$ and $C$. And so on. Additive models are easier to interpret than models with interactions. This is why additive models are frequently used (often even without checking whether this is realistic). We will discuss in the next section how to test whether an additive model suffices; here we just formulate the two-factor additive model and give expressions for the least squares estimators of its parameters.

The two-factor additive model is given by

$$\Omega : \begin{cases} Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}, \\ \mathrm{E}\, e_{ijk} = 0, \\ \mathrm{Cov}\,(e_{ijk}, e_{lmr}) = \begin{cases} \sigma^2, & (i,j,k) = (l,m,r) \\ 0, & (i,j,k) \neq (l,m,r) \end{cases}, \end{cases} \tag{1.25}$$

$k = 1, \ldots, K_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$ with side conditions

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0. \tag{1.26}$$

For the rest of this chapter we consider only the special case in which $K_{ij} = K$ for all $i, j$ (balanced treatments). It is not difficult to see that solving the corresponding normal equations we find the same least squares estimators $\hat{\mu}$, $\hat{\alpha}_i$ and $\hat{\beta}_j$ for $\mu$, $\alpha_i$ and $\beta_j$, respectively, as given in (1.24) for the two-factor model *with* interactions.

**Example 1.2** Blood pressure (continued):
We recall the model of Example 1.1

$$\Omega : Y_i = \mu + x_{i1}\alpha_1 + x_{i2}\alpha_2 + x_{i3}\alpha_3 + x_{i4}\beta_1 + x_{i5}\beta_2 + x_{i6}\beta_3 + e_i,$$

where the $e_i$ satisfy the required assumptions. We now define $Y_{ijk}$ as the systolic blood pressure of the $k$-th individual in age class $i$ and socio-economic class $j$, $(i, j = 1, 2, 3, k = 1, \ldots, 4)$. If $y_{ijk}$ denotes the realization of $Y_{ijk}$, then, for instance, $y_{231} = 120$ and $y_{123} = 112$. The model can be rewritten in the form of (1.21) as

$$\Omega : Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk} \qquad i, j = 1, 2, 3, k = 1, \ldots, 4$$

with side conditions

$$\alpha_1 + \alpha_2 + \alpha_3 = \beta_1 + \beta_2 + \beta_3 = 0$$

As explained above, the parameters $\mu, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2$ and $\beta_3$ can be estimated by the least squares estimators subject to the side conditions. We find

$$\begin{array}{llll} \hat{\mu} = & 138.5 & & \\ \hat{\alpha}_1 = & -16.4 & \hat{\alpha}_2 = -1.1 & \hat{\alpha}_3 = 17.5 \\ \hat{\beta}_1 = & -6.2 & \hat{\beta}_2 = 1.6 & \hat{\beta}_3 = 4.6. \end{array}$$

## 1.3 Hypothesis Testing

Let us now consider the problem of which effects to put in a model. For this we need some knowledge about a couple of probability distributions. This we will treat first. Then we derive some general results which we will apply to the one- and two-factor models afterwards.

### 1.3.1 Some Probability Distributions

Let $X$ be a $k$-dimensional random vector, that is $X = (X_1, \ldots, X_k)^T$, where each element $X_i$ is a random variable. Its expectation is defined as $\mathrm{E}\, X = (\mathrm{E}\, X_1, \ldots, \mathrm{E}\, X_k)^T$ and its covariance matrix $\mathrm{Cov}\, X$ as the $k \times k$ matrix having $\mathrm{Cov}(X_i, X_j)$ as its $(i,j)$th element. We can now define a multivariate version of the normal distribution.

**Definition 1.1** *A $k$-dimensional random vector $X$ has a multivariate normal distribution, if every linear combination $\sum_{i=1}^k a_i X_i$ is normally distributed.*

If a $k$-dimensional random vector $X$ has a multivariate normal distribution with expectation vector $\mu$ and covariance matrix $\Sigma$, we write $X \sim \mathcal{N}_k(\mu, \Sigma)$. From the definition of multivariate normality follow several useful results.

**Corollary 1.1** *If $X$ has a multivariate normal distribution, then all its elements are normally distributed.*

**Corollary 1.2** *If $X_1, \ldots, X_k$ are independent random variables with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $i = 1, \ldots, k$, then $X = (X_1, \ldots, X_k)^T \sim \mathcal{N}_k(\mu, diag(\sigma_1^2, \ldots, \sigma_k^2))$ with $\mu = (\mu_1, \ldots, \mu_k)^T$ and $diag(\sigma_1^2, \ldots, \sigma_k^2)$ the $k \times k$ diagonal matrix with $\sigma_1^2, \ldots, \sigma_k^2$ on the diagonal.*

This is true because a linear combination of *independent*, normally distributed random variables is normally distributed again.

**Corollary 1.3** *If $X \sim \mathcal{N}_k(\mu, \Sigma)$ and $A$ is a nonrandom $p \times k$ matrix, then $AX \sim \mathcal{N}_p(A\mu, A\Sigma A^T)$.*

This follows from the definition of a multivariate normal distribution and the linearity and bilinearity of the expectation and the covariance of random variables, respectively.

We will also need another distribution, the so-called *chisquared* distribution.

**Definition 1.2**
*(i) If $X_1, \ldots, X_k$ are independent and identically distributed (i.i.d.) with $X_i \sim \mathcal{N}(0,1)$, then $X = \sum_{i=1}^k X_i^2$ has a (central) chisquared distribution with $k$ degrees of freedom, notation $X \sim \chi_k^2$.*
*(ii) If $X_1, \ldots, X_k$ are independent and $X_i \sim \mathcal{N}(\mu_i, 1)$, $i = 1, \ldots, k$, then $X = \sum_{i=1}^k X_i^2$ has a non-central chisquared distribution with $k$ degrees of freedom and non-centrality parameter $\nu = \sum_{i=1}^k \mu_i^2$, notation $X \sim \chi_k^2(\nu)$.*

From this definition expectations and variances are easily obtained.

**Corollary 1.4**
*(i) The expectation of a random variable with a (central) chisquared distribution with $k$ degrees of freedom is equal to $k$, and the variance equals $2k$.*
*(ii) The expectation of a non-central chisquared distributed random variable with $k$ degrees of freedom and non-centrality parameter $\nu$ equals $k + \nu$, and its variance is $2k + 4\nu$.*

Since the non-centrality parameter of a non-central chisquared distribution is always positive, we see that the expectation of a random variable with a non-central chisquared distribution is larger than the expectation of a random variable with a central chisquared distribution with the same number of degrees of freedom.

If $X \sim \mathcal{N}_k(0, I_{k\times k})$, then $X^T X = \sum_{i=1}^k X_i^2 \sim \chi_k^2$, because $X_1, \ldots, X_k$ are i.i.d. with a standard normal distribution. This result can be generalized as follows.

**Lemma 1.1** *Suppose $X \sim \mathcal{N}_k(\mu, \sigma^2 I_{k\times k})$ and $A$ is a symmetric, idempotent [3] $k\times k$ matrix with $r = rank(A)$, then*

$$X^T A X/\sigma^2 \sim \chi_r^2(\nu)$$

*with $\nu = \mu^T A \mu/\sigma^2$.*

**Proof.** Matrix $A$ is symmetric, so there is an orthogonal [4] matrix $Q$ such that $A = Q^T \Lambda Q$ with $\Lambda$ a diagonal matrix with the eigenvalues of $A$ on the diagonal. Then $\Lambda = QAQ^T = QAAQ^T = QAQ^TQAQ^T = \Lambda^2$. Thus the eigenvalues of $A$ are zero or one. Since the rank of a symmetric, idempotent matrix is equal to the sum of its eigenvalues and $rank(A) = r$, matrix $\Lambda$ has $r$ diagonal entries equal to one. Without loss of generality the first $r$ entries can be assumed to be equal to one. Define $Y = \sigma^{-1}QX$, then $Y \sim \mathcal{N}_k(\sigma^{-1}Q\mu, I_{r\times r})$, since $\sigma^{-1}Q\sigma^2 I_{k\times k}(\sigma^{-1}Q)^T = I_{k\times k}$. The assertion now follows from

$$X^T A X = X^T Q^T \Lambda Q X = (QX)^T \Lambda (QX) = \sigma^2 Y^T \Lambda Y = \sigma^2 \sum_{i=1}^r Y_i^2 \sim \sigma^2 \chi_r^2(\nu),$$

where in the last step we used that $Y_1, \ldots, Y_n$ are independent and normally distributed, and the equalities

$$\sum_{i=1}^r \sigma^{-2}(Q\mu)_i^2 = \sigma^{-2}(Q\mu)^T \Lambda (Q\mu) = \mu^T A \mu/\sigma^2 = \nu.$$

□

Finally, the *F*- or *Fisher-distribution* can be defined in terms of two independent chisquared distributed random variables.

**Definition 1.3**
*(i) If $X_1 \sim \chi_{k_1}^2$, $X_2 \sim \chi_{k_2}^2$, and $X_1$ and $X_2$ are independent, then the ratio $X = \frac{X_1/k_1}{X_2/k_2}$ has a (central) F-distribution with $k_1$ and $k_2$ degrees of freedom, notation $X \sim F_{k_1,k_2}$.*
*(ii) If $X_1 \sim \chi_{k_1}^2(\nu)$, $X_2 \sim \chi_{k_2}^2$, and $X_1$ and $X_2$ are independent, then the ratio $X = \frac{X_1/k_1}{X_2/k_2}$ has a non-central F-distribution with $k_1$ and $k_2$ degrees of freedom and non-centrality parameter $\nu$, notation $X \sim F_{k_1,k_2}(\nu)$.*

[3] A square matrix $A$ is called idempotent if $AA = A$.
[4] A square matrix $Q$ is called orthogonal if $QQ^T = Q^TQ = I$.

Again, although the formulas are a little more complex than for the chi-squared distribution, it is intuitively clear that the expectation of a random variable with a non-central $F$-distribution is larger than the expectation of a random variable with a central $F$-distribution with the same degrees of freedom.

### 1.3.2 The Residual Sum of Squares

Information on which effects to include in a model can be obtained by testing hypotheses concerning the parameters. To perform such tests some assumptions about the distribution of the errors have to be made. We assume from now on that the errors $e_1, \ldots, e_n$ are independent and identically distributed with a $\mathcal{N}(0, \sigma^2)$-distribution, as is often done in practical situations too. Thus the observations $Y_1, \ldots, Y_n$ are independent and each have a normal distribution with expectation $\mathrm{E}\, Y_i = \eta_i = \sum_{j=0}^{p} x_{ij}\beta_j$ (with $x_{i0} = 1$) and variance $\mathrm{Var}\, Y_i = \sigma^2, i = 1, \ldots, n$.

One of the most important quantities in the context of hypothesis testing is the minimum sum of squared differences between the observations and their expectations, $S(\hat{\beta})$ (cf. (1.5)). It is known as the *residual sum of squares* or the *error sum of squares*, and is often denoted by $S_\Omega$. It is given by

$$\begin{aligned} S_\Omega = S(\hat{\beta}) &= \sum_{i=1}^{n}(Y_i - \sum_{j=0}^{p} x_{ij}\hat{\beta}_j)^2 \\ &= (Y - X\hat{\beta})^T(Y - X\hat{\beta}). \end{aligned} \tag{1.27}$$

Although this is not immediately clear from its definition, $S_\Omega$ does not depend on the choice of the generalized inverse, and hence not on the choice of $\hat{\beta}$. Instead $S_\Omega$ is uniquely determined. Concerning its distribution we have the following result.

**Lemma 1.2** *The random variable $S_\Omega/\sigma^2$ has a central chisquared distribution with number of degrees of freedom equal to $rank(I - X(X^TX)^-X^T) = n - rank(X)$.*

**Proof.** First we remark that it is easily checked that $I - X(X^TX)^-X^T$ is symmetric and idempotent.

To find the distribution of $S_\Omega$, we rewrite $S_\Omega$ as a quadratic form in $Y$:

$$\begin{aligned} S_\Omega &= (Y - X\hat{\beta})^T(Y - X\hat{\beta}) \\ &= (Y - X(X^TX)^-X^TY)^T(Y - X(X^TX)^-X^TY) \\ &= Y^T(I - X(X^TX)^-X^T)^T(I - X(X^TX)^-X^T)Y \\ &= Y^T(I - X(X^TX)^-X^T)Y, \end{aligned} \tag{1.28}$$

where in the first step we used the fact that $\hat{\beta} = (X^TX)^-X^TY$ (see (1.8)) and in the last step that $I - X(X^TX)^-X^T$ is symmetric and idempotent. From the normality assumption

for the measurement errors, (1.28) and Lemma 1.1 it follows that $S_\Omega/\sigma^2$ has a chisquared distribution with number of degrees of freedom equal to $rank(I - X(X^TX)^-X^T)$. We merely state here without proof that $rank(I - X(X^TX)^-X^T)$ equals $n - rank(X)$. The distribution is central since

$$\begin{aligned}
&(\mathrm{E}\,Y)^T(I - X(X^TX)^-X^T)\,\mathrm{E}\,Y \\
&= (X\beta)^T(I - X(X^TX)^-X^T)X\beta \\
&= \beta^T(X^TX - X^TX(X^TX)^-X^TX)\beta \\
&= 0,
\end{aligned} \tag{1.29}$$

where once more we used the defining properties of a generalized inverse. We see that indeed the distribution of $S_\Omega$ does not depend on the choice of the generalized inverse. □ As a result we have

$$\mathrm{E}\,S_\Omega/\sigma^2 = n - rank(X),$$

and hence $S_\Omega/(n - rank(X))$ is an unbiased estimator of $\sigma^2$. We can now derive several tests for specific models.

### 1.3.3 A Test for One-factor Models

Let us consider the one-factor model with intercept equal to zero:

$$\Omega : \begin{cases} Y_{ij} = \alpha_i + e_{ij}, \\ \mathrm{E}\,e_{ij} = 0, \\ \mathrm{Cov}\,(e_{ij}, e_{kl}) = \begin{cases} \sigma^2, & (i,j) = (k,l) \\ 0, & (i,j) \neq (k,l) \end{cases}, \end{cases} \tag{1.30}$$

for $j = 1, \ldots, J_i,\ i = 1, \ldots, I$. The interesting question here is whether all levels have the same expectation. This means that we would like to test the hypothesis

$$\mathrm{H}_0 : \alpha_1 = \cdots = \alpha_I,$$

or

$$\mathrm{H}_0 : \text{the smaller model } \omega \text{ holds}$$

with $\omega$ given by

$$\omega : \begin{cases} Y_{ij} = \alpha + e_{ij}, \\ \mathrm{E}\,e_{ij} = 0, \\ \mathrm{Cov}\,(e_{ij}, e_{kl}) = \begin{cases} \sigma^2, & (i,j) = (k,l) \\ 0, & (i,j) \neq (k,l) \end{cases}, \end{cases} \tag{1.31}$$

for $j = 1, \ldots, J_i,\ i = 1, \ldots, I$. We have seen that under the model $\Omega$, $\hat{\alpha}_i = Y_{i\cdot}$ and $rank(X) = I$, so that

$$S_\Omega = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - Y_{i\cdot})^2, \tag{1.32}$$

$S_\Omega/\sigma^2 \sim \chi^2_{n-I}$ and $S_\Omega/(n-I)$ is an unbiased estimator of $\sigma^2$. Under the model $\omega$ all $Y_{ij}$ have the same distribution, $rank(X) = 1$, and

$$\hat{\alpha} = Y_{\cdot\cdot}, \qquad S_\omega = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - Y_{\cdot\cdot})^2, \tag{1.33}$$

$S_\omega/\sigma^2 \sim \chi^2_{n-1}$ and $S_\omega/(n-1)$ is an unbiased estimator of $\sigma^2$. Since $\omega$ is a smaller model, $S_\omega$ will generally be larger than $S_\Omega$. In fact, the larger $S_\omega - S_\Omega$, the less plausible it will be that $\omega$ is the better model of the two. More precisely, we have that $(S_\omega - S_\Omega)/\sigma^2$ also has a chisquared distribution. It has $I-1$ degrees of freedom and non-centrality parameter $\nu = \sum_{i=1}^{I} J_i(\alpha_i - \alpha_\cdot)^2/\sigma^2$, which equals zero under $H_0$. That $(S_\omega - S_\Omega)/\sigma^2$ has $I-1$ degrees of freedom can be intuitively argued from the fact that compared to $\omega$, in $\Omega$ there are $I-1$ additional parameters parameters that can be chosen freely.

Moreover, $(S_\omega - S_\Omega)$ and $S_\Omega$ are independent, so that we find that under $H_0$

$$\mathcal{F} = \frac{(S_\omega - S_\Omega)/(I-1)}{S_\Omega/(n-I)} \sim F_{I-1,n-I}, \tag{1.34}$$

and we can test $H_0$ by rejecting $H_0$ for large values of $\mathcal{F}$.

Apart from serving in the test statistic, the sums of squares are of separate interest. Due to the fact that cross products cancel out, we have

$$\begin{aligned} S_\omega - S_\Omega &= \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - Y_{\cdot\cdot})^2 - \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - Y_{i\cdot})^2 \\ &= \sum_{i=1}^{I} J_i (Y_{i\cdot} - Y_{\cdot\cdot})^2. \end{aligned} \tag{1.35}$$

From (1.32), (1.33) and (1.35) we see that the sums of squares and their differences are measures for the spread of the observations around certain means. This explains the following terminology.

**Definition 1.4**
*(i) $S_\Omega$ is the sum of squares (of variation) within groups.*
*(ii) $S_\omega$ is the sum of squares of total variation around the general mean.*
*(iii) $S_\omega - S_\Omega$ is the sum of squares (of variation) between groups.*

The above results are usually summarized in an *analysis of variance table*, abbreviated as ANOVA table, like Table 1.2. In such tables the abbreviation $SS$ stands for sum of squares, $DF$ for degrees of freedom, $MS$ for the mean square $SS/DF$, and E($MS$) for the expected mean square. An ANOVA table for $\Omega$ being the one-factor model (1.30) is given below (Table 1.2).

**Table 1.2**

| Source | $SS$ | $DF$ | $MS$ | E($MS$) |
|---|---|---|---|---|
| Between groups | $S_\omega - S_\Omega$ | $I-1$ | $\frac{S_\omega - S_\Omega}{I-1}$ | $\sigma^2 + \frac{1}{I-1}\sum_{i=1}^{I} J_i(\alpha_i - \alpha_\cdot)^2$ |
| Within groups | $S_\Omega$ | $n-I$ | $\frac{S_\Omega}{n-I}$ | $\sigma^2$ |
| Total | $S_\omega$ | $n-1$ | — | — |

### 1.3.4 Tests for Two-factor Models

Let us first consider the additive two-factor model (1.25). Here the important questions are whether factors $A$ and/or $B$ should be included in the model. In other words, the hypotheses of interest are:

$$\mathrm{H}_A : \alpha_1 = \cdots = \alpha_I = 0,$$

and

$$\mathrm{H}_B : \beta_1 = \cdots = \beta_J = 0.$$

We only discuss testing of $\mathrm{H}_A$; testing of $\mathrm{H}_B$ can be done in exactly the same way.

Under model (1.25), the residual sum of squares (1.27) turns into

$$\begin{aligned} S_\Omega = S(\hat{\beta}) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - Y_{i\cdot\cdot} - Y_{\cdot j\cdot} + Y_{\cdots})^2. \end{aligned} \tag{1.36}$$

We wish to test whether the factor $A$ has to be included in the model or not. Intuitively, if factor $A$ does not have much influence, then $\mathrm{H}_A$ should not be rejected, and then

including factor $A$ in the model anyway would not improve the model much. Hence this would not decrease the residual sum of squares much. Therefore for testing $\mathrm{H}_A$ we may again consider $S_\omega - S_\Omega$, where $S_\omega$ now is the residual sum of squares under the more restricted, smaller model $\omega = \Omega \cap \mathrm{H}_A$, with only factor $B$ in it. Under $\mathrm{H}_A$ the difference $S_\omega - S_\Omega$ should be small. Since we know $S_\Omega$ from (1.36), we have to find $S_\omega$.

In order to derive an expression in terms of the observations for $S_\omega$ we need to know the least squares estimators of $\mu$ and $\beta_1, \ldots, \beta_J$ under $\omega$. They can be found by minimizing $S_\omega(\beta)$ with respect to the parameters. Obviously, $S_\omega(\beta)$ is given by

$$S_\omega(\beta) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \mu - \beta_j)^2.$$

The normal equations are given by

$$\begin{aligned} \frac{\partial}{\partial \mu} S_\omega(\beta) &= -2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \mu - \beta_j) = 0, \\ \frac{\partial}{\partial \beta_j} S_\omega(\beta) &= -2\sum_{i=1}^{I}\sum_{k=1}^{K}(Y_{ijk} - \mu - \beta_j) = 0, \qquad j = 1, \ldots, J. \end{aligned} \tag{1.37}$$

Solving these equations under the side condition $\sum_{j=1}^{J}\beta = 0$ yields the least square estimators for $\omega$

$$\begin{aligned} \hat{\mu} &= Y_{\cdot\cdot\cdot} \\ \hat{\beta}_j &= Y_{\cdot j \cdot} - Y_{\cdot\cdot\cdot}, \end{aligned} \tag{1.38}$$

which are the same as under the larger model $\Omega$ (cf. (1.24). The residual sum of squares in the model $\omega, S_\omega$, is thus given by

$$\begin{aligned} S_\omega &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \hat{\mu} - \hat{\beta}_j)^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - Y_{\cdot\cdot\cdot} - (Y_{\cdot j \cdot} - Y_{\cdot\cdot\cdot}))^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - Y_{\cdot j \cdot})^2. \end{aligned} \tag{1.39}$$

But this means that $S_\omega - S_\Omega$, which is often denoted by $SS_A$, equals

$$SS_A = S_\omega - S_\Omega = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - Y_{\cdot j \cdot})^2 - S_\Omega$$

$$
\begin{aligned}
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}((Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...}) + (Y_{i..} - Y_{...}))^2 - S_\Omega \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{i..} - Y_{...})^2 \qquad (1.40) \\
&= JK\sum_{i=1}^{I}(Y_{i..} - Y_{...})^2 \\
&= JK\sum_{i=1}^{I}\hat{\alpha}_i^2,
\end{aligned}
$$

where the fourth equality results from the fact that cross products cancel out. Similarly to the procedure for one-factor models, an appropriate test statistic can now be constructed based on the magnitude of $SS_A = S_\omega - S_\Omega$ compared to $S_\Omega$. Under $\Omega$ given by (1.25), the design matrix $X$ has rank $I + J - 1$, under $\omega$ this rank equals $J$. Hence we have under $\Omega$

$$S_\Omega/\sigma^2 \sim \chi^2_{n-(I+J-1)},$$

and under $\omega$

$$S_\omega/\sigma^2 \sim \chi^2_{n-J}.$$

Also in this situation we have that $(S_\omega - S_\Omega)/\sigma^2$ has a chisquared distribution. It has $I - 1$ degrees of freedom and non-centrality parameter $JK\sum_{i=1}^{I}\alpha_i^2/\sigma^2$, which equals zero under $\mathrm{H}_A$. (Compared to $\omega$, in $\Omega$ there are $I$ additional parameters $\alpha_1, \ldots, \alpha_I$, of which $I - 1$ can be chosen freely.) Moreover, $S_\omega - S_\Omega$ and $S_\Omega$ are independent, and we see that under $\mathrm{H}_A$

$$\mathcal{F}_A = \frac{SS_A/(I-1)}{S_\Omega/(n-(I+J-1))} \sim F_{I-1,n-(I+J-1)}. \qquad (1.41)$$

Therefore $\mathcal{F}_A$ can be used for testing the hypothesis $\mathrm{H}_A$. The hypothesis $\mathrm{H}_A$ will be rejected if $\mathcal{F}_A > F_{I-1,n-(I+J-1),1-\alpha}$, where $\alpha$ is the level of the test.

