Ulf Olsson

Generalized Linear Models

An Applied Approach

Studentlitteratur

The papers and inks used in this product are environment-friendly.

Art. No 31023
eISBN10 91-44-03141-6
eISBN13 978-91-44-03141-5

Cover design: Henrik Hast

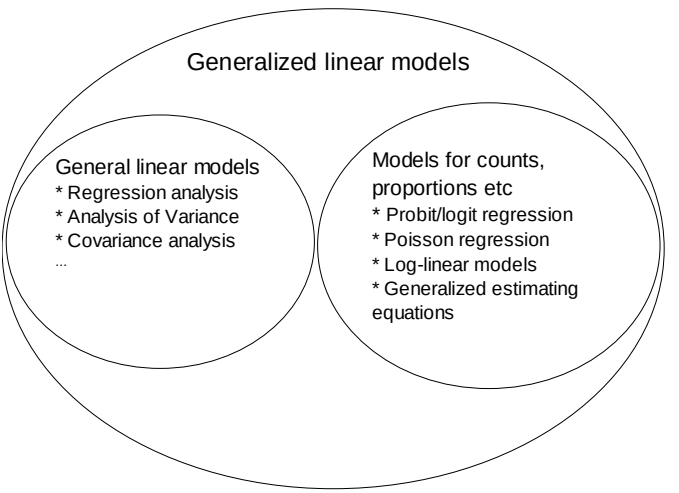
Printed in Sweden
Studentlitteratur, Lund
Web-address: www.studentlitteratur.se

Printing/year 1 2 3 4 5 6 7 8 9 10 | 2006 05 04 03 02

Contents

Preface

Generalized Linear Models (GLIM:s) is a very general class of statistical models that includes many commonly used models as special cases. For example the class of General Linear Models (GLM:s) that includes linear regression, analysis of variance and analysis of covariance, is a special case of GLIM:s. GLIM:s also include log-linear models for analysis of contingency tables, probit/logit regression, Poisson regression, and much more.



In this book we will make an overview of generalized linear models and present examples of their use. We assume that the reader has a basic understanding of statistical principles. Particularly important is a knowledge of statistical model building, regression analysis and analysis of variance. Some knowledge of matrix algebra (which is summarized in Appendix A), and knowledge of basic calculus, are mathematical prerequisites. Since many of the examples are based on analyses using SAS, some knowledge of the SAS system is recommended.

In Chapter 1 we summarize some results on general linear models, assuming equal variances and normal distributions. The models are formulated in

matrix terms. Generalized linear models are introduced in Chapter 2. The exponential family of distributions is discussed, and we discuss Maximum Likelihood estimation and ways of assessing the fit of the model. This chapter provides the basic theory of generalized linear models. Chapter 3 covers model checking, which includes systematic ways of assessing whether the data deviates from the model in some systematic way. In chapters 4–7 we consider applications for different types of response variables. Response variables as continuous variables, as binary/binomial variables, as counts and as ordinal response variables are discussed, and practical examples using the Genmod software of the SAS package are given. Finally, in Chapter 8 we discuss theory and applications of a more complex nature, like quasi-likelihood procedures, repeated measures models, mixed models and analysis of survival data.

Terminology in this area of statistics is a bit confused. In this book we will let the acronym GLM denote "General Linear Models", while we will let GLIM denote "Generalized Linear Models". This is also a way of paying homage to two useful computer procedures, the GLM procedure of the SAS package, and the pioneering GLIM software.

Several students and colleagues have read and commented on earlier versions of the book. In particular, I would like to thank Gunnar Ekbohm, Jan-Eric Englund, Carolyn Glynn, Anna Gunsjö, Esbjörn Ohlsson, Tomas Pettersson and Birgitta Vegerfors for giving many useful comments.

Most of the data sets for the examples and exercises are available on the Internet. They can be downloaded from the publishers home page which has address http://www.studentlitteratur.se.

1. General Linear Models

1.1 The role of models

Many of the methods taught during elementary statistics courses can be collected under the heading *general linear models*, GLM. Statistical packages like SAS, Minitab and others have standard procedures for general linear models. GLM:s include regression analysis, analysis of variance, and analysis of covariance. Some applied researchers are not aware that even their simplest analyses are, in fact, model based.

Models play an important role in statistical inference. A model is a mathematical way of describing the relationships between a response variable and a set of independent variables. Some models can be seen as a theory about how the data were generated. Other models are only intended to provide a convenient summary of the data. Statistical models, as opposed to deterministic models, account for the possibility that the relationship is not perfect. This is done by allowing for unexplained variation, in the form of residuals.

A way of describing a frequently used class of statistical models is

$$\text{Response} = \text{Systematic component} + \text{Residual component} \tag{1.1}$$

Models of type (1.1) are, at best, approximations of the actual conditions. A model is seldom "true" in any real sense. The best we can look for may be a model that can provide a reasonable approximation to reality. However, some models are certainly better than others. The role of the statistician is to find a model that is reasonable, while at the same time it is simple enough to be interpretable.

1.2 General Linear Models

In a general linear model (GLM), the observed value of the dependent variable y for observation number i $(i = 1, 2, ..., n)$ is modeled as a linear function of $(p-1)$ so called independent variables $x_1, x_2, \ldots, x_{p-1}$ as

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i(p-1)} + e_i \tag{1.2}$$

or in matrix terms

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \tag{1.3}$$

In (1.3),

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

is a vector of observations on the dependent variable;

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & & \\ \vdots & & \ddots & \\ 1 & x_{n1} & & x_{n(p-1)} \end{pmatrix}$$

is a known matrix of dimension $n \times p$, called a *design matrix* that contains the values of the independent variables and one column of 1:s corresponding to the intercept;

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$$

is a vector containing p parameters to be estimated (including the intercept); and

$$\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

is a vector of residuals. It is common to assume that the residuals in $\mathbf{e}$ are independent, normally distributed and that the variances are the same for all e_i. Some models do not contain any intercept term β_0. In such models, the leftmost column of the design matrix $\mathbf{X}$ is omitted.

The purpose of the analysis may be model building, estimation, prediction, hypothesis testing, or a combination of these. We will briefly summarize some results on estimation and hypothesis testing in general linear models. For a more complete description reference is made to standard textbooks in regression analysis, such as Draper and Smith (1998) or Sen and Srivastava (1990); and textbooks in analysis of variance, such as Montgomery (1984) or Christensen (1996).

1.3 Estimation

Estimation of parameters in general linear models is often done using the method of least squares. For normal theory models this is equivalent to Maximum Likelihood estimation. The parameters are estimated with those values for which the sum of the squared residuals, $\sum_i e_i^2$, is minimal. In matrix terms, this sum of squares is

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) . \tag{1.4}$$

Minimizing (1.4) with respect to the parameters in $\boldsymbol{\beta}$ gives the normal equations

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}. \tag{1.5}$$

If the matrix $\mathbf{X}'\mathbf{X}$ is nonsingular, this yields, as estimators of the parameters of the model,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}. \tag{1.6}$$

Throughout this text we will use a "hat", $\widehat{\ }$, to symbolize an estimator. If the inverse of $\mathbf{X}'\mathbf{X}$ does not exist, we can still find a solution, although the

solution may not be unique. We can use generalized inverses (see Appendix A) and find a solution as

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}. \tag{1.7}$$

Alternatively we can restrict the number of parameters in the model by introducing constraints that lead to a nonsingular $\mathbf{X}'\mathbf{X}$.

1.4 Assessing the fit of the model

1.4.1 Predicted values and residuals

When the parameters of a general linear model have been estimated you may want to assess how well the model fits the data. This is done by subdividing the variation in the data into two parts: systematic variation and unexplained variation. Formally, this is done as follows.

We define the predicted value (or fitted value) of the response variable as

$$\widehat{y}_i = \sum_{j=0}^{p-1} \widehat{\beta}_j x_{ij} \tag{1.8}$$

or in matrix terms

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}. \tag{1.9}$$

The predicted values are the values that we would get on the dependent variable if the model had been perfect, i.e. if all residuals had been zero. The difference between the observed value and the predicted value is the observed residual:

$$\widehat{e}_i = y_i - \widehat{y}_i. \tag{1.10}$$

1.4.2 Sums of squares decomposition

The total variation in the data can be measured as the total sum of squares,

$$SS_T = \sum_i (y_i - \overline{y})^2 .$$

This can be subdivided as

$$\begin{aligned} \sum_i (y_i - \overline{y})^2 &= \sum_i (y_i - \widehat{y}_i + \widehat{y}_i - \overline{y})^2 \\ &= \sum_i (y_i - \widehat{y}_i)^2 + \sum_i (\widehat{y}_i - \overline{y})^2 + 2\sum_i (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) . \end{aligned} \tag{1.11}$$

The last term can be shown to be zero. Thus, the total sum of squares SS_T can be subdivided into two parts:

$$SS_{Model} = \sum_i (\widehat{y}_i - \overline{y})^2$$

and

$$SS_e = \sum_i (y_i - \widehat{y}_i)^2 .$$

SS_e, called the residual (or error) sum of squares, will be small if the model fits the data well.

The sum of squares can also be written in matrix terms. It holds that

$$\begin{aligned} SS_T &= \sum_i (y_i - \overline{y})^2 = \mathbf{y}'\mathbf{y} - n\overline{y}^2 \text{ with } n-1 \text{ degrees of freedom } (df). \\ SS_{Model} &= \sum_i (\widehat{y}_i - \overline{y})^2 = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\overline{y}^2 \text{ with } p-1 \ df. \\ SS_e &= \sum_i (y_i - \widehat{y}_i)^2 = \mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \text{ with } n-p \ df. \end{aligned}$$

The subdivision of the total variation (the total sum of squares) into parts is often summarized as an analysis of variance table:

Source	Sum of squares (SS)	df	$MS = SS/df$
Model	$SS_{Model} = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\overline{y}^2$	$p-1$	MS_{Model}
Residual	$SS_e = \mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$	$n-p$	$MS_e = \widehat{\sigma}^2$
Total	$SS_T = \mathbf{y}'\mathbf{y} - n\overline{y}^2$	$n-1$	

These results can be used in several ways. MS_e provides an estimator of σ^2, which is the variance of the residuals. A descriptive measure of the fit of the model to data can be calculated as

$$R^2 = \frac{SS_{Model}}{SS_T} = 1 - \frac{SS_e}{SS_T}. \tag{1.12}$$

R^2 is called the coefficient of determination. It holds that $0 \leq R^2 \leq 1$. For data where the predicted values $\widehat{y}_i$ all are equal to the corresponding observed values y_i, R^2 would be 1. It is not possible to judge a model based on R^2 alone. In some applications, for example econometric model building, models often have values of R^2 very close to 1. In other applications models can be valuable and interpretable although R^2 is rather small. When several models have been fitted to the same data, R^2 can be used to judge which model to prefer. However, since R^2 increases (or is unchanged) when new terms are

added to the model, model comparisons are often based on the adjusted R^2. The adjusted R^2 decreases when irrelevant terms are added to the model. It is defined as

$$R^2_{adj} = 1 - \frac{n-1}{n-p}\left(1 - R^2\right) = 1 - \frac{MS_e}{SS_T/\left(n-1\right)}. \tag{1.13}$$

This can be interpreted as

$$R^2_{adj} = 1 - \frac{\text{Variance estimated from the model}}{\text{Variance estimated without any model}}.$$

A formal test of the full model (i.e. a test of the hypothesis that $\beta_1, ..., \beta_{p-1}$ are all zero) can be obtained as

$$F = \frac{MS_{Model}}{MS_e}. \tag{1.14}$$

This is compared to appropriate percentage points of the F distribution with $(p-1,\ n-p)$ degrees of freedom.

1.5 Inference on single parameters

Parameter estimators in general linear models are linear functions of the observed data. Thus, the estimator of any parameter β_j can be written as

$$\widehat{\beta}_j = \sum_i w_{ij} y_i \tag{1.15}$$

where w_{ij} are known weights. If we assume that all y_i:s have the same variance σ^2, this makes it possible to obtain the variance of any parameter estimator as

$$Var\left(\widehat{\beta}_j\right) = \sum_i w_{ij}^2 \sigma^2. \tag{1.16}$$

The variance σ^2 can be estimated from data as

$$\widehat{\sigma}^2 = \frac{\sum_i \widehat{e}_i^2}{n-p} = MS_e. \tag{1.17}$$

The variance of a parameter estimator $\widehat{\beta}_j$ can now be estimated as

$$\widehat{Var}\left(\widehat{\beta}_j\right) = \sum_i w_{ij}^2 \widehat{\sigma}^2. \tag{1.18}$$

This makes it possible to calculate confidence intervals and to test hypotheses about single parameters. A test of the hypothesis that the parameter β_j is zero can be made by comparing

$$t = \frac{\widehat{\beta}_j}{\sqrt{\widehat{Var}\left(\widehat{\beta}_j\right)}} \tag{1.19}$$

with the appropriate percentage point of the t distribution with $n-p$ degrees of freedom. Similarly,

$$\widehat{\beta}_j \pm t_{(1-\alpha/2,n-p)}\sqrt{\widehat{Var}\left(\widehat{\beta}_j\right)} \tag{1.20}$$

would provide a $(1-\alpha)\cdot 100\%$ confidence interval for the parameter β_j.

1.6 Tests on subsets of the parameters

In some cases it is of interest to make simultaneous inference about several parameters. For example, in a model with p parameters one may wish to simultaneously test if q of the parameters are zero. This can be done in the following way:

Estimate the parameters of the full model. This will give an error sum of squares, SS_{e1}, with $(n-p)$ degrees of freedom. Now estimate the parameters of the smaller model, i.e. the model with fewer parameters. This will give an error sum of squares, SS_{e2}, with $(n-p-q)$ degrees of freedom, where q is the number of parameters that are included in model 1, but not in model 2. The difference $SS_{e2}-SS_{e1}$ will be related to a χ^2 distribution with q degrees of freedom. We can now test hypotheses of type H_0: $\beta_1 = \beta_2 =, ..., \beta_q = 0$ by the F test

$$F = \frac{\left(SS_{e2} - SS_{e1}\right)/q}{SS_{e1}/\left(n-p\right)} \tag{1.21}$$

with (q, $n-p$) degrees of freedom.

1.7 Different types of tests

Tests of single parameters in general linear models depend on the order in which the hypotheses are tested. Tests in balanced analysis of variance designs are exceptions; in such models the different parameter estimates are

independent. In other cases there are several ways to test hypotheses. SAS handles this problem by allowing the user to select among four different types of tests.

Type 1 means that the test for each parameter is calculated as the change in SS_e when the parameter is added to the model, in the order given in the `MODEL` statement. If we have the model `Y = A B A*B`, SS_A is calculated first as if the experiment had been a one-factor experiment. (model: `Y=A`). Then $SS_{B|A}$ is calculated as the reduction in SS_e when we run the model `Y=A B`, and finally the interaction $SS_{AB|A,B}$ is obtained as the reduction in SS_e when we also add the interaction to the model. This can be written as $SS(A)$, $SS(B|A)$ and $SS(AB|A,B)$. Type I SS are sometimes called sequential sums of squares.

Type 2 means that the SS for each parameter is calculated as if the factor had been added last to the model except that, for interactions, all main effects that are part of the interaction should also be included. For the model `Y = A B A*B` this gives the SS as $SS(A|B)$; $SS(B|A)$ and $SS(AB|A,B)$.

Type 3 is, loosely speaking, an attempt to calculate what the SS would have been if the experiment had been balanced. These are often called partial sums of squares. These SS cannot in general be computed by comparing model SS from several models. The Type 3 SS are generally preferred when experiments are unbalanced. One problem with them is that the sum of the SS for all factors and interactions is generally not the same as the Total SS. Minitab gives the Type 3 SS as "Adjusted Sum of Squares".

Type 4 differs from Type 3 in the method of handling empty cells, i.e. incomplete experiments.

If the experiment is balanced, all these SS will be equal. In practice, tests in unbalanced situations are often done using Type 3 SS (or "Adjusted Sum of Squares" in Minitab). Unfortunately, this is not an infallible method.

1.8 Some applications

1.8.1 Simple linear regression

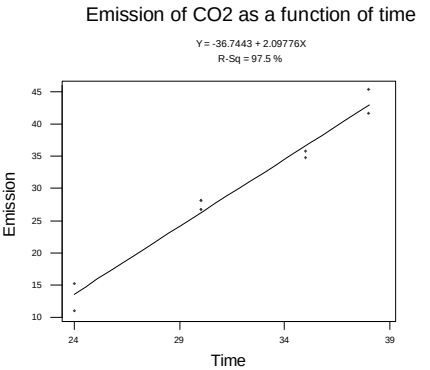
In regression analysis, the design matrix $\mathbf{X}$ often contains one column that only contains 1:s (corresponding to the intercept), while the remaining co-

lumns contain the values of the independent variables. Thus, the small regression model $y_i = \beta_0 + \beta_1 x_i + e_i$ with $n = 4$ observations can be written in matrix form as

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}. \tag{1.22}$$

Example 1.1 An experiment has been made to study the emission of CO_2 from the root zone of Barley (Zagal et al, 1993). The emission of CO_2 was measured on a number of plants at different times after planting. A small part of the data is given in the following table and graph:

Emission	Time
11.069	24
15.255	24
26.765	30
28.200	30
34.730	35
35.830	35
41.677	38
45.351	38



One purpose of the experiment was to describe how y=CO_2-emission develops over time. The graph suggests that a linear trend may provide a reasonable approximation to the data, over the time span covered by the experiment. The linear function fitted to these data is $\widehat{y} = -36.7 + 2.1x$. A SAS regression output, including ANOVA table, is given below. It can be concluded that the emission of CO_2 increases significantly with time, the rate of increase being about 2.1 units per time unit.

```
Dependent Variable: EMISSION
                                      Sum of           Mean
Source                    DF         Squares         Square  F Value    Pr > F
Model                      1     992.3361798     992.3361798   234.63    0.0001
Error                      6      25.3765201      4.2294200
Corrected Total            7    1017.7126999

                  R-Square             C.V.       Root MSE      EMISSION Mean

                  0.975065         6.887412       2.056555           29.85963

                                        T for H0:    Pr > |T|   Std Error of
Parameter                 Estimate    Parameter=0                 Estimate

INTERCEPT            -36.74430710          -8.33      0.0002     4.40858691
TIME                   2.09776164          15.32      0.0001     0.13695161
```

□

1.8.2 Multiple regression

Generalization of simple linear regression models of type (1.1) to include more than one independent variable is rather straightforward. For example, suppose that y may depend on two variables, and that we have made $n = 6$ observations. The regression model is then $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$, $i = 1, \ldots, 6$. In matrix terms this model is

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \\ 1 & x_{51} & x_{52} \\ 1 & x_{61} & x_{62} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{pmatrix}. \tag{1.23}$$

Example 1.2 Professor Orley Ashenfelter issues a wine magazine, "Liquid assets", giving advice about good years. He bases his advice on multiple regression of

$$y = \text{Price of the wine at wine auctions}$$

with meteorological data as predictors. The New York Times used the headline "Wine Equation Puts Some Noses Out of Joint" on an article about Prof. Ashenberger. Base material was taken from "Departures" magazine, September/October 1990, but the data are invented. The variables in the data set below are:

- Rain_W=Amount of rain during the winter.
- Av_temp=Average temperature.