An ANOVA table for $\Omega$ being the additive two-factor model (1.25) is given below (Table 1.3).

**Table 1.3**

| Source | $SS$ | $DF$ | $MS$ | E($MS$) |
|---|---|---|---|---|
| A | $SS_A$ | $I-1$ | $\frac{SS_A}{I-1}$ | $\sigma^2 + \frac{JK}{I-1}\sum_{i=1}^{I}\alpha_i^2$ |
| B | $SS_B$ | $J-1$ | $\frac{SS_B}{J-1}$ | $\sigma^2 + \frac{IK}{J-1}\sum_{j=1}^{J}\beta_j^2$ |
| Residual | $S_\Omega$ | $n-I-J+1$ | $\frac{S_\Omega}{n-I-J+1}$ | $\sigma^2$ |
| Total | $\sum_{i,j,k}^{I,J,K}(Y_{ijk}-Y_{...})^2$ | $n-1$ | — | — |

**Example 1.3** Blood pressure (continued):
Suppose that we wish to test whether the factor age has to be included in the model, that is we wish to test the hypothesis $\mathrm{H}_A : \alpha_1 = \alpha_2 = \alpha_3 = 0$. We have $S_\Omega/\sigma^2 \sim \chi^2_{31}$ , $SS_A/\sigma^2 = (S_\omega - S_\Omega)/\sigma^2 \sim \chi^2_2(12\sum_{i=1}^{3}\alpha_i^2/\sigma^2)$,

$$\mathcal{F}_A = \frac{SS_A/2}{S_\Omega/31} \sim F_{2,31}(12\sum_{i=1}^{3}\alpha_i^2/\sigma^2).$$

Under $\mathrm{H}_A$ the test statistic $\mathcal{F}_A$ has a $F_{2,31}$ distribution. Hence, we reject $\mathrm{H}_A$ if $\mathcal{F}_A > F_{2,31,1-\alpha}$, with $\alpha$ the level of the test. The results are

**Table 1.4**

| Source | $SS$ | $DF$ | $MS$ |
|---|---|---|---|
| age | 6890.39 | 2 | 3445.19 |
| socio | 747.72 | 2 | 373.86 |
| Residual | 16320.86 | 31 | 526.48 |

The test statistic $\mathcal{F}_A$ in testing hypothesis $\mathrm{H}_A$ takes the value 6.54. The probability $P_{\mathrm{H}_A}(\mathcal{F}_A \geq 6.54) = 0.004$. Thus for any reasonable value of $\alpha$, the level of the test, we have to reject the hypothesis $\mathrm{H}_A$. If hypothesis $\mathrm{H}_B$ is tested, we find for the test statistic 0.71 and $P_{\mathrm{H}_B}(T_B \geq 0.71) = 0.50$. Hence, we do not reject $\mathrm{H}_B$, in a reasonable model the factor socio-economic class should not be included. It seems that the factor age suffices to explain the variability in blood pressure under an additive linear model. Moreover, $s_\Omega = 16320.86$ can be seen as a realization of the chisquared distributed $S_\Omega$ with $3 \times 3 \times 4 - (3 + 3 - 1) = 31$ degrees of freedom. The estimate of the standard deviation $\sigma$ of the measurement errors based on this is 22.95.

Let us now turn to the general two-factor linear model *with* interactions as given by (1.21). The residual sum of squares under this model $\Omega$ is given by

$$\begin{aligned} S_\Omega &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\gamma}_{ij})^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - Y_{ij.})^2, \end{aligned}$$

and the rank of the design matrix $X$ equals IJ. In view of the results at the beginning of this section this means that $S_\Omega/\sigma^2$ has a central chisquared distribution with $n - rank(X) = n - IJ$ degrees of freedom.

The first thing one usually wants to know is whether the interaction terms should be included or not. A test for additivity thus tests the hypothesis

$$\mathrm{H}_{AB} : \gamma_{ij} = 0 \qquad \text{for every } i \text{ and } j.$$

The procedure is similar to the one for testing $\mathrm{H}_A$ under the additive model. In this case we consider the restricted model $\omega = \mathrm{H}_{AB} \cap \Omega$ which is in fact the additive model (1.25). This means that $S_\omega$ equals $S_\Omega$ of model (1.25). From (1.36) we obtain

$$S_\omega = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...})^2. \tag{1.42}$$

The difference between the two residual sum of squares is

$$\begin{aligned} SS_{AB} = S_\omega - S_\Omega &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...})^2 - S_\Omega \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}((Y_{ijk} - Y_{ij.}) + (Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...}))^2 - S_\Omega \\ &= K\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2 \qquad (1.43) \\ &= K\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{\gamma}_{ij}^2, \end{aligned}$$

where for the fourth equality it is used once again that cross products cancel out. Under $\Omega$ the statistic $SS_{AB}/\sigma^2$ has a chisquared distribution with non-centrality parameter $K\sum_{i=1}^{I}\sum_{j=1}^{J}\gamma_{ij}^2/\sigma^2$. Under $\omega$ this parameter is equal to zero. The number of degrees of freedom is $(I-1)(J-1)$ and can be found by counting the number of the $\gamma_{ij}$ which can be estimated freely under the assumption $\sum_{i=1}^{I}\gamma_{ij} = \sum_{j=1}^{J}\gamma_{ij} = 0$ for every $i$ and $j$. Since

$SS_{AB}/\sigma^2$ and $S_\Omega/\sigma^2$ are chisquared distributed and are also independent, we get under $\mathrm{H}_{AB}$

$$\mathcal{F}_{AB} = \frac{SS_{AB}/(I-1)(J-1)}{S_\Omega/(n-IJ)} \sim F_{(I-1)(J-1),n-IJ}. \tag{1.44}$$

We reject $\mathrm{H}_{AB}$ if $\mathcal{F}_{AB} > F_{(I-1)(J-1),n-IJ,1-\alpha}$, with $\alpha$ the level of the test. If the hypothesis $\mathrm{H}_{AB}$ is rejected, at least one $\gamma_{ij}$ is unequal to zero and there is interaction between the factors $A$ and $B$. If the hypothesis $\mathrm{H}_{AB}$ is not rejected, we say that the model is additive or that the main effects do not interact. In practice one usually continues with the investigation of the influence of the main effects under the additive model.

Again a summarizing ANOVA table can be made for $\Omega$ being the general two-factor model (1.21) (see Table 1.5). From this table it can be seen that additivity cannot be tested when $K = 1$, because then $S_\Omega$ has $n - IJ = IJ(K-1) = 0$ degrees of freedom.

**Table 1.5**

| Source | $SS$ | $DF$ | $MS$ | $\mathrm{E}(MS)$ |
|---|---|---|---|---|
| $A$ main effect | $SS_A$ | $I-1$ | $\frac{SS_A}{I-1}$ | $\sigma^2 + \frac{JK}{I-1}\sum_{i=1}^{I}\alpha_i^2$ |
| $B$ main effect | $SS_B$ | $J-1$ | $\frac{SS_B}{J-1}$ | $\sigma^2 + \frac{IK}{J-1}\sum_{j=1}^{J}\beta_j^2$ |
| $AB$ interaction | $SS_{AB}$ | $(I-1)(J-1)$ | $\frac{SS_{AB}}{(I-1)(J-1)}$ | $\sigma^2 + \frac{K}{(I-1)(J-1)}\sum_{i,j}^{I,J}\gamma_{ij}^2$ |
| Residual | $S_\Omega$ | $n-IJ$ | $\frac{S_\Omega}{IJ(K-1)}$ | $\sigma^2$ |
| Total | $\sum_{i,j,k}^{I,J,K}(Y_{ijk}-Y_{...})^2$ | $n-1$ | — | — |

Testing whether or not the factor $A$ should be included is under the general two-factor linear model with interactions (1.21) completely analogous to that under the additive model (1.25), except that the test statistic $\mathcal{F}_A$ for testing $\mathrm{H}_A$ is now given by

$$\mathcal{F}_A = \frac{SS_A/(I-1)}{S_\Omega/(n-IJ)}, \tag{1.45}$$

because $S_\Omega/\sigma^2$ in this case has $(n-IJ)$ degrees of freedom instead of $(n-(I+J-1))$. The test statistic $\mathcal{F}_A$ has a $F_{I-1,n-IJ}$ distribution under $\mathrm{H}_A$. Likewise for testing $\mathrm{H}_B$.

We remark that, if $\mathrm{H}_{AB}$ is and $\mathrm{H}_A$ is not rejected, one should not conclude that effect $A$ has no influence on our observations. Since the effects $A$ and $B$ interact, the influence of $A$ is possibly hidden in the interaction terms!

**Example 1.4** Blood pressure (continued):
We now investigate the influence of age and socio-economic class on blood pressure under a two-factor model with interaction terms, denoted by $\gamma_{ij}, i, j = 1, 2, 3$. The results are summarized in Table 1.6.

**Table 1.6**

| Source | $SS$ | $DF$ | $MS$ | F | $p$ value |
|---|---|---|---|---|---|
| Main Effects | | | | | |
| age | 6890.39 | 2 | 3445.19 | 6.36 | 0.005 |
| socio | 747.72 | 2 | 373.86 | 0.69 | 0.510 |
| Interaction | | | | | |
| age-socio | 1690.11 | 4 | 422.53 | 0.780 | 0.548 |
| Residual | 14630.75 | 27 | 541.88 | | |

In conclusion, the hypothesis $H_{AB}$ is not rejected, and $H_B$ is not rejected either. The observations do not convince us to conclude that socio-economic class has any effect on blood pressure. However, the hypothesis concerning age is rejected: age does have an effect on blood pressure. The final model will only include the main effect 'age'. Its parameters are the estimates given in Example 1.2.

In this chapter we have assumed that the observations $Y$ are random and the effects are not. An analysis taking randomness of the effects into account is more involved but can be performed in a similar way. Furthermore, we have assumed that every combination of the effects is observed at least once. If this assumption is not fulfilled, the analysis will be more complicated. However in the last decades a lot of work is done and one can handle many problems like those just mentioned. This chapter describes the basic theory which can be found in more detail in Scheffé (1959). A somewhat more recent reference is Searle (1971). For more specific problems we refer to recent statistical journals or specialized books.

## Chapter 2

# Nonlinear Regression

## 2.1 Introduction

In this chapter we discuss *parametric nonlinear regression models*. The structure of the nonlinear models that we consider is quite similar to that of the linear models introduced in Chapter 1. The main difference is—as the name indicates—that the expectation of an observation is no longer a linear function of some unknown parameters, but a nonlinear one. We thus use similar notation as for linear models: we have $n$ random observations $Y_i$, $i = 1, \ldots, n$ and a $k$-dimensional vector $x_i$ denotes the vector of explanatory variables for the $i$-th observation. All observations are assumed to be independent. It is assumed that the true regression relationship between $Y_i$ and $x_i$ is the sum of a systematic part, described by a function $f$ of $x_i$, and a random part $\varepsilon_i$. Usually $f$ depends on some unknown parameters. We write

$$Y_i = f(x_i, \theta) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

where $\theta = (\theta_1, \ldots, \theta_p)^T$ is the unknown parameter vector. The function $f$ is called the *regression function*. The random variable $\varepsilon_i$ is called the measurement error. Its expectation is assumed to be zero and its unknown variance is denoted by $\sigma_i^2$. The goal is to find an appropriate function $f$ and estimate the values of its unknown parameters.

Obviously, linear models are a special case of (2.1), namely when $f$ is assumed to be a linear function of $\theta$ (called $\beta$ in the regression context). Although the family of linear models is quite flexible because all kind of transformations of the explanatory variables can be incorporated in the model, there are many situations where based on theoretical considerations and/or experimental evidence the linear assumption is found to be inadequate. In some of those cases the exact mathematical form for $f$ is known. For example, when the relationship between $Y_i$ and $x_i$ is determined by a known underlying physical law. The variable $\varepsilon_i$ is then real measurement error. However, in most situations we have little or no idea about the form of $f$ and our task is to find some function $f$ for which (2.1) is as close to the real nonlinear relationship as possible. Clearly, in this case it is important to choose $f$ from a family that, by varying its parameters, is large and flexible enough to contain a sufficiently wide variety of functions. If several models fit the

data equally well, we choose the simplest one. Despite its name, the variable $\varepsilon_i$ is now of course no longer just measurement error, but equals the discrepancy between $Y_i$ and $f(x_i, \theta)$ *including* measurement error. If there is evidence that the differences between the $\sigma_i^2$ are small, then one may assume $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for all $i$. We shall do so in the sequel. In fact, we assume that the $\varepsilon_i$ are independent and identically distributed, so that in most practical situations asymptotic normality of $\hat{\theta}$ holds. In practice normality of $\varepsilon_i$ is often assumed, but this is not necessary for what follows.

In the next section we show how to estimate the unknown parameters of (2.1). Then we describe how to determine the accuracy of the estimators. Finally we discuss some methods for checking whether the assumptions on which the statistical analysis is based are valid. A thorough treatment of nonlinear regression can be found in Gallant (1987), Bates and Watts (1988), or Seber and Wild (1989).

**Example 2.1** Puromycin:
Data on the velocity of an enzyme reaction were obtained. The number of counts per minute of radioactive product from the reaction was measured as a function of substrate concentration in parts per million (ppm). From these counts the initial rate, or velocity, of the reaction was calculated (counts/minute$^2$). The experiment was conducted once with the enzyme treated with Puromycin, and once with the enzyme untreated. The experimenter expected a so-called Michaelis-Menten relationship between the reaction velocity and the concentration.

The well-known Michaelis-Menten model for enzyme kinetics relates the initial velocity $V$ of an enzymatic reaction to the substrate concentration $[S]$ through the nonlinear equation

$$V = \frac{V_{max}[S]}{K_s + [S]}, \tag{2.2}$$

where $V_{max}$ is the asymptotic (maximum) velocity and $K_s$ is the Michaelis-Menten constant which equals the substrate concentration for which the velocity is half of its maximum. In the notation of this chapter, this would mean that the data should be modelled by a nonlinear model with regression function $f$ given by

$$f(x, \theta) = \frac{\theta_1 x}{\theta_2 + x}, \tag{2.3}$$

or

$$Y_i = \frac{\theta_1 x_i}{\theta_2 + x_i} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.4}$$

where $x$ is the substrate concentration, $\theta_1$ is the asymptotic (maximum) velocity and $\theta_2$ is the Michaelis-Menten constant. The experimenter had the idea that the asymptotic velocity parameter $\theta_1$ should be affected by the introduction of Puromycin but that the Michaelis-Menten parameter $\theta_2$ should not, and wanted to

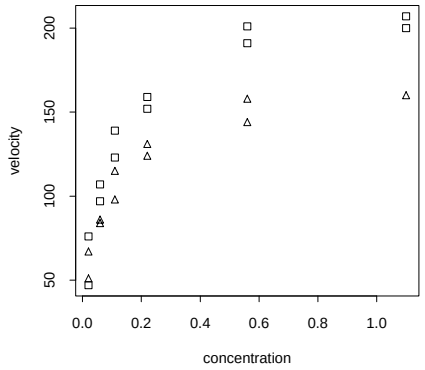


Figure 2.1: Puromycin data: squares represent data from the experiment with treated enzyme, triangles from that with untreated enzyme.

investigate this. The plot of the data in Figure 2.1 suggests that the experimenter's conjecture might be correct.

This clearly is an example where the relationship between the observations and the explanatory variable is determined by a known underlying physical process and where the parameters have physical interpretation.

## 2.2 Parameter Estimation

Like for linear models, for nonlinear models the most frequently used method for parameter estimation is least squares. In the nonlinear case the least squares estimator $\hat{\theta}$ minimizes the sum of squares

$$S(\theta) = \sum_{i=1}^{n} (Y_i - f(x_i, \theta))^2 \tag{2.5}$$

with respect to $\theta$. However, unlike in the linear least squares situation, $S(\theta)$ may have several local minima in addition to the absolute minimum $S(\hat{\theta})$. If the $\varepsilon_i$ are assumed to be independent and identically distributed with variance $\sigma^2$, it can be shown that, under certain regularity conditions on $f$, $\hat{\theta}$ and $\hat{\sigma}^2 = S(\hat{\theta})/(n-p)$ are consistent estimators of $\theta$ and $\sigma^2$, respectively. With further regularity assumptions on $f$, $\hat{\theta}$ is also asymptotically normally distributed as $n \to \infty$. If, in addition, it is assumed that the $\varepsilon_i$ are normally distributed, then $\hat{\theta}$ is also the maximum likelihood estimator of $\theta$.

As in the linear case, minimization of $S(\theta)$ is performed by solving the set of $p$ equations

$$\sum_{i=1}^{n} \frac{\partial f}{\partial \theta_l}(x_i, \theta)(Y_i - f(x_i, \theta)) = 0, \qquad l = 1, \ldots, p, \tag{2.6}$$

where $\frac{\partial f}{\partial \theta_l}$ is the partial derivative of $f$ with respect to $\theta_l$. The equations (2.6) are the *normal equations* for the nonlinear model (2.1). Because $f$ is nonlinear in $\theta$, there is in general no explicit expression for the solution of (2.6) available and iterative procedures need to be used to approximate the solution.

The iterative procedure which is mostly employed in this context is the one known as the *Gauss-Newton method*. For its use some assumptions, like differentiability with respect to $\theta$, have to be imposed on $f$. The Gauss-Newton method starts with an initial guess $\theta^0$ for $\theta$. It keeps improving this estimate in a clever way using a linear approximation of the regression function $f$. The procedure stops when the estimate does not change anymore.

More precisely, the regression function $f(x_i, \theta)$ is expanded in a first order Taylor series about $\theta^0$ yielding

$$f(x_i, \theta) \approx f(x_i, \theta^0) + \sum_{l=1}^{p} \frac{\partial f}{\partial \theta_l}(x_i, \theta^0)(\theta_l - \theta_l^0), \qquad i = 1, \ldots, n. \tag{2.7}$$

With this the measurement errors $R_i(\theta) = Y_i - f(x_i, \theta)$ (seen as a function of $\theta$) can be approximated as

$$\begin{aligned} R_i(\theta) &\approx Y_i - \left( f(x_i, \theta^0) + \sum_{l=1}^{p} \frac{\partial f}{\partial \theta_l}(x_i, \theta^0)(\theta_l - \theta_l^0) \right) \\ &= R_i(\theta^0) - \sum_{l=1}^{p} \frac{\partial f}{\partial \theta_l}(x_i, \theta^0)\delta_l, \qquad i = 1, \ldots, n, \end{aligned} \tag{2.8}$$

and the sum of squares $S(\theta)$ as

$$S(\theta) \approx \sum_{i=1}^{n} (R_i(\theta^0) - \sum_{l=1}^{p} \frac{\partial f}{\partial \theta_l}(x_i, \theta^0)\delta_l)^2. \tag{2.9}$$

Here $R_i(\theta^0) = Y_i - f(x_i, \theta^0)$ and $\delta_l = \theta_l - \theta_l^0$. The approximate residual sum of squares is then minimized with respect to $\delta_l$. Its solution $\delta^0 = (\delta_1^0, \ldots, \delta_p^0)^T$ is called the *Gauss increment*. Let $\theta^1 = \theta^0 + \delta^0$, $f(x, \theta) = (f(x_1, \theta), \ldots, f(x_n, \theta))^T$ and $Y = (Y_1, \ldots, Y_n)^T$. Then the point $f(x, \theta^1)$ should now be closer to $Y$ than $f(x, \theta^0)$ in the sense that $S(\theta^1) < S(\theta^0)$. Hence the better parameter value $\theta^1 = \theta^0 + \delta^0$ will be our next estimate of $\theta$. We notice that $\delta^0$ is easily computed. This is because the form of (2.9) is identical to that of the usual residual sum of squares for linear regression with $Y_i$ replaced by $R_i(\theta^0)$, $X$ by the derivative matrix $V^0$ with $(i,l)$-th element $\frac{\partial f}{\partial \theta_l}(x_i, \theta^0)$, and $\beta_l$ by $\delta_l$. This means that—assuming $V^0$ to be of full rank—we have

$$\delta^0 = (V^{0T} V^0)^{-1} V^{0T} R(\theta^0), \tag{2.10}$$

where $R(\theta^0) = (R_1(\theta^0), \ldots, R_n(\theta^0))^T$.

Next, another iteration is performed by calculating new residuals $R_i(\theta^1) = Y_i - f(x_i, \theta^1)$, a new derivative matrix $V^1$ with $(i,l)$-th element $\frac{\partial f}{\partial \theta_l}(x_i, \theta^1)$, and a new increment $\delta^1$. This process is repeated until convergence is obtained, that is until the increment is so small that there is practically no change in the elements of the parameter vector. The estimated value of $\hat{\theta}$ is now taken to be equal to the final estimate of $\theta$ resulting from this procedure. Nowadays most statistical packages contain a procedure for numerical computation of $\hat{\theta}$ in this way.

**Example 2.2** Puromycin (continued):
Although we see from Figure 2.1 that there is a nonlinear relationship between the concentration and the velocity, we can derive from (2.4) that the reciprocals of the variables satisfy the linear relationship

$$\frac{1}{Y_i} = \beta_0 + \beta_1 \frac{1}{x_i} + e_i, \qquad i = 1, \ldots, n, \tag{2.11}$$

where $\beta_0 = 1/\theta_1$ and $\beta_1 = \theta_2/\theta_1$. Hence we could perform a linear regression analysis on the reciprocals. However, one needs to be cautious with this, since the reciprocal transformation will also affect the measurement errors. This means that if it would be plausible to model the $\varepsilon_i$ in the original model (2.4) as being normally distributed, the normal assumption would not be adequate any more for the $e_i$ in model (2.11) for the reciprocals. Still, in such situations where a linear relationship between transformed variables exists, it is recommended to perform a linear regression anyway, since this may yield an idea of good starting values for the parameters in the nonlinear regression problem.

Let us perform a linear regression analysis on the reciprocals of the treated data. We obtain $\hat{\beta} = (0.005107, 0.000247)^T$, leading to an estimate $\tilde{\theta} = (195.8, 0.04841)^T$ for $\theta$. Figures 2.2.a–2.2.c show the results. We see that the linear regression on the reciprocals looks quite acceptable at first sight (Figure 2.2.a). However, the QQ-plot of the residuals (Figure 2.2.c) indicates that the normality assumption is probably not very adequate here. Also we see from the derived expected curve on the original scale (Figure 2.2.b) that the predicted asymptotic velocity is too small. This is probably due to the fact that the variance of the replicates has been distorted by the transformation: the cases with low concentration (high reciprocal concentration) dominate the determination of the parameters, the curve does not fit the data well at the high concentrations.