Table 1.1: *Data for prediction of the quality of wine.*

Year	Rain_W	Av_temp	Rain_H	Quality
1975	123	23	23	89
1976	66	21	100	70
1977	58	20	27	77
1978	109	26	33	87
1979	46	22	102	73
1980	40	19	77	70
1981	42	18	85	60
1982	167	25	14	92
1983	99	28	17	87
1984	48	24	47	79
1985	85	24	28	84
1986	177	27	11	93
1987	80	22	45	75
1988	64	25	40	82
1989	75	25	16	88

- Rain_H=Rain in the harvest season.
- y=Quality, which is an index based on auction prices.

A set of data of this type is reproduced in Table 1.1.

A multiple regression output from Minitab based on these data is as follows:

Regression Analysis

```
The regression equation is
Quality = 48.9 + 0.0594 Rain_W + 1.36 Av_temp - 0.118 Rain_H

Predictor        Coef       StDev          T        P
Constant        48.91       10.41       4.70    0.001
Rain_W        0.05937     0.02767       2.15    0.055
Av_temp        1.3603      0.4187       3.25    0.008
Rain_H       -0.11773     0.04010      -2.94    0.014

S = 3.092       R-Sq = 91.6%     R-Sq(adj) = 89.4%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3     1152.43      384.14     40.18    0.000
Residual Error    11      105.17        9.56
Total             14     1257.60
```

The output indicates that the three predictor variables do indeed have a relationship to the wine quality, as measured by the price. The variable Rain_W is not quite significant but would be included in a predictive model. The size and direction of this relationship is given by the estimated coefficients of the regression equation. It appears that years with much winter rain, a high average temperature, and only a small amount of rain at harvest time, would produce good wine. □

1.8.3 t tests and dummy variables

Classification variables (non-numeric variables), such as treatments, groups or blocks can be included in the model as so called dummy variables, i.e. as variables that only take on the values 0 or 1. For example, a simple t test on data with two groups and three observations per group can be formulated as

$$\begin{aligned} y_{ij} &= \mu + \beta d_i + e_{ij} \\ i &= 1, 2;\ j = 1, 2, 3. \end{aligned}$$

Here, μ is a general mean value, d_i is a dummy variable that has value $d_i = 1$ if observation i belongs to group 1 and $d_i = 0$ if it belongs to group 2, and e_{ij} is a residual. According to this model, the population mean value for group 1 is $\mu_1 = \mu + \beta$ and the population mean value for group 2 is simply $\mu_2 = \mu$. In the t test situation we want to examine whether μ_1 is different from μ_2, i.e. whether β is different from 0. This model can be written in matrix terms as

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \beta \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \end{pmatrix}. \tag{1.24}$$

Example 1.3 In a pharmacological study (Rea et al, 1984), researchers measured the concentration of Dopamine in the brains of six control rats and of six rats that had been exposed to toluene. The concentrations in the striatum region of the brain are given in Table 1.2.

The interest lies in comparing the two groups with respect to average Dopamine level. This is often done as a two sample t test. To illustrate that the t test is actually a special case of a general linear model, we analyzed these data with Minitab using regression analysis with Group as a dummy variable. Rats in the toluene group were given the value 1 on the dummy variable, while rats in the control group were coded as 0. The Minitab output of the regression analysis is:

Table 1.2: *Dopamine levels in the brains of rats under two treatments.*

Dopamine, ng/kg	
Toluene group	Control group
3.420	1.820
2.314	1.843
1.911	1.397
2.464	1.803
2.781	2.539
2.803	1.990

Regression Analysis

```
The regression equation is
Dopamine level = 1.90 + 0.717 Group

Predictor        Coef       StDev          T        P
Constant       1.8987      0.1830      10.38    0.000
Group          0.7168      0.2587       2.77    0.020

S = 0.4482      R-Sq = 43.4%     R-Sq(adj) = 37.8%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      1.5416      1.5416      7.68    0.020
Residual Error    10      2.0084      0.2008
Total             11      3.5500
```

The output indicates a significant Group effect ($t = 2.77$, $p = 0.020$). The size of this group effect is estimated as the coefficient $\widehat{\beta}_1 = 0.7168$. This means that the toluene group has an estimated mean value that is 0.7168 units higher than the mean value in the control group. The reader might wish to check that this calculation is correct, and that the t test given by the regression routine does actually give the same results as a t test performed according to textbook formulas. Also note that the F test in the output is related to the t test through $t^2 = F$: $2.77^2 = 7.68$. These two tests are identical. □

1.8.4 One-way ANOVA

The generalization of models of type (1.24) to more than two groups is rather straightforward; we would need one more column in $\mathbf{X}$ (one new dummy variable) for each new group. This leads to a simple oneway analysis of variance (ANOVA) model. Thus, a one-way ANOVA model with three treatments,

each with two observations per treatment, can be written as

$$\begin{aligned} y_{ij} &= \mu + \beta_i + e_{ij}, \\ i &= 1, 2, 3,\ j = 1, 2 \end{aligned} \tag{1.25}$$

We can introduce three dummy variables d_1, d_2 and d_3 such that $d_i = \begin{cases} 1 & \text{for group } i \\ 0 & \text{otherwise} \end{cases}$. The model can now be written as

$$\begin{aligned} y_{ij} &= \mu + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + e_{ij} \\ &= \mu + \beta_i d_i + e_{ij}, \\ i &= 1, 2, 3,\ j = 1, 2 \end{aligned} \tag{1.26}$$

Note that the third dummy variable d_3 is not needed. If we know the values of d_1 and d_2 the group membership is known so d_3 is redundant and can be removed from the model. In fact, any combination of two of the dummy variables is sufficient for identifying group membership so the choice to delete one of them is to some extent arbitrary. After removing d_3, the model can be written in matrix terms as

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{pmatrix} \tag{1.27}$$

Although there are three treatments we have only included two dummy variables for the treatments, i.e. we have chosen the restriction $\beta_3 = 0$.

Follow-up analyses

One of the results from a one-way ANOVA is an over-all F test of the hypothesis that all group (treatment) means are equal. If this test is significant, it can be followed up by various types of comparisons between the groups. Since the ANOVA provides an estimator $\widehat{\sigma}_e^2 = MS_e$ of the residual variance σ_e^2, this estimator should be used in such group comparisons if the assumption of equal variance seems tenable.

A pairwise comparison between two group means, i.e. a test of the hypothesis that two groups have equal mean values, can be obtained as

$$t = \frac{\overline{y}_i - \overline{y}_{i'}}{\sqrt{MS_e \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)}}$$

with degrees of freedom taken from MS_e. A confidence interval for the difference between the mean values can be obtained analogously.

In some cases it may be of interest to do comparisons which are not simple pairwise comparisons. For example, we may want to compare treatment 1 with the average of treatements 2, 3 and 4. We can then define a contrast in the treatment means as $L = \mu_1 - \frac{\mu_2+\mu_3+\mu_4}{3}$. A general way to write a contrast is

$$L = \sum_i h_i \mu_i, \tag{1.28}$$

where we define the weights h_i such that $\sum_i h_i = 0$. The contrast can be estimated as

$$\widehat{L} = \sum_i h_i \overline{y}_i, \tag{1.29}$$

and the estimated variance of $\widehat{L}$ is

$$\widehat{Var}\left(\widehat{L}\right) = MS_e \sum_i \frac{h_i^2}{n_i}. \tag{1.30}$$

This can be used for tests and confidence intervals on contrasts.

Problems when the number of comparisons is large

After you have obtained a significant F test, there may be many pairwise comparisons or other contrasts to examine. For example, in a one-way ANOVA with seven treatments you can make 21 pairwise comparisons. If you make many tests at, say, the 5% level you may end up with a number of significant results even if all the null hypotheses are true. If you make 100 such tests you would expect, on the average, 5 significant results. Thus, even if the significance level of each individual test is 5% (the so called comparisonwise error rate), the over-all significance level of all tests (the experimentwise error rate), i.e. the probability to get at least one significant result given that all null hypotheses are true, is larger. This is the problem of mass significance.

There is some controversy whether mass significance is a real problem. For example, Nelder (1971) states "In my view, multiple comparison methods have no place at all in the interpretation of data". However, other authors have suggested various methods to protect against mass significance. The general solution is to apply a stricter limit on what we should declare "significant". If a single t test would be significant for $|t| > 2.0$, we could use the limit 2.5 or 3.0 instead. The SAS procedure GLM includes 16 different

Table 1.3: *Change in urine production following treatment with different contrast media ($n = 57$).*

Medium	Diff	Medium	Diff	Medium	Diff
Diatrizoate	32.92	Isovist	2.44	Ringer	0.10
Diatrizoate	25.85	Isovist	0.87	Ringer	0.40
Diatrizoate	20.75	Isovist	−0.22	Mannitol	9.19
Diatrizoate	20.38	Isovist	1.52	Mannitol	0.79
Diatrizoate	7.06	Omnipaque	8.51	Mannitol	10.22
Hexabrix	6.47	Omnipaque	16.11	Mannitol	4.78
Hexabrix	5.63	Omnipaque	7.22	Mannitol	14.64
Hexabrix	3.08	Omnipaque	9.03	Mannitol	6.98
Hexabrix	0.96	Omnipaque	10.11	Mannitol	7.51
Hexabrix	2.37	Omnipaque	6.77	Mannitol	9.55
Hexabrix	7.00	Omnipaque	1.16	Mannitol	5.53
Hexabrix	4.88	Omnipaque	16.11	Ultravist	12.94
Hexabrix	1.11	Omnipaque	3.99	Ultravist	7.30
Hexabrix	4.14	Omnipaque	4.90	Ultravist	15.35
Isovist	2.10	Ringer	0.07	Ultravist	6.58
Isovist	0.77	Ringer	−0.03	Ultravist	15.68
Isovist	−0.04	Ringer	0.34	Ultravist	3.48
Isovist	4.80	Ringer	0.08	Ultravist	5.75
Isovist	2.74	Ringer	0.51	Ultravist	12.18

methods for deciding which limit to use. A simple but reasonably powerful method is to use Bonferroni adjustment. This means that each individual test is made at the significance level α/c, where α is the desired over-all level and c is the number of comparisons you want to make.

Example 1.4 Liss et al (1996) studied the effects of seven contrast media (used in X-ray investigations) on different physiological functions of 57 rats. One variable that was studied was the urine production. Table 1.3 shows the change in urine production of each rat before and after treatment with each medium. It is of interest to compare the contrast media with respect to the change in urine production.

This analysis is a oneway ANOVA situation. The procedure GLM in SAS produced the following result:

```
                        General Linear Models Procedure

Dependent Variable: DIFF   DIFF
                                        Sum of            Mean
Source                DF          Squares          Square   F Value     Pr > F
Model                  6     1787.9722541     297.9953757     16.46     0.0001
Error                 50      905.1155428      18.1023109
Corrected Total       56     2693.0877969

               R-Square             C.V.        Root MSE            DIFF Mean
               0.663912         61.95963       4.2546811            6.8668596

Source                DF      Type III SS     Mean Square   F Value     Pr > F
MEDIUM                 6     1787.9722541     297.9953757     16.46     0.0001
```

There are clearly significant differences between the media ($p < 0.0001$). To find out more about the nature of these differences we requested Proc GLM to print estimates of the parameters, i.e. estimates of the coefficients β_i for each of the dummy variables. The following results were obtained:

```
                                                T for H0:    Pr > |T|   Std Error of
Parameter                      Estimate          Parameter=0                Estimate

INTERCEPT                  9.90787500 B                6.59      0.0001     1.50425691
MEDIUM     Diatrizoate    11.48412500 B                4.73      0.0001     2.42554139
           Hexabrix       -5.94731944 B               -2.88      0.0059     2.06740338
           Isovist        -8.24365278 B               -3.99      0.0002     2.06740338
           Mannitol       -2.21920833 B               -1.07      0.2882     2.06740338
           Omnipaque      -1.51817500 B               -0.75      0.4554     2.01817243
           Ringer         -9.69787500 B               -4.40      0.0001     2.20200665
           Ultravist       0.00000000 B                   .           .              .

NOTE: The X'X matrix has been found to be singular and a generalized inverse
      was used to solve the normal equations.   Estimates followed by the
      letter 'B' are biased, and are not unique estimators of the parameters.
```

Note that Proc GLM reports the $\mathbf{X}'\mathbf{X}$ matrix to be singular. This is as expected for an ANOVA model: not all dummy variables can be included in the model. The procedure excludes the last dummy variable, setting the parameter for Ultravist to 0. All other estimates are comparisons of the estimated mean value for that medium, with the mean value for Ultravist. Least squares estimates of the mean values for the media can be calculated and compared. Since this can result in a large number of pairwise comparisons (in this case, $7 \cdot 6/2 = 21$ comparisons), some method for protection against mass significance might be considered. The least squares means are given in Table 1.4 along with indications of significant pairwise differences using Bonferroni adjustment.

Before we close this example, we should take a look at how the data behave. For example, we can prepare a boxplot of the distributions for the different

Table 1.4: *Least squares means, and pairwise comparisons between treatments, for the contrast media experiment.*

	Diatri-zoate	Ultra-vist	Omni-paque	Manni-tol	Hexa-brix	Isovist	Ringer
Mean	21.39	9.91	8.39	7.69	3.96	1.66	0.21
Diatrizoate	–						
Ultravist	*	–					
Omnipaque	*	n.s.	–				
Mannitol	*	n.s.	n.s.	–			
Hexabrix	*	n.s.	n.s.	n.s.	–		
Isovist	*	*	*	n.s.	n.s.	–	
Ringer	*	*	*	*	n.s.	n.s.	–

media. This boxplot is given in Figure 1.1. The plot indicates that the variation is quite different for the different media, with a large variation for Diatrizoate and a small variation for Ringer (which is actually a placebo). This suggests that one assumption underlying the analysis, the assumption of equal variance, may be violated. We will return to these data later to see if we can make a better analysis. □

1.8.5 ANOVA: Factorial experiments

The ideas used above can be extended to factorial experiments that include more than one factor and possible interactions. The dummy variables that correspond to the interaction terms would then be constructed by multiplying the corresponding main effect dummy variables with each other.

This feature can be illustrated by considering a factorial experiment with factor A (two levels) and factor B (three levels), and where we have two observations for each factor combination. The model is

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, \\ i &= 1, 2,\ j = 1, 2, 3,\ \ k = 1, 2 \end{aligned} \tag{1.31}$$

The number of dummy variables that we have included for each factor is equal to the number of factor levels minus one, i.e. the last dummy variable for each factor has been excluded. The number of non-redundant dummy variables equals the number of degrees of freedom for the effect. In matrix terms,

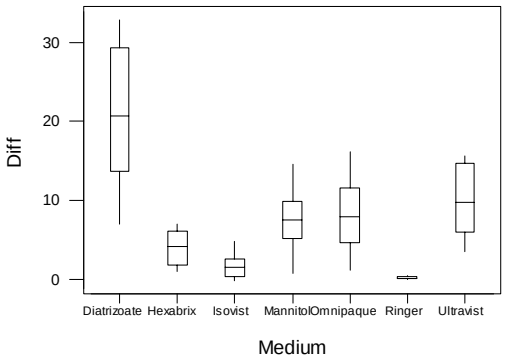


Figure 1.1: *Boxplot of change in urine production for different contrast media.*

$$\begin{pmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{131} \\ y_{132} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \\ y_{231} \\ y_{232} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \end{pmatrix} + \begin{pmatrix} e_{111} \\ e_{112} \\ e_{121} \\ e_{122} \\ e_{131} \\ e_{132} \\ e_{211} \\ e_{212} \\ e_{221} \\ e_{222} \\ e_{231} \\ e_{232} \end{pmatrix}. \quad (1.32)$$

Example 1.5 Lindahl et al (1999) studied certain reactions of fungus myceliae on pieces of wood by using radioactively labeled ^{32}P. In one of the experiments, two species of fungus (*Paxillus involutus* and *Suillus variegatus*) were used, along with two sizes of wood pieces (Large and Small); the response was a certain chemical measurement denoted by C. The data are reproduced in Table 1.5.

These data were analyzed as a factorial experiment with two factors. Part of the Minitab output was:

Table 1.5: *Data for a two-factor experiment.*

Species	Size	C	Species	Size	C
H	Large	0.0010	S	Large	0.0021
H	Large	0.0011	S	Large	0.0001
H	Large	0.0017	S	Large	0.0016
H	Large	0.0008	S	Large	0.0046
H	Large	0.0010	S	Large	0.0035
H	Large	0.0028	S	Large	0.0065
H	Large	0.0003	S	Large	0.0073
H	Large	0.0013	S	Large	0.0039
H	Small	0.0061	S	Small	0.0007
H	Small	0.0010	S	Small	0.0011
H	Small	0.0020	S	Small	0.0019
H	Small	0.0018	S	Small	0.0022
H	Small	0.0033	S	Small	0.0011
H	Small	0.0015	S	Small	0.0012
H	Small	0.0040	S	Small	0.0009
H	Small	0.0041	S	Small	0.0040

General Linear Model: C versus Species; Size

```
Analysis of Variance for C, using Adjusted SS for Tests

Source          DF     Seq SS     Adj SS     Adj MS       F      P
Species          1  0.0000025  0.0000025  0.0000025    0.93  0.342
Size             1  0.0000002  0.0000002  0.0000002    0.09  0.772
Species*Size     1  0.0000287  0.0000287  0.0000287   10.82  0.003
Error           28  0.0000742  0.0000742  0.0000027
Total           31  0.0001056
```

The main conclusion from this analysis is that the interaction Species × Size is highly significant. This means that the effect of Size is different for different species. In such cases, interpretation of the main effects is not very meaningful. As a tool for interpreting the interaction effect, a so called interaction plot can be prepared. Such a plot for these data is as given in Figure 1.2. The mean value of the response for species S is higher for large wood pieces than for small wood pieces. For species H the opposite is true: the mean value is larger for small wood pieces. This is an example of an interaction. □

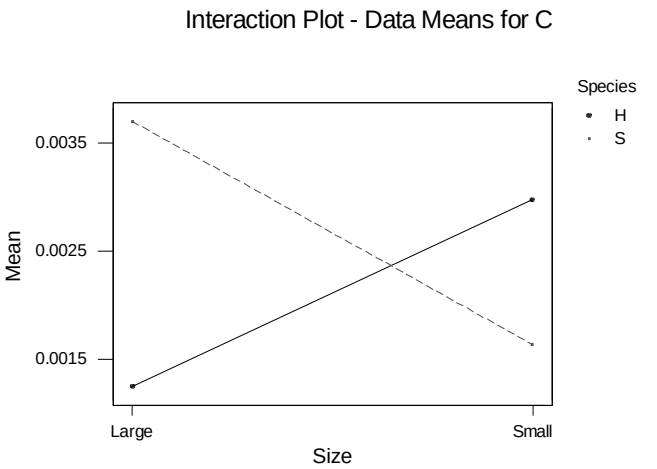


Figure 1.2: *Interaction plot for the 2-factor experiment.*

1.8.6 Analysis of covariance

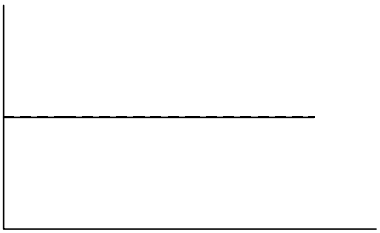
In regression analysis models the design matrix $\mathbf{X}$ contains quantitative variables. In ANOVA models, the design matrix only contains dummy variables corresponding to treatments, design structure and possible interactions. It is quite possible to include a mixture of quantitative variables and dummy variables in the design matrix. Such models are called covariance analysis, or ANCOVA, models.

Let us look at a simple case where there are two groups and one covariate. Several different models can be considered for the analysis of such data even in the simple case where we assume that all relationships are linear:

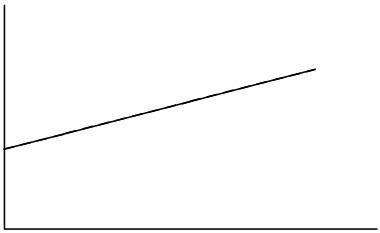
1. There is no relationship between x and y in any of the groups and the groups have the same mean value.
2. There is a relationship between x and y; the relationship is the same in the groups.
3. There is no relationship between x and y but the groups have different levels.
4. There is a relationship between x and y; the lines are parallel but at different levels.

5. There is a relationship between x and y; the lines are different in the groups.

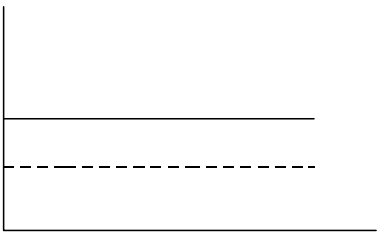
These five cases correspond to different models that can be represented in graphs or in formulas:



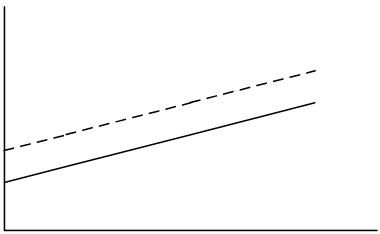
Model 1: $y_{ij} = \mu + e_{ij}$



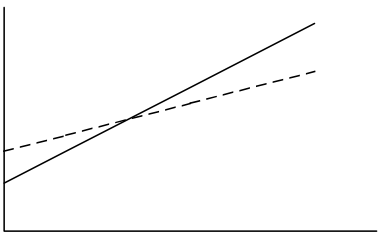
Model 2: $y_{ij} = \mu + \beta x + e_{ij}$



Model 3: $y_{ij} = \mu + \alpha_i + e_{ij}$



Model 4: $y_{ij} = \mu + \alpha_i + \beta x + e_{ij}$



Model 5: $y_{ij} = \mu + \alpha_i + \beta x + \gamma \cdot d_i \cdot x + e_{ij}$

Model 5 is the most general of the models, allowing for different intercepts $(\mu + \alpha_i)$ and different slopes $\beta + \gamma d_i$, where d is a dummy variable indicating group membership. If it can be assumed that the term γd_i is zero for all i, then we are back at model 4. If, in addition, all α_i are zero, then model 2 is correct. If, on the other hand. β is zero, we would use model 3. If finally β is zero in model 2, then model 1 describes the situation. This is an example of a set of models where some of the models are nested within other models. The model choice can be made by comparing any model to a simpler model which only differs in terms of one factor.

1.8.7 Non-linear models

Models can be non-linear in different ways. A model can contain non-linear functions of the parameters, like $y = \beta_0 + \beta_1 e^{\beta_2 x} + e$. We will not consider such models, which are called intrinsically nonlinear, or nonlinear in the parameters. Some models can be transformed into a linear form by a suitable choice of transformation. For example, the model $y = e^{\beta_0 + \beta_1 x}$ can be made linear by using a log transformation: $\log(y) = \beta_0 + \beta_1 x$. Other models can be linear in the parameters, but nonlinear in the variables, like

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 e^{x_i} + e_i. \tag{1.33}$$

Such models are simple to analyze using general linear models. Formally, each transformation of x is treated as a new variable. Thus, if we denote $u_i = x_i^2$ and $v_i = e^{x_i}$ then the model (1.33) can be written as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 u_i + \beta_3 v_i + e_i \tag{1.34}$$

which is a standard multiple regression model. Models of this type can be handled using standard GLM software.

1.9 Estimability

In some types of general linear models it is impossible to estimate all model parameters. It is then necessary to restrict some parameters to be zero, or to use some other restriction on the parameters.

As an example, a two-factor ANOVA model with two levels of factor A, three levels of factor B and two replications can be written as

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, & (1.35)\\ i &= 1, 2,\ j = 1, 2, 3,\ \ k = 1, 2. & (1.36) \end{aligned}$$

In this model it would be possible to replace μ with $\mu + c$ and to replace each α_i with $\alpha_i - c$, where c is some constant. The same kind of ambiguity holds also for other parameters of the model. This model contains a total of 12 parameters: μ, α_1, α_2, β_1, β_2, β_3, $(\alpha\beta)_{11}$, $(\alpha\beta)_{12}$, $(\alpha\beta)_{13}$, $(\alpha\beta)_{21}$, $(\alpha\beta)_{22}$, and $(\alpha\beta)_{23}$, but only 6 of the parameters can be estimated. As noted above, computer programs often solve this problem by restricting some parameters to be zero.

However, it may be possible to estimate certain functions of the parameters in a unique way. Such functions, if they exist, are called estimable functions. A linear combination of model parameters is estimable if it can be written as a linear combination of expected values of the observations.

Let us denote with $\mu_{ij\cdot}$ the mean value for the treatment combination that has factor A at level i and factor B at level j. It holds that

$$\mu_{ij\cdot} = E\left(y_{ijk}\right) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \tag{1.37}$$

which is a linear function of the parameters. This function is estimable. In addition, any linear function of the μ_{ij}:s is also estimable. For example, the expected value of all observations with factor A at level i can be written as

$$\mu_{i\cdot\cdot} = \frac{\mu_{11\cdot} + \mu_{12\cdot} + \mu_{13\cdot}}{3}.$$

This is a linear function of cell means. Since the cell means are estimable, $\mu_{i\cdot\cdot}$ is also estimable.

1.10 Assumptions in General linear models

The classical application of general linear models rests on the following set of assumptions:

The model used for the analysis is assumed to be correct.

The residuals are assumed to be independent.

The residuals are assumed to follow a Normal distribution.

The residuals are assumed to have the same variance σ_e^2, independent of $\mathbf{X}$, i.e. the residuals are homoscedastic.

Different diagnostic tools have been developed to detect departures from these assumptions. Since similar tools are used for generalized linear models, reference is made to Chapter 3 for details.

1.11 Model building

1.11.1 Computer software for GLM:s

There are many options for fitting general linear models to data. One option is to use a regression package and leave it to the user to construct appropriate dummy variables for class variables. However, most statistical packages have routines for general linear models that automatically construct the appropriate set of dummy variables.

Let us use letters at the end of the alphabet (X, Y, Z) to denote numeric variables. Y will be used for the dependent variable. Letters in the beginning of the alphabet (A, B) will symbolize class variables (groups, treatments, blocks, etc.)

Computer software requires the user to state the model in symbolic terms. The model statement contains operators that specify different aspects of the model. In the following table we list the operators used by SAS. Examples of the use of the operators are given below.

Operator	Explanation, SAS example
*	Interaction: A*B. Also used for polynomials: X*X
(none)	Both effects present: A B
|	All main effects and interactions: A|B=A B A*B
()	Nested factor: A(B). "A nested within B"
@	Order operator: A|B|C @ 2 means that all main effects and all interaction up to and including second order interactions are included.

The kinds of models that we have discussed in this chapter can symbolically be written in SAS language as indicated in the following table.

Model	Computer model (SAS)
Simple linear regression	Y = X
Multiple regression	Y = X Z
t tests, oneway ANOVA	Y = A
Two-way ANOVA with interaction	Y = A B A*B or Y=A|B
Covariance analysis model 1	Y =
Covariance analysis model 2	Y = X
Covariance analysis model 3	Y = A
Covariance analysis model 4	Y = A X
Covariance analysis model 5	Y = A X A*X

1.11.2 Model building strategy

Statistical model building is an art as much as it is a science. There are many requirements on models: they should make sense from a subject-matter point of view, they should be simple, and at the same time they should capture most of the information in the data. A good model is a compromise between parsimony and completeness. This means that it is impossible to state simple rules for model building: there will certainly be cases where the rules are not

relevant. However, the following suggestions, partially based on McCullagh and Nelder (1989, p. 89), are useful in many cases:

- Include all relevant main effects in the model, even those that are not significant.
- If an interaction is included, the model should also include all main effects and interactions it comprises. For example, if the interaction A*B*C is included, the model should also include A, B, C, A*B, A*C and B*C.
- A model that contains polynomial terms of type x^a should also contain the lower-degree terms x, x^2, ... , x^{a-1}.
- Covariates that do not have any detectable effect should be excluded.
- The conventional 5% significance level is often too strict for model building purposes. A significance level in the range 15-25% may be used instead.
- Alternatively, criteria like the Akaike information criterion can be used. This is discussed on page 46 in connection with generalized linear models.

1.11.3 A few SAS examples

In SAS terms, grouping variables (classification variables) are called CLASS variables. As examples of SAS programs for a few of the models discussed above we can consider the regression model (1.1) using Proc GLM. The analysis could be done with a program that does not include any CLASS variables:

```
PROC GLM DATA=Regression;
   MODEL y = x;
RUN;
```

The t test (or the oneway ANOVA) can be modelled as

```
PROC GLM DATA=Anova;
   CLASS group;
   MODEL y = group;
RUN;
```

The difference between the two programs is that in the t test, the independent variable ("group") is given as a CLASS variable. This asks SAS to build appropriate dummy variables.

1.12 Exercises

Exercise 1.1 Cicirelli et al (1983) studied protein synthesis in developing egg cells of the frog *Xenopus laevis*. Radioactively labeled leucine was injected into egg cells. At various times after injection, radioactivity measurements were made. From these measurements it was possible to calculate how much of the leucine had been incorporated into protein. The following data, quoted from Samuels and Witmer (1999), are mean values of two egg cells. All egg cells were taken from the same female.

Time	Leucine (ng)
0	0.02
10	0.25
20	0.54
30	0.69
40	1.07
50	1.50
60	1.74

A. Use linear regression to estimate the rate of incorporation of the labeled leucine.

B. Plot the data and the regression line.

C. Prepare an ANOVA table.

Exercise 1.2 The level of cortisol has been measured for three groups of patients with different syndromes: a) adenoma b) bilateral hyperplasia c) cardinoma. The results are summarized in the following table:

a	b	c
3.1	8.3	10.2
3.0	3.8	9.2
1.9	3.9	9.6
3.8	7.8	53.8
4.1	9.1	15.8
1.9	15.4	
	7.7	
	6.5	
	5.7	
	13.6	

A. Make an analysis of these data that can answer the question whether there are any differences in cortisol level between the groups. A complete solution should contain hypotheses, calculations, test statistic, and a conclusion. A

graphical display (for example a boxplot) may help in the interpretation of the results.

B. There are some indications that the assumptions underlying the analysis in A. are not fulfilled. Examine this, indicate what the problems are, and suggest what can be done to improve the analysis. No new ANOVA is needed.

Exercise 1.3 Below are some data on the emission of carbon dioxide from the root system of plants (Zagal et al, 1993). Two levels of nitrogen were used, and samples of plants were analyzed 24, 30, 35 and 38 days after germination. The data were as follows:

	Days from germination			
Level of Nitrogen	24	30	35	38
High	8.220	19.296	25.479	31.186
	12.594	31.115	34.951	39.237
	11.301	18.891	20.688	21.403
Low	15.255	28.200	32.862	41.677
	11.069	26.765	34.730	43.448
	10.481	28.414	35.830	45.351

A. Analyze the data in a way that treats Days from germination as a quantitative factor. Treat level of nitrogen as a dummy variable, and assume that all regressions are linear.

i) Fit a model that assumes that the two regression lines are parallel.

ii) Fit a model that does not assume that the regression lines are parallel.

iii) Test the hypothesis that the regressions are parallel.

B. What is the expected rate of CO_2 emission for a plant with a high level of nitrogen, 35 days after germination? The same question for a plant with a low level of nitrogen? Use the model you consider the best of the models you have fitted under A. and B. above. Make the calculation by hand, using the computer printouts of model equations.

C. Graph the data. Include both the observed data and the fitted Y values in your graph.

D. According to your best analysis above, is there any significant effect of:

i) Interaction

ii) Level of nitrogen

ii) Days from germination

Exercise 1.4 Gowen and Price, quoted from Snedecor and Cochran (1980), counted the number of lesions of Aucuba mosaic virus after exposure to X-rays for various times. The results were:

Exposure	Count
0	271
15	108
30	59
45	29
60	12

It was assumed that the Count (y) depends on the exposure time (x) through an exponential relation of type $y = Ae^{-Bx}$. A convenient way to estimate the parameters of such a function is to make a linear regression of $\log(y)$ on x.

A. Perform a linear regression of $\log(y)$ on x.

B. What assumptions are made regarding the residuals in your analysis in A.?

C. Plot the data and the fitted function in the same graph.

2. Generalized Linear Models

2.1 Introduction

In Chapter 1 we briefly summarized the theory of general linear models (GLM:s). GLM:s are very useful for data analysis. However, GLM:s are limited in many ways. Formally, the classical applications of GLM:s rest on the assumptions of normality, linearity and homoscedasticity.

The generalization of GLM:s that we will present in this chapter will allow us to model our data using other distributions than the Normal. The choice of distribution affects the assumptions we make regarding variances, since the relation between the variance and the mean is known for many distributions. For example, the Poisson distribution has the property that $\mu = \sigma^2$.

This chapter is the most theoretical chapter in the book. It builds on the theory of Maximum Likelihood estimation (see Appendix B), and on the class of distributions called the exponential family. In later chapters we will apply the theory in different situations.

2.1.1 Types of response variables

This book is concerned with statistical models for data. In these models, the concept of a response variable is crucial. In general linear models, the response variable Y is often assumed to be quantitative and normally distributed. But this is by no means the only type of response variables that we might meet in practice. Some examples of different types of response variables are:

- Continuous response variables.
- Binary response variables.
- Response variables in the form of proportions.
- Response variables in the form of counts.
- Response in the form of rates.
- Ordinal response.

We will here give a few examples of these types of response variables.

2.1.2 Continuous response

Models where the response variable is considered to be continuous are common in many application areas. In fact, since measurements cannot be made to infinite precision, few response variables are truly continuous, but continuous models are still often used as approximations. Many response variables of this type are modeled as general linear models, often assuming normality and homoscedasticity. It is common for response variables to be restricted to positive values. Physical measurements in cm or kg are examples of this. Since the Normal distribution is defined on $[-\infty, \infty]$, the normality assumption cannot hold exactly for such data, and one has to revert to approximations.

We may illustrate the concept of continuous response using data of a type often used in general linear models; other examples will be discussed in later chapters.

Example 2.1 In the pharmacological study discussed in Example 1.3 the concentration of Dopamine was measured in the brains of six control rats and of six rats that had been exposed to toluene. The results were given on page 13. In this example the response variable may be regarded as essentially continuous. □

2.1.3 Response as a binary variable

Binary response, often called quantal response in earlier literature, is the result of measurements where it has only been recorded whether an event has occurred ($Y = 1$) or not ($Y = 0$). A common approach to modeling this type of data is to model the probability that the event will occur. Since a probability p is limited by $0 \leq p \leq 1$, models for the data should use this restriction. Binary data are often modeled using the Bernoulli distribution,

which is a special case of the Binomial distribution where $n = 1$. The binomial distribution is further discussed on page 88.

Example 2.2 Collett (1991), quoting data from Brown (1980), reports some data on the treatment of prostatic cancer. The issue of concern was to find indicators whether the cancer had spread to the surrounding lymph nodes. Surgery is needed to ascertain the extent of nodal involvement. Some variables that can be measured without surgery may be indicators of nodal involvement. Thus, one purpose of the modeling is to formulate a model that can predict whether or not the lymph nodes have been affected. The data are of the type given in the following table. Only a portion of the data is listed; the actual data set contained 53 patients.

Age	Acid level	X-ray result	Tumour size	Tumour grade	Nodal involvement
66	0.48	0	0	0	0
65	0.46	1	0	0	0
61	0.50	0	1	0	0
58	0.48	1	1	0	1
65	0.84	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮

In this type of data, the response Y has value 1 if nodal involvement has occurred and 0 otherwise. This is called a binary response. Even some of the independent variables (X-ray results, Tumour size and Tumour grade) are binary variables, taking on only the values 0 or 1. These data will be analyzed in Chapter 5. □

2.1.4 Response as a proportion

Response in the form of proportions (binomial response) is obtained when a group of n individuals is exposed to the same conditions. f out of the n individuals respond in one way ($Y = 1$) while the remaining $n-f$ individuals respond in some other way ($Y = 0$). The response is the proportion $\widehat{p} = \frac{f}{n}$. The response of the individuals might be to improve from a certain medical treatment; to die from a specified dose of an insecticide; or for a piece of equipment to fail. A proportion corresponds to a probability, and modeling of the response probability is an important part of the data analysis. In such models the fact that $0 \leq p \leq 1$ should be allowed to influence the choice of model. Binary response is a special case of binomial response with $n = 1$.

Example 2.3 Finney (1947) reported on an experiment on the effect of Rotenone, in different concentrations, when sprayed on the insect *Macrosi-*

phoniella sanborni, in batches of about fifty. The results are given in the following table.