Estimation of the parameters in the nonlinear model (2.4) with least squares using the Gauss-Newton method and starting values $\tilde{\theta}$ as obtained from the linear regression yields $\hat{\theta} = (212.7, 0.06412)^T$. The expected curve based on this in Figure 2.2.d looks better than the one in Figure 2.2.b. The residual sum of squares is indeed reduced from 1920.6 to 1195.4. Since the number of treated observations is 12 and the number of parameters is 2, an estimate for the variance $\sigma^2$ of the measurement errors is $1195.4/(12-2) = 119.54$ yielding an estimated standard error

of 10.93. In view of the order of magnitude of the observations this is a reasonable value. Figure 2.2.e shows that, although not perfect, the assumption of normality for the residuals in the nonlinear model may be adequate. Furthermore, the plot of the residuals against the fitted expected values in Figure 2.2.f does not show any particular structure, except perhaps the one relatively large residual which should be further investigated. So we have no obvious reasons to doubt this analysis.

For the untreated data, 11 in total, we find with the Gauss-Newton method $\hat{\theta} = (160.3, 0.04771)^T$, residual sum of squares=859.6, estimated variance $\hat{\sigma}^2 = 95.51$, estimated standard error=9.77. At first sight these estimates suggest that the conjecture of the experimenter—that the first parameter is and the second is not influenced by the treatment—is correct. However, the relative differences between the two cases are similar for the two parameters. Thus this needs further investigation.

## 2.3 Accuracy of Estimators; Confidence Regions

We first discuss the accuracy of estimators of $\theta$ or a function of $\theta$. Next we provide some results for the expected response.

### 2.3.1 Classical Asymptotic Results

Recall that in the linear regression situation, an estimator of the covariance matrix of $\hat{\beta}$ is given by

$$\widehat{\mathrm{Cov}}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}.$$

Furthermore, in the linear regression situation, a $(1-\alpha)100\%$ confidence interval for $\beta_l$, $l = 0, \ldots, p$, is given by

$$\hat{\beta}_l \pm t_{(n-p-1);(1-\alpha/2)} \hat{\sigma} \sqrt{((X^T X)^{-1})_{ll}},$$

and a $(1-\alpha)100\%$ confidence region for the $p+1$-dimensional parameter vector $\beta$ is given by

$$\{\beta : \frac{(\beta - \hat{\beta})^T (X^T X)(\beta - \hat{\beta})}{(p+1)\hat{\sigma}^2} < F_{(p+1),(n-p-1);1-\alpha}\}.$$

Analogously, if $f$ satisfies the regularity assumptions for $\hat{\theta}$ to be asymptotically normally distributed as mentioned in Section 2.2, then for $n \to \infty$, $\hat{\theta} - \theta \sim N_p(0, \sigma^2 (V^T V)^{-1})$, where $\theta$ is the true value of the parameter vector and $V$ is the derivative matrix evaluated at $\theta$, that is with $(i,l)$-th element $\frac{\partial f}{\partial \theta_l}(x_i, \theta)$, and this normal distribution can be used to construct an estimator of the covariance matrix of $\hat{\theta}$ and *approximate* $(1-\alpha)100\%$

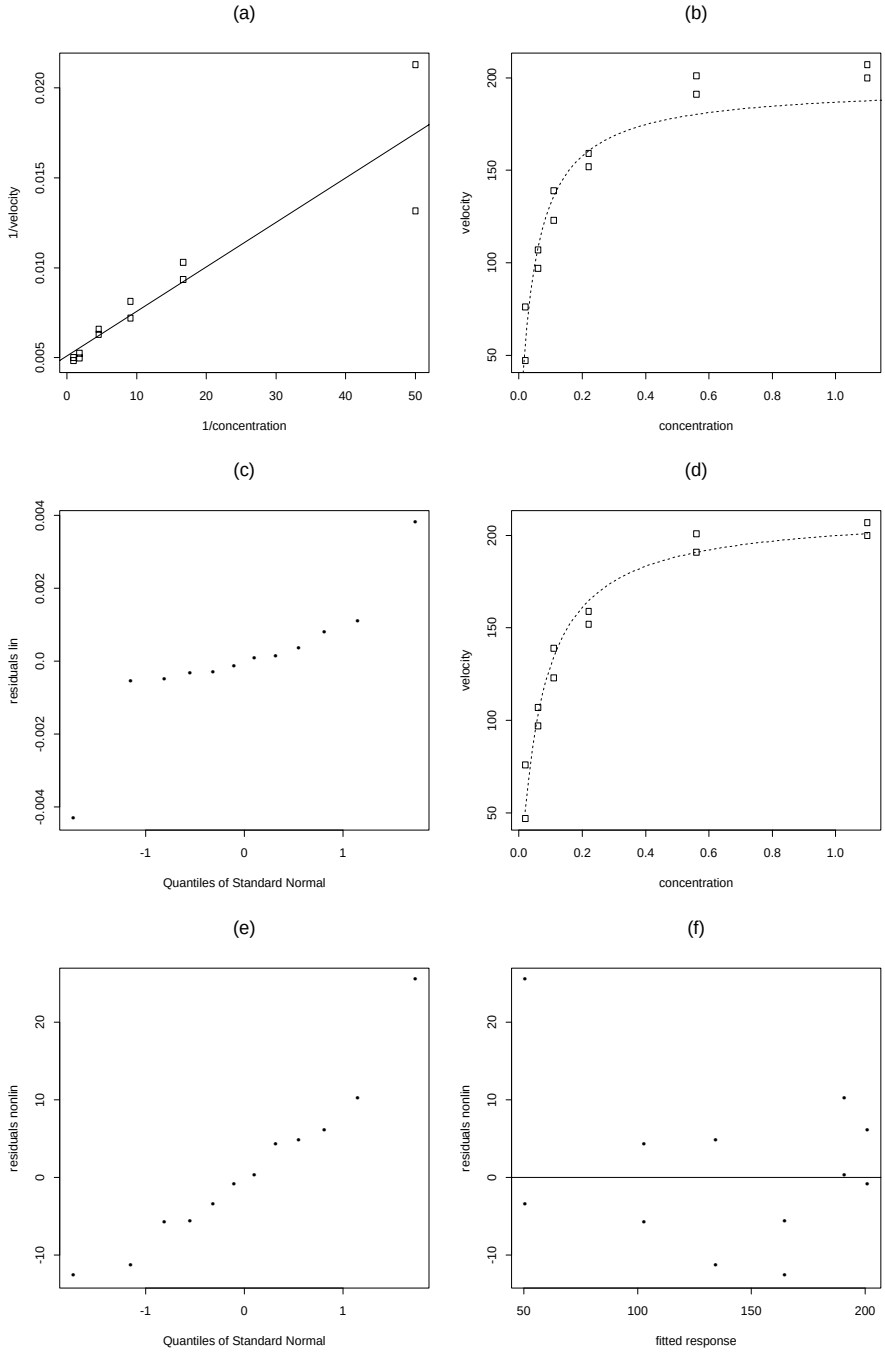


Figure 2.2: Puromycin data: a) Linear regression on treated reciprocals. b) Derived expected curve and data for treated experiment. c) Normal QQ-plot for residuals of linear regression on treated reciprocals. d) Expected curve based on nonlinear regression and data for treated experiment. e) Normal QQ-plot for residuals of nonlinear regression on treated data. f) Scatterplot of residuals of nonlinear regression on treated data versus the fitted expected values.

confidence intervals or regions. Indeed, we see that an estimator for the covariance matrix of $\hat{\theta}$ is given by

$$\widehat{\mathrm{Cov}}(\hat{\theta}) = \hat{\sigma}^2(\hat{V}^T\hat{V})^{-1}, \tag{2.12}$$

where $\hat{V}$ is the derivative matrix evaluated at $\hat{\theta}$ having $(i,l)$-th element $\frac{\partial f}{\partial \theta_l}(x_i, \hat{\theta})$. An approximate $(1-\alpha)100\%$ confidence interval for the $\theta_l$, $l = 1, \ldots, p$, is defined as

$$\hat{\theta}_l \pm t_{(n-p);(1-\alpha/2)}\hat{\sigma}\sqrt{((\hat{V}^T\hat{V})^{-1})_{ll}}, \tag{2.13}$$

and an approximate $100(1-\alpha)\%$ confidence region for the $p$-dimensional parameter vector $\theta$ in the nonlinear case is given by

$$\{\theta : \frac{(\theta-\hat{\theta})^T(\hat{V}^T\hat{V})(\theta-\hat{\theta})}{p\hat{\sigma}^2} < F_{p,(n-p);1-\alpha}\}. \tag{2.14}$$

These have, when $n$ is large, approximate coverage of $100(1-\alpha)\%$.

### 2.3.2 Using the Bootstrap

A completely different approach is to use the bootstrap for constructing variance estimates or confidence intervals. As we know, also for these estimates to be useful or these intervals to yield approximately the correct coverage, $n$ has to be large. Let $\lambda(\theta)$ be a general function of $\theta$. We discuss the construction of a bootstrap estimator of the variance of $\hat{\lambda} = \lambda(\hat{\theta})$, and of a bootstrap confidence interval for $\lambda(\theta)$ based on $\hat{\lambda} = \lambda(\hat{\theta})$. Note that $\lambda = \theta_j$ is a special case, $j = 1, \ldots, p$.

Let $\hat{\varepsilon}_i = R_i(\hat{\theta}) = Y_i - f(x_i, \hat{\theta})$, and $\tilde{\varepsilon}_i = \hat{\varepsilon}_i - 1/n\sum_{i=1}^n \hat{\varepsilon}_i$, i.e. the $i$-th centered residual. We generate a sample of $B$ bootstrap values $\lambda^*$ according to the following bootstrap scheme.

- Generate a bootstrap sample $\varepsilon_1^*, \ldots, \varepsilon_n^*$ from the empirical distribution of $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n$.
- Compute $Y_1^*, \ldots, Y_n^*$ from

$$Y_i^* = f(x_i, \hat{\theta}) + \varepsilon_i^*, \qquad i = 1, \ldots, n.$$

- Compute the value $\hat{\theta}^*$ which minimizes

$$S^*(\theta) = \sum_{i=1}^n (Y_i^* - f(x_i, \theta))^2.$$

- The bootstrap value $\lambda^*$ is $\lambda^* = \lambda(\hat{\theta}^*)$.

Repetition of this scheme $B$ times yields $\lambda_1^*, \ldots, \lambda_B^*$. The empirical distribution of $\lambda_1^*, \ldots, \lambda_B^*$ is a bootstrap estimator of the distribution of $\hat{\lambda}$. This means that a bootstrap estimator of the variance of $\hat{\lambda} = \lambda(\hat{\theta})$ is given by

$$\mathcal{S}^2(\lambda^*) = \frac{1}{B-1}\sum_{i=1}^{B}(\lambda_i^* - \frac{1}{B}\sum_{i=1}^{B}\lambda_i^*)^2, \tag{2.15}$$

and a bootstrap confidence interval for $\lambda(\theta)$ with approximately $100(1-\alpha)\%$ coverage is for instance

$$\left[2\hat{\lambda} - \lambda^*_{([(1-\alpha/2)B]+1)}, 2\hat{\lambda} - \lambda^*_{([\alpha/2B]+1)}\right]. \tag{2.16}$$

We remark that this bootstrap scheme does not need the assumption that $\hat{\theta}$ is asymptotically normally distributed. If $\hat{\theta}$ is asymptotically normally distributed, however, then the results obtained by the bootstrap will be close to those obtained by the classical methods, provided $n$ and $B$ are large.

### 2.3.3 Approximate Confidence Regions for the Expected Response

Again based on the asymptotic normality of $\hat{\theta}$, approximate confidence regions for the expected response can be generated using the analogues of linear regression. Let, as before, $\hat{V}$ denote the derivative matrix evaluated at $\hat{\theta}$, and let for an arbitrary explanatory vector $x$, the derivative vector $\hat{v}_x$ be defined by $\hat{v}_x = (\frac{\partial f}{\partial \theta_1}(x,\hat{\theta}), \ldots, \frac{\partial f}{\partial \theta_p}(x,\hat{\theta}))^T$. Then an approximate $100(1-\alpha)\%$ confidence interval for the expected response at the value $x$ for the explanatory vector is given by

$$f(x,\hat{\theta}) \pm \hat{\sigma}\sqrt{\hat{v}_x^T(\hat{V}^T\hat{V})^{-1}\hat{v}_x}\; t_{(n-p);1-\alpha/2}\;, \tag{2.17}$$

and an approximate $100(1-\alpha)\%$ confidence band for the expected response at *any* explanatory vector $x$ is

$$f(x,\hat{\theta}) \pm \hat{\sigma}\sqrt{\hat{v}_x^T(\hat{V}^T\hat{V})^{-1}\hat{v}_x}\;\sqrt{pF_{p,(n-p);\alpha}}\;. \tag{2.18}$$

The expressions in (2.17) and (2.18) differ, because (2.17) concerns an interval at a single specific "point", whereas (2.18) concerns the band produced by the intervals at all the values of $x$ considered simultaneously.

If the assumption of asymptotic normality of $\hat{\theta}$ is not realistic, one can for an approximate confidence interval for the expected response at a specific value $x_0$ resort to the bootstrap: taking $\lambda(\hat{\theta}) = f(x_0,\hat{\theta})$ in (2.16) yields a bootstrap confidence interval for $f(x_0,\theta)$—which is the expected response at $x_0$—with approximately $100(1-\alpha)\%$ coverage.

**Example 2.3** Puromycin (continued):
In Example 2.2 we have seen that the normality assumption for the $\varepsilon_i$ may be plausible. This suggests that we can obtain estimates of the variance of the parameter estimators and construct approximate confidence regions for the parameters according to the classical asymptotic results using (2.13) and (2.14). For the treated observations we find 6.95 and 0.00828 for the estimated standard deviations of $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively. With the bootstrap the estimates are 6.15 and 0.0076, so both somewhat smaller. The 95% confidence intervals are (197.2, 228.2) and (0.0457, 0.0826), respectively; using the bootstrap and (2.16) we find (199.6, 223.4) and (0.0485, 0.0781). Since the two sets of intervals do not differ too much, normality of the $\varepsilon_i$ is once more not contradicted. Moreover, the histograms of the bootstrap values in Figure 2.3 also indicate that the distributions of $\hat{\theta}_1$ and $\hat{\theta}_2$ may be approximately normal.

For the untreated observations the results based on the classical theory are: estimates of standard deviation 6.48 and 0.0078; intervals (145.6, 174.9) and (0.0301, 0.0653).

An approximate 95% confidence interval for $\theta_1^{treated} - \theta_1^{untreated}$ based on the classical theory is (31.9, 72.9), which does not contain the value zero, indicating that for the two cases the parameters $\theta_1$ are different. An approximate 95% confidence interval for $\theta_2^{treated} - \theta_2^{untreated}$ is (-0.0081, 0.0410) which contains the value zero, and we may not conclude that the parameters $\theta_2$ for the two cases are different. This means that the experimenter's conjecture seems to be correct.

An approximate 95% confidence region for $\theta$ according to (2.14) is shown in Figure 2.4 (left). On the right we see an approximate 95% confidence band for the expected response based on (2.18). It is clear from this picture that confidence bands for nonlinear models may behave quite differently from those for linear models. In this example, because the regression function is constrained to go through the origin, the band reduces to 0 here. Also, because the regression function approaches an asymptote, the band approaches an asymptote.

## 2.4 Assessment of Fit and Model Choice

An important question is whether the chosen model is adequate. The model may not fit too well globally or in a few points. This is why we need to assess the fit of the model to the data and to check the appropriateness of the model assumptions. For this we can use the same techniques as in linear regression, namely inspect several plots and compare sums of squares. Especially important for nonlinear regression is to check the parameter values. All three can detect global deviations of the model from the data; for detection of outliers especially graphical techniques are useful. The sums of squares comparison is used when a choice has to be made between different models.

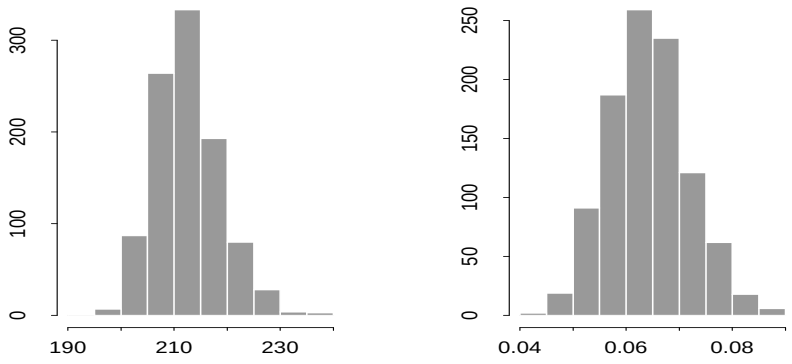


Figure 2.3: Puromycin data: histograms of 1000 bootstrap values for $\hat{\theta}_1$ (left) and $\hat{\theta}_2$ (right) for treated data.

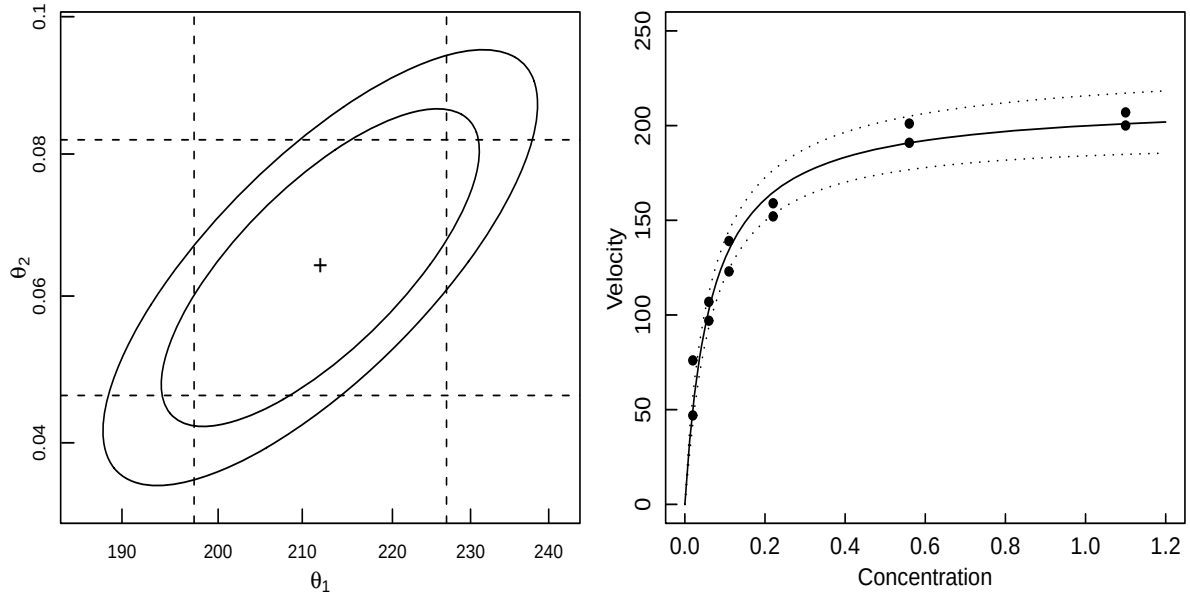


Figure 2.4: Puromycin treated data: on the left the least squares estimate (+), the joint approximate 95% and 99% confidence regions (solid lines), and the marginal approximate 95% confidence intervals (dashed lines) for $\theta$ are shown; on the right the fitted expected response (solid line), and an approximate 95% confidence band for the expected response (pair of dotted lines) are shown.

### 2.4.1 The Estimated Parameter Values

In nonlinear estimation of $\theta$, it is possible that the numerical algorithm turns out to have converged to parameter values which are obviously, or suspiciously, wrong. The convergence may have been to a local minimum, or the procedure got stalled because of some unpleasant properties of the expectation surface. Assessment of any fitted model should therefore begin with a careful consideration of the parameter estimates. If the parameter values do not make sense, one should check whether indeed there was actual convergence, and if there was, whether the convergence occurred smoothly. It should also be investigated whether correct and appropriate starting values were used. It is recommendable to try different sets of starting values. If these checks are all satisfactory, but the parameter estimate is not, it may be that the regression function is not appropriate. It is then advisable to discuss things with the researcher who collected the data.

When convergence to reasonable values has been obtained, check the estimated standard errors and $t$-ratios (calculated as (parameter estimate)/(estimated standard error)). If a $t$-ratio is not significantly different from zero, one should consider deleting that parameter from the regression function and refitting the model.

Also check the estimated correlation matrix for the parameter estimator. If any parameters are excessively highly correlated, this may indicate overparametrization of the model. (What is a high correlation is dependent on the type of data and on the model; correlations above 0.9 in absolute value should definitely be investigated). Correlated variables should be deleted or variables and parameters transformed to reduce collinearity. In general, one should always try to choose the simplest regression function that still makes sense.

### 2.4.2 Plots

When a simple, adequate regression function has been chosen, several plots may be used to verify the model assumptions. Apart from the global fit, here are three specific aspects of the model that need inspection:

- the choice of the regression function;
- the assumptions on the errors (constant variance and normality);
- the independence assumption for the observations.

The following graphs can give insight in these matters.

A plot of the fitted expected values $f(x_i, \hat{\theta})$ together with the responses $Y_i$ against $i$ or against one of the explanatory variables is very helpful to assess the fit of the model, globally as well as locally. A scatter plot of the $f(x_i, \hat{\theta})$ or observed $Y_i$ alone versus each of the explanatory variables may also be used for this, as well as a scatter plot of the

fitted versus the observed values. If the plots show any strange structures or outliers, one should investigate possible causes.

Another class of frequently used graphs are plots of the (standardized) residuals

$$R_i(\hat{\theta}) = Y_i - f(x_i, \hat{\theta}), \qquad i = 1, \ldots, n,$$

$$R_{i,stand}(\hat{\theta}) = \frac{Y_i - f(x_i, \hat{\theta})}{\hat{\sigma}}, \qquad i = 1, \ldots, n.$$

The residuals can be plotted against the explanatory variables or against $f(x_i, \hat{\theta})$. If the plots show any particular structure, that is a nonrandom behaviour, this may be an indication of a bad choice of the regression function. In such cases one could try to expand the model in a sensible way.

If the residuals tend to increase or decrease with increasing values of an explanatory variable or with increasing values of $f(x_i, \hat{\theta})$, this may indicate that the assumption of constant variance is not realistic. Of course the latter assumption may be checked more adequately if there are replications, that is if there are several observations, instead of only one $Y_i$, for the explanatory vector $x_i$, and this for each $i = 1, \ldots, n$. If there is nonconstant variance, one could transform the data to induce constant variance and change the regression function such that the model still makes sense, or try using weighted least squares.

Furthermore, a normal $QQ$-plot of the residuals may indicate whether or not normality for the $\varepsilon_i$ may be assumed. If there is pronounced lack of normality, try to decide whether this is due to a small number of outliers or whether this is due to inadequacy of the regression function.

If there are no replications, a so-called *lag plot* of $R_i(\hat{\theta})$ against $R_{i-1}(\hat{\theta})$ is convenient to detect some correlation between errors, and hence raise doubts about the independence assumption. This is especially recommended if time or distance is involved as an explanatory variable. If correlations are found, the model must be altered to account for dependencies. For this, concepts of time series analysis, discussed in one of the following chapters, could be used.