Conc	Log(Conc)	No. of insects	No. affected	% affected
10.2	1.01	50	44	88
7.7	0.89	49	42	86
5.1	0.71	46	24	52
3.8	0.58	48	16	33
2.6	0.41	50	6	12

One aim with this experiment was to find a model for the relation between the probability p that an insect is affecteded and the dose, i.e. the concentration. Such a model can be written, in general terms, as

$$g(p) = f(\text{Concentration}).$$

The functions g and f should be chosen such that the model cannot produce a predicted probability that is larger than 1. These data will be discussed later on page 89. □

2.1.5 Response as a count

Counts are measurements where the response indicates how many times a specific event has occurred. Counts are often recorded in the form of frequency tables or crosstabulations. Count data are restricted to integers ≥ 0. Models for counts should take this limitation into account.

Example 2.4 Sokal and Rohlf (1973) reported some data on the color of Tiger beetles (*Cicindela fulgida*) collected during different seasons. The results are:

Season	Red	Other	Total
Early spring	29	11	40
Late spring	273	191	464
Early summer	8	31	39
Late summer	64	64	128
Total	374	297	671

The data may be used to study how the color of the beetle depends on season. A common approach is to test whether there is independence between season and color through a χ^2 test. We will return to the analysis of these data later (page 117). □

2.1.6 Response as a rate

In some cases, the response can be assumed to be proportional to the size of the object being measured. For example, the number of birds of a certain species that have been sighted may depend on the area of the habitat that has been surveyed. In this case the response may be measured as "number of sightings per km^2", which we will call a rate. In the analysis of data of this type, one has to account for differences in size between objects.

Example 2.5 The data below, quoted from Agresti (1996), are accident rates for elderly drivers, subdivided by sex. For each sex the number of person years (in thousands) is also given. The data refer to 16262 Medicaid enrollees.

	Females	Males
No. of accidents	175	320
No. of person years ('000)	17.3	21.4

Accident data can often be modeled using the Poisson distribution. In this case, we have to account for the fact that males and females have different observation periods, in terms of number of person years. Accident rate can be measured as (no. of accidents)/(no. of person years). In a later chapter (page 131), we will discuss how this type of data can be modelled. □

2.1.7 Ordinal response

Response variables are sometimes measured on an ordinal scale, i.e. on a scale where the categories are ordered but where the distance between scale steps is not constant. Examples of such variables are ratings of patients; answers to attitude items; and school marks.

Example 2.6 Norton and Dunn (1985) studied the relation between snoring and heart problems for a sample of 2484 patients. The data were obtained through interviews with the patients. The amount of snoring was assessed on a scale ranging from "Never" to "Always", which is an ordinal variable. An interesting question is whether there is any relation between snoring and heart problems. The data are:

	Snoring				
Heart problems	Never	Some-times	Often	Always	Total
Yes	24	35	21	30	110
No	1355	603	192	224	2374
Total	1379	638	213	254	2484

The main interest lies in studying possible dependence between snoring and heart problems. Analysis of ordinal data is discussed in Chapter 7. □

2.2 Generalized linear models

Generalized linear models provide a unified approach to modelling of all the types of response variables we have met in the examples above. In this section we will summarize the theory of generalized linear models. In later sections we will return to the examples and see how the theory can be applied in specific cases.

Let us return to the general linear model (1.3):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{2.1}$$

Let us denote

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \tag{2.2}$$

as the *linear predictor* part of the model (1.3). Generalized linear models are a generalization of general linear models in the following ways:

1. An assumptions often made in a GLM is that the components of $\mathbf{y}$ are independently normally distributed with constant variance. We can relax this assumption to permit the distribution to be any distribution that belongs to the exponential family of distributions. This includes distributions such as Normal, Poisson, gamma and binomial distributions.

2. Instead of modeling $\boldsymbol{\mu} = E(\mathbf{y})$ directly as a function of the linear predictor $\mathbf{X}\boldsymbol{\beta}$, we model some function $g(\boldsymbol{\mu})$ of $\boldsymbol{\mu}$. Thus, the model becomes

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}. \tag{2.3}$$

The function $g(\cdot)$ in (2.3), is called a *link function*.

The specification of a generalized linear model thus involves:

1. specification of the distribution
2. specification of the link function $g(\cdot)$
3. specification of the linear predictor $\mathbf{X}\boldsymbol{\beta}$.

We will discuss these issues, starting with the distribution.

2.3 The exponential family of distributions

The exponential family is a general class of distributions that includes many well known distributions as special cases. It can be written in the form

$$f(y;\theta,\phi) = \exp\left[\frac{(y\theta - b(\theta))}{a(\phi)} + c(y,\phi)\right] \tag{2.4}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are some functions. The so called canonical parameter θ is some function of the location parameter of the distribution. Some authors differ between exponential family, which is (2.4) assuming that $a(\phi)$ is unity, and exponential dispersion family, which include the function $a(\phi)$ while assuming that the so called dispersion parameter ϕ is a constant; see Jørgensen (1987); Lindsey (1997, p. 10f). As examples of the usefulness of the exponential family, we will demonstrate that some well-known distributions are, in fact, special cases of the exponential family.

2.3.1 The Poisson distribution

The Poisson distribution can be written as a special case of an exponential family distribution. It has probability function

$$\begin{aligned} f(y;\mu) &= \frac{\mu^y e^{-\mu}}{y!} \\ &= \exp\left[y\log(\mu) - \mu - \log(y!)\right]. \end{aligned} \tag{2.5}$$

We can compare this expression with (2.4). We note that $\theta = \log(\mu)$ which means that $\mu = \exp(\theta)$. We insert this into (2.5) and get

$$f(y;\mu) = \exp\left[y\theta - \exp(\theta) - \log(y!)\right]$$

Thus, (2.5) is a special case of (2.4) with $\theta = \log(\mu)$, $b(\theta) = \exp(\theta)$, $c(y,\phi) = -\log(y!)$ and $a(\phi) = 1$.

2.3.2 The binomial distribution

The binomial distribution can be written as

$$\begin{aligned} f(y;p) &= \binom{n}{y} p^y (1-p)^{n-y} \\ &= \exp\left[y\log\left(\frac{p}{1-p}\right) + n\log(1-p) + \log\binom{n}{y}\right]. \end{aligned} \tag{2.6}$$

We use $\theta = \log\left(\frac{p}{1-p}\right)$ i.e. $p = \frac{\exp(\theta)}{1+\exp(\theta)}$. This can be inserted into 2.6 to give

$$f(y;p) = \exp\left[y\theta + n\log\left(\frac{1}{1+\exp(\theta)}\right) + \log\binom{n}{y}\right].$$

It follows that the binomial distribution is an exponential family distribution with $\theta = \log\left(\frac{p}{1-p}\right)$, $b(\theta) = n\log[1+\exp(\theta)]$, $c(y,\phi) = \log\binom{n}{y}$ and $a(\phi) = 1$.

2.3.3 The Normal distribution

The Normal distribution can be written as

$$\begin{aligned} f\left(y;\mu,\sigma^2\right) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-\mu)^2}{2\sigma^2}} \\ &= \exp\left[\frac{\left(y\mu - \frac{\mu^2}{2}\right)}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right]. \end{aligned} \tag{2.7}$$

This is an exponential family distribution with $\theta = \mu$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \theta^2/2$, and $c(y,\phi) = -\left[y^2/\phi + \log(2\pi\phi)\right]/2$. (In fact, it is an exponential dispersion family distribution; see above.)

2.3.4 The function $b(\cdot)$

The function $b(\cdot)$ is of special importance in generalized linear models because $b(\cdot)$ describes the relationship between the mean value and the variance in the distribution. To show how this works we consider Maximum Likelihood estimation of the parameters of the model. For a brief introduction to Maximum Likelihood estimation reference is made to Appendix B.

The first derivative: b'

We denote the log likelihood function with $l(\theta,\phi;y) = \log f(y;\theta,\phi)$. According to likelihood theory it holds that

$$E\left(\frac{\partial l}{\partial \theta}\right) = 0 \tag{2.8}$$

and that

$$E\left(\frac{\partial^2 l}{\partial \theta^2}\right) + E\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right] = 0. \tag{2.9}$$

From (2.4) we obtain that $l(\theta;\phi,y) = (y\theta - b(\theta))/a(\phi) + c(y,\phi)$. Therefore,

$$\frac{\partial l}{\partial \theta} = [y - b'(\theta)]/a(\phi) \tag{2.10}$$

and

$$\frac{\partial^2 l}{\partial \theta^2} = -b''(\theta)/a(\phi) \tag{2.11}$$

where b' and b'' denote the first and second derivative, respectively, of b with respect to θ. From (2.8) and (2.10) we get

$$E\left(\frac{\partial l}{\partial \theta}\right) = E\{[y - b'(\theta)]/a(\phi)\} = 0 \tag{2.12}$$

so that

$$E(y) = \mu = b'(\theta). \tag{2.13}$$

Thus the mean value of the distribution is equal to the first derivative of b with respect to θ. For the distributions we have discussed so far, these derivatives are:

$$\begin{aligned}
\text{Poisson} &: b(\theta) = \exp(\theta) \text{ gives } b'(\theta) = \exp(\theta) = \mu \\
\text{Binomial} &: b(\theta) = n\log(1+\exp(\theta)) \text{ gives } b'(\theta) = n\frac{\exp(\theta)}{1+\exp(\theta)} = np \\
\text{Normal} &: b(\theta) = \frac{\theta^2}{2} \text{ gives } b'(\theta) = \theta = \mu
\end{aligned}$$

For each of the distributions the mean value is equal to $b'(\theta)$.

The second derivative: b''

From (2.9) and (2.11) we get

$$-\frac{b''(\theta)}{a(\phi)} + \frac{Var(y)}{a^2(\phi)} = 0 \tag{2.14}$$

so that

$$Var(y) = a(\phi)\cdot b''(\theta). \tag{2.15}$$

We see that the variance of y is a product of two terms: the second derivative of $b(\cdot)$, and the function $a(\phi)$ which is independent of θ. The parameter ϕ is called the dispersion parameter and $b''(\theta)$ is called the variance function.

For the distributions that we have discussed so far the variance functions are as follows:

$$\begin{aligned}
\text{Poisson} \quad &: \quad b''(\theta) = \exp(\theta) = \mu \\
\text{Binomial} \quad &: \quad b''(\theta) = \frac{n\exp(\theta)(1+\exp(\theta)) - (\exp(\theta))^2}{(1+\exp(\theta))^2} \\
&= \quad n\frac{\exp(\theta)}{(1+\exp(\theta))^2} = np(1-p) \\
\text{Normal} \quad &: \quad a(\phi)\, b''(\theta) = \phi \cdot 1 = \sigma^2
\end{aligned}$$

The variance function is often written as $V(\mu) = b''(\theta)$. The notation $V(\mu)$ does not mean "the variance of μ"; rather, $V(\mu)$ indicates how the variance depends on the mean value μ in the distribution, where μ in turn is a function of θ. In the table on page 41 we summarize some characteristics of a few distributions in the exponential family; see also McCullagh and Nelder (1989).

2.4 The link function

The link function $g(\cdot)$ is a function relating the expected value of the response Y to the predictors $X_1 \ldots X_p$. It has the general form $g(\mu) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$. The function $g(\cdot)$ must be monotone and differentiable. For a monotone function we can define the inverse function $g^{-1}(\cdot)$ by the relation $g^{-1}(g(\mu)) = \mu$. The choice of link function depends on the type of data. For continuous normal-theory data an identity link may be appropriate. For data in the form of counts, the link function should restrict μ to be positive, while data in the form of proportions should use a link that restricts μ to the interval $[0, 1]$. Some commonly used link functions and their inverses are:

The identity link: $\eta = \mu$. The inverse is simply $\mu = \eta$.

The logit link: $\eta = \log[\mu/(1-\mu)]$. The inverse $\mu = \frac{\exp(\eta)}{1+\exp(\eta)}$ is restricted to the interval $[0, 1]$.

The probit link: $\eta = \Phi^{-1}(\mu)$, where Φ is the standard Normal distribution function. The inverse $\mu = \Phi(\eta)$ is restricted to the interval $[0, 1]$.

The complementary log-log link: $\eta = \log[-\log(1-\mu)]$. The inverse $\mu = 1 - \exp(-\exp(\eta))$ is restricted to the interval $[0, 1]$.

Power links: $\eta = \left(\mu^\lambda - 1\right)/\lambda$ where we take $\eta = \log(\mu)$ for $\lambda = 0$. Examples of power links are $\eta = \mu^2$; $\eta = \frac{1}{\mu}$; $\eta = \sqrt{\mu}$; and $\eta = \log(\mu)$. These all belong to the Box-Cox family of transformations. For $\lambda \neq 0$, the inverse

Table 2.1: *Properties of some distributions in the exponential family*

	Poisson	Binomial	Normal	Gamma	Inverse Gaussian	Negative Binomial
$a(\phi)$	1	1	σ^2	ν^{-1}	σ^2	1
$b(\cdot)$	e^{θ}	$n\log\left(1+e^{\theta}\right)$	$\theta^2/2$	$-\log(-\theta)$	$-(-2\theta)^{1/2}$	$-\frac{\log(1-e^{\theta})}{k}$
$c(y;\phi)$	$-\log y!$	$\log\binom{n}{y}$	$-\frac{1}{2}\left(\frac{y^2}{\phi}+\log(2\pi\phi)\right)$	$\nu\log(\nu y)$ $-\log y-\log\Gamma(\nu)$	$-\frac{1}{2}\log\left(2\pi\phi y^3+\frac{1}{\phi y}\right)$	
$\mu(\theta)=E(Y;\theta)$	e^{θ}	$n\cdot e^{\theta}/\left(1+e^{\theta}\right)$	θ	$-\frac{1}{\theta}$	$-\frac{1}{\theta^2}$	$-\frac{e^{\theta}}{(-1+e^{\theta})k}$
Canonical link	log	logit	identity	$-\frac{1}{\mu}$	$-\frac{1}{\mu^2}$	$\log\left(\frac{k\mu}{1+k\mu}\right)$
$V(\mu)$	μ	$np(1-p)$	σ^2	μ^2	μ^3	$\mu+k\mu^2$
$Var(y)$	μ	$n\mu(1-\mu)$	σ^2	μ^2/ν	$\sigma^2\mu^3$	$\mu+k\mu^2$

Figure 2.1:

link is $\mu = e^{\frac{\ln(\lambda\eta+1)}{\lambda}}$. For the log link with $\lambda = 0$, the inverse link is $\mu = \exp(\eta)$ which is restricted to the interval $0, \infty$.

2.4.1 Canonical links

Certain link functions are, in a sense, "natural" for certain distributions. These are called *canonical links*. The canonical link is that function which transforms the mean to a canonical location parameter of the exponential dispersion family member (Lindsey, 1997). This means that the canonical link is that function $g(\cdot)$ for which $g(\mu) = \theta$. It holds that:

$$\begin{aligned} \text{Poisson} &: \theta = \log(\mu) \text{ so the canonical link is log.} \\ \text{Binomial} &: \theta = \log\frac{p}{1-p} \text{ which is the logit link.} \\ \text{Normal} &: \theta = \mu \text{ so the canonical link is the identity link.} \end{aligned}$$

The canonical links for a few distributions are listed in the table on page 41. Computer procedures such as Proc Genmod in SAS use the canonical link by default once the distribution has been specified. It should be noted, however, that there is no guarantee that the canonical links will always provide the "best" model for a given set of data. In any particular application the data may exhibit peculiar behavior, or there may be theoretical justification for choosing links other than the canonical links.

2.5 The linear predictor

The linear predictor $\mathbf{X}\boldsymbol{\beta}$ plays the same role in generalized linear models as in general linear models. In regression settings, $\mathbf{X}$ contains values of independent variables. In ANOVA settings, $\mathbf{X}$ contains dummy variables corresponding to qualitative predictors (treatments, blocks etc). In general, the model states that some function of the mean of $\mathbf{y}$ is a linear function of the predictors: $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$. As noted in Chapter 1, $\mathbf{X}$ is called a design matrix.

2.6 Maximum likelihood estimation

Estimation of the parameters of generalized linear models is often done using the Maximum Likelihood method. The estimates are those parameter values

that maximize the log likelihood, which for a single observation can be written

$$l = \log\left[L\left(\theta, \phi; y\right)\right] = \frac{y\theta - b\left(\theta\right)}{a\left(\phi\right)} + c\left(y, \phi\right). \tag{2.16}$$

The parameters of the model is a $p \times 1$ vector of regression coefficients $\boldsymbol{\beta}$ which are, in turn, functions of θ. Differentiation of l with respect to the elements of $\boldsymbol{\beta}$, using the chain rule, yields

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta}\frac{d\theta}{d\mu}\frac{d\mu}{d\eta}\frac{\partial \eta}{\partial \beta_j}. \tag{2.17}$$

We have shown earlier that $b'\left(\theta\right) = \mu$, and that $b''\left(\theta\right) = V$, the variance function. Thus, $\frac{\partial \mu}{\partial \theta} = V$. From the expression for the linear predictor $\eta = \sum_j x_j\beta_j$ we obtain $\frac{\partial \eta}{\partial \beta_j} = x_j$. Putting things together,

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} &= \frac{\left(y - \mu\right)}{a\left(\phi\right)}\frac{1}{V}\frac{d\mu}{d\eta}x_j \\ &= \frac{W}{a\left(\phi\right)}\left(y - \mu\right)\frac{d\eta}{d\mu}x_j. \end{aligned} \tag{2.18}$$

In 2.18, W is defined from

$$W^{-1} = \left(\frac{d\eta}{d\mu}\right)^2 V. \tag{2.19}$$

So far, we have written the likelihood for one single observation. By summing over the observations, the likelihood equation for one parameter β_j is given by

$$\sum_i \frac{W_i\left(y_i - \mu_i\right)}{a\left(\phi\right)}\frac{d\eta_i}{d\mu_i}x_{ij} = 0. \tag{2.20}$$

We can solve (2.20) with respect to β_j since the μ_i:s are functions of the parameters β_j. Asymptotic variances and covariances of the parameter estimates are obtained through the inverse of the Fisher information matrix (see Appendix B). Thus,

$$\begin{pmatrix} Var\left(\widehat{\beta}_0\right) & Cov\left(\widehat{\beta}_0,\widehat{\beta}_1\right) & & Cov\left(\widehat{\beta}_0,\widehat{\beta}_{p-1}\right) \\ Cov\left(\widehat{\beta}_1,\widehat{\beta}_0\right) & Var\left(\widehat{\beta}_1\right) & & \\ & & \ddots & \vdots \\ Cov\left(\widehat{\beta}_{p-1},\widehat{\beta}_0\right) & & \cdots & Var\left(\widehat{\beta}_{p-1}\right) \end{pmatrix} = \\ = -E\begin{pmatrix} \frac{\partial^2 l}{\partial\beta_0^2} & \frac{\partial l}{\partial\beta_0}\frac{\partial l}{\partial\beta_1} & & \frac{\partial l}{\partial\beta_0}\frac{\partial l}{\partial\beta_{p-1}} \\ \frac{\partial l}{\partial\beta_1}\frac{\partial l}{\partial\beta_0} & \frac{\partial^2 l}{\partial\beta_1^2} & & \\ & & \ddots & \vdots \\ \frac{\partial l}{\partial\beta_{p-1}}\frac{\partial l}{\partial\beta_0} & & \cdots & \frac{\partial^2 l}{\partial\beta_{p-1}^2} \end{pmatrix}^{-1}. \tag{2.21}$$

2.7 Numerical procedures

Maximization of the log likelihood (2.16), which is equivalent to solving the likelihood equations (2.20), is done using numerical methods. A commonly used procedure is the iteratively reweighted least squares approach; see McCullagh and Nelder (1989). Briefly, this algorithm works as follows:

1. Linearize the link function $g(\cdot)$ by using the first order Taylor series approximation $g(y) \approx g(\mu) + (y-\mu)\, g'(\mu) = z$.
2. Let $\widehat{\eta}_0$ be the current estimate of the linear predictor, and let $\widehat{\mu}_0$ be the corresponding fitted value derived from the link function $\eta = g(\mu)$. Form the adjusted dependent variate $z_0 = \widehat{\eta}_0 + (y - \widehat{\mu}_0)\left(\frac{d\eta}{d\mu}\right)_0$ where the derivative of the link is evaluated at $\widehat{\mu}_0$.
3. Define the weight matrix W from $W_0^{-1} = \left(\frac{d\eta}{d\mu}\right)^2 V_0$, where V is the variance function.
4. Perform a weighted regression of dependent variable z on predictors x_1, x_2, ... , x_p using weights W_0. This gives new estimates $\widehat{\boldsymbol{\beta}}_1$ of the parameters, from which a new estimate $\widehat{\eta}_1$ of the linear predictor is calculated.
5. Repeat steps 1−4 until the changes are sufficiently small.

2.8 Assessing the fit of the model

2.8.1 The deviance

The fit of a generalized linear model to data may be assessed through the deviance. The deviance is also used to compare nested models.

Different models can have different degrees of complexity. The *null model* has only one parameter that represents a common mean value μ for all observations. In contrast, the *full* (or *saturated*) model has n parameters, one for each observation. For the saturated model, each observation fits the model perfectly, i.e. $y = \widehat{y}$. The full model is used as a benchmark for assessing the fit of any model to the data. This is done by calculating the *deviance*. The deviance is defined as follows:

Let $l(\widehat{\mu}, \phi; y)$ be the log likelihood of the current model at the Maximum Likelihood estimate, and let $l(y, \phi; y)$ be the log likelihood of the full model. The deviance D is defined as

$$D = 2\left(l(y, \phi; y) - l(\widehat{\mu}, \phi; y)\right). \tag{2.22}$$

It can be noted that for a Normal distribution, the deviance is just the residual sum of squares. The Genmod procedure in SAS presents two deviance statistics: the deviance and the scaled deviance. For distributions that have a scale parameter ϕ, the scaled deviance is $D^* = D/\phi$. It is actually the scaled deviance that is used for inference. For distributions such as Binomial and Poisson, the deviance and the scaled deviance are identical.

If the model is true, the deviance will asymptotically tend towards a χ^2 distribution as n increases. This can be used as an over-all test of the adequacy of the model. The degree of approximation to a χ^2 distribution is different for different types of data.

A second, and perhaps more important use of the deviance is in comparing competing models. Suppose that a certain model gives a deviance D_1 on df_1 degrees of freedom (df), and that a simpler model produces deviance D_2 on df_2 degrees of freedom. The simpler model would then have a larger deviance and more df. To compare the two models we can calculate the difference in deviance, $(D_2 - D_1)$, and relate this to the χ^2 distribution with $(df_2 - df_1)$ degrees of freedom. This would give a large-sample test of the significance of the parameters that are included in model 1 but not in model 2. This, of course, requires that the parameters included in model 2 is a subset of the parameters of model 1, i.e. that the models are nested.

2.8.2 The generalized Pearson χ^2 statistic

An alternative to the deviance for testing and comparing models is the Pearson χ^2, which can be defined as

$$\chi^2 = \sum_i (y_i - \widehat{\mu})^2 / \widehat{V}(\widehat{\mu}). \tag{2.23}$$

Here, $\widehat{V}(\widehat{\mu})$ is the estimated variance function. For the Normal distribution, this is again the residual sum of squares of the model, so in this case, the deviance and Pearson's χ^2 coincide. In other cases, the deviance and Pearson's χ^2 have different asymptotic properties and may produce different results. Maximum likelihood estimation of the parameters in generalized linear models seeks to minimize the deviance, which may be one reason to prefer the deviance over the Pearson χ^2. Another reason is that the Pearson χ^2 does not have the same additive properties as the deviance for comparing nested models. Computer packages for generalized linear models often produce both the deviance and the Pearson χ^2. Large differences between these may be an indication that the χ^2 approximation is bad.

2.8.3 Akaike's information criterion

An idea that has been put forward by several authors is to "penalize" the likelihood functions such that simpler models are being preferred. A general expression of this idea is to measure the fit of a model to data by a measure such as

$$D_C = D - \alpha q \phi. \tag{2.24}$$

Here, D is the deviance, q is the number of parameters in the model, and ϕ is the dispersion parameter. If ϕ is constant, it can be shown that $\alpha \approx 4$ is roughly equivalent to testing one parameter at the 5% level. It can be shown that $\alpha \approx 2$ would lead to prediction errors near the minimum. This is the information criterion (AIC) suggested by Akaike (1973). Akaike's information criterion is often used for model selection: the model with the smallest value of D_C would then be the preferred model. Note that some computer program report the AIC with the opposite sign; large values of AIC would then indicate a good model. The AIC is not very useful in itself since the scale is somewhat arbitrary. The main use is to compare the AIC of competing models in order to decide which model to prefer.

2.8.4 The choice of measure of fit

The deviance, and the Pearson χ^2, can provide large-sample tests of the fit of the model. The usefulness of these tests depends on the kind of data being analyzed. For example, Collett (1991) concludes that for binary data with all $n_i = 1$, the deviance cannot be used to assess the over-all fit of the model (p. 65). For Normal models the deviance is equal to the residual sum of squares which is not a model test by itself.

The advantage of the deviance, as compared to the Pearson χ^2, is that it is a likelihood-based test that is useful for comparing nested models. Akaike's information criterion is often used as a way of comparing several competing models, without necessarily making any formal inference.

2.9 Different types of tests

Hypotheses on single parameters or groups of parameters can be tested in different ways in generalized linear models.

2.9.1 Wald tests

Maximum likelihood estimation of the parameters of some model results in estimates of the parameters and estimates of the standard errors of the estimators. The estimates of standard errors are often asymptotic results that are valid for large samples. Let us denote the asymptotic standard error of the estimator $\widehat{\beta}$ with $\widehat{\sigma}_{\widehat{\beta}}$. If the large-sample conditions are valid, we can test hypotheses about single parameters by using

$$z = \frac{\widehat{\beta}}{\widehat{\sigma}_{\widehat{\beta}}} \tag{2.25}$$

where z is a standard Normal variate. This is called a Wald test. In normal theory models, tests based on (2.25), but with z replaced by t, are exact. In other cases the Wald tests are asymptotic tests that are valid only in large samples. In some cases the Wald tests are presented as χ^2 tests. This is based on the fact that if z is standard Normal, then z^2 follows a χ^2 distribution on 1 degree of freedom. Note that the Wald tests for single parameters are related to the Pearson χ^2.

2.9.2 Likelihood ratio tests

Likelihood ratio tests are based on the following principle. Denote with L_1 the likelihood function maximized over the full parameter space, and denote with L_0 the likelihood function maximized over parameter values that correspond to the null hypothesis being tested.

The likelihood ratio statistic is

$$-2\log\left(L_0/L_1\right) = -2\left[\log\left(L_0\right) - \log\left(L_1\right)\right] = -2\left(l_0 - l_1\right). \tag{2.26}$$

Under rather general conditions, it can be shown that the distribution of the likelihood ratio statistic approaches a χ^2 distribution as the sample size grows. Generally, the number of degrees of freedom of this statistic is equal to the number of parameters in model 1 minus the number of parameters in model 0. Exceptions occur if there are linear constraints on the parameters. In the same way as the Wald tests are related to the Pearson χ^2, the likelihood ratio tests are related to the deviance.

2.9.3 Score tests

We will illustrate score tests (also called efficient score tests) based on arguments taken from Agresti (1996). In Figure 2.2, we illustrate a hypothetical likelihood function. $\widehat{\beta}$ is the Maximum Likelihood estimator of some parameter β. We are testing a hypothesis H_0: $\beta = 0$. L_1 and L_0 denote the likelihood under H_1 and H_0, respectively.

The Wald test uses the behavior of the likelihood function at the ML estimate $\widehat{\beta}$. The asymptotic standard error of $\widehat{\beta}$ depends on the curvature of the likelihood function close to $\widehat{\beta}$.

The score test is based on the behavior of the likelihood function close to 0, the value stated in H_0. If the derivative at H_0 is "large", this would be an indication that H_0 is wrong, while a derivative close to 0 would be a sign that we are close to the maximum. The score test is calculated as the square of the ratio of this derivative to its asymptotic standard error. It can be treated as an asymptotic χ^2 variate on 1 *df*.

The likelihood ratio test uses information on the log likelihood both at $\widehat{\beta}$ and at 0. It compares the likelihoods L_1 and L_0 using the asymptotic χ^2 distribution of $-2\left(\log L_0 - \log L_1\right)$. Thus, in a sense, the LR statistic uses more information than the Wald and score statistics. For this reason, Agresti (1996) suggests that the likelihood ratio statistic may be the most reliable of the three.

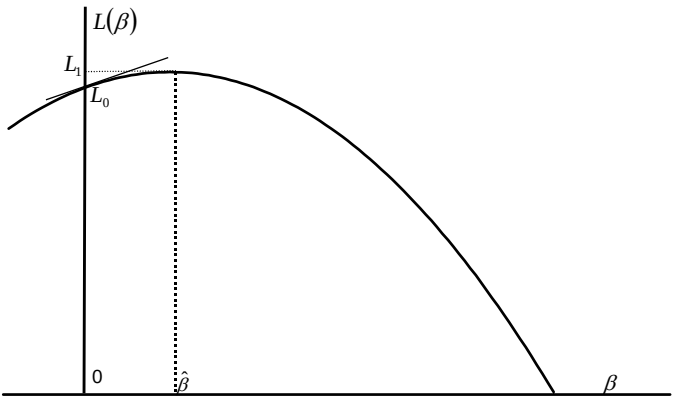


Figure 2.2: *A likelihood function indicating information used in Wald, LR and score tests.*

2.9.4 Tests of Type 1 or 3

Tests in generalized linear models have the same sequential property as tests in general linear models. Proc Genmod in SAS offers Type 1 or Type 3 tests. The interpretation of these tests is the same as in general linear models. In a Type 1 analysis the result of a test depends on the order in which terms are included in the model. A type 3 analysis does not depend on the order in which the Model statement is written: it can be seen as an attempt to mimic the analysis that would be obtained if the data had been balanced. In general linear models the Type 1 and Type 3 tests are obtained through sums of squares. In Generalized linear models the tests are Likelihood ratio tests, but there is an option in Genmod to use Wald tests instead. See Chapter 1 for a discussion on Type 1 and Type 3 tests.

2.10 Descriptive measures of fit

In general linear models, the fit of the model to data can be summarized as $R^2 = \frac{SS_{Model}}{SS_T}$. It holds that $0 \leq R^2 \leq 1$. A value of R^2 close to 1 would indicate a good fit. An adjusted version of R^2 has been proposed to account for the fact that R^2 increases even when irrelevant factors are added to the model; see Chapter 1.

Similar measures of fit have been proposed also for generalized linear models.

Cox and Snell (1989) suggested to use

$$R^2 = 1 - \left(\frac{L_0}{L_{Max}}\right)^{2/n} \tag{2.27}$$

where L is the likelihood. This measure equals the usual R^2 for Normal models, but has the disadvantage that it is always smaller than 1. In fact,

$$R^2_{\max} = 1 - (L_0)^{2/n}. \tag{2.28}$$

For example, in a binomial model with half of the observations in each category, this maximum equals 0.75 even if there is perfect agreement between the variables. Nagelkerke (1991) therefore suggested the modification

$$\overline{R}^2 = R^2/R^2_{\max}. \tag{2.29}$$

Similar coefficients have been suggested by Ben-Akiva and Lerman (1985), and by Horowitz (1982). The coefficients by Cox and Snell, and by Nagelkerke, are available in the Logistic procedure in SAS.

2.11 An application

Example 2.7 Samuels and Witmer (1999) report on a study of methods for producing sheep's milk for use in the manufacture of cheese. Ewes were randomly assigned to either mechanical or manual milking. It was suspected that the mechanical method might irritate the udder and thus producing a higher concentration of somatic cells in the milk. The data in the following table show the counts of somatic cells for each animal.

	Mechanical	Manual
	2966	186
	269	107
	59	65
	1887	126
	3452	123
	189	164
	93	408
	618	324
	130	548
	2493	139
$\overline{y}$	1216	219
s	1343	156

This is close to a textbook example that may be used for illustrating two-sample t tests. However, closer scrutiny of the data reveals that the variation is quite different in the two groups. In fact, some kind of relationship between the mean and the variance may be at hand. □

We will illustrate the analysis of these data by attempting several different analyses. An analysis similar to a standard two-sample t test can be obtained by using a generalized linear model with a dummy variable for group membership, a Normal distribution and an identity link.

Cell counts often conform to Poisson distributions. This means that a Poisson distribution with the canonical (log) link is another option.

The SAS programs for analysis of these data were of the following type:

```
PROC GENMOD DATA=sheep;
CLASS milking;
MODEL count = milking /
              dist = normal;
RUN;
```

A model assuming a Normal distribution and with milking method as a factor was fitted. By default, the program then chooses the canonical link which, for the Normal distribution, is the identity link. In this model, the Wald test of group differences is significant ($p = 0.014$). If we use a standard t test assuming equal variances, this gives $p = 0.044$. The difference between these p-values is explained by the fact that the Wald test essentially approximates the t distribution with a Normal distribution.

The Poisson model gives a deviance of 14863 on 18 *df*. The group difference is highly significant: $\chi^2 = 5451.12$ on 1 *df*, $p < 0.0001$. The output for this model is as follows:

```
                    Model Information

     Description                    Value

     Data Set                       WORK.SHEEP
     Distribution                   POISSON
     Link Function                  LOG
     Dependent Variable             COUNT
     Observations Used              20
       Criteria For Assessing Goodness Of Fit

Criterion             DF         Value      Value/DF

Deviance              18    14862.7182      825.7066
Scaled Deviance       18    14862.7182      825.7066
Pearson Chi-Square    18    14355.5077      797.5282
Scaled Pearson X2     18    14355.5077      797.5282
Log Likelihood         .    83800.0507             .
```

```
                    Analysis Of Parameter Estimates

Parameter           DF    Estimate     Std Err   ChiSquare  Pr>Chi

INTERCEPT            1      7.1030      0.0091  613300.717  0.0001
MILKING    Man       1     -1.7139      0.0232   5451.1201  0.0001
MILKING    Mech      0      0.0000      0.0000           .       .
SCALE                0      1.0000      0.0000           .       .

NOTE:  The scale parameter was held fixed.
```

However, since the ratio deviance/df is so large, a second analysis was made where the program estimated the dispersion parameter ϕ. This approach, which is related to a phenomenon called overdispersion, is discussed in the next chapter. In the analysis where the scale parameter was used, the scaled deviance was 18.0 and the p value for milking was 0.010.

Two other distributions were also tested: the gamma distribution and the inverse gaussian distribution. These were used with their respective canonical links. In addition, a Wilcoxon-Mann-Whitney test was made. The results for all these models can be summarized as follows:

Model	p value
Normal, Glim	0.0140
Normal, t test	0.0440
Log normal	0.0405
Gamma	0.0086
Inverse Gaussian	0.0610
Poisson	<0.0001
Poisson with ϕ	0.0102
Wilcoxon	0.1620

Although most models seem to indicate significant group differences, the p values are rather different. The first Poisson model gives a strongly significant result while the standard t test is only just below the magical 0.05 limit. The non-parametric Wilcoxon test is not significant. This illustrates the fact that significance testing is not a mechanical procedure. To decide which of the results to use we need to assess the different models, based both on statistical consideration and on subject-matter knowledge. Methods for statistical model diagnostics in generalized linear models is the topic of the next chapter.

2.12 Exercises

Exercise 2.1 The distribution of waiting times (for example the time you wait in line at a bank) can sometimes be approximated by an exponential distribution. The density of the exponential distribution is $f(x) = \lambda e^{-\lambda x}$ for $x > 0$. Does the exponential distribution belong to the exponential family? If so, what are the functions $b(\cdot)$ and $c(\cdot)$? What is the variance function?

Exercise 2.2 Sometimes data are collected which are "essentially Poisson", but where it is impossible to observe the value $y = 0$. For example, if data are collected by interviewing occupants in houses on "how many occupants are there in this house", it would be impossible to get an answer from houses that are not occupied.

The truncated Poisson distribution can sometimes be used to model such data. It has probability function

$$p(y_i|\lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{(1 - e^{-\lambda})\, y_i!}$$

for $y_i = 1, 2, 3, ...$

A. Investigate whether the truncated Poisson distribution is a member of the Exponential family.

B. Derive the variance function.

Exercise 2.3 Aanes (1961) studied the effect of a certain type of poisoning in sheep. The survival time and the weight were recorded for 13 sheep that had been poisoned. Results:

Weight	Survival	Weight	Survival
46	44	59	44
55	27	64	120
61	24	67	29
75	24	60	36
64	36	63	36
75	36	66	36
71	44		

Find a generalized linear model for the relationship between survival time and weight. Try several different distributions and link functions. Do not use only the canonical link for each distribution. Plot the data and the fitted models.

3. Model diagnostics

3.1 Introduction

In general linear models, the fit of the model to data can be explored by using residual plots and other diagnostic tools. For example, the normality assumption can be examined using normal probability plots, the assumption of homoscedasticity can be checked by plotting residuals against $\widehat{y}$, and so on. Model diagnostics in Glim:s can be performed in similar ways.

In this chapter we will discuss different model diagnostic tools for generalized linear models. Our discussion will be fairly general. We will return to these issues in later chapters when we consider analysis of different types of response variables.