All plots mentioned above may help finding outliers. For obvious outliers, one should check whether the data are correctly recorded and correctly entered into the computer. If they are, there may be good nonstatistical reasons for removing them. One should discuss this with the experimenter, as well as the consequences of such removal for parameter estimates, standard deviations, fitted values, etcetera. Outliers should never be deleted for statistical convenience only!

**Example 2.4** Puromycin (continued):
From the foregoing examples and corresponding plots we may conclude that the nonlinear regression model for the treated Puromycin data is quite adequate. The

$t$-ratios being 30.6 and 7.7 for the two parameters are highly significant ($t_{10;0.957} = 2.23$) and we do not want to delete any of them. We already have inspected several plots giving us no reason for suspicion against the model or the parameter estimates. The one large residual should be further investigated. Finally, since there are replications in the data, it is not useful to make a lag plot.

### 2.4.3 Model Comparison

In some situations there may be several candidates for the regression function. It is recommendable to always discuss this issue with the person who collected the data. If there are practical or scientific reasons for preferring one model over the others, strong weight should be given to the experimenter's reasons because the primary aim of data analysis is to explain or account for the behaviour of the data, not simply to get the best fit. If the experimenter cannot provide convincing reasons for choosing one model over the others, then statistical analyses can be used. As mentioned above, comparison of sums of squares is usually used to make a choice. We distinguish two cases, where the comparison is between *nested* models and where it is between *non-nested* models. For example, the next three models can be obtained from each other by putting one or more of the parameters equal to zero:

$$f(x, \theta) = \theta_1 x + \theta_2 \exp^{-\theta_3 x},$$

$$f(x, \theta) = \theta_1 x + \theta_2,$$

$$f(x, \theta) = \theta_1 x.$$

These are nested models. Two non-nested models are, for instance,

$$f(x, \theta) = \theta_1 (1 - \exp^{-\theta_2 x}),$$

$$f(x, \theta) = \frac{\theta_1 x}{\theta_2 + x},$$

both of which start at $f = 0$ when $x = 0$ and approach the asymptote $\theta_1$ as $x \to \infty$, so they both could very well be candidates for a regression function for the same data. We discuss how to perform model comparison in the two different situations.

Comparison of nested models is usually performed under the assumption of normally distributed errors. We can then proceed as in the linear regression case and use an (approximate) $F$ test, which is, of course, in fact a comparison of residual sums of squares. Let $RSS_p = S_p(\hat{\theta})$ and $RSS_q = S_q(\hat{\theta})$ be the residual sums of squares of the models with $p$ and $q$ ($p < q$) parameters, respectively. Let the smaller model be obtained from the larger model by putting $\theta_{p+1}, \ldots, \theta_q$ equal to zero. We reject

$H_0$: The smaller model with $p$ parameters holds

or, equivalently, we reject

$H_0$: $\theta_{p+1} = \cdots = \theta_q = 0$
if

$$\mathcal{F}^{pq} = \frac{(RSS_p - RSS_q)/(q-p)}{RSS_q/(n-q)} \geq F_{(q-p),(n-q);1-\alpha} \ , \tag{2.19}$$

where $\alpha$ is the level of the test. This can be presented in a sort of ANOVA table:

**Table 2.1**

| source | sum of squares | $DF$ | $MS$ | $F$-ratio |
|---|---|---|---|---|
| extra parameters | $RSS_p - RSS_q$ | $q-p$ | $MS_{q-p} = \frac{RSS_p - RSS_q}{q-p}$ | $MS_{q-p}/MS_q$ |
| large model | $RSS_q$ | $n-q$ | $MS_q = \frac{RSS_q}{n-q}$ | |
| small model | $RSS_p$ | $n-p$ | $MS_p = \frac{RSS_p}{n-p}$ | |

For comparison of nested models without the normality assumption for the errors sometimes a similar approach can be taken. For this we refer to the literature.

In many situations where two or more cases are being compared, like in the Puromycin example, we can construct nested models in such way that comparison of these models tells us whether or not the different cases differ with respect to their parameter values. Such nested models contain so-called *incremental parameters* which account for a change in a parameter between the different cases. An advantage of using incremental parameters is that a preliminary evaluation of the need for the full model (with all the parameters different for the different situations) can be made directly from the regression output without having to do additional computation. We illustrate the use of incremental parameters by means of the Puromycin example.

**Example 2.5** Puromycin (continued):
In the Puromycin experiment we have two situations, treated and untreated. It was hypothesized that the Puromycin treatment should affect the maximum velocity parameter $\theta_1$, but not the half-velocity parameter $\theta_2$. To determine if the parameter $\theta_2$ is unchanged, we can use a comparison of residual sums of squares in the following way. We consider a full model and a partial model. The full model corresponds to different sets of parameters for the treated and untreated data. The partial model

corresponds to different $\theta_1$ parameters, but identical $\theta_2$ parameters. To make the two models nested models, we introduce the *indicator variable* $x_2$, defined by

$$x_2 = \begin{cases} 0, & \text{for untreated case,} \\ 1, & \text{for treated case.} \end{cases} \tag{2.20}$$

The substrate concentration we now denote by $x_1$, so that we have two explanatory variables in the model instead of one. The full model is described by

$$f(x, \theta) = \frac{(\theta_1 + \phi_1 x_2)x_1}{(\theta_2 + \phi_2 x_2) + x_1}, \tag{2.21}$$

or

$$Y_i = \frac{(\theta_1 + \phi_1 x_{i2})x_{i1}}{(\theta_2 + \phi_2 x_{i2}) + x_{i1}} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.22}$$

where $\theta_1$ is the maximum velocity for the untreated case, $\phi_1$ is the incremental maximum velocity due to the treatment, $\theta_2$ is the (possibly common) Michaelis-Menten parameter, and $\phi_2$ the incremental parameter for a change in the Michaelis-Menten constant due to the treatment. We expect $\phi_1$ to be nonzero, and are interested to test whether $\phi_2$ could be zero. That means that the partial model is the model (2.22) in which $\phi_2 = 0$. We therefore can fit the full and the partial model to all data, treated as well as untreated, and test the hypothesis $H_0$: $\phi_2 = 0$, by means of an $F$-test or, equivalently (since there is only one additional parameter in the full model compared to the partial model), by means of a $t$-test. The results of the fit of the full model are shown in Table 2.2. It appears that the $t$ ratio for the parameter $\phi_2$ is nonsignificant ($t_{19;0.975} = 2.09$) and hence $\phi_2$ could be zero.

**Table 2.2**

| parameter | estimate | stdev | $t$ ratio | corr. matrix | | | |
|---|---|---|---|---|---|---|---|
| $\theta_1$ | 160.3 | 6.90 | 23.2 | 1.00 | | | |
| $\theta_2$ | 0.0477 | 0.00828 | 5.8 | 0.78 | 1.00 | | |
| $\phi_1$ | 52.4 | 9.55 | 5.5 | −0.72 | −0.56 | 1.00 | |
| $\phi_2$ | 0.0164 | 0.0114 | 1.4 | −0.56 | −0.73 | 0.77 | 1.00 |

We thus fit the partial model. These results are given in Table 2.3 and the sum of squares analysis for testing $H_0$ is presented in Table 2.4. We conclude from Table 2.4 that the $p$-value for the $F$-test is larger than $\alpha = 0.05$ and we do not reject $H_0$. Based on this test the partial model is a good model for all data. We note that this is a very efficient way to perform the nonlinear analysis: we do the fitting for both cases *and* the comparison of the two cases in one strike.

**Table 2.3**

| parameter | estimate | stdev | $t$ ratio | corr. matrix | | |
|---|---|---|---|---|---|---|
| $\theta_1$ | 166.6 | 5.81 | 28.7 | 1.00 | | |
| $\theta_2$ | 0.0580 | 0.00591 | 9.8 | 0.61 | 1.00 | |
| $\phi_1$ | 42.0 | 6.27 | 6.7 | −0.54 | −0.06 | 1.00 |

**Table 2.4**

| source | sum of squares | $DF$ | $MS$ | $F$-ratio | $p$-value |
|---|---|---|---|---|---|
| extra parameters | 186 | 1 | 186.0 | 1.7 | 0.21 |
| full model | 2055 | 19 | 108.2 | | |
| partial model | 2241 | 20 | | | |

When trying to decide which of several non-nested models is to be preferred, the experimenter's advice is even more important, since the use of $F$ tests is not appropriate in this case. One can still base a decision on an analysis of the residuals, however. Generally the model with the smallest residual mean square and the most random-looking residuals should be chosen. Inspection of the plots mentioned above can sustain a decision.

# Chapter 3

# Generalized Linear Models

## 3.1 Introduction

The models considered in the preceding chapters can be thought of as extensions of linear regression models. Various *nonlinear* and *nonnormal* regression models have also been studied on an individual basis for many years. However, only in 1972 did Nelder and Wedderburn provide a unified and accessible theoretical and computational framework for a whole class of such models, called *generalized linear models* (GLMs), which have been of enormous influence in applied statistics. In this chapter we consider this class of models. A detailed treatment is given in McCullagh and Nelder (1989).

### 3.1.1 Basic Concepts and Examples

As usual we have $n$ observed data $y_1, \ldots, y_n$, which are realizations of the independent random response variables $Y_1, \ldots, Y_n$. Our goal is to investigate the relationship between the response variables and $p$ non-random explanatory variables or *covariates*. The $(p+1)$-dimensional vector $x_i$ denotes the vector of explanatory variables for the $i$-th observation, along with the intercept coefficient. The classical example of a model for the responses is the ordinary linear regression model, which can be represented by

$$\begin{aligned} &(i) \quad && Y_i \sim \mathrm{N}(\mu_i, \sigma^2), \\ &(ii) \quad && \eta_i = x_i^T \beta, \\ &(iii) \quad && \eta_i = g(\mu_i) = \mu_i, \end{aligned} \tag{3.1}$$

for $i = 1, \ldots, n$, where $\beta = (\beta_0, \ldots, \beta_p)^T$ is a vector of $p+1$ unknown constants, $\beta_0$ belonging to the "explanatory variable" $(1, \ldots, 1)^T$ called *intercept*. This representation may look somewhat strange at first sight. However, by choosing this representation we are able to illustrate the concept of a generalized linear model. In (3.1) we have split the model for the $Y_1, \ldots, Y_n$ into a *random component* (i) and a *systematic component* (ii). The random component of the model specifies the distribution of $Y_i$. The systematic component consists of a vector of so-called *predictors*, $\eta_i$, one for each observation. It

specifies the way in which the explanatory variables come into the model. In (iii) the so-called *link function*, denoted by $g$, specifies the connection between the random and the systematic component. More precisely, $g$ expresses $\eta_i$ as a function of $\mathrm{E}\,Y_i$, that is $g(\mathrm{E}\,Y_i) = \eta_i$. In linear regression $g$ is the identity function. Like in (3.1), in the sequel $\mathrm{E}\,Y_i$ will be denoted by $\mu_i$. The idea is to generalize the classical linear regression model by allowing other distributions than a normal one for the response variables and by allowing other link functions than the identity function for the relationship between the random and the systematic component. Part (ii) of (3.1) is not generalized: the predictors $\eta_i$ are still linear, hence the name generalized *linear* model.

**Example 3.1** Ship Damage:
An insurance company wishes to investigate the dependence of the number of reported ship damages for different ship types on the variables year of construction, aggregate months service, and period of operation. The number of reported ship damages is nonnegative and not necessarily very large. It would therefore not be realistic to use a classical linear regression model, because this would assume this number to be normally distributed and hence its expectation to be on the whole real line. In this case the assumption of a Poisson distribution could be more adequate. Moreover, since the predictor is assumed to be linear as in (3.1)(ii), the choice of the identity link function as in (3.1)(iii) is not appropriate either because it would mean that the expected number of ship damages may become negative. To avoid this a link function should be chosen which maps the positive half line on the whole real line.

**Example 3.2** Brochure:
In order to decide to whom a brochure about a new type of savings account should be sent, a bank wants to make a profile of its customers. A study is performed on the relationship between the customers' current savings and their age, income, level of education, number of money transfers over the last 10 years. A response variable $Y_i$ is used, which takes the value 0 if the $i$-th customer has less than a certain amount in her or his savings account, and 1 otherwise. Obviously, also in this case a normal distribution for the $Y_i$ is not appropriate; one would rather think of an alternative distribution. If the idea is that the probability of having less than a certain amount in one's saving account depends on the variables age, income, level of education, and number of money transfers over the last 10 years, then this relationship would not usually be linear since a straight-line relationship would imply probabilities to lie outside the legitimate range of 0 to 1 for some values of the explanatory variables. Therefore the identity link function is not suitable, and a link function should be chosen which maps the interval (0,1) onto the whole real line.

**Example 3.3** Kyphosis:
A surgeon has collected measurements on a number of children after they had undergone corrective spinal surgery. The purpose is to investigate whether and how the age of a child, the number of vertebrae involved in the operation, and the beginning of the range of the involved vertebrae affect the occurrence of a postoperative deformity called *Kyphosis*. The response variable is here the absence or presence of Kyphosis, clearly a 0-1 variable. Like the response variable in Example 3.2, the identity link function is not suitable, and a link function should be chosen which maps the interval (0,1) onto the whole real line.

**Example 3.4** HIV:
In order to quantify the spread of HIV infections in a certain region, the region is divided in several areas and for each area the number of HIV infected people per area is registered. For a homogeneous region the number of HIV infected people in each area would be distributed according to a binomial distribution with everywhere the same probability of a person being infected. However, most likely the region will not be homogeneous, but the areas will differ on a number of variables—such as age distribution of the inhabitants, total number of inhabitants, degree of urbanisation, proportion of homosexual inhabitants, etcetera—associated with the probability of an inhabitant of the area being infected. That is, there would be a relationship between the probability of a person being infected and a set of explanatory variables. As in the previous example this relationship would not be assumed to be linear, but a link function will be chosen such that the probability of a person being infected belongs to the interval (0,1) , and hence the expected number of HIV infected people per area will be between zero and its total number of inhabitants.

We note that for convenience binomially distributed observations are usually scaled, so that they are proportions instead of the numbers themselves.

As mentioned above, the link function relates the linear predictor $\eta_i$ to the expected value of $Y_i$. In classical linear models, the identity link is sensible in the sense that both $\eta_i$ and $Y_i$ can take any value on the real line. However, when we are dealing with counts and the distribution is assumed to be Poisson as in Example 3.1, we must have $\mathrm{E}\, Y_i > 0$, so that the identity link is less attractive because $\eta_i$ may then be negative. A more realistic model is obtained by using the log link,

$$g(\mu) = \log \mu. \tag{3.2}$$

With this link function additive effects contributing to $\eta_i$ become multiplicative effects contributing to $\mathrm{E}\, Y_i$. The model for the independent responses $Y_1, \ldots, Y_n$ in this case

becomes

$$\begin{aligned} &(i) && Y_i \sim \text{Poisson}(\mu_i), \\ &(ii) && \eta_i = x_i^T \beta, \\ &(iii) && \eta_i = g(\mu_i) = \log(\mu_i), \end{aligned} \tag{3.3}$$

for $i = 1, \ldots, n$. Its link function gives this model its name: it is called a *log-linear model*. We remark that the name log-linear model is not used here for a model in which $\log Y_i$ satisfies the classical linear regression model!

For 0-1 random variables or, more general, as in Example 3.4, *proportions* (of "successes") associated with binomial random variables we have $0 < \mathrm{E}\, Y_i < 1$ and, as noted in Examples 3.2 and 3.3, a link should satisfy the condition that it maps the interval (0,1) onto the whole real line. Two link functions are often used to accomplish this, namely the *logit function*

$$g(\mu) = \log[\mu/(1-\mu)], \tag{3.4}$$

and the *probit function*

$$g(\mu) = \Phi^{-1}(\mu). \tag{3.5}$$

Here $\Phi$ denotes the standard normal distribution function. The model

$$\begin{aligned} &(i) && n_i Y_i \sim Bin(n_i, \mu_i), \\ &(ii) && \eta_i = x_i^T \beta, \\ &(iii) && \eta_i = g(\mu_i) = \log[\mu_i/(1-\mu_i)], \end{aligned} \tag{3.6}$$

for $i = 1, \ldots, n$, is called a *logistic regression model*. When the model

$$\begin{aligned} &(i) && n_i Y_i \sim \text{Bin}(n_i, \mu_i), \\ &(ii) && \eta_i = x_i^T \beta, \\ &(iii) && \eta_i = g(\mu_i) = \Phi^{-1}(\mu_i) \end{aligned} \tag{3.7}$$

for $i = 1, \ldots, n$, is used, one usually speaks of a *probit regression model*. Both models are frequently used in economic and medical applications, but the logistic models seems to be somewhat more popular. For 0-1 or binomial data, we therefore restrict ourselves in this chapter to logistic regression models.

Other examples of GLMs are the one with the gamma distribution for the $Y_i$ and the reciprocal $\mu^{-1}$ as its link function, and the one with the inverse Gaussian distribution for the $Y_i$ and the squared reciprocal $\mu^{-2}$ as its link function.

**Example 3.5** Ship Damage (continued):
The insurance company used for their investigation the following type of data.

**Table 3.1**

| Ship type | Year of construction | Period of operation | Aggregate months service | Number of damage incidents |
|---|---|---|---|---|
| A | 1960–64 | 1960–74 | 127 | 0 |
| A | 1960–64 | 1960–79 | 63 | 0 |
| . | . | . | . | . |
| . | . | . | . | . |
| B | 1960–64 | 1960–74 | 44882 | 39 |
| B | 1960–64 | 1975–79 | 17176 | 29 |
| . | . | . | . | . |
| . | . | . | . | . |

There are five ship types, each represented eight times in the study. For these data a loglinear model like (3.3) would be adequate with four explanatory variables.

**Example 3.6** Kyphosis (continued):
There were 81 children examined for Kyphosis after their surgeries. The first couple of observations is

**Table 3.2**

| | Kyphosis | Age (Months) | Number | Beginning |
|---|---|---|---|---|
| 1 | absent | 71 | 3 | 5 |
| 2 | absent | 158 | 3 | 14 |
| 3 | present | 128 | 4 | 5 |
| 4 | absent | 2 | 5 | 1 |
| 5 | absent | 1 | 4 | 15 |
| . | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |

For these data we adopt the logistic regression model (3.6) with three explanatory variables, and $n_i = 1$, $i = 1, \ldots, n$. Boxplots of these data are shown in Figure 3.1. We see quite different distributions for the groups with and without Kyphosis, which means that different values of the explanatory variables may indeed yield a different response. It therefore makes sense to investigate a logistic regression model with explanatory variables age, number and beginning. We have to stress here that the boxplots are merely used here for a descriptive, summarizing purpose; it is of course not the case—as making boxplots suggests—that the explanatory variables in a GLM are random variables!

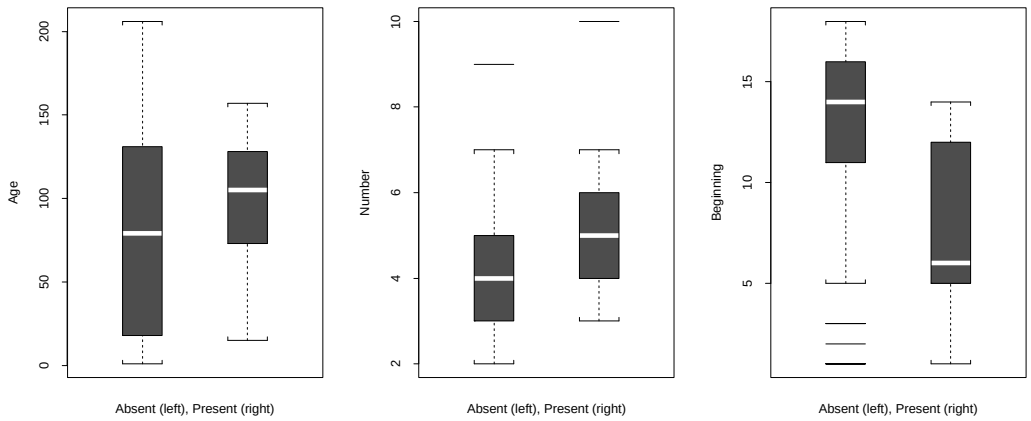


Figure 3.1: Boxplots of the Kyphosis data.

### 3.1.2 A Uniform Framework

Analogously to the models above, we can now present the general set up for the GLMs. Let there be $n$ independent random response variables $Y_1, \ldots, Y_n$ and $p$ non-random covariates. The $p$-vector $x_i$ denotes the vector of covariates for $Y_i$. For $i = 1, \ldots, n$ the relationship between $Y_i$ and $x_i$ is given by

$$\begin{aligned} &(i) && Y_i \text{ has probability density function } f_i, \\ &(ii) && \eta_i = x_i^T \beta, \\ &(iii) && \eta_i = g(\mu_i), \end{aligned} \tag{3.8}$$

where $f_i$ is the probability density function of a *one*-parameter exponential family of distributions of the form

$$f_i(y) = f_i(y, \theta_i) = \exp\left( \frac{y\theta_i - b(\theta_i)}{\phi/A_i} + c(y, \phi/A_i) \right), \tag{3.9}$$

and $b$ is an *arbitrary* monotonic, differentiable function. We remark that $\theta_i$ in (3.9) is the one parameter of the exponential family specific to $Y_i$. Since $\phi$ is a (possibly known) scale parameter common to all $Y_i$ (like $\sigma^2$ in (3.1)(i)), we are not interested to estimate it in the first place. Hence, it does not count as a parameter for the exponential family. The symbol $A_i$ in (3.9) denotes a known weight constant. The specific form of $f_i$ is determined by the functions $b$ and $c$. It can be shown that

$$\begin{aligned} &\mathrm{E}\, Y_i = \mu_i = b'(\theta_i), && i = 1, \ldots, n, \\ &\mathrm{Var}\, Y_i = b''(\theta_i)\phi/A_i, && i = 1, \ldots, n. \end{aligned} \tag{3.10}$$

Of course $b$ and $c$ have to satisfy some conditions for $f_i$ to be a density and for (3.10) to hold.

We note that (3.8) means that a GLM is completely specified by two choices, namely the choice of the exponential family and the choice of the link function, since the systematic component is the same for all GLMs.

**Example 3.7**
(a) If we set $b(\theta_i) = \frac{1}{2}\theta_i^2$, $\phi = \sigma^2$, $A_i = 1$ and $c(y_i, \phi/A_i) = -\frac{1}{2}(y_i/\sigma)^2 - \log\sigma\sqrt{2\pi}$ in (3.9), then it is easy to see that we recover the normal distribution (3.1)(i) with one parameter, the expectation $\mu_i = \theta_i$, since the scale factor $\phi = \sigma^2$ is fixed.
(b) With $b(\theta_i) = \mathrm{e}^{\theta_i}$, $\phi$ and $A_i$ being equal to 1, and $c(y_i, \phi/A_i) = -\log(y_i!)$ we find the Poisson distribution with expectation $\mathrm{e}^{\theta_i}$.
(c) If we put $b(\theta_i) = \log(1 + \mathrm{e}^{\theta_i})$, $\phi = 1$, $A_i = n_i$, $c(y_i, \phi/A_i) = -\log\binom{n_i}{n_i y_i}$, then we obtain the distribution of proportions associated with binomial random variables with parameters $n_i$ and $\frac{\mathrm{e}^{\theta_i}}{(1+\mathrm{e}^{\theta_i})}$. In the context of GLMs the parameter $n_i$ is the total number out of which the number of "successes" is counted for the $i$-th case, and thus can be considered to be fixed. This leaves one parameter. Notice that the alternative distribution is a special case of this.