The purpose of model diagnostics is to examine whether the model provides a reasonable approximation to the data. If there are indications of systematic deviations between data and model, the model should be modified.

The diagnostic tools that we consider are the following:

- Residual plots (similar to the residual plots used in GLM:s) can be used to detect various deviations between data and model: outliers, problems with distributions or variances, dependence, and so on.
- Some observations may have an unusually large impact on the results. We will discuss tools to identify such influential observations.
- Overdispersion means that the variance is larger than would be expected for the chosen distribution. We will discuss ways to detect over-dispersion and to modify the model to account for over-dispersion.

3.2 The Hat matrix

In general linear models, a residual is the difference between the observed value of y and the fitted value $\widehat{y}$ that would be obtained if the model were

perfectly true: $e = y - \widehat{y}$. The concept of a "residual" is not quite as clear-cut in generalized linear models.

The estimated expected value ("fitted value") of the response in a general linear model is $\widehat{E(Y_i)} = \widehat{\mu}_i = \widehat{y}_i$. The fitted values are linear functions of the observed values. For linear predictors, it holds that

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y} \tag{3.1}$$

where $\mathbf{H}$ is known as the "hat matrix". $\mathbf{H}$ is idempotent (i.e. $\mathbf{HH} = \mathbf{H}$) and symmetric.

Example: In simple linear regression, $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$. In this case, the hat matrix is $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$.

The hat matrix may have a more complex form in other models than GLM:s. Except in degenerate cases, it is possible to compute the hat matrix. We will not give general formulae here; however, most computer software for Glim:s have options to print the hat matrix.

3.3 Residuals in generalized linear models

In general linear models, the observed residuals are simply the difference between the observed values of y and the values $\widehat{y}$ that are predicted from the model: $\widehat{e} = y - \widehat{y}$. In generalized linear models, the variance of the residuals is often related to the size of $\widehat{y}$. Therefore, some kind of scaling mechanism is needed if we want to use the residuals for plots or other model diagnostics. Several suggestions have been made on how to achieve this.

3.3.1 Pearson residuals

The raw residual for an observation y_i can be defined as $\widehat{e}_i = y_i - \widehat{y}_i$. The Pearson residual is the raw residual standardized with the standard deviation of the fitted value:

$$e_{i,Pearson} = \frac{y_i - \widehat{y}_i}{\sqrt{\widehat{Var}\left(\widehat{y}_i\right)}}. \tag{3.2}$$

The Pearson residuals are related to the Pearson χ^2 through $\chi^2 = \sum_i e^2_{i,Pearson}$.

Example: In a Poisson model, $e_{i,Pearson} = \frac{(y_i - \widehat{y}_i)}{\sqrt{\widehat{y}_i}}$.

Example: In a binomial model, $e_{i,Pearson} = \frac{(y_i - n_i\widehat{p}_i)}{\sqrt{n_i\widehat{p}_i(1-\widehat{p}_i)}}$.

If the model holds, Pearson residuals can often be considered to be approximately normally distributed with a constant variance, in large samples. However, even when they are standardized with the standard error of $\widehat{y}$, the variance of Pearson residuals cannot be assumed to be 1. This is since we have standardized the residuals using estimated standard errors. Still, the standard errors of Pearson residuals can be estimated. It can be shown that adjusted Pearson residuals can be obtained as

$$e_{i,adj,P} = \frac{e_{i,Pearson}}{\sqrt{1-h_{ii}}} \tag{3.3}$$

where h_{ii} are diagonal elements from the hat matrix. The adjusted Pearson residuals can often be considered to be standard Normal, which means that e.g. residuals outside ± 2 will occur in about 5% of the cases. This can be used to detect possible outliers in the data.

3.3.2 Deviance residuals

Observation number i contributes an amount d_i to the deviance, as a measure of fit of the model: $D = \sum_i d_i$. We define the deviance residuals as

$$e_{i,Deviance} = sign\,(y_i - \widehat{y}_i)\,\sqrt{d_i}. \tag{3.4}$$

The deviance residuals can also be written in standardized form, i.e. such that their variance is close to unity. This is obtained as

$$e_{i,adj,D} = \frac{e_{i,Deviance}}{\sqrt{1-h_{ii}}} \tag{3.5}$$

where h_{ii} are again diagonal elements from the hat matrix.

3.3.3 Score residuals

The Wald tests, Likelihood ratio tests and Score tests, presented in the previous chapter, provide different ways of testing hypotheses about parameters of the model. The score residuals are related to the score tests.

In Maximum Likelihood estimation, the parameter estimates are obtained by solving the score equations, which are of type

$$U = \frac{\partial l}{\partial \theta} = 0 \tag{3.6}$$

where θ is some parameter. The score equations involve sums of terms U_i, one for each observation. These terms can, properly standardized, be interpreted

as residuals, i.e. as the contribution from each observation to the score. The standardized score residuals are obtained from

$$e_{i,adj,S} = \frac{U_i}{\sqrt{(1-h_{ii})\, v_i}} \tag{3.7}$$

where h_{ii} are diagonal elements of the hat matrix, and v_i are elements of a certain weight matrix.

3.3.4 Likelihood residuals

Theoretically it would be possible compare the deviance of a model that comprises all the data with the deviance of a model with observation i excluded. However, this procedure would require heavy computations. An approximation to the residuals that would be obtained using this procedure is

$$e_{i,Likelihood} = sign\left(y_i - \widehat{y}_i\right) \sqrt{h_{ii}\left(e_{i,Score}\right)^2 + \left(1-h_{ii}\right)\left(e_{i,Deviance}\right)^2} \tag{3.8}$$

where h_{ii} are diagonal elements of the hat matrix. This is a kind of weighted average of the deviance and score residuals.

3.3.5 Anscombe residuals

The types of residuals discussed so far have distributions that may not always be close to Normal, in samples of the sizes we often meet in practice. Anscombe (1953) suggested that the residuals may be defined based on some transformation of observed data and fitted values. The transformation would be chosen such that the calculated residuals are approximately standard Normal. Anscombe defined the residuals as

$$e_{i,Anscombe} = \frac{A\left(y_i\right) - A\left(\widehat{y}_i\right)}{\sqrt{\widehat{Var}\left(A\left(y_i\right) - A\left(\widehat{y}_i\right)\right)}} \tag{3.9}$$

The function $A(\cdot)$ is chosen depending on the type of data. For example, for Poisson data the Anscombe residuals take the form $e_{i,Anscombe} = \frac{3}{2}\left(y^{2/3} - \widehat{y}^{2/3}\right)/\widehat{y}^{1/6}$. In general, the Anscombe residuals are rather difficult to calculate, which may explain why they have not reached widespread use.

3.3.6 The choice of residuals

The types of residuals discussed above are related to the types of tests and other model building tools that are used:

The deviance residuals are related to the deviance as a measure of fit of the model and to Likelihood ratio tests.

The Pearson residuals are related to the Pearson χ^2 and to the Wald tests.

The score residuals are related to score tests.

The likelihood residuals are a compromise between score and deviance residuals

The Anscombe residuals, although theoretically appealing, are not often used in practice in programs for fitting generalized linear models.

In the previous chapter we suggested that the likelihood ratio tests may be preferred over Wald tests and score tests for hypothesis testing in Glim:s. By extending this argument, the deviance residuals may be the preferred type of residuals to use for model diagnostics. Collett (1991) suggested that either the deviance residuals or the likelihood residuals should be used.

3.4 Influential observations and outliers

Some of the observations in the data may have an unduly large impact on the parameter estimates. If such so called influential observations are changed by a small amount, or if they are deleted, the estimates may change drastically. An outlier is an observation for which the model does not give a good approximation. Outliers can often be detected using different types of plots. Note that influential observations are not necessarily outliers. An observation can be influential and still be close to the main bulk of the data. Diagnostic tools are needed to detect influential observations and outliers.

3.4.1 Leverage

The leverage of observation i on the fitted value $\widehat{\mu}_i$ is the derivative of $\widehat{\mu}_i$ with respect to y_i. This derivative is the corresponding diagonal element h_{ii} of the hat matrix $\mathbf{H}$. Since $\mathbf{H}$ is idempotent it holds that $tr\,(\mathbf{H}) = p$, i.e. the number of parameters. The average leverage of all observations is therefore p/n. Observations with a leverage of, say, twice this amount may need to be examined. Computer software like the Insight procedure in SAS (2000b), and the related JMP program (SAS, 2000a) have options to store the hat matrix in a file for further processing and plotting.

3.4.2 Cook's distance and Dfbeta

Dfbeta is the change in the estimate of a parameter when observation i is deleted. The Dfbetas can be combined over all parameters as

$$D_i = \frac{1}{p}\left(\widehat{\beta} - \widehat{\boldsymbol{\beta}}_{(i)}\right)' \mathbf{X}'\mathbf{X}\left(\widehat{\beta} - \widehat{\boldsymbol{\beta}}_{(i)}\right).$$

It can be shown that this yields the so called Cook's distance C_i. In principle, the calculation of C_i (or D_i) requires extensive re-fitting of the model which may take time even on fast computers. However, an approximation to C_i can be obtained as

$$C_i \approx \frac{h_{ii}\left(e_{i,Pearson}\right)^2}{p\left(1 - h_{ii}\right)} \tag{3.10}$$

where p is the number of parameters and h_{ii} are elements of the hat matrix.

3.4.3 Goodness of fit measures

Another type of measure of the influence of an observation is to compute the change in deviance, or the change in Pearson's χ^2, when the observation is deleted. A large change in the measure of fit may indicate an influential observation.

3.4.4 Effect on data analysis

Computer programs for generalized linear models often include options to calculate the measures of influence discussed above, and others. It belongs to good data analytic practice to use such program options to investigate influential observations and outliers. A statistical result that may be attributed to very few observations should, of course, be doubted. Thus, data analysis in generalized linear models should contain both an analysis of the residuals, discussed above, and an analysis of influential observations.

3.5 Partial leverage

In models with several explanatory variables it may be of interest to study the impact of a variable, say variable x_j, on the results. The partial leverage of variable j can be obtained in the following way. Let $\mathbf{X}_{[j]}$ be the design matrix with the column corresponding to variable x_j removed. Fit the generalized

linear model to this design matrix and calculate the residuals. Also, fit a model with variable x_j as the response and the remaining variables $\mathbf{X}_{[j]}$ as regressors. Calculate the residuals from this model as well. A partial leverage plot is a plot of these two sets of residuals. It shows how much the residuals change between models with and without variable x_j. Partial leverage plots can be produced in procedure Insight (SAS, 2000b).

3.6 Overdispersion

A generalized linear model can sometimes give a good summary the data, in the sense that both the linear predictor and the distribution are correctly chosen, and still the fit of the full model may be poor. One possible reason for this may be a phenomenon called over-dispersion.

Over-dispersion occurs when the variance of the response is larger than would be expected for the chosen distribution. For example, if we use a Poisson distribution to model the data we would expect the variance to be equal to the mean value: $\mu = \sigma^2$. Similarly, for data that are modelled using a binomial distribution, the variance is a function of the response probability: $\sigma^2 = np\,(1-p)$. Thus, for many distributions it is possible to infer what the variance "should be", given the mean value. In Chapter 2 we noted that for distributions in the exponential family, the variance is some function of the mean: $\sigma^2 = V\,(\mu)$.

Under-dispersion, i.e. a "too small" variance, is theoretically possible but rather unusual in practice. Interesting examples of under-dispersion can be found in the analysis of Mendel's classical genetic data; these data are better than would be expected by chance.

In models that do not contain any scale parameter, over-dispersion can be detected as a poor model fit, as measured by deviance/df. Note, however, that a poor model fit can also be caused by the wrong choice of linear predictor or wrong choice of distribution or link. Thus, a poor fit does not necessarily mean that we have over-dispersion.

Over-dispersion may have many different reasons. However, the main reason is often some type of lack of homogeneity. This lack of homogeneity may occur between groups of individuals; between individuals; and within individuals.

As an example, consider a dose-response experiment where the same dose of an insecticide is given to two batches of insects. In one of the batches, 50 out of 100 insects die, while in the other batch 65 out of 100 insects die. Formally, this means that the response probabilities in the two batches are significantly different (the reader may wish to confirm that a "textbook" Chi-square test gives $\chi^2 = 4.6$, $p = 0.032$). This may indicate that the batches of insects are

not homogenous with respect to tolerance to the insecticide. If these data are part of some larger dose-response experiment, using more batches of animals and more doses, this type of inhomogeneity would result in a poor model fit because of overdispersion.

3.6.1 Models for overdispersion

Before any attempts are made to model the over-dispersion, you have to examine all other possible reasons for poor model fit. These include:

- Wrong choice of linear predictor. For example, you may have to add terms to the predictor, such as new covariates, interaction terms or nonlinear terms.
- Wrong choice of link function.
- There may be outliers in the data.
- When the data are sparse, the assumptions underlying the large-sample theory may not be fulfilled, thus causing a poor model fit.

A common effect of over-dispersion is that estimates of standard errors are under-estimates. This leads to test statistics which are too large: it becomes too easy to get a significant result. A simple way to model over-dispersion is to introduce a scale parameter ϕ into the variance function. Thus, we would assume that $Var(Y) = \phi\sigma^2$. For binomial data this means that we would use the variance $np(1-p)\phi$, and for Poisson data we would use $\phi\mu$ as variance. The parameter ϕ is often called the over-dispersion parameter. A simple, but somewhat rough, way to estimate ϕ is to fit a "maximal model"[1] to the data, and to use the mean deviance (i.e. Deviance/df), or Pearson χ^2/df, from that model as an estimator of ϕ. We can then re-fit the model, using the obtained value of the over-dispersion parameter. Williams (1982) suggested a more sophisticated iterative procedure for estimating ϕ; see Collett (1991) for details.

A more satisfactory approach would be to model the over-dispersion based on some specific model. One possible model is to assume that the mean parameter has a separate value for each individual. Thus, the mean parameter would be assumed to follow some random distribution over individuals while the response follows a second distribution, given the mean value. This would

[1] Note that this "maximal model" is not the same as the saturated model, which has $\phi = 0$. Instead, the "maximal model" is a somewhat subjectively chosen "large" model which includes all effects that can reasonably be included.

lead to compound distributions. A few examples of compound distributions are discussed in Chapters 5 and 6. See also Lindsey (1997) for details.

We will return to the topic of over-dispersion as we discuss fitting of generalized linear models to different types of data. Another approach to over-dispersion, based on so-called Quasi-likelihood estimation, is discussed in Chapter 8.

3.7 Non-convergence

When using packages like Genmod for fitting generalized linear models, it may happen that the program reports that the procedure has not converged. Sometimes the convergence is slow and the procedure reports estimates of standard errors that are very large. Typical error messages might be

```
WARNING: The negative of the Hessian is not positive definite. The
convergence is questionable.
WARNING: The procedure is continuing but the validity of the model fit is
questionable.
WARNING: The specified model did not converge.
```

Note that in SAS, the error messages are given in the program log. You can get some output even if these warnings have been given.

Non-convergence occurs because of the structure of the data in relation to the model that is being fitted. A common problem is that the number of observed data values is small relative to the number of parameters in the model. The model is then under-identified. This can easily happen in the analysis of multidimensional crosstables. For example, a crosstable of dimension $4 \cdot 3 \cdot 3 \cdot 3$ contains 108 cells. If the sample size is moderate, say $n = 100$, the average number of observations per cell will be less than 1. It is then easy to imagine that many of the cells will be empty. Convergence problems are likely in such cases.

When the data are binomial, the procedure may fail to converge when it tries to fit estimated proportions close to 0 or 1. This may happen when many observed proportions are 0 or 1.

As a general advice: when the procedure does not converge, try to simplify the model as much as possible by removing, in particular, interaction terms. Make tables and other summaries of the data to find out the reasons for the failure to converge.

3.8 Applications

3.8.1 Residual plots

In this section we will discuss a number of useful ways to check models, using the statistics we have discussed in this chapter. As illustrations of the various plots we use the example on somatic cells in the milk of sheep, discussed in the previous chapter (page 50). For the illustrations we use a model with a Normal distribution and a unit link, and a model with a Poisson distribution and a log link. The following types of residual plots are often useful:

1. A plot of residuals against the fitted values $\widehat{\eta}$ should show a pattern where the residuals have a constant mean value of 0 and a constant range. Deviations from this "random" pattern may arise because of incorrect link function; wrong choice of scale of the covariates; or omission of non-linear terms in the linear predictor.

2. A plot of residuals against covariates should show the same pattern as the previous plot. Deviations from this pattern may indicate the wrong link function, incorrect choice of scale or omission of non-linear terms.

3. Plotting the residuals in the order the observations are given in the data may help to detect possible dependence between observations.

4. A normal probability plot of the residuals plots the sorted residuals against their expected values. These are given by

$$\Phi^{-1}\left[(i - 3/8) \,/\, (n + 1/4)\right]$$

   where Φ^{-1} is the inverse of the standard Normal distribution function, i is the order of the observation, and n is the sample size. This plot should yield a straight line, as long as we can assume that the residuals are approximately Normal.

5. The residuals can also be plotted to detect an omitted covariate u. This is done as follows: fit a model with u as response, using the same model as for y. Obtain unstandardized residuals from both these models, and plot these against each other. Any systematic pattern in this plot may indicate that u should be used as a covariate.

Plots of residuals against predicted values for the data in the example on Page 50 are given in Figure 3.1 and Figure 3.2 for Normal and Poisson distributions, respectively. The plots of residuals against predicted values indicate that the variation is larger for larger predicted values. This tendency is strongest for the Normal model.