## 3.2 Parameter Estimation

The expression (3.9) may falsely suggest that we have $n$ unknown parameters $\theta_1, \ldots, \theta_n$ to estimate. However, there are only $p + 1$ parameters to estimate, namely the unknown constants $\beta_0, \ldots, \beta_p$. Since the $\beta_0, \ldots, \beta_p$ are linked to the $\theta_1, \ldots, \theta_n$ through the equations (3.8)(ii), (3.8)(iii), and (3.10), the density functions $f_1, \ldots, f_n$ implicitly depend on $\beta_0, \ldots, \beta_p$. This is why a natural method for the estimation of the $\beta_0, \ldots, \beta_p$ is the maximum likelihood method.

We denote the maximum likelihood estimator (m.l.e.) of $\beta$ by $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T$. It is the value obtained by maximizing the log-likelihood

$$l(\theta) = l(\beta) = \sum_{i=1}^{n} \left( \frac{Y_i\theta_i - b(\theta_i)}{\phi/A_i} + c(Y_i, \phi/A_i) \right) \tag{3.11}$$

with respect to $\beta$. Except for the classical linear regression model, where they are the same, the least squares estimator of $\beta$ will generally have inferior performance compared to the m.l.e. $\hat{\beta}$.

In most cases an explicit expression of the m.l.e. is not available, and $\hat{\beta}$ needs to be computed numerically. Nelder and Wedderburn (1972) proposed *Fisher scoring* as a general method for the numerical evaluation of $\hat{\beta}$ in GLMs. That is, given a trial estimate $\beta^0$, update to $\beta^1$ given by

$$\beta^1 = \beta^0 + \left\{ \mathrm{E}_{\beta^0} \left( -\frac{\partial^2 l}{\partial\beta\partial\beta^T} \right) \right\}^{-1} \frac{\partial l}{\partial\beta} \tag{3.12}$$

where both derivatives are evaluated at $\beta^0$, and the expectation is evaluated as if $\beta^0$ were the true parameter value. Then $\beta^0$ is replaced by $\beta^1$ and the updating is repeated yielding $\beta^0, \beta^1, \beta^2, \ldots$ with $\beta^m$ tending to $\hat{\beta}$ when $m$ tends to infinity. The updating is stopped when $\beta^m - \beta^{m-1}$ is small enough and $\hat{\beta}$ is taken to be the final $\beta^m$. We notice that this method is very similar to the well-known Newton-Raphson method for finding a zero of the function $\frac{\partial l}{\partial \beta}$. Except here the expected value of the derivative of $\frac{\partial l}{\partial \beta}$ is used instead of the derivative itself as Newton-Raphson does.

It turns out that for a GLM, the Fisher scoring updating equations (3.12) can be written as

$$\beta^1 = (X^T W^0 X)^{-1} X^T W^0 z^0, \tag{3.13}$$

where $X$ is the matrix with $x_i^T$ as its $i$-th row, $z^0$ is the $n$-vector with $i$-th component

$$z_i^0 = (Y_i - \mu_i^0) g'(\mu_i^0) + x_i^T \beta^0, \tag{3.14}$$

and $W^0$ the $n \times n$ diagonal matrix with $i$-th diagonal element

$$w_i^0 = A_i \{g'(\mu_i^0)^2 b''(\theta_i^0)\}^{-1}. \tag{3.15}$$

Thus, for the iteration $z^0$ and $W^0$ are evaluated as if $\beta^0$ were the true parameter value. (For instance, $\mu_i^0 = \mathrm{E}_{\beta^0} Y_i$.) Expression (3.13) means that each iteration of the Fisher scoring method for numerical evaluation of the m.l.e. is in fact a weighted least squares regression for the "working response vector" $z^0$ on the model matrix $X$ with a "working weights matrix" $W^0$ (cf. weighted linear regression using a normal linear model with unequal variances). Since both $z^0$ and $W^0$ are functions of the current estimate of $\beta$, they need to be re-evaluated each iteration. From the computational perspective this is therefore an example of an *iteratively weighted least squares calculation*. Most statistical packages and subroutine libraries provide the basic routines for this computation or even for a complete GLM analysis.

Notice how the scale parameter $\phi$ has cancelled out: its value is not addressed during the iterative estimation of $\beta$.

In the normal linear model, it is readily seen that $z$ evaluated at the true parameter $\beta$ is the same as $Y$, and that $W$ evaluated at the true parameter $\beta$ is the identity matrix, so no iteration is needed and (3.13) merely confirms that the maximum likelihood and least squares coincide in this case.

***

To see the truth of (3.13), we need to evaluate the derivatives in (3.12). While doing this we will for notational simplicity omit the superscripts for the parameters, for $z$ and for $W$ when no confusion is possible. We first consider the derivatives with

respect to the linear predictor $\eta_i$:

$$\begin{aligned} \frac{\partial l}{\partial \eta_i} &= \frac{\partial l}{\partial \theta_i}\frac{d\theta_i}{d\eta_i} = \frac{\partial l}{\partial \theta_i} \Big/ \left(\frac{d\eta_i}{d\mu_i}\frac{d\mu_i}{d\theta_i}\right) \\ &= \left(\frac{Y_i - \mu_i}{\phi/A_i}\right) \Big/ \{g'(\mu_i)b''(\theta_i)\}, \end{aligned} \tag{3.16}$$

where we used (3.10) and (3.8 (iii)). Clearly $\frac{\partial^2 l}{\partial \eta_i \eta_j} = 0$ if $i \neq j$, and while $\frac{\partial^2 l}{\partial \eta_i^2}$ involves higher derivatives of $g$ and $b$, we see that its (negative) expectation does not:

$$\begin{aligned} \mathrm{E}_{\beta^0}\left(-\frac{\partial^2 l}{\partial \eta_i^2}\right) &= \frac{d\mu_i}{d\eta_i} \Big/ \frac{\phi}{A_i}\{g'(\mu_i)b''(\theta_i)\} \\ &= A_i\{\phi g'(\mu_i)^2 b''(\theta_i)\}^{-1}. \end{aligned} \tag{3.17}$$

Let $z^*$ be the $n$-vector with $z_i^* = (Y_i - \mu_i)g'(\mu_i)$. Then, in summary, we have from (3.16)

$$\phi\frac{\partial l}{\partial \eta} = Wz^* \tag{3.18}$$

and from (3.17)

$$\phi\,\mathrm{E}_{\beta^0}\left(-\frac{\partial^2 l}{\partial \eta \partial \eta^T}\right) = W. \tag{3.19}$$

But by the chain rule, since $\eta = X\beta$, we have

$$\frac{\partial l}{\partial \beta} = X^T\frac{\partial l}{\partial \eta} \tag{3.20}$$

and

$$\mathrm{E}_{\beta^0}\left(-\frac{\partial^2 l}{\partial \beta \partial \beta^T}\right) = X^T\,\mathrm{E}_{\beta^0}\left(-\frac{\partial^2 l}{\partial \eta \partial \eta^T}\right)X. \tag{3.21}$$

Thus the Fisher scoring equations (3.12) become

$$\beta^1 = \beta^0 + (X^TWX)^{-1}X^TWz^* \tag{3.22}$$

which we can more simply express in the required form (3.13).
***

## 3.3 Inference in GLMs

Analysis of data based on a generalized linear model means more than merely estimating the regression coefficients. To judge the quality of a GLM among other things, we need estimates of standard deviations of estimators, confidence intervals for the parameters, measures of goodness-of-fit, and methods for model selection and other hypothesis testing.

In order to assess the accuracy of the maximum likelihood estimate $\hat{\beta}$ one could consider the asymptotic variance matrix of the estimator $\hat{\beta}$, which is given by the inverse of the Fisher information matrix

$$\left\{ \mathrm{E}_\beta \left( -\frac{\partial^2 l}{\partial\beta\partial\beta^T} \right) \right\}^{-1} = \phi(X^T W X)^{-1}, \tag{3.23}$$

where $W$ is the $n \times n$ diagonal matrix with $i$-th diagonal element

$$w_i = A_i \{ g'(\mu_i)^2 b''(\theta_i) \}^{-1}. \tag{3.24}$$

Of course, when using this variance, $W$ will be evaluated at the maximum likelihood estimate $\hat{\beta}$, i.e. it will be replaced by $\hat{W}$, and $\phi$ will be replaced by an estimate also. For the latter one could for instance use

$$\hat{\phi} = \frac{\chi^2}{(n-p-1)}, \tag{3.25}$$

where $\chi^2$ is the so-called *(generalized) Pearson chisquared statistic* given by

$$\chi^2 = \phi \sum_{i=1}^{n} \frac{\{Y_i - \mathrm{E}_{\hat{\beta}}\, Y_i\}^2}{\mathrm{Var}_{\hat{\beta}}(Y_i)} = \sum_{i=1}^{n} \frac{(Y_i - b'(\hat{\theta}_i))^2}{b''(\hat{\theta}_i)/A_i}. \tag{3.26}$$

where $\mathrm{E}_{\hat{\beta}}\, Y_i$ and $\mathrm{Var}_{\hat{\beta}}(Y_i)$, as the notation suggests, are evaluated at the maximum likelihood estimate $\hat{\beta}$ (cf. (3.10)). It can be easily checked that this estimator coincides with the usual estimator for $\sigma^2$ in the classical linear regression model.

Based on the asymptotic normality of maximum likelihood estimators approximate confidence intervals for the parameters can be constructed in the usual way. The estimates of the asymptotic variance of the estimators described above are used for this. An approximate $(1-\alpha)100\%$ confidence interval for $\beta_j$ is given by

$$\hat{\beta}_j \pm t_{(n-p-1);(1-\alpha/2)} \sqrt{\hat{\phi}((X^T \hat{W} X)^{-1})_{jj}}. \tag{3.27}$$

In order to check whether an explanatory variable should be included in the model in practice one computes, like for the nonlinear models, the corresponding $t$-ratio, which is the estimate divided by the estimated standard deviation, and compares this $t$-ratio with the plus or minus $(1-\alpha/2)$-point of the $t$-distribution with $(n-p-1)$ degrees of freedom. If the $t$-ratio is not significant, this is an indication that the corresponding explanatory variable could perhaps be deleted from the model.

**Example 3.8** Kyphosis (continued):
For the logistic model for the Kyphosis data we find the vector of estimates $\hat{\beta} = (-2.037, 0.0109, 0.4106, -0.2065)^T$ for the intercept and the variables age, number and beginning. The corresponding vector of estimated standard deviations of the estimators based on (3.23) is $\sqrt{\widehat{\mathrm{Var}}(\hat{\beta})} = (1.4492, 0.0064, 0.2248, 0.0677)^T$; $\phi$ is taken to be 1. The approximate 95% confidence intervals for $\beta_0, \ldots, \beta_3$ are (-4.9226, 0.8488), (-0.0019, 0.0238), (-0.0370, 0.8582), and (-0.3413, -0.0717), respectively. The $t$-ratios for the intercept, age, number and beginning are -1.4056, 1.6962, 1.8266, and -3.0510. The 0.975-point of the $t$-distribution with $(81 - 4) = 77$ degrees of freedom is 1.99. Hence only the $t$-ratio for the variable 'beginning' is significant at the 0.05 level. Of course, this could also be concluded from the confidence intervals: all intervals contain the value zero, but the one for the parameter corresponding to the variable 'beginning' does not. This suggests that the only important explanatory variable is the variable 'beginning'.

We now briefly summarize some other basic methods for checking the quality of the model.

The global quality of a GLM with the m.l.e. $\hat{\beta}$ as its parameter value can for instance be assessed by considering the difference between this model and the largest possible, full or *saturated*, model. The saturated model is the model with $n$ explanatory variables—that is one parameter value for each observation—which of course will fit perfectly: for the saturated model holds that

1. $\hat{Y} = Y$,
2. the residual sum of squares equals zero,
3. the log-likelihood is maximized for those values of the parameters,
   $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_n)$ say, that yield $Y$ as predicted response,
   or $b'(\tilde{\theta}_i) = Y_i,\ i = 1, \ldots, n$.

A quantity that measures the difference between these two models is the so-called *deviance* $D$, which is defined as the scaled log-likelihood-ratio statistic

$$\begin{aligned} D &= 2\phi[l(\tilde{\theta}) - l(\hat{\theta})] \\ &= 2\sum_{i=1}^{n} A_i\{Y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\}. \end{aligned} \tag{3.28}$$

The contribution of the $i$-th observation to the deviance, $d_i = A_i\{Y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\}$, is sometimes called the $i$-th *deviance increment*. The deviance $D$ is a global measure of the closeness of the fit of the model to the data, and can be interpreted very much like

the residual sum of squares in the classical linear regression model, which is indeed what it reduces to for that case. A smaller model will generally have a larger deviance than a larger model. An alternative measure for the quality of fit of a model to the data is provided by the Pearson chisquared statistic $\chi^2$ given by (3.26).

Both the deviance and the Pearson chisquared statistic have, after being divided by $\phi$, exact (central) chisquared distributions with $n-p-1$ degrees of freedom under the classical linear regression model. For other GLMs only asymptotic results are available. For example, under a logistic regression model, both the deviance and the Pearson chisquared statistic have an asymptotic (for $n$ tending to infinity) chisquared distribution with $n-p-1$ degrees of freedom (here $\phi = 1$). For other examples and references see McCullagh and Nelder (1989). It has to be stressed, however, that these asymptotic results may be useless when $D$ or $\chi^2$ are calculated from a limited amount of data.

The main merit of the deviance and Pearson chisquared statistic lies in their usefulness for comparing two *nested* models. For model selection between nested models we may use $D$ or $\chi^2$ in a similar manner as the sums of squares in an analysis of variance. For $D$ this procedure is known as the *analysis of deviance*. If the deviances devided by $\phi$ are asymptotically chisquared distributed, then $1/\phi$ times the difference of the deviances of two nested models again is asymptotically chisquared distributed with number of degrees of freedom equal to the difference in number of parameters of the models; this distribution is central under the smaller model. This means that in this case an approximate test can be performed to test the hypothesis that the smaller model is adequate. This test uses the difference of the deviances as its test statistic, which has under the hypothesis that the smaller model suffices a (central) chisquared distribution with number of degrees of freedom equal to the difference in the number of parameters of the large and the small model. In general, even if asymptotic chisquared distribution results do not apply, a large reduction in $D$ or $\chi^2$ upon using the larger model is an indication that the larger model should be preferred.

**Example 3.9** Kyphosis (continued):
Since the number of observations is not very small, we can assume that the deviance for this data approximately follows a chisquared distribution. We can thus compare the full model containing the three explanatory variables and smaller models, which are nested in the full one. Let us consider the following analysis of deviance table [1].

**Table 3.3**

[1] We note that in the 'Test' column the terms comprising the difference between each model and the one above it are named, in the 'Deviance' column the difference in deviances is reported, and 'Df' denotes the difference in degrees of freedom. This is often done when adjacent models are nested, as is the case here.

Response: Kyphosis

| Terms | Resid. Df | Resid. Dev | Test | Df | Deviance |
|---|---|---|---|---|---|
| Age + Number + Beginning | 77 | 61.37993 | | | |
| Number + Beginning | 78 | 64.53647 | -Age | -1 | -3.156541 |
| Beginning | 79 | 68.07218 | -Number | -1 | -3.535712 |

We see from the table that if we compare the largest model with all variables to the smaller one with age deleted, or compare the model with two variables to the smallest one with only the variable 'beginning', in both cases we do not reject the hypothesis that the smaller model is better ($\chi_{1;0.95} = 3.84$). However, if we would have directly compared the largest model with all variables to the smallest one with only the variable 'beginning', the hypothesis that the smaller model is better would have been rejected ($\chi_{2;0.95} = 5.99$; the test statistic being $3.156541 + 3.535712 = 6.692253 > 5.99$). This contradiction probably has to do with the fact that the test statistics are only *approximately* chisquared distributed. Although it seems that the model with only the variable 'beginning' is to be preferred, from a statistical point of view one has to be somewhat careful with this conclusion.

Actually, it turns out that the model can even be reduced further: the intercept can be left out, since its $t$-ratio in the model with 'beginning' is not significant and also the approximate test for the small model without intercept does not reject. The final model could thus be chosen to be a logistic model without intercept, with the variable 'beginning' and estimated parameter value for $\beta_1$ equal to -0.1446.

Apart from global measures of fit, we also need diagnostic checks. For a local assessment of goodness-of-fit, to be used in checking both model and data adequacy (e.g. absence of outliers), an informative plot for some of the GLMs is the fitted expected responses versus the responses. The points should more or less lay on the line $y = x$. However, this does not hold for 0-1 response variables.

Next, it is useful to examine the residuals $Y_i - \hat{Y}_i = Y_i - \hat{\mu}_i = Y_i - b'(\hat{\theta}_i)$ or, preferably, some standardization of them. Two commonly used definitions for the $i$-th standardized residual are

$$R_i^D = \text{sign}(Y_i - \hat{\mu}_i)\sqrt{d_i}, \tag{3.29}$$

and

$$R_i^P = \frac{Y_i - \hat{\mu}_i}{\sqrt{b''(\hat{\theta}_i)/A_i}}, \tag{3.30}$$

where $d_i$ is the deviance increment defined above. These are known as the *deviance residuals* and *Pearson residuals* respectively.

To conclude this chapter we mention three types of plots involving residuals that give information about possible systematic departures from the chosen model. Interpretation of these plots, however, is quite difficult and needs a lot of experience. In particular, it depends on the specific GLM that is used. For more details on this subject we refer to

the literature.

(i) Residuals are plotted against the corresponding fitted values which are transformed to such a scale that the error distribution has constant variance. This means that we use for instance

$$\begin{aligned} \hat{\mu}_i & \quad \text{for normal errors,} \\ 2\sqrt{\hat{\mu}_i} & \quad \text{for Poisson errors,} \\ 2\sin^{-1}\sqrt{\hat{\mu}_i} & \quad \text{for binomial errors.} \end{aligned}$$

The plot should yield a more or less straight line for the normal distribution. For other distributions the contours are curves, but give a slope $-1$ at the point zero for the residual, and the curvature should be very small. Larger curvature may arise from several causes, including the wrong choice of link function, wrong scale of one or more covariates (that is a transformation of this/these covariates would perform better), or omission of a quadratic or higher power term in a covariate. Ways of distinguishing these can be found in McCullagh and Nelder (1989). It is recommended to use the deviance residuals for this plot. Note that this plot is generally uninformative for binary data because all the points will more or less lie on one of two curves depending on whether $y_i = 0$ or $y_i = 1$.

(ii) A second plot for informal model checking plots the residuals against an explanatory variable. When the chosen model is correct, the pattern of the graph should look like the ones described under the first plot. Again the appearance of systematic trend may indicate the wrong choice of link function, the wrong scale of the explanatory variable, or point to a missing quadratic or higher power term in a covariate.

(iii) A third plot, known as an *added variable plot*, gives a check on whether an omitted covariate, $u$ say, should be included in the linear predictor. It is not adequate to plot the residuals against $u_i$ itself for this purpose. First we must obtain the unstandardized residuals for $u$ as response, using the same linear predictor as for $Y$. The unstandardized residuals for $Y$ are then plotted against the unstandardized residuals for $u$. If $u$ is correctly omitted, no trend should be apparent.

**Example 3.10** Kyphosis (continued):
For the Kyphosis data a plot of the fitted values against the responses is uninformative. Also the first two residual plots are difficult to interpret. In Figure 3.2 we see an added variable plot for the variable 'beginning' for the model that only contains the number and age (left). The appearance of a somewhat strange, nonrandom pattern, could indicate that the variable 'beginning' was incorrectly omitted from the

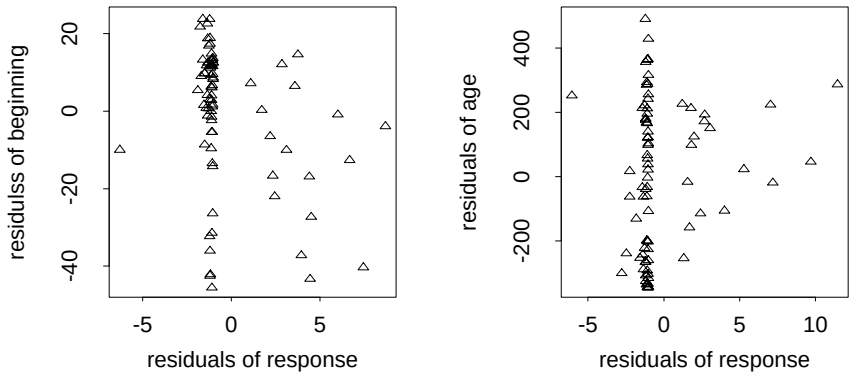


Figure 3.2: Added variable plots for the Kyphosis data: beginning on age and number (left), and age on number and beginning (right).

model. On the right of Figure 3.2, however, we see a similar pattern in the added variable plot for the variable 'age' for the model that only contains the number and beginning. Since we had earlier concluded that is was justified to omit age, we have to investigate what is going on here. It turns out that both iteration procedures for the explanatory variables on the other two were terminated prematurely, because of computational difficulties (taking logarithms of 0). These plots therefore cannot be trusted. This illustrates that one never should carelessly accept all computational output!

# Chapter 4

# Time Series

## 4.1 Introduction

A stochastic *time series* is a sequence $\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$, also denoted by $\{X_t\}$, of random variables[1] which represents a set of observations made sequentially in time. The subscript $t$ ($t = \ldots, -2, -1, 0, 1, 2, \ldots$) indicates the time at which the variable $X_t$ was observed. This notation suggests that the observations were made at equally spaced time points, which we shall assume to be the case indeed[2]. Often a time series consists of repeated observations on the same object. This is one reason why the usual assumption of independent random variables is in general not adequate for time series: time series analysis deals with the statistical analysis of a *sequence* of *dependent* random variables. Thus the order of the random variables is essential. For instance, it often happens that observations made "closely together in time" are dependent, whereas observations made at time points far apart are almost independent. Figures 4.1.a–4.1.d show some typical examples of time series.

The goal of a time series analysis could be to give a concise summary of the data, for example in order to be able to detect changes in a process. A summary of a time series usually consists of one or more graphical representations of the data, rather than of a number of summary statistics. This is because the nature of a time series, as opposed to a random sample from a distribution, is often such that it requires a *function* rather than a single number to summarize each of the series' essential features. For example, the description of the mean value of a series by means of a function of time makes more sense than one by a single quantity. (Of course, apart from the purpose of summarizing, making plots is as important for getting insight in the data as it is for independent observations.) Another goal could be to predict future values based on observed values, for instance for sales data. One could also be interested to control future values by means of adjusting parameters, like in the case of global warming, one may want to generate simulations of a process, or one's aim could be to separate noise from signals in processes where noise

[1] The $X_t$'s can also be random vectors. In this chapter we only consider random variables.