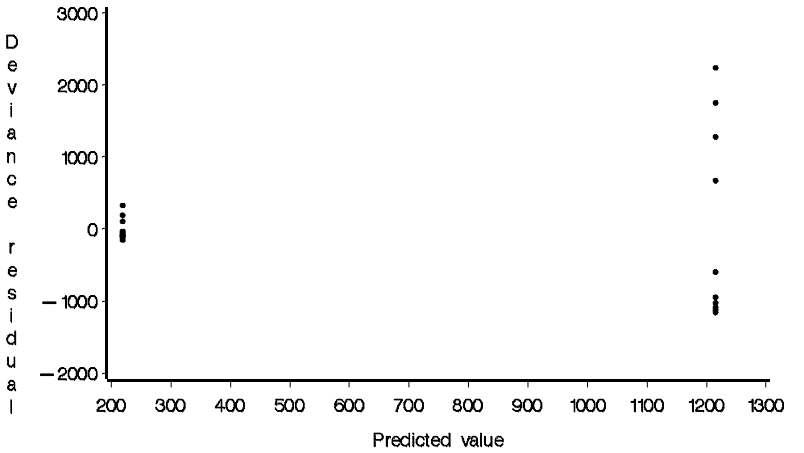


Figure 3.1: *Plot of residuals against predicted values for example data. Normal distribution and identity link.*

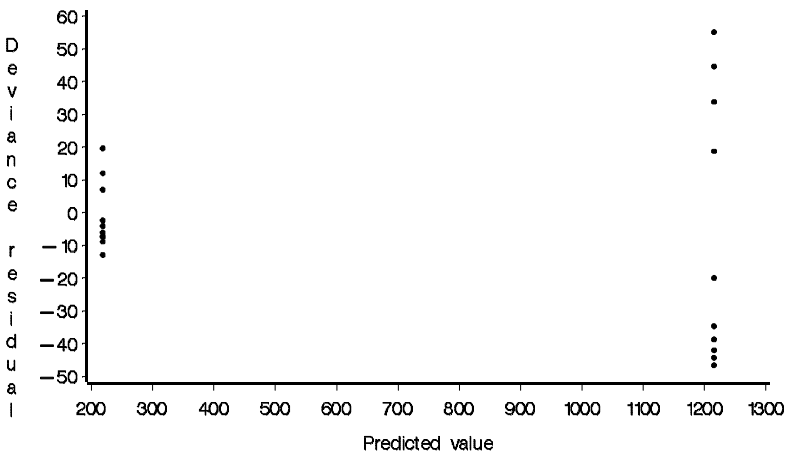


Figure 3.2: *Plot of residuals against predicted values for example data. Poisson distribution and log link.*

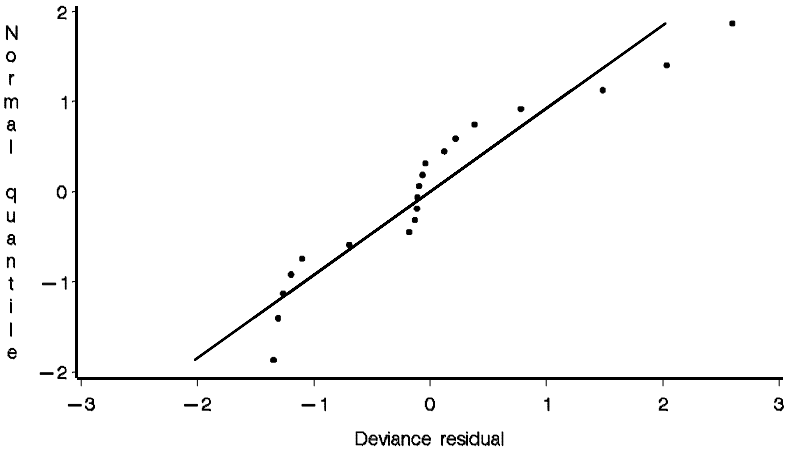


Figure 3.3: *Normal probability plot for the example data. Normal distribution with an identity link.*

A Normal probability plot is a plot of the residuals against their normal quantiles. Normal probability plots can be produced i.a. by Proc Univariate in SAS. SAS code for the normal probability plots presented here was as follows. The deviance residuals were stored in the file `ut` under the name `resdev`.

```
PROC UNIVARIATE normal data=ut;
  VAR resdev;
  PROBPLOT resdev /
      NORMAL (MU=est SIGMA=est color=black w=2 ) height=4;
  LABEL resdev="Deviance residual";
RUN;
```

Normal probability plots for these data are given in Figures 3.3 and 3.4, for the Normal and Poisson models, respectively. The distribution of the residuals is closer to Normal for the Poisson model, but the fit is not perfect.

3.8.2 Variance function diagnostics

McCullagh and Nelder (1989) suggest the following procedure for checking the variance function. Assume that the variance is proportional to μ^{ζ}, where ζ is some constant. Fit the model for different values of ζ, and plot the deviance against ζ. The value of ζ for which the deviance is as small as possible is suggested by the data.

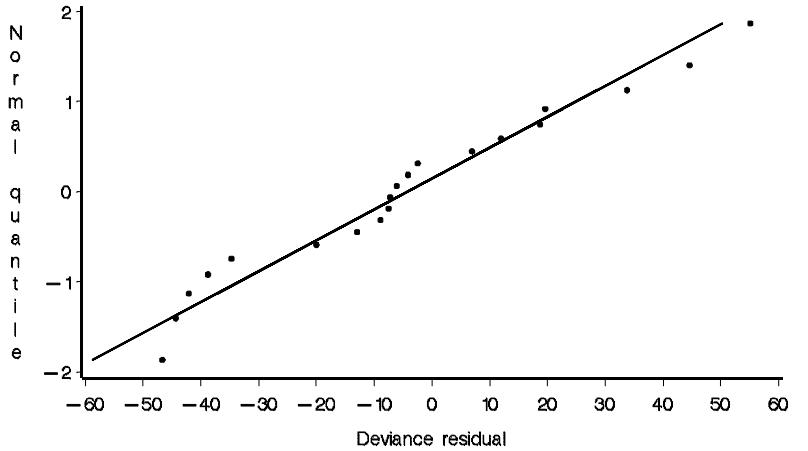


Figure 3.4: *Normal probability plot for the example data. Poisson distribution with a log link.*

3.8.3 Link function diagnostics

To check the link function we need the so called adjusted dependent variable z. This is defined as $z_i = g(\widehat{\mu}_i)$. This can be plotted against $\widehat{\eta}$. If the link is correct this should result in an essentially linear plot.

3.8.4 Transformation of covariates

So called partial residual plots can be used to detect whether any of the covariates need to be transformed. The partial residual is defined as $u = z - \widehat{\eta} + \widehat{\gamma}x$, where z is the adjusted dependent variable, $\widehat{\eta}$ is the fitted linear predictor, x is a covariate and $\widehat{\gamma}$ is the parameter estimate for the covariate. The partial residuals can be plotted against x. The plot should be approximately linear if no transformation is needed. Curvature in the plot is an indication that x may need to be transformed.

3.9 Exercises

Exercise 3.1 For the data in Exercise 1.1:

A. Calculate predicted values and residuals

B. Plot the residuals against the predicted values

C. Prepare a Normal probability plot

D. Calculate the leverage of all observations

Comment on the results

Exercise 3.2 For the data in Exercise 1.3:

A. Calculate predicted values and residuals

B. Plot the residuals against the predicted values

C. Prepare a Normal probability plot

D. Calculate the leverage of all observations

Comment on the results

Exercise 3.3 Use one or two of your "best" models for the data in Exercise 2.3 to:

A. Calculate predicted values and residuals

B. Plot the residuals against the predicted values

C. Prepare a Normal probability plot

D. Calculate the leverage of all observations

Comment on the results

4. Models for continuous data

4.1 GLM:s as GLIM:s

General linear models, such as regression models, ANOVA, t tests etc. can be stated as generalized linear models by using a Normal distribution and an identity link. We will illustrate this on some of the GLM examples we discussed in Chapter 1.

Throughout this chapter, we will use the SAS (2000b) procedure Genmod for data analysis.

4.1.1 Simple linear regression

A simple linear regression model can be written in Genmod as

```
PROC GENMOD;
  MODEL y = x /
  DIST=Normal
  LINK=Identity ;
RUN;
```

The identity link is the default link for the Normal distribution. We used this program on the regression data given on page 9. The results are:

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.EMISSION
Distribution	NORMAL
Link Function	IDENTITY
Dependent Variable	EMISSION
Observations Used	8

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	6	25.3765	4.2294
Scaled Deviance	6	8.0000	1.3333
Pearson Chi-Square	6	25.3765	4.2294
Scaled Pearson X2	6	8.0000	1.3333
Log Likelihood	.	-15.9690	.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-36.7443	3.8179	92.6233	0.0001
TIME	1	2.0978	0.1186	312.8360	0.0001
SCALE	1	1.7810	0.4453	.	.

NOTE: The scale parameter was estimated by maximum likelihood.

The regression model is estimated as $\widehat{y} = -36.7443 + 2.0978 \cdot Time$. This is the same estimate as given by a standard regression routine. Note that the deviance reported by the Genmod procedure is equal to the error sum of squares in the output on page 10. Also, the scaled deviance is 8, which is equal to the sample size. The tests, however, are not the same as in a standard regression analysis: the Genmod tests are Wald tests while the tests in the regression output are t tests. These tests are equivalent only in large samples. The Wald test of the hypothesis that the parameter β_j is zero is given by

$$z = \frac{\widehat{\beta}_j - 0}{\sqrt{\widehat{Var}\left(\widehat{\beta}_j\right)}},$$

where z is a standard Normal variate. The t test in the regression output was obtained as

$$t = \frac{\widehat{\beta}_j - 0}{\sqrt{\widehat{Var}\left(\widehat{\beta}_j\right)}}$$

with $n - p$ degrees of freedom.

SCALE gives the estimated scale parameter as 1.7810. This is the ML estimate of σ. Note that the ML estimator of σ^2 is biased. An unbiased estimate of σ^2 is given by $\widehat{\sigma}^2 = \text{Deviance}/df = 4.2294$ giving $\widehat{\sigma} = \sqrt{4.2294} = 2.0566$. The relation between these two estimates is that the ML estimate does not account for the degrees of freedom: $\widehat{\sigma}^2_{ML} = \frac{n-p}{n}\,\widehat{\sigma}^2$. For these data we get $\widehat{\sigma}^2_{ML} = \frac{6}{8} \cdot 4.2294 = 3.1721$ so $\widehat{\sigma}_{ML} = \sqrt{3.1721} = 1.781$.

4.1.2 Simple ANOVA

A Genmod program for an ANOVA model (of which the simple t test is a special case) can be written as

```
PROC GENMOD DATA=lissdata;
  CLASS medium;
  MODEL diff = medium /
     DIST=normal
     LINK=identity ;
RUN;
```

The output from this program, using the data on page 16, contains the following information:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	50	905.1155	18.1023
Scaled Deviance	50	57.0000	1.1400
Pearson Chi-Square	50	905.1155	18.1023
Scaled Pearson X2	50	57.0000	1.1400
Log Likelihood	.	-159.6823	.

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	9.9079	1.4089	49.4563	0.0001
MEDIUM	Diatrizoate	1	11.4841	2.2717	25.5554	0.0001
MEDIUM	Hexabrix	1	-5.9473	1.9363	9.4340	0.0021
MEDIUM	Isovist	1	-8.2437	1.9363	18.1257	0.0001
MEDIUM	Mannitol	1	-2.2192	1.9363	1.3136	0.2518
MEDIUM	Omnipaque	1	-1.5182	1.8902	0.6451	0.4219
MEDIUM	Ringer	1	-9.6979	2.0624	22.1116	0.0001
MEDIUM	Ultravist	0	0.0000	0.0000	.	.
SCALE		1	3.9849	0.3732	.	.

NOTE: The scale parameter was estimated by maximum likelihood.

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
MEDIUM	6	62.1517	0.0001

The scale parameter, which is the ML estimator of σ, is estimated as 3.98, while an unbiased estimate of σ^2 is given by Deviance/df=18.1023. The scaled deviance is 57. As noted above, the scaled deviance is equal to n and the deviance is equal to the residual sum of squares in Normal theory models. The effect of Medium is significant (p<0.0001). The parameter estimates for the different media are given, which makes it possible to make e.g. pairwise comparisons between the media. Note that the parameter estimates are the same as for the ANOVA output given on page 17. However, the ANOVA gives the tests as an over-all F test and as t tests for single parameters, while the Genmod analysis gives the type 3 test as a χ^2 approximations to the likelihood ratio test, while the tests of single parameters are Wald tests.

The examples given in this section show that many analyses that can be run as general linear models, using e.g. Proc GLM in SAS, can alternatively be run using Proc Genmod. In fact, the JMP program (SAS Institute, 2000a), as well as the related procedure Insight in SAS, take a generalized linear model approach to all model fitting. This also holds for the pioneering Glim software (Francis et al, 1993).

4.2 The choice of distribution

One advantage of the generalized linear model approach is that it is not necessary to limit the models to Normal distributions. In many cases there are theoretical justifications for assuming other distributions than the Normal. Experience with the type of data at hand can often suggest a suitable distribution. Figure 4.1 (based on Leemis, 1986) summarizes the relation-

ships between some common distributions. Note, however, that not all these distributions are members of the exponential family.

4.3 The Gamma distribution

Among all distributions in the exponential family, a particularly useful class of distributions is the gamma distribution.

If α is positive, then the integral

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt \tag{4.1}$$

is called a gamma function. For the gamma function, it holds that

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha) \text{ for } \alpha > 0 \tag{4.2}$$

and that

$$\Gamma(n) = (n-1)! \tag{4.3}$$

where $(n-1)! = (n-1)\cdot(n-2)\cdot\ldots\cdot 2\cdot 1$. The gamma distribution is defined, using the gamma function, as

$$f(y;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\,\beta^\alpha} y^{\alpha-1} e^{-y/\beta}. \tag{4.4}$$

The gamma distribution has two parameters, α and β. The parameter α describes the shape of the distribution, mainly the peakedness and is often called the shape parameter. The parameter β mostly influences the spread of the distribution and is called the scale parameter. For the gamma distribution it holds that $E(y) = \alpha\beta$ and $Var(y) = \alpha\beta^2$.

Note that the gamma distribution is a member of the exponential family. It has a reciprocal canonical link; in fact, $g(y) = -\frac{1}{y}$. The variance function of the gamma distribution is $V(\mu) = \mu^2$. These relations also hold for the special cases of the gamma distribution that are described below. A few examples of gamma distributions are illustrated in Figure 4.2.

4.3.1 The Chi-square distribution

The χ^2 distribution is a special case of the gamma distribution. A χ^2 distribution with p degrees of freedom can be obtained as a gamma distribution

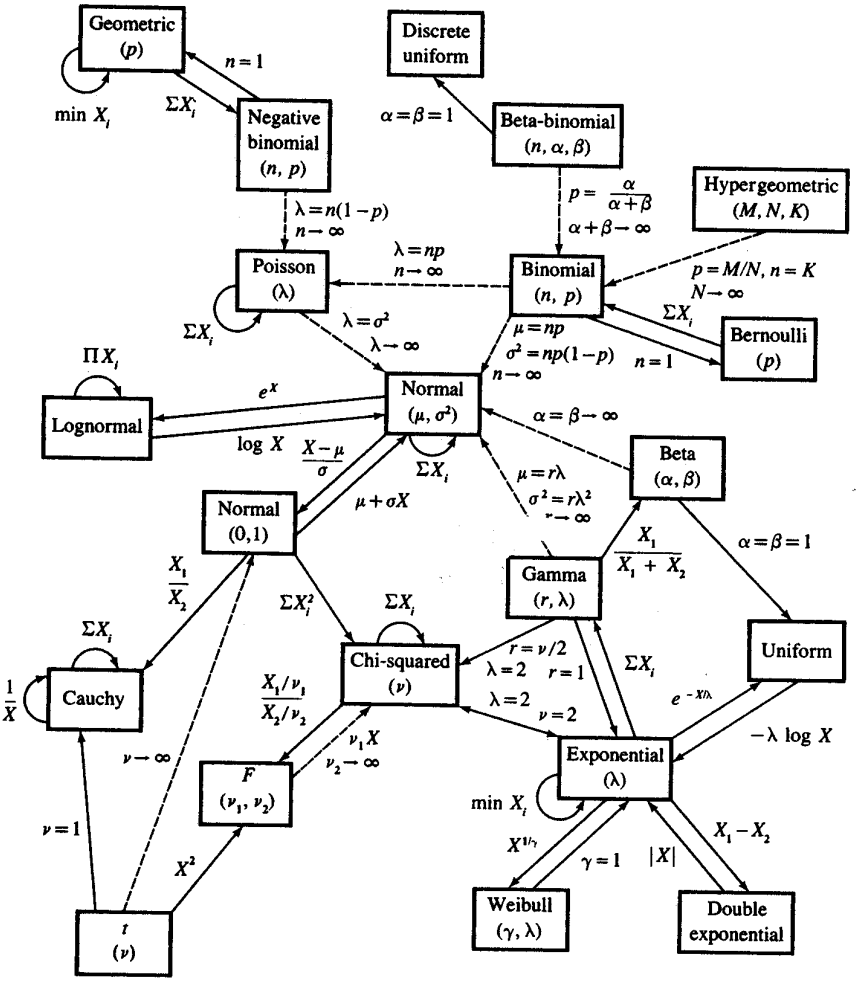


Figure 4.1: *Relationships among common distributions. Adapted from Leemis (1986).*

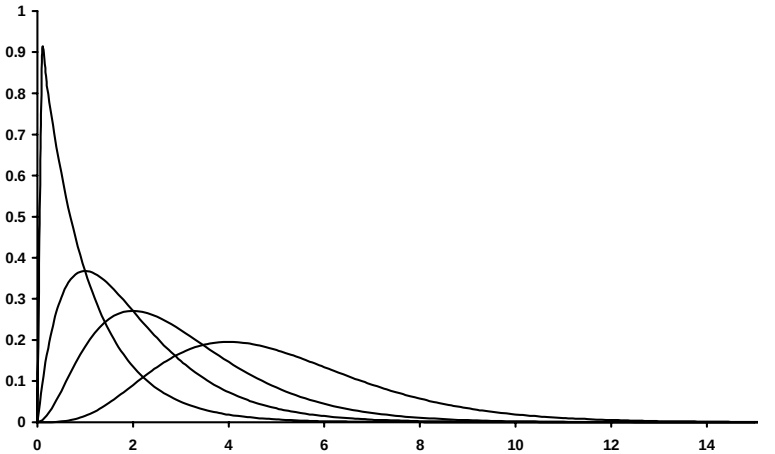


Figure 4.2: *Gamma distributions with parameters* $\alpha = 1$ *and* β *=1, 2, 3 and 5, respectively.*

with parameters $\alpha = p/2$ and $\beta = 2$. The χ^2 distribution has mean value $E\left(\chi^2\right) = p$ and variance $Var\left(\chi^2\right) = 2p$. Note that, for data from a Normal distribution, $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2$ with $(n-1)$ degrees of freedom. For this reason, the gamma distribution is sometimes used for modelling of variances. An example of this is given on page 157.

4.3.2 The Exponential distribution

The exponential distribution has density

$$f\left(y;\beta\right) = \frac{1}{\beta} e^{-y/\beta} \tag{4.5}$$

It can be obtained as a gamma distribution with $\alpha = 1$. The exponential distribution is sometimes used as a simple model for lifetime data.