[2] If the observation times are not equally spaced in time, a better notation would be $X_{t_1}, X_{t_2}, \ldots$.

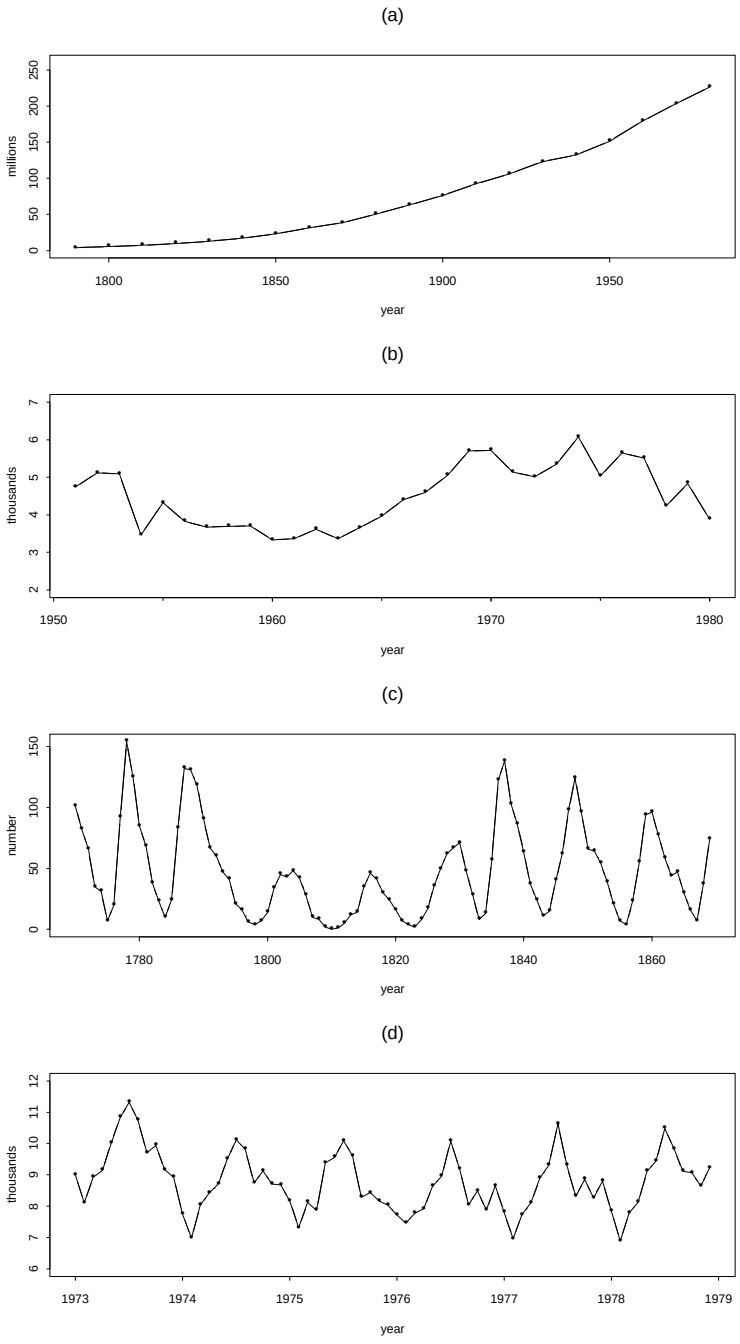


Figure 4.1: a) Population of the U.S.A. at ten-year intervals, 1970–1980 (U.S. Bureau of the Census). b) Strikes in the U.S.A., 1951–1980 (Bureau of Labor Statistics, U.S. Labor Department). c) The Wölfer sunspot numbers. d) Monthly accidents deaths in the U.S.A., 1973–1978 (National Safety Council).

blurs the signal of interest. Numerous examples of different objectives could be given.

For purposes more complex than just giving a suitable summary, one has to make some assumptions about the structure of the data, that is one has to assume some kind of model, in order to be able to perform a proper time series analysis. Indeed, if the $X_t$ could be completely arbitrary, then it would be impossible to infer something about the distribution of the observations $(X_1, \ldots, X_t)$, let alone of that of the future $X_{t+1}, X_{t+2}, \ldots$. Once a model is chosen, unknown model parameters can be estimated, the goodness of fit to the data can be investigated, and—hopefully—the model can help understanding the mechanisms that generated the time series.

## 4.2 Stationarity, Trend, Seasonality

One of the most important concepts in the context of modelling time series is *stationarity*, which occurs in two forms.

**Definition 4.1** *A time series is strictly stationary if, for any value of $k$, the joint distribution of $(X_{t+1}, \ldots, X_{t+k})$ does not depend on $t$.*

**Definition 4.2** *A time series is second order stationary or weakly stationary if, for any value of $k$, $\mathrm{E}\, X_t$ and $\mathrm{E}\, X_t X_{t+k}$ exist and do not depend on $t$.*

Strict stationarity is a strong and, in practice, uncheckable assumption. For most practical purposes the assumption of weak stationarity is sufficient. When we use in what follows the term stationary, we mean weakly stationary.

Since most time series show dependence for at least some pairs $(X_s, X_t)$, a concept that describes the degree of dependence is a useful modelling tool. The so-called autocovariance function can play this role.

**Definition 4.3** *If $\{X_t\}$ is a process with $\mathrm{Var}\, X_t < \infty$ for each $t$, then its autocovariance function $\gamma_X(\cdot, \cdot)$ is defined by*

$$\gamma_X(s,t) = \mathrm{Cov}(X_s, X_t) = \mathrm{E}[(X_s - \mathrm{E}\, X_s)(X_t - \mathrm{E}\, X_t)]. \tag{4.1}$$

It is easy to see that for a stationary process $\{X_t\}$ the autocovariance function satisfies $\gamma_X(s,t) = \gamma_X(s-t, 0)$ for all $s$ and $t$. It is therefore convenient to redefine the autocovariance function of a stationary process $\{X_t\}$ as a function of only one variable,

$$\gamma_X(h) = \gamma_X(h,0) = \mathrm{Cov}(X_{t+h}, X_t), \qquad h \in \mathbb{Z}, \tag{4.2}$$

which is independent of $t$. The function $\gamma_X(.)$ will be referred to as the autocovariance function of $\{X_t\}$ and $\gamma_X(h)$ as its value at lag $h$. The *autocorrelation function* of $\{X_t\}$ is a scaled version of the autocovariance function. It is defined as follows.

**Definition 4.4** *The autocorrelation function at lag $h$ of a stationary process $\{X_t\}$ is*

$$\rho_X(h) = \mathrm{Cor}(X_{t+h}, X_t) = \frac{\gamma_X(h)}{\gamma_X(0)}, \qquad h \in \mathbb{Z}, \tag{4.3}$$

*independently of $t$.*

Note that for a stationary process $\{X_t\}$, $\gamma_X(0)$ is the variance of $X_t$ and $\rho_X(0) = 1$.

Another useful tool is the *partial autocorrelation function*, denoted by $\alpha(\cdot)$. We shall not give a formal definition. The partial autocorrelation function at lag $h$, $\alpha(h)$, may be regarded as the correlation between $X_t$ and $X_{t+h}$ adjusted for the intervening observations $X_{t+1}, \ldots, X_{t+h-1}$. As we shall see later on, its natural estimator may help deciding which class of models to choose.

**Example 4.1** The simplest example of a stationary random sequence is *white noise*. It consists of a sequence of independent random variables[3] $\ldots, Z_{-1}, Z_0, Z_1, \ldots$ each with mean zero and finite variance $\sigma^2$. Its autocovariance function is

$$\gamma_Z(h) = \begin{cases} \sigma^2, & h = 0, \\ 0, & h \neq 0. \end{cases} \tag{4.4}$$

Because its random variables are independent, white noise itself is not very useful as a model for time series. However, it is often used as a building block to construct time series of which the random variables are dependent. Such series are more interesting for practical purposes since they are more realistic and with one of them as a model the future can to some extent be predicted from the past. Figure 4.2 shows two examples of realizations of a white noise process.

**Example 4.2** Suppose that $\{Z_t\}$ is a white noise process with $\sigma^2$ as the variance of $Z_t$, and let $\beta$ be a constant. The process $\{X_t\}$ defined by

$$X_t = Z_t + \beta Z_{t-1}, \tag{4.5}$$

[3]White noise is sometimes defined as a sequence of random variables that are uncorrelated each with mean zero and finite variance $\sigma^2$, or as a sequence of random variables that are independent and identically distributed each with mean zero and finite variance $\sigma^2$. Both definitions yield the same autocovariance function as given in (4.4). Since stationarity was defined in terms of means and second moments, the use of one of these alternative definitions makes no difference for the theory and definitions later on in this chapter.

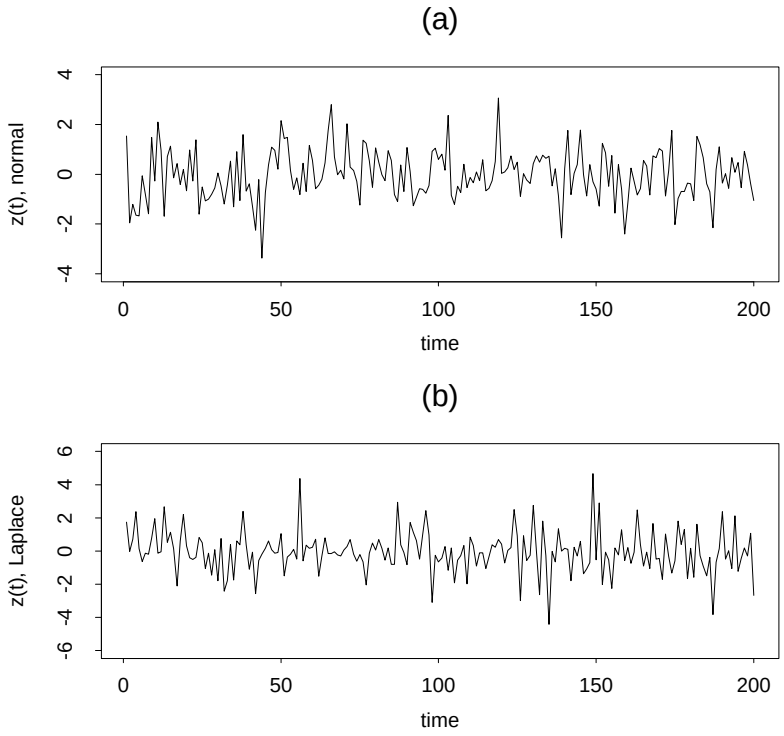


Figure 4.2: Simulated realizations of two white noise processes. a) Normally distributed. b) Laplace distributed.

is called a *moving average* process of order one. This process is stationary with $\mathrm{E}\, X_t = 0$ and autocovariance function

$$\begin{aligned} \gamma_X(h) &= \operatorname{Cov}(Z_{t+h} + \beta Z_{t+h-1}, Z_t + \beta Z_{t-1}) \\ &= \begin{cases} \sigma^2(1+\beta^2), & h = 0, \\ \sigma^2\beta, & h = \pm 1, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{4.6}$$

Thus $X_s$ and $X_t$ are uncorrelated whenever $s$ and $t$ are at least two time units apart. This is an example of short range dependence. Usually the constant $\beta$ is unknown and needs to be estimated from the data.

**Example 4.3** Let

$$X_t = \begin{cases} Y_t, & \text{if } t \text{ is odd,} \\ Y_t + 1, & \text{if } t \text{ is even,} \end{cases} \tag{4.7}$$

where $\{Y_t\}$ is a stationary time series. Then $\operatorname{Cov}(X_{t+h}, X_t) = \gamma_Y(h)$, which does not depend on $t$ because $\{Y_t\}$ is stationary. However, since $\{X_t\}$ does not have a constant expectation, it is not a stationary process.

**Example 4.4** Let $X_0$ be identically zero and define the random sequence $\{X_t\}$ by the recursion

$$X_t = X_{t-1} + Z_t, \qquad t = 1, 2, \ldots, \tag{4.8}$$

where $\{Z_t\}$ is a white noise process with $\operatorname{Var} Z_t = \sigma^2$. The sequence $\{X_t\}$ so defined is an example of the type of stochastic process known as *random walk*. Since (4.8) can also be written as

$$X_t = X_0 + \sum_{i=1}^{t} Z_i, \tag{4.9}$$

we see that $\mathrm{E}\, X_t = 0$ and $\gamma_X(0) = \operatorname{Var} X_t = t\sigma^2$. This means that $\{X_t\}$ is not a stationary process. We remark that the process of (first order) differences $\{X_t - X_{t-1}\} = \{Z_t\}$ is stationary, since $\{Z_t\}$ was assumed to be a white noise process.

It is easiest to analize stationary series. This is why one usually tries in some way or another to first obtain a stationary series out of an originally non-stationary one. Two important causes for non-stationarity are *trend* and *seasonality*. They are both defined in terms of the behavior of the mean level of the process. Trend is a long-term steady increase or decrease in the mean level of the process. For example, the process in Figure 4.1.a shows a trend. A time series contains a seasonal component if it shows a cyclic change in the mean level. This can, for instance, consist of annual or monthly variation (see

Figure 4.1.d). Thus before the theory of stationary stochastic processes can be used for a time series analysis, the first thing to do is to remove from the data trend and seasonal components, if present. However, it has to be kept in mind that in some time series deterministic trend and/or seasonality are more important than stochastic variations. In that case it could be better to fit a curve to the series—for instance by least squares while using a nonlinear model—instead of carrying out a stochastic time series analysis. Since it is in general much more difficult to model a time series than data from replicated experiments, the results of a time series analysis should not be overrated. Conclusions are often already obvious by just using common sense.

### 4.2.1 Estimation and Elimination of Trend and Seasonal Components

After getting background information and defining the goal of the analysis, the first and most important step in any time series analysis is to plot the observations against time: a *time plot* is a plot of the observed values of $X_t$ against their observation times $t$. This graph should show the important features of the series such as trend, seasonality, outliers and discontinuities like sharp changes in global behavior. Plotting a time series is not as easy as it sounds. The choice of scales, the size of the intercept, and the way the points are plotted (e.g. as a continuous line or as separated dots) may substantially affect the way the plot looks.

Let us assume that the time plot suggests that the data are a realization of the (classical decomposition) model

$$X_t = m_t + s_t + Y_t, \tag{4.10}$$

where $m_t$ is a (deterministic) trend component, $s_t$ is a (deterministic) seasonal component with known period $d$ and $Y_t$ is the random noise component which is assumed to be stationary. Without loss of generality we may assume that $\mathrm{E}\, Y_t = 0$. If the seasonal and noise fluctuations appear to increase with the level of the process, then a preliminary transformation (such as taking logarithms or square roots) of the data is often used to make the transformed data compatible with the model (4.10).

The aim is to estimate the deterministic components $m_t$ and $s_t$, so that after subtraction of the estimates from $X_t$, the residual noise component $Y_t$ will be a stationary random process. The theory of stationary processes can then be used to find a suitable model for the process $\{Y_t\}$, to analyse its properties, and to use it together with $\{m_t\}$ and $\{s_t\}$ for purposes of prediction, control, simulation, etcetera, of $\{X_t\}$.

An alternative approach to obtain a stationary process is, as was already illustrated in Example 4.4, to take differences repeatedly to the data $\{X_t\}$ until the differenced observations resemble a realization of some stationary process. This is in general faster and simpler than first determining estimates for trend and seasonal components. For

some applications, however, explicit knowledge of these estimates is needed and then taking differences is less suitable.

### 4.2.2 Estimation and Elimination of Trend in Absence of Seasonality

In the absence of a seasonal component the model (4.10) becomes

$$X_t = m_t + Y_t, \tag{4.11}$$

where again we assume that $\mathrm{E}\, Y_t = 0$. We discuss three methods for the elimination of $m_t$ from the data, $X_1, \ldots, X_n$ say. The first two methods provide estimators $\hat{\mu}_t$ for $m_t$, so that $\hat{\mu}_t$ can be subtracted from $X_t$ to obtain estimated values $\hat{Y}_t$ of the stationary noise process $\{Y_t\}$; the last method is a differencing method which yields a stationary process straight away.

*(i) Least squares estimation of $m_t$*
In this procedure one fits a parametric family of functions $m_t = f(a, t)$ to the data by choosing the parameter (vector) $a$ such that it minimizes $\sum_{t=1}^{n}(X_t - m_t)^2$. For example, $f(a, t)$ is a polynomial like $m_t = a_0 + a_1 t + a_2 t^2$, where $a = (a_0, a_1, a_2)^T$ is the unknown parameter vector. If the resulting estimated function of $m_t$ is $\hat{m}_t$, then the residuals $\hat{Y}_t = X_t - \hat{m}_t$, $t = 1, \ldots n$, should look like realizations from a stationary process. Moreover, $\{\hat{m}_t\}$, $t > n$, is a natural predictor of the future of the process $\{X_t\}$. For instance, if we estimate $Y_{n+1}$ by its mean value, i.e. zero, then $\hat{m}_{n+1}$ is an estimator for $X_{n+1}$. However, if the residuals $\{\hat{Y}_t\}$ are highly correlated, it may be possible to obtain a better estimate of $Y_{n+1}$, and, with that, for $X_{n+1}$.

**Example 4.5** Fitting a function of degree two to the population data of Figure 4.1.a, $1790 \le t \le 1980$, gives the estimated parameter values $\hat{a}_0 = 2.097911 \times 10^{10}$, $\hat{a}_1 = -2.334962 \times 10^7$, and $\hat{a}_2 = 6.498591 \times 10^3$. A graph of the fitted function is shown with the original data in Figure 4.3. The estimated values of the noise process $\{Y_t\}$ are the residuals $\{\hat{Y}_t\}$ obtained by subtraction of $\hat{m}_t = \hat{a}_0 + \hat{a}_1 t + \hat{a}_2 t^2$ from $\{X_t\}$.

*(ii) Smoothing by means of a moving average*
Let $q$ be a nonnegative integer, and let $W_t$ be defined by

$$W_t = (2q+1)^{-1} \sum_{j=-q}^{q} X_{t+j}, \tag{4.12}$$

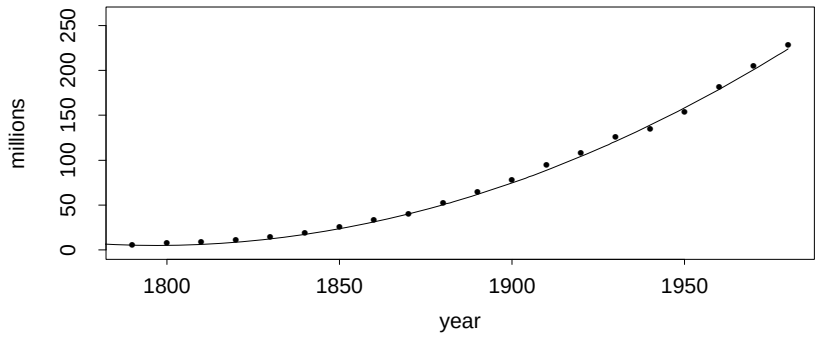


Figure 4.3: Population of the U.S.A., showing the parabola fitted by least squares.

where $\{X_t\}$ is as in (4.11). The process $\{W_t\}$ is called a two-sided *moving average*. Combining (4.11) and (4.12) we find that for $q+1 \le t \le n-q$,

$$W_t = (2q+1)^{-1} \sum_{j=-q}^{q} m_{t+j} + (2q+1)^{-1} \sum_{j=-q}^{q} Y_{t+j}. \tag{4.13}$$

Assuming that $m_t$ is approximately linear over the interval $[t-q, t+q]$ and that the average of the random noise components $Y_{t+j}$ over this interval is close to zero, we see that $W_t$ is approximately equal to $m_t$. The moving average thus provides us with the estimators

$$\hat{m}_t = (2q+1)^{-1} \sum_{j=-q}^{q} X_{t+j}, \qquad q+1 \le t \le n-q \tag{4.14}$$

for the trend function $m_t$. Since $X_t$ is not observed for $t \le 0$ or $t > n$, (4.14) cannot be used for $t \le q$ or $t > n-q$, which makes this method less suitable for prediction or simulation purposes.

**Example 4.6** The results of applying this procedure to the strike data of Figure 4.1.b are shown in Figure 4.4.a. The estimated noise terms, $\hat{Y}_t = X_t - \hat{m}_t$, are shown in Figure 4.4.b. As expected, they show no apparent trend.

*(iii) Differencing to generate stationary data*
Let the difference operator $\nabla$ be defined by

$$\nabla X_t = X_t - X_{t-1}. \tag{4.15}$$

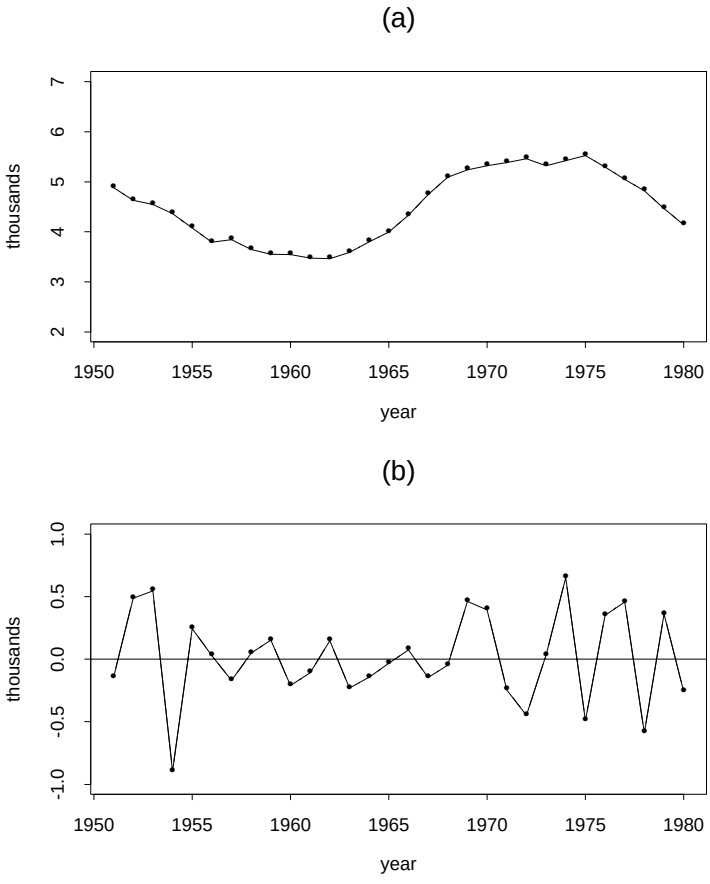


Figure 4.4: Strike data from Figure 4.1.b. a) Simple 5-term moving average $\hat{m}_t$. b) Residuals, $Y_t - X_t - \hat{m}_t$, after subtracting the 5-term moving average from the data.

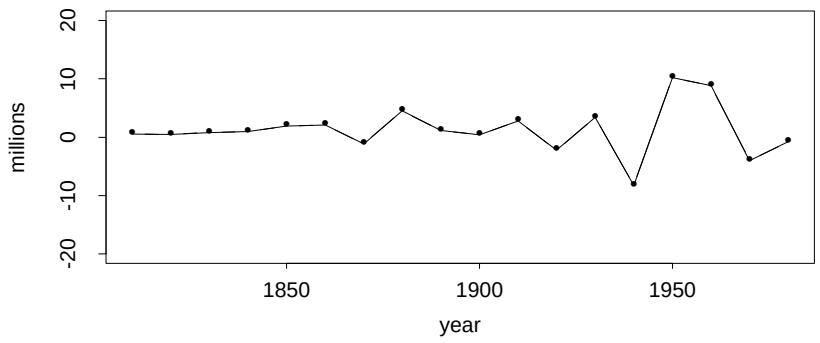


Figure 4.5: The twice differenced series derived from the population data of Figure 4.1a.

If, for instance, we have a linear trend function $m_t = at + b$, then $\{X_t\} = \{m_t + Y_t\}$ is not stationary. However, the sequence $\{\nabla X_t\} = \{a + Y_t - Y_{t-1}\}$ is stationary. In the same way any polynomial trend of degree $k$ can be reduced to a constant by application of the difference operator $k$ times. This suggests the application of the operator $\nabla$ to any sequence $\{X_t\}$ repeatedly until a sequence $\{\nabla^k X_t\}$ is found which can plausibly be modelled as a realization of a stationary process. It is often experienced in practice that the required order $k$ of differencing is rather small, frequently one or two.

**Example 4.7** Applying this technique to the twenty population values of Figure 4.1a. we see that two differencing operations are sufficient to produce a series with no clear trend. The differenced data, $\nabla^2 X_n = X_n - 2X_{n-1} + X_{n-2}$, are plotted in Figure 4.5. Notice that the fluctuations in $\nabla^2 X_n$ increase with the value of $X_n$. This phenomenon can be eliminated by first taking logarithms and then applying the operator $\nabla^2$ to the series $\{\log X_t\}$.

### 4.2.3 Estimation and Elimination of both Trend and Seasonality

The methods described above for removal of the trend can be adapted in a natural way to eliminate both trend and seasonality in the general model (4.10), where we assume additionally that $\mathrm{E}\, Y_t = 0$, and where $s_{t+d} = s_t$, and $\sum_{j=1}^{d} s_j = 0$ necessarily have to hold. We illustrate these methods by applying them to the accident data shown in Figure 4.1.d for which the period $d$ of the seasonal component is clearly 12.

*(I) The small trend method*
For this method it will be convenient to index the data by year and by month. Thus $X_{j,k}$,

$j = 1, \ldots, 6$, $k = 1, \ldots, 12$, will denote the number of accidental deaths reported for the $k$-th month of the $j$-th year, (1972+$j$), that is $X_{j,k} = X_{k+12(j-1)}$.

If the trend is small—as is the case in the accident data—it is not unreasonable to assume that the trend is constant (does not increase or decrease) within a year. Let us say that the trend equals $m_j$ for the $j$-th year. Since $\sum_{k=1}^{12} s_k = 0$, we have for $m_j$ the natural, unbiased, estimator

$$\hat{m}_j = \frac{1}{12} \sum_{k=1}^{12} X_{j,k}, \qquad j = 1, \ldots, 6. \tag{4.16}$$

Note that this is precisely the least squares estimator for the expectation of the series for each year separately, with the polynomial taken to be a constant.

For $s_k$ we can use the estimate

$$\hat{s}_k = \frac{1}{6} \sum_{j=1}^{6} (X_{j,k} - \hat{m}_j), \qquad k = 1, \ldots, 12, \tag{4.17}$$

which indeed satisfies $\sum_{k=1}^{12} \hat{s}_k = 0$. The estimated random noise component for month $k$ of the $j$-th year is then

$$\hat{Y}_{j,k} = X_{j,k} - \hat{m}_j - \hat{s}_k, \qquad j = 1, \ldots, 6, \ k = 1, \ldots, 12. \tag{4.18}$$

For periods other than 12 the method can be applied analogously.

**Example 4.8** In Figures 4.6.a, 4.6.b, 4.6.c are plotted the detrended observations $(X_{j,k} - \hat{m}_j)$, the estimated seasonal components $\hat{s}_k$, and the detrended, deseasonalized observations $\hat{Y}_{j,k}$, respectively. The latter have no apparent trend or seasonality.

*II) Moving average estimation*

The following method is in general preferable to method (I) because it does not rely on the assumption that $m_t$ is nearly constant over each cycle.

Suppose that we have observations $X_1, \ldots, X_n$. The first step is to estimate the trend by applying a moving average which is chosen such that the seasonal component is eliminated and the noise is dampened down. If the period $d$ is even, say $d = 2q$, then we use

$$\hat{m}_t = (0.5X_{t-q} + X_{t-q+1} + \ldots + X_{t+q-1} + 0.5X_{t+q})/d, \tag{4.19}$$

for $q+1 \leq t \leq n-q$. If the period is odd, say $d = 2q+1$, then we use the moving average (4.14). The use of (4.14) and (4.19) is intuitively clear: the average is taken of variables around $X_t$ in one period.

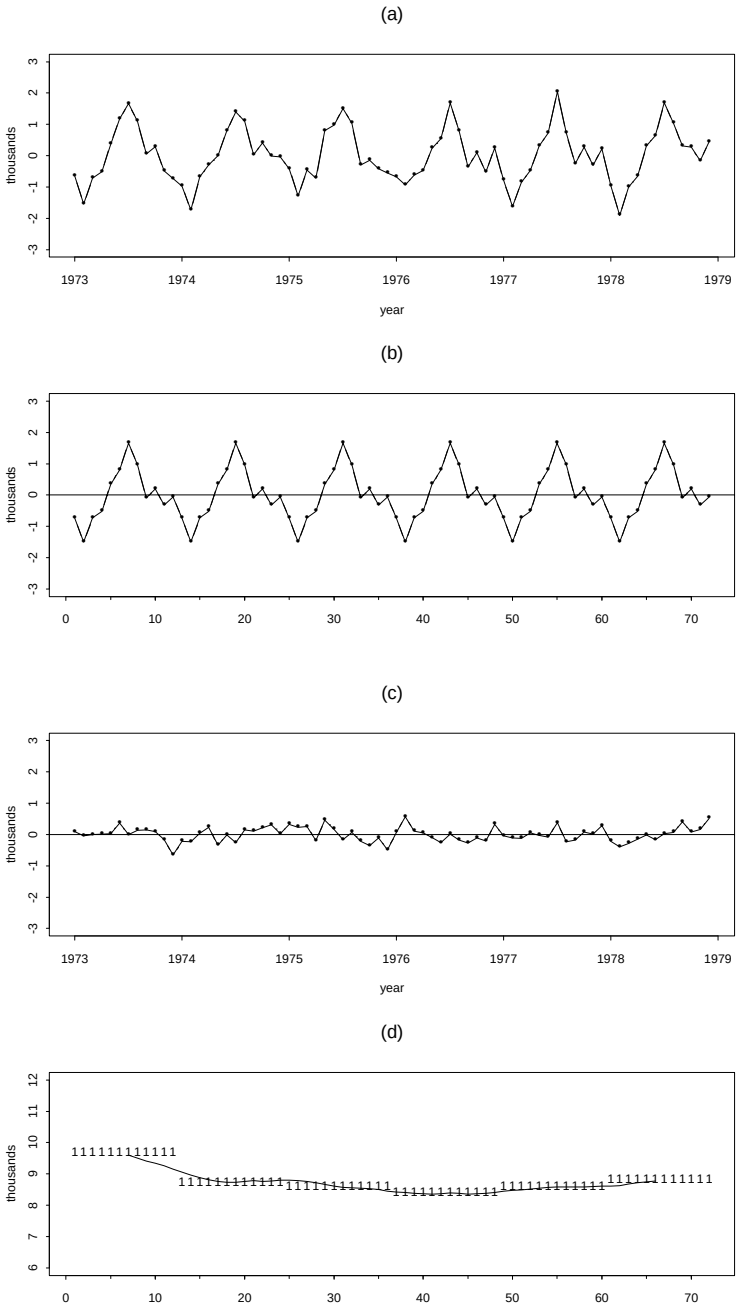


Figure 4.6: Accident data from Figure 4.1.a a) Accidental deaths after subtracting the trend estimated by (I). b) The seasonal component of the accidental deaths, estimated by (I). c) The detrended and deseasonalized accidental deaths (I). d) Comparison of the piecewise constant (I) and moving averages (II) estimates of trend for the monthly accidental deaths.

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{s}_k$ (I) | −744 | −1504 | −724 | −523 | 338 | 808 | 1665 | 961 | −87 | 197 | −321 | −67 |
| $\hat{s}_k$ (II) | −804 | −1522 | −737 | −526 | 343 | 746 | 1680 | 987 | −109 | 258 | −259 | −57 |

Table 4.1: Estimated seasonal components for the accidental deaths data

The second step is to estimate the seasonal component. For each $k = 1, \ldots, d$ we define the average $V_k$ over $j$ of the deviations $\{(X_{k+jd} - \hat{m}_{k+jd})\}$, $q < k + jd \leq n - q$. Since these average deviations do not necessarily sum exactly to zero, we have to estimate the seasonal component $s_k$ by $\hat{s}_k$ given by

$$\hat{s}_k = V_k - d^{-1} \sum_{i=1}^{d} V_i, \qquad k = 1, \ldots, d, \tag{4.20}$$

which do sum to zero indeed, and for $k > d$, we use $\hat{s}_k = \hat{s}_{k-d}$. We now can construct deseasonalized data $D_1, \ldots, D_n$ as follows:

$$D_t = X_t - \hat{s}_t, \qquad t = 1, \ldots, n. \tag{4.21}$$

Finally, we may obtain a better estimate of the trend from the deseasonalized data than the one defined in (4.19). To this end, we can reestimate the trend from $\{D_t\}$ by means of method (i) or (ii) of the foregoing section. If this reestimation is done in order to have a parametric form for the trend, so that it can be extrapolated for the purpose of prediction or simulation, then of course the least squares method (i) should be used for this.

The results of applying methods (I) and (II) to the accidental deaths data are quite similar, since in this case the piecewise constant and moving average estimates of $m_t$ are reasonably close (see Figure 4.6.d).

**Example 4.9** In Figure 4.6.d the trend estimate $\hat{m}_t$, $6 < t \leq 66$, for the accidental deaths data (Figure 4.1.a) is shown. Also plotted is the piecewise constant estimate obtained from method (I). A comparison of the estimates of $s_k$, $k = 1, \ldots, 12$, obtained by methods (I) and (II) is made in Table 4.1. We see that the results of applying methods (I) and (II) to the accidental deaths data are quite similar. This is due to the fact that in this case the piecewise constant and moving average estimates of $m_t$ are rather close (see Figure 4.6.d).

*(III) Differencing at lag d*

The technique of differencing, explained as method (iii) for non-seasonal data, can be adapted to deal with seasonality of period $d$ by introducing the lag-$d$ difference operator $\nabla_d$ (not to be confused with $\nabla^d$) defined by

$$\nabla_d X_t = X_t - X_{t-d}, \qquad t > d. \tag{4.22}$$

Application of $\nabla_d$ to the model (4.10), where $\{s_t\}$ has period $d$, yields

$$\nabla_d X_t = m_t - m_{t-d} + Y_t - Y_{t-d}, \tag{4.23}$$

which gives a decomposition of the difference $\nabla_d X_t$ into a trend component $(m_t - m_{t-d})$ and a noise term $(Y_t - Y_{t-d})$. Clearly the process $\{(Y_t - Y_{t-d})\}$ is stationary, since $\{(Y_t\}$ is. It follows that for data satisfying model (4.10), a stationary sequence can be obtained by first applying $\nabla_d$ to the data, and next eliminating the trend using one of the methods (i)–(iii) described in the foregoing section.

**Example 4.10** Figure 4.7.a shows the result of applying the operator $\nabla_{12}$ to the accidental deaths data. The seasonal component evident in Figure 4.1.d is absent from the graph of $\nabla_{12} X_t$. There still appears to be a non-decreasing trend, however. If we now apply the operator $\nabla$ to $\nabla_{12} X_t$ and plot the resulting differences, $\nabla\nabla_{12} X_t$, we obtain the graph shown in Figure 4.7.b, which has no apparent trend or seasonal component.

### 4.2.4 Filtering

Taking moving averages as in (4.14) and (4.19) and differencing are examples of the use of a linear operator or *linear filter* which converts one time series $\{X_t\}$ into another, $\{\tilde{X}_t\}$ say, by the linear operation

$$\tilde{X}_t = \sum_{j=-\infty}^{+\infty} a_j X_{t-j}. \tag{4.24}$$

The filter (4.14) is an example of a so-called *low-pass filter*, which takes away from the data $\{X_t\}$ the rapidly fluctuating (or high frequency) component $\{\hat{Y}_t\}$ to leave the slowly varying estimated trend $\{\hat{m}_t\}$ (see Figure 4.6.d). In general, a filter to remove trend would have coefficients that add up to zero, whereas seasonality would be removed by a filter with coefficients that add up to one (over the period of the seasonality).

## 4.3 Models for Stationary Time Series

After possible trend and seasonality components are removed, we assume that we are left with realizations from a stationary process. In the Examples 4.1 and 4.2 we have already seen some specific models for a stationary process. We now discuss some frequently used general types of models for stationary processes.

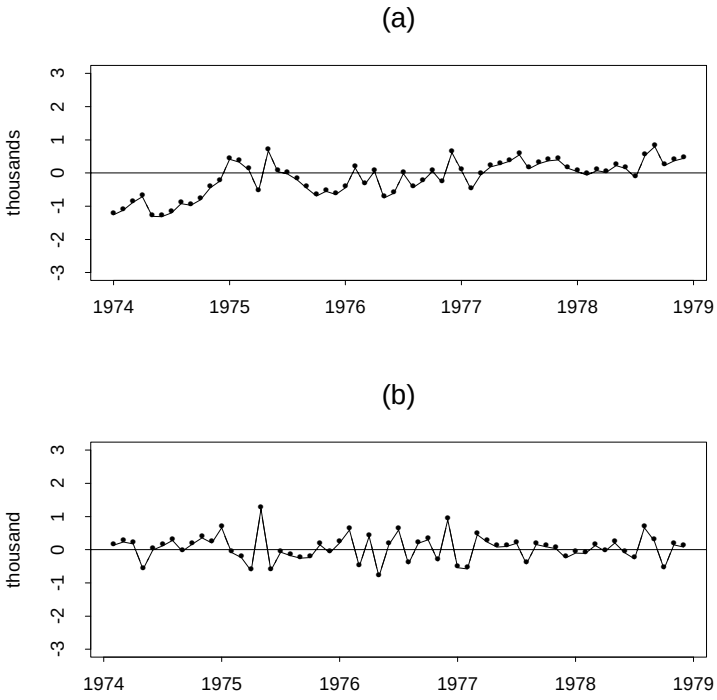


Figure 4.7: Accident data from Figure 4.1.a a) The differenced series $\{\nabla_{12} X_t,\ t = 13, \ldots, 72\}$. b) The differenced series $\{\nabla\nabla_{12} X_t,\ t = 14, \ldots, 72\}$.

### 4.3.1 Moving Average Processes

Suppose that $\{Z_t\}$ is a white noise process with $\sigma^2$ as the variance of $Z_t$. Let $q$ be a positive number and let $\beta_0, \beta_1, \ldots, \beta_q$ be constants.

**Definition 4.5** *The process $\{X_t\}$ defined by*

$$X_t = \beta_0 Z_t + \beta_1 Z_{t-1} + \cdots + \beta_q Z_{t-q} \tag{4.25}$$

*is a moving average process of order $q$ (abbreviated: MA(q) process).*

The $Z_t$ are usually scaled so that $\beta_0 = 1$. In Example 4.2 we had a MA(1) process. Likewise we find for the general MA($q$) process

$$\mathrm{E}\, X_t = 0 \tag{4.26}$$

$$\mathrm{Var}\, X_t = \sigma^2 \sum_{i=0}^{q} \beta_i^2 \tag{4.27}$$

$$\gamma_X(h) = \begin{cases} \sigma^2 \sum_{i=0}^{q-h} \beta_i \beta_{i+h}, & h = 0, 1, \ldots, q, \\ \gamma_X(-h), & h = -1, \ldots, -q, \\ 0, & \text{otherwise} \end{cases} \tag{4.28}$$

$$\rho_X(h) = \begin{cases} 1, & h = 0, \\ \sum_{i=0}^{q-h} \beta_i \beta_{i+h} / \sum_{i=0}^{q} \beta_i^2, & h = 1, \ldots, q, \\ \rho_X(-h), & h = -1, \ldots, -q, \\ 0, & \text{otherwise}, \end{cases} \tag{4.29}$$

since the $Z_t$ are independent. We conclude that MA processes are stationary. Note that the autocovariance function 'cuts off' at lag $q$, which is a special feature of MA processes.

MA processes are used in many areas, particularly econometrics. For example, economic variables or quantities are affected by a variety of random events such as strikes, government decisions, shortages of key materials and so on. Such events will not only have an immediate effect, but may also affect to a lesser extent economic variables in several subsequent periods, and so it is at least plausible that a MA process may be appropriate.

### 4.3.2 Autoregressive Processes

Let again $\{Z_t\}$ be a white noise process with variance $\sigma^2$, let $p$ be a positive number and $\alpha_1, \ldots, \alpha_p$ constants.

**Definition 4.6** *If $\{X_t\}$ is stationary and satisfies*

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + Z_t \tag{4.30}$$

*then $\{X_t\}$ is an autoregressive process of order $p$ (abbreviated: AR(p) process).*

An autoregressive process is rather like a multiple regression model, but $X_t$ is not regressed on independent variables, but on past values of itself; hence the prefix 'auto'.

**Example 4.11** Suppose that $\{Z_t\}$ is a white noise process with $\sigma^2$ as the variance of $Z_t$, and let $\alpha$ be a constant. Let $\{X_t\}$ be stationary and assume that it satisfies

$$X_t = \alpha X_{t-1} + Z_t, \tag{4.31}$$

so that $\{X_t\}$ is an autoregressive process of order one. By iteration we find

$$\begin{aligned} X_t &= \alpha(\alpha X_{t-2} + Z_{t-1}) + Z_t = \ldots = \\ &= \alpha^k X_{t-k} + \alpha^{k-1} Z_{t-k+1} + \alpha^{k-2} Z_{t-k+2} + \ldots + \alpha Z_{t-1} + Z_t, \end{aligned} \tag{4.32}$$

which suggests that $X_t$ can be written as

$$X_t = Z_t + \alpha Z_{t-1} + \alpha^2 Z_{t-2} + \ldots,$$

which is an *infinite order* moving average process. It can be shown that for $|\alpha| < 1$ this is true indeed, and in this case $\mathrm{E}\, X_t = 0$ and autocovariance function

$$\gamma_X(h) = \frac{\alpha^{|h|}}{1-\alpha^2}\sigma^2.$$

This means that $\gamma_X(h) \to 0$ as $|h| \to \infty$. The dependence decreases, but never becomes zero.

As in the first order case we can write a general AR process as a MA process of infinite order under some conditions on the $\alpha_j$'s. In that case the AR process has expectation $\mathrm{E}\, X_t = 0$. We can find an expression for the autocovariance function of the process in terms of the $\alpha_j$'s and $\sigma^2$ by multiplying (4.30) by $X_{t-h}$. This gives

$$X_t X_{t-h} = \alpha_1 X_{t-1} X_{t-h} + \cdots + \alpha_p X_{t-p} X_{t-h} + Z_t X_{t-h} \tag{4.33}$$

Since $\mathrm{E}\, X_t = 0$, $\mathrm{E}\, X_t Z_t = \sigma^2$, and since $Z_t$ and $X_{t-h}$ are independent for $h > 0$, taking expectations on both sides of (4.33) yields

$$\begin{aligned} \gamma_X(h) &= \alpha_1 \gamma_X(h-1) + \cdots + \alpha_p \gamma_X(h-p) \\ &= \sum_{j=1}^{p} \alpha_j \gamma_X(h-j), \qquad h > 0, \\ \gamma_X(0) &= \sum_{j=1}^{p} \alpha_j \gamma_X(j) + \sigma^2, \end{aligned} \tag{4.34}$$

and hence

$$\rho_X(h) = \sum_{j=1}^{p} \alpha_j \rho_X(h-j), \qquad h > 0, \tag{4.35}$$

$$\sigma^2 = \gamma_x(0)\left(1 - \sum_{j=1}^{p} \alpha_j \rho_X(j)\right),$$

where the latter set of equations is known as the Yule–Walker equations. Although this is not immediately clear from (4.35), for the general AR process the autocovariance function behaves similarly as for the AR(1) process in Example 4.11: $\gamma_X(h) \to 0$ as $|h| \to \infty$. The dependence decreases, but never becomes zero.

AR processes are applied to many situations in which it is reasonable to assume that the present value of a time series depends on the immediate past values together with random error.

### 4.3.3 ARMA Models

Another useful class of models for time series is formed by combining MA and AR processes. A mixed autoregressive/moving average process containing $p$ AR terms and $q$ MA terms is said to be an ARMA process of order $(p, q)$. It is given by

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + Z_t + \beta_1 Z_{t-1} + \ldots + \beta_q Z_{t-q} \tag{4.36}$$

with $\{X_t\}$ a stationary process. Note that an ARMA(0,$q$) process is a MA($q$) process and an ARMA($p$,0) process is an AR($p$) process. The importance of ARMA processes lies in the fact that a stationary time series may often be described by an ARMA model involving fewer parameters than if a pure MA or AR process would be used.

**Example 4.12** In Figure 4.8 simulations are shown of an AR(2), a MA(2) and an ARMA(1,2) process.

### 4.3.4 ARIMA Models

For completeness, and because in practice most time series are non-stationary, we also mention the class of so-called ARIMA models for non-stationary processes. Since, as we have seen, differencing of a series can produce a stationary one if the observed series is non-stationary in the mean, the variable $X_t$ in (4.36) may be replaced by $\nabla^d X_t$ in order to obtain a model capable of describing certain types of non-stationary time series. Such a model is called an autoregressive integrated moving average process of order $(p, d, q)$ (ARIMA$(p, d, q)$). The term 'integrated' stems from the fact that the stationary model

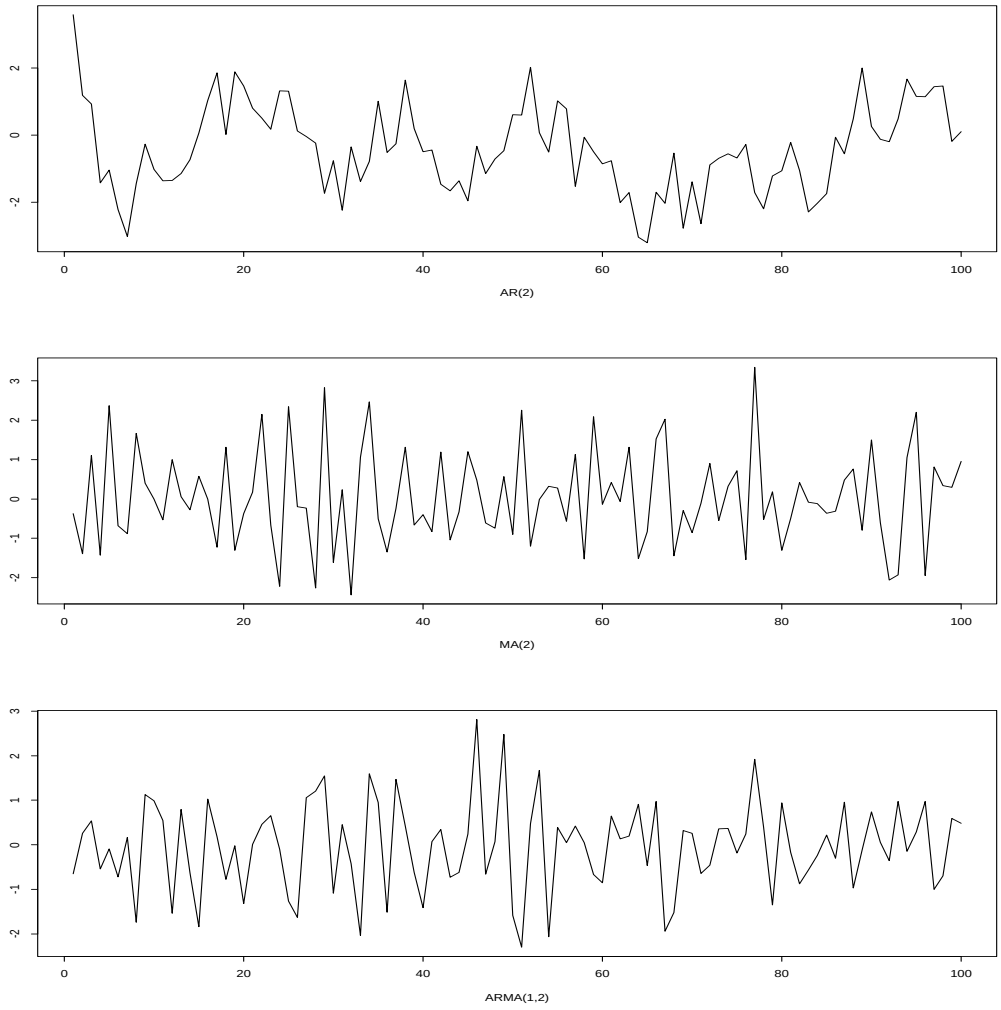


Figure 4.8: a) Simulated AR(2) process. b) Simulated MA(2) process. c) Simulated ARMA(1,2) process.

which is fitted to the differenced data has to be summed or 'integrated' to provide a model for the non-stationary data. ARIMA models are frequently used to describe econometric time series.

## 4.4 Estimation of Mean and (Partial) Autocorrelation Function

For simplicity the models we have seen for stationary processes all had expectation zero. Of course we could have added a constant $\mu$ to obtain a process with expectation equal to $\mu$ while this would not affect stationarity of the process. Let our observed data be realizations of $X_1, \ldots, X_n$ and let us assume that $\{X_t\}$ is a stationary process. A natural estimator for the expectation of this stationary process is the sample mean

$$\hat{\mu}_n = \bar{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t. \tag{4.37}$$

This is the moment estimator and it is unbiased.

In order to select an appropriate model we need to be able to make plots of the (partial) autocorrelation function. Since these functions are unknown, they need to be estimated from the data. The most frequently used estimator for the autocovariance function of $\{X_t\}$ is the sample autocovariance function

$$\hat{\gamma}_n(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_{t+h} - \bar{X}_n)(X_t - \bar{X}_n), \qquad h \geq 0, \tag{4.38}$$

and, hence, one estimates the autocorrelation function with the sample autocorrelation function

$$\hat{\rho}_n(h) = \hat{\gamma}_n(h)/\hat{\gamma}_n(0). \tag{4.39}$$

Both $\hat{\gamma}_n(h)$ and $\hat{\rho}_n(h)$ are biased, but under general assumptions nearly unbiased for large sample sizes. A plot of $\rho_n(h)$ against $h$ is called a *correlogram*. It might seem more natural to use the divisor $(n-h)$ instead of $n$ in definition (4.38), but the definition given here is the conventional one; it is used for mathematical convenience.

One way to construct a *partial* correlogram is to estimate the partial autocorrelation coefficients $\alpha(k), k = 1, 2, \ldots$ in the following way. Fit an autoregressive process of order $k$ on the observed process, and estimate the partial autocorrelation coefficient $\alpha(k)$ by means of the estimate $\hat{\alpha}_{k,n}$ of the last autoregressive coefficient $\alpha_k$ of the fitted AR($k$) process. Do this for every $k = 1, 2, \ldots$. Clearly, if the underlying process is AR($p$), then $\alpha(k) = 0$ for all $k > p$, and the partial autocorrelogram should show a cut-off after lag $p$.

If the time series is *Gaussian*, that is if all random variables are normal, then $\bar{X}_n$ is normally distributed. The distributions of $\hat{\gamma}_n(h)$, $\hat{\rho}_n(h)$, and $\hat{\alpha}_{k,n}$ are complicated, even

under normality. However, frequently the estimators are asymptotically normally distributed when $n \to \infty$. This knowledge can be used to determine approximate confidence intervals.

## 4.5 Choice of Model and Determination of its Order

How to choose a model for a stationary series? Often the shapes of the sample autocorrelation and partial autocorrelation function are used to determine whether a MA, AR or ARMA model should be used. For instance, we have seen that for a MA process the autocovariance function becomes zero at lags with absolute value greater than its order, whereas this is not the case for an AR process. Hence, if a plot of the autocorrelation function shows a sharp cut-off at some point after which all autocorrelations are (not significantly different from) zero, then MA should be chosen. Similarly, if a plot of the partial autocorrelation function shows such cut-off, then an AR model should be preferred. If neither of these two plots shows a sharp cut-off to zero, we choose an ARMA model.

After a class of models is selected, the order of the process needs to be determined. As may be clear from the above, if a MA process is thought to be appropriate for a given set of data, then the order of the process is usually evident from the correlogram. It can be taken as the largest value of $h$ for which the estimated autocorrelation is significantly different from zero.

The order of an AR process can be similarly estimated by eye from the partial correlogram. It can be taken as the largest value of $k$ for which the estimated partial autocorrelation is significantly different from zero.

To aid the judgement of when an estimated (partial) autocorrelation is significantly different from zero, most statistical packages automatically draw approximate 95% confidence limits around zero on the autocorrelation and partial autocorrelation plots.

A more formal method, which is frequently implemented in software for time series analysis, is order selection with the aid of the Akaike criterion (AIC). This criterion can not only be used for the determination of the order of AR(I)MA processes. It is much more general and is used in a variety of other contexts in which the number of parameters needs to be established too. It is based on the principle that increasing the number of parameters—here the order(s) of the model—will increase the likelihood on the one hand, but will make the model more complex and introduce additional errors due to parameter estimation on the other hand. Loosely speaking this criterion balances the reduction of estimated error variance with the number of parameters to be fitted. More precisely, the number of parameters should be chosen in such a way that AIC($k$) defined by

$$\text{AIC}(k) = -2 \log \ \mathit{likelihood} + 2k, \tag{4.40}$$

where $k$ is the number of parameters, is minimized with respect to $k$. In order to be able to compute the likelihood for AR(I)MA processes, the distribution of $\{Z_t\}$ needs to be

specified. It is usually assumed to be normal. Several other criteria have been proposed to improve upon AIC. For these we refer to the literature.

**Example 4.13** In Figure 4.9 the estimated autocorrelation and partial autocorrelation functions of the AR, MA and ARMA process in Figure 4.8 are shown. Is this what you would expect?

## 4.6 Estimation of the Model Parameters

After a class of models is selected and the order of the process is determined, the model parameters have to be estimated.

The simplest way to obtain estimates $\hat{\alpha}_{k,n}$ of the coefficients $\alpha_k$, $k = 1, \ldots, p$, and $\hat{\sigma}^2$ of $\sigma^2$ of an AR($p$) process is to solve the sample equivalent of the Yule-Walker equations (4.35):

$$\begin{aligned} \hat{\rho}_n(k) &= \sum_{j=1}^{p} \hat{\alpha}_{j,n}\hat{\rho}_n(k-j), \qquad k = 1, \ldots, p \\ \hat{\sigma}^2 &= \hat{\gamma}(0)(1 - \sum_{j=1}^{p} \hat{\alpha}_{j,n}\hat{\rho}_n(j)) \end{aligned} \tag{4.41}$$

for $\hat{\alpha}_{k,n}$ , $k = 1, \ldots, p$, and $\hat{\sigma}^2$, where the $\hat{\rho}_n(h)$ in (4.41) can be obtained from the data according to (4.38) and (4.39). Most statistical packages contain efficient routines for this.

For the estimation of the coefficients of MA processes the so-called *innovations algorithm* can be used. We refer to the literature for a description of this method.

The estimation of the coefficients of general AR(I)MA processes is more difficult than of AR and MA processes, because efficient explicit estimators cannot be found. Instead some form of numerical iteration must be performed. Many statistical packages contain a procedure for numerical approximation of the maximum likelihood estimates of the parameters of ARIMA models under the assumption of normality for the $Z_t$'s. Since such computation uses iterative methods, starting values for the parameters are needed as input values; finding good starting values can be a problem.

An alternative is to compute some kind of least squares estimates. The Hannan-Rissanen estimation procedure first estimates the $Z_t$'s by putting

$$\tilde{Z}_t = X_t - \tilde{\alpha}_1 X_{t-1} - \ldots - \tilde{\alpha}_m X_{t-m}, \qquad t = m+1, \ldots, n, \tag{4.42}$$

where $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_m$ are the Yule-Walker estimates resulting from fitting an AR($m$) model for some $m > \max(p, q)$. Then $\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q$ and $\sigma^2$ are estimated by least squares via linear regression of $X_t$ on $X_{t-1}, \ldots, X_{t-p}, \tilde{Z}_t, \ldots, \tilde{Z}_{t-q}$, $t = m+1+q, \ldots, n$. Once more we refer to the literature for a more detailed description.

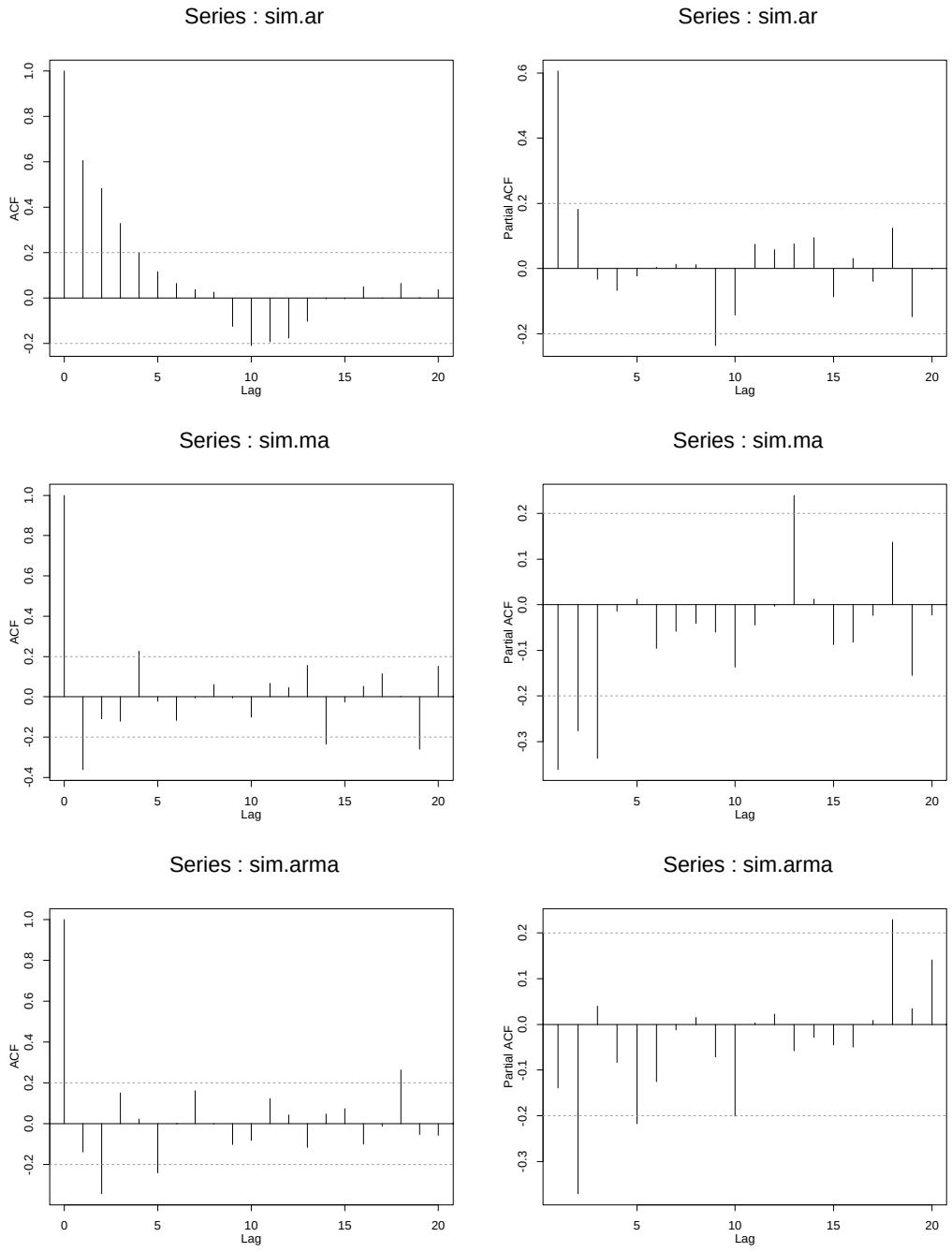


Figure 4.9: Autocorrelation and partial autocorrelation functions of the simulated processes of Figure 4.8. The dashed lines represent the approximate 95% confidence limits around zero.

## 4.7 Diagnostic Checking

When a model is chosen and its parameters estimated, the appropriateness of the fitted model should be investigated. The adequacy of an ARMA model can be checked in various ways. The usual approach is to extract from the data and the model a sequence of *residuals* to correspond to the underlying, but unobservable, white noise sequence, and to check that the statistical properties of these residuals are indeed consistent with white noise.

All ARMA processes have in their definition a white noise sequence $\{Z_t\}$. One way to obtain the residuals $\{R_t, t = 1, \ldots, n\}$ is to substitute into the defining equations (4.25), (4.30) and (4.36) (or their counterparts for the case $\mu \neq 0$) the estimated values of all parameters and the observed time series $\{X_t, t = 1, \ldots, n\}$, and solving the resulting set of equations for $\{R_t, t = 1, \ldots, n\}$.

For AR($p$) processes extraction of the residual sequence in this way is easy. We find in this case

$$R_t = (X_t - \hat{\mu}_n) - \sum_{j=1}^{p} \hat{\alpha}_{j,n}(X_{t-j} - \hat{\mu}_n), \qquad t = p+1, \ldots, n. \tag{4.43}$$

We see that $R_t$ is undefined for $t \leq p$.

For a MA($q$) process, the residuals are obtained recursively. Note that the defining equation can be written in the form

$$Z_t = (X_t - \mu) - \sum_{j=1}^{q} \beta_j Z_{t-j}. \tag{4.44}$$

Set $R_t = 0$ for $t \leq 0$, and compute

$$\begin{aligned} R_1 &= (X_1 - \hat{\mu}_n), \\ R_2 &= (X_2 - \hat{\mu}_n) - \hat{\beta}_{1,n} R_1, \end{aligned} \tag{4.45}$$

and so on until, for $t > q$,

$$R_t = (X_t - \hat{\mu}_n) - \sum_{j=1}^{q} \hat{\beta}_{j,n} R_{t-j}. \tag{4.46}$$

It is recommended to discard the $R_t$ for $t \leq q$.

Finally, in the case of an ARMA($p, q$) process, the defining equation gives

$$Z_t = (X_t - \mu) - \sum_{i=1}^{p} \alpha_i (X_{t-i} - \mu) - \sum_{j=1}^{q} \beta_j Z_{t-j}. \tag{4.47}$$

Setting $R_t = 0$ for $t \leq p$, we can extract the later $R_t$ as

$$R_t = (X_t - \hat{\mu}_n) - \sum_{i=1}^{p} \hat{\alpha}_{i,n}(X_{t-i} - \hat{\mu}_n) - \sum_{j=1}^{q} \hat{\beta}_{j,n} R_{t-j}. \tag{4.48}$$

Here a cautious strategy would be to discard the $R_t$ for $t \leq \max(p, q)$.

A first impression of the behavior of the residuals can be obtained by inspecting its autocorrelation function. If the residuals are close to white noise, then their correlogram should look like that of white noise too. For large $n$, the sample autocorrelations of a white noise sequence $Z_1, \ldots, Z_n$ are approximately normally distributed with mean zero and variance $1/n$. Thus, the values $\hat{\rho}_{R,n}(h)$ in the correlogram of $R_1, \ldots, R_n$ that are in absolute value greater than $2/\sqrt{n}$, twice the standard deviation, can be regarded as significantly different from zero about the 5% level. Note that, if a large number of $\hat{\rho}_{R,n}(h)$ is calculated, it is likely that some will exceed this threshold, even if $\{R_t\}$ is a white noise sequence. We remark that interpretation is somewhat complicated by the fact that the $\hat{\rho}_{R,n}(h)$ are dependent, and that the effects of the parameter estimation are ignored.

The residual sequence $\{R_t\}$ can also be subjected to several tests for white noise. The so-called *Portmanteau test* uses the statistic

$$Q_m = n \sum_{h=1}^{m} \hat{\rho}_{R,n}(h)^2, \tag{4.49}$$

which is, if $\{R_t\}$ is a white noise process, for large $n$, and $m << n$, approximately chi-squared distributed with $m$ degrees of freedom. Another test for white noise is the *Ljung-Box test*, which is a refinement of the Portmanteau test, where $Q_m$ is replaced by

$$Q_m^{LB} = n(n+2) \sum_{h=1}^{m} (n-h)^{-1} \hat{\rho}_{R,n}(h)^2. \tag{4.50}$$

Both tests ignore estimation effects too, and their results should be considered accordingly.

**Example 4.14** In the top panel of Figure 4.10 the AR(2) process of Figure 4.8 is shown, together with the estimated AR(2) process based on the Yule-Walker equations. From Figure 4.9 we may have guessed that the order should have been 3, or perhaps 4. The choice according to the above mentioned Akaike criterion would indeed be 3. The bottom panel of Figure 4.10 displays the correlogram of the corresponding residual sequence. From this we find no reason to doubt that the residuals resemble a white noise process, as they should.

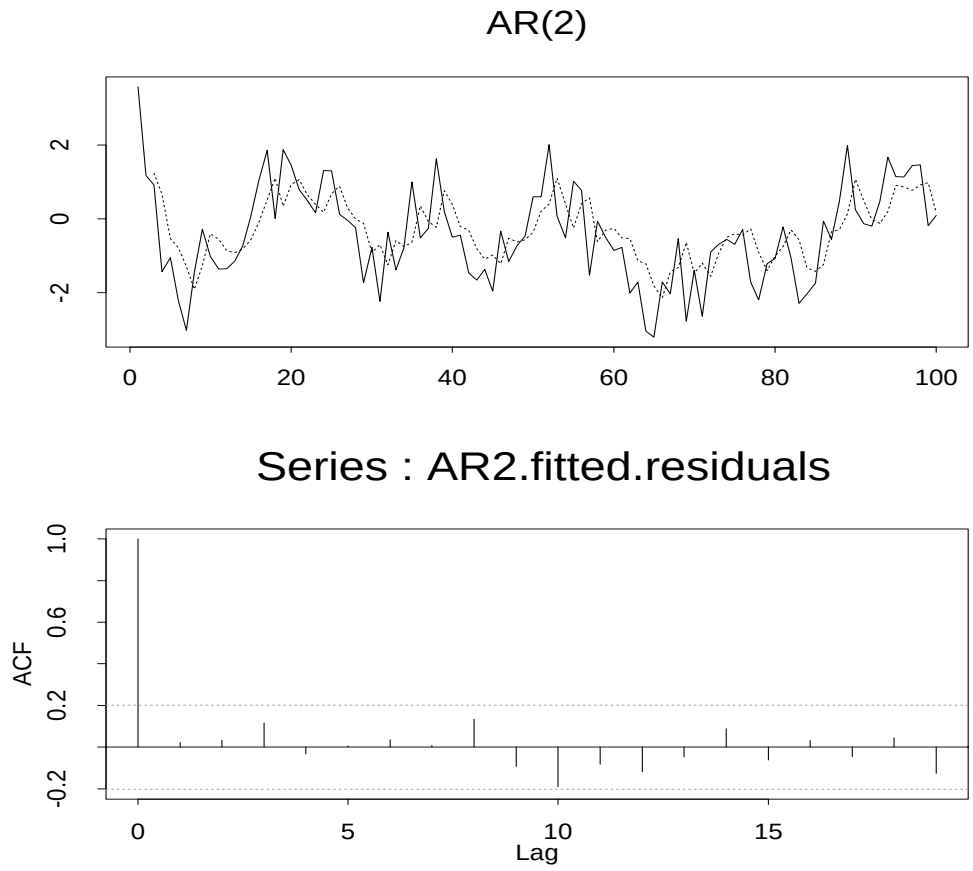


Figure 4.10: Top panel: the AR(2) process of Figure 4.8 (solid line) together with the estimated AR(2) process based on the Yule-Walker equations (dotted line). Bottom panel: estimated autocorrelation function of the corresponding residuals.

## 4.8 Final Remarks

In this chapter we have discussed the basics of the analysis of time series based on its evolution in time. The autocovariance function is the main tool for this. This type of analysis is often called an analysis in the time domain. Another way of looking at a time series is based on the so-called spectral density function, which describes how the variation in a time series may be accounted for by cyclic components at different frequencies (see for instance Brockwell and Davis (1996) and Chatfield (1989) for an introduction or Brockwell and Davis (1993) for a more detailed and more mathematical approach). Inference based on the spectral density is often called an analysis in the frequency domain. The two approaches are equivalent, but may shed light on different aspects of a series. The frequency domain approach is used in particular by engineers and physicists.

Also for other, more advanced topics, we refer to the above mentioned references.

## Chapter 5

# References


Bates, D.M. and Watts, D.G., (1988), *Nonlinear Regression Analysis and its Applications*, Wiley, New York.

Brockwell, P.J. and Davis, R.A., (1993), *Time Series: Theory and Methods*, Springer, New York.

Brockwell, P.J. and Davis, R.A., (1996), *Introduction to Time Series and Forecasting* Springer, New York.

Chatwin, C., (1989), *The Analysis of Time Series* (4th ed.), Chapman and Hall, London.

Gallant, A.R., (1987), *Nonlinear Statistical Models*, Wiley, New York.

Nelder, J.A., and Wedderburn, R.W.M., (1972), Generalized Linear Models, *Journal of the Royal Statistical Society* A, 135, p. 370–384.

McCullagh, P., and Nelder, J.A., (1989), *Generalized Linear Models* (2nd ed.), Chapman and Hall, London.

Searle, S.R., (1971), *Linear Models*, Wiley, New York.

Seber, G.A.F., and Wild, C.J., (1989), *Nonlinear Regression*, Wiley, New York.

Scheffé, H., (1959), *The Analysis of Variance* Wiley, New York.