4.3.3 An application with a gamma distribution

Example 4.1 Hurn et al (1945), quoted from McCullagh and Nelder (1989), studied the clotting time of blood. Two different clotting agents were com-

pared for different concentrations of plasma. The data are:

	Clotting time	
Conc	Agent 1	Agent 2
5	118	69
10	58	35
15	42	26
20	35	21
30	27	18
40	25	16
60	21	13
80	19	12
100	18	12

Duration data can often be modeled using the gamma distribution. The canonical link of the gamma distribution is minus the inverse link, $-1/\mu$. Preliminary analysis of the data suggested that the relation between clotting time and concentration was better approximated by a linear function if the concentrations were log-transformed. Thus, the models that were fitted to the data were of type

$$\frac{1}{\mu} = \beta_0 + \beta_1 d + \beta_2 x + \beta_3 dx$$

where $x = \log(\text{conc})$ and d is a dummy variable with $d = 1$ for lot 1 and $d = 0$ for lot 2. This is a kind of covariance analysis model (see Chapter 1). A Genmod analysis of the full model gave the following output:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	14	0.0294	0.0021
Scaled Deviance	14	17.9674	1.2834
Pearson Chi-Square	14	0.0298	0.0021
Scaled Pearson X2	14	18.2205	1.3015
Log Likelihood	.	-26.5976	.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-0.0239	0.0013	359.9825	0.0001
AGENT	1	1	0.0074	0.0015	24.9927	0.0001
AGENT	2	0	0.0000	0.0000	.	.
LC		1	0.0236	0.0005	1855.0452	0.0001
LC*AGENT	1	1	-0.0083	0.0006	164.0704	0.0001
LC*AGENT	2	0	0.0000	0.0000	.	.
SCALE		1	611.1058	203.6464	.	.

NOTE: The scale parameter was estimated by maximum likelihood.

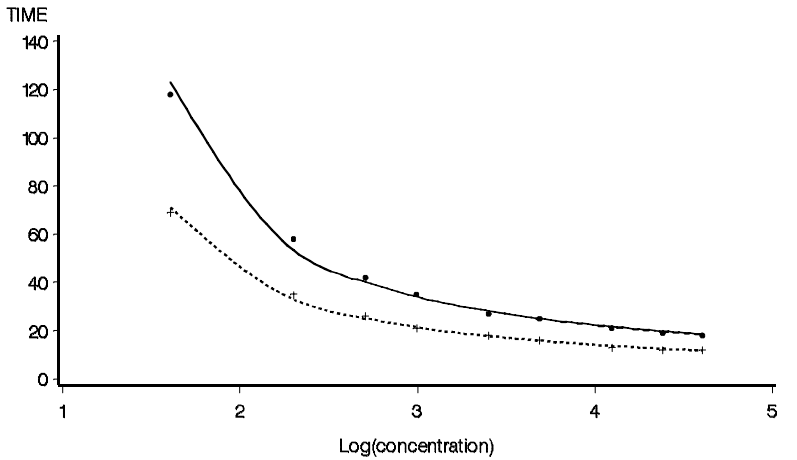


Figure 4.3: *Relation between clotting time and log(concentration)*

We can see that all parameters are significantly different from zero, which means that we cannot simplify the model any further. The scaled deviance is 17.97 on 14 *df*. A plot of the fitted model, along with the data, is given in Figure 4.3. The fit is good, but McCullagh and Nelder note that the lowest concentration value might have been misrecorded. □

4.4 The inverse Gaussian distribution

The inverse Gaussian distribution, also called the Wald distribution, has its roots in models for random movement of particles, called Brownian motion after the British botanist Robert Brown. The density function is

$$f(x;\mu,\lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left[\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right] \tag{4.6}$$

The distribution has two parameters, μ and λ. It has mean value μ and variance μ^3/λ. It belongs to the exponential family and is available in procedure Genmod. The distribution is skewed to the right, and resembles the log-normal and gamma distributions. A graph of the shape of inverse Gaussian distributions is given in Figure 4.4.

In a so called Wiener process for a particle, the time T it takes for the particle to reach a barrier for the first time has an inverse Gaussian distribution. The

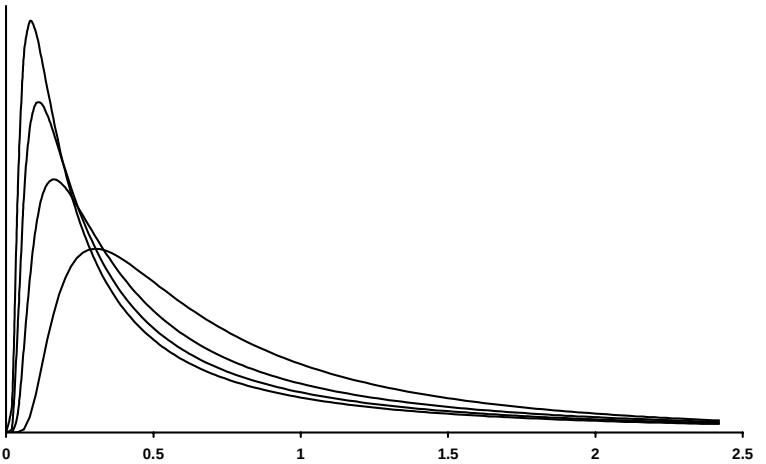


Figure 4.4: *Inverse Gaussian distributions with* $\lambda = 1$ *and mean values* $\mu = 1$ *(lowest curve),* $\mu = 2$*,* $\mu = 3$ *and* $\mu = 4$*, respectively.*

distribution has also been used to model the length of time a particle remains in the blood; maternity data; crop field size; and length of stay in hospitals. See Armitage and Colton (1998) for references.

4.5 Model diagnostics

For the example data on page 76, the deviance residuals and the predicted values were stored in a file for further analysis. In this section we will present some examples of model diagnostics based on these data.

4.5.1 Plot of residuals against predicted values

The residuals can be plotted against the predicted values. In Normal theory models this kind of plot can be used to detect heteroscedasticity. Such a plot for our example data is given in Figure 4.5.

The plot does not show the even, random pattern that would be expected. Two observations in the lower right corner are possible outliers.

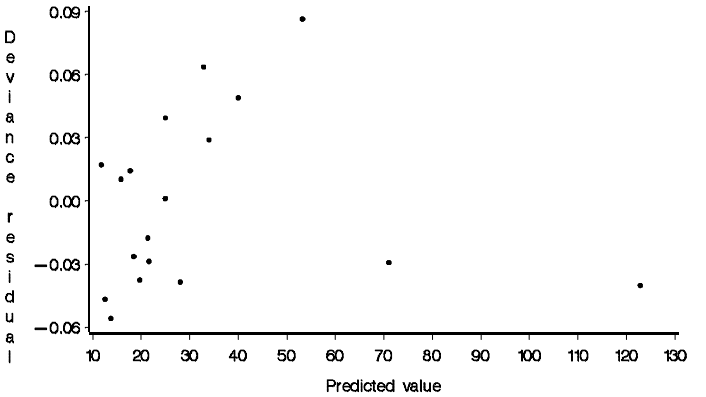


Figure 4.5: *Plot of residuals against predicted values for the Gamma regression data*

4.5.2 Normal probability plot

The Normal probability plot can be used to assess the distributional properties of the residuals. For most generalized linear models the residuals can be regarded as asymptotically Normal. However, the distributional properties of the residuals in finite samples depend upon the type of model. Still, the normal probability plot is a useful tool for detecting anomalies in the data. A Normal probability plot for the gamma regression data is given in Figure 4.6.

4.5.3 Plots of residuals against covariates

A plot of residuals against quantitative covariates can be used to detect whether the assumed model is too simple. In simple linear models, systematic patterns in this kind of plot may indicate that non-linear terms are needed, or that some observations may be outliers. A plot of deviance residuals against log(concentration) for the gamma regression data is given in Figure 4.7.

The figure may indicate that the two observations in the lower left corner are outliers. These are the same two observations that stand out in figure 4.5.

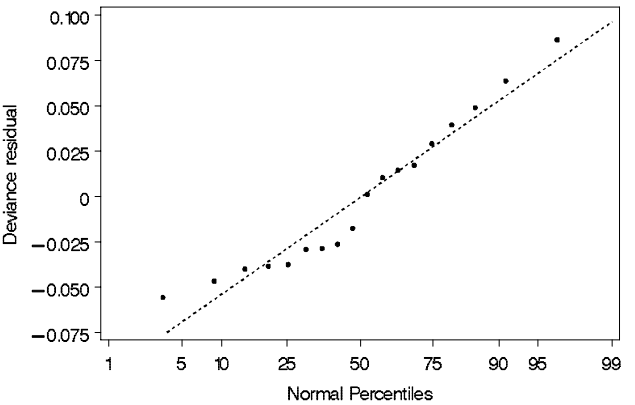


Figure 4.6: *Normal probability plot for the gamma regression data*

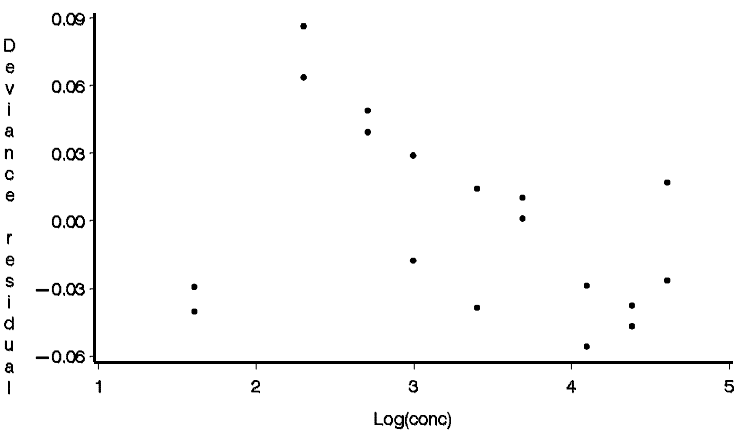


Figure 4.7: *Plot of deviance residuals against Log(conc) for the gamma regression data*

4.5.4 Influence diagnostics

The value of Dfbeta with respect to log(conc) was calculated for all observations and plotted against log(conc). The resulting plot is given in Figure 4.8. The figure shows that observations with the lowest value of log(conc) have the largest influence on the results. These are the same observations that were noted in other diagnostic plots above; the two possible outliers noted earlier are actually placed on top of each other in this plot.

The diagonal elements of the Hat matrix were computed using Proc Insight (SAS, 2000b). These values are plotted against the sequence numbers of the observations in Figure 4.9. Since there are four parameters and $n = 18$, the average leverage is $4/18 = 0.222$. As noted in Chapter 3, observation with a leverage above twice that amount, i.e. here above $2 \cdot 0.22 = 0.44$, should be examined. For these data the first two observations have a high leverage; these are the observations that have been noted in the other diagnostic plots.

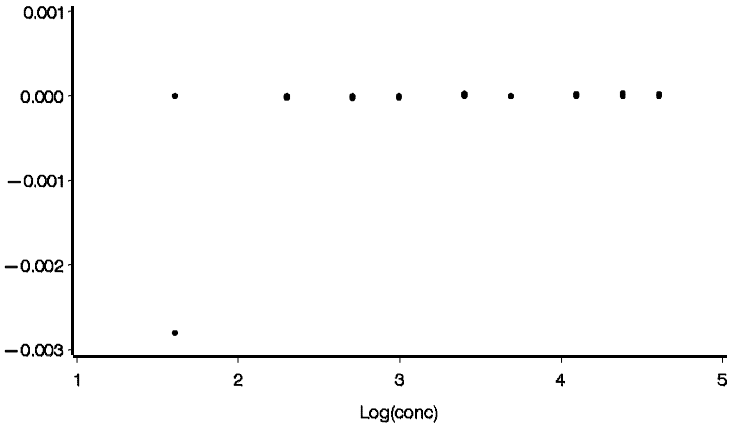


Figure 4.8: *Dfbeta plotted against log(conc) for the gamma regression data.*

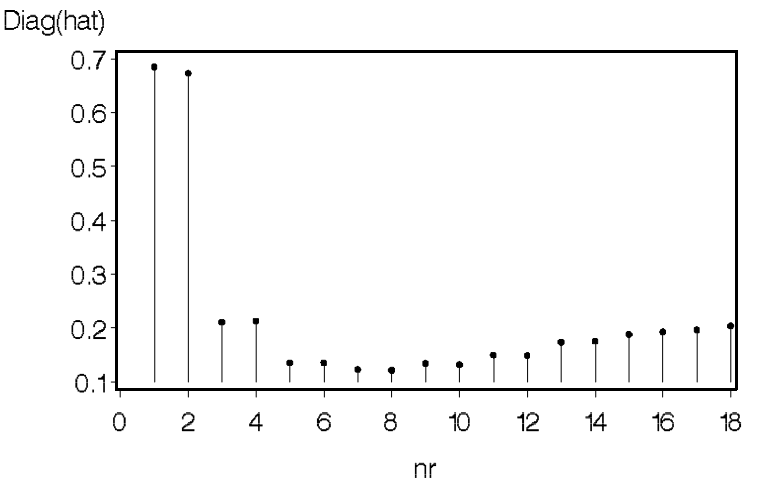


Figure 4.9: *Leverage plot, Gamma regression data.*

4.6 Exercises

Exercise 4.1 The following data, taken from Box and Cox (1964), show the survival times (in 10 hour units) of a certain variety of animals. The experiment is a two-way factorial experiment with factors Poison (three levels) and Treatment (four levels).

	Treatment			
Poison	A	B	C	D
I	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
II	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
III	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.36
	0.18	0.38	0.24	0.31
	0.23	0.29	0.22	0.33

Analyze these data to find possible effects of poison, treatment, and interactions. The analysis suggested by Box and Cox was a standard twoway ANOVA on the data transformed as $z = 1/y$. Make this analysis, and also make a generalized linear model analysis assuming that the data can be approximated with a gamma distribution. In both cases, make residual diagnostics and influence diagnostics.

Exercise 4.2 The data given below are the time intervals (in seconds) between successive pulses along a nerve fibre. Data were extracted from Cox and Lewis (1966), who gave credit to Drs. P. Fatt and B. Katz. The original data set consists of 799 observations; we use the first 200 observations only. If pulses arrive in a completely random fashion one would expect the distribution of waiting times between pulses to follow an exponential distribution. Fit an exponential distribution to these data by applying a generalized linear model with an appropriate distribution and link, and where the linear predictor only contains an intercept. Compare the observed data with the fitted distribution using different kinds of plots. The data are as follows:

0.21	0.03	0.05	0.11	0.59	0.06
0.18	0.55	0.37	0.09	0.14	0.19
0.02	0.14	0.09	0.05	0.15	0.23
0.15	0.08	0.24	0.16	0.06	0.11
0.15	0.09	0.03	0.21	0.02	0.14
0.24	0.29	0.16	0.07	0.07	0.04
0.02	0.15	0.12	0.26	0.15	0.33
0.06	0.51	0.11	0.28	0.36	0.14
0.55	0.28	0.04	0.01	0.94	0.73
0.05	0.07	0.11	0.38	0.21	0.49
0.38	0.38	0.01	0.06	0.13	0.06
0.01	0.16	0.05	0.10	0.16	0.06
0.06	0.06	0.06	0.11	0.44	0.05
0.09	0.04	0.27	0.50	0.25	0.25
0.08	0.01	0.70	0.04	0.08	0.16
0.38	0.08	0.32	0.39	0.58	0.56
0.74	0.15	0.07	0.26	0.25	0.01
0.17	0.64	0.61	0.15	0.26	0.03
0.05	0.34	0.07	0.10	0.09	0.02
0.30	0.07	0.12	0.01	0.16	0.14
0.49	0.07	0.11	0.35	1.21	0.17
0.01	0.35	0.45	0.07	0.93	0.04
0.96	0.14	1.38	0.15	0.01	0.05
0.23	0.31	0.05	0.05	0.29	0.01
0.74	0.30	0.09	0.02	0.19	0.47
0.01	0.51	0.12	0.12	0.43	0.32
0.09	0.20	0.03	0.05	0.13	0.15
0.05	0.08	0.04	0.09	0.10	0.10
0.26	0.07	0.68	0.15	0.01	0.27
0.05	0.03	0.40	0.04	0.21	0.29
0.24	0.08	0.23	0.10	0.19	0.20
0.26	0.06	0.40	0.51	0.15	1.10
0.16	0.78	0.04	0.27	0.35	0.71
0.15	0.29				

5. Binary and binomial response variables

In binary and binomial models, we model the response probabilities as functions of the predictors. A probability has range $0 \leq p \leq 1$. Since the linear predictor $\mathbf{X}\boldsymbol{\beta}$ can take on any value on the real line, we would like the model to use a link $g(p)$ that transforms a probability to the range $(-\infty, \infty)$. Three different functions are often used for this purpose: the probit link; the logit link; and the complementary log-log link. We will briefly discuss some arguments related to the choice of link for binary and binomial data.

5.1 Link functions

5.1.1 The probit link

The probit link transforms a probability by applying the function $\Phi^{-1}(p)$, where Φ is the standard Normal distribution function. One way to justify the probit link is as follows. Suppose that underlying the observed binary response Y is a continuous variable ξ that is normally distributed. If the value of ξ is larger than some threshold τ, then we observe $Y = 0$, else we observe $Y = 1$. The Normal distribution, used in this context, is called a tolerance distribution. This situation is illustrated in Figure 5.1.

In mathematical terms, the probit is that value of τ for which

$$p = \Phi(\tau) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\tau} e^{-u^2/2} du. \tag{5.1}$$

This is the integral of the standard Normal distribution. Thus, $\tau = \Phi^{-1}(p)$. In the original work leading to the probit (see Finney, 1947), the probit was defined as $\text{probit}(p) = 5 + \Phi^{-1}(p)$, to avoid working with negative numbers. However, most current computer programs define the probit without addition of the constant 5.

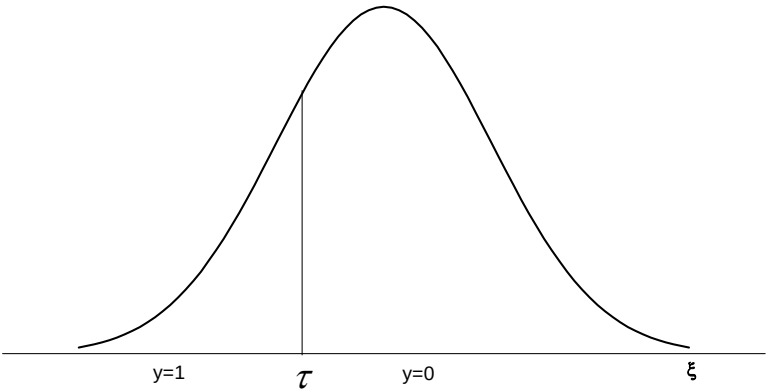


Figure 5.1: *The relation between* y *and* ξ *for a probit model*

5.1.2 The logit link

The logit link, or logistic transformation, transforms a probability as

$$\text{logit}\,(p) = \log \frac{p}{1-p}. \tag{5.2}$$

The ratio $\frac{p}{1-p}$ is the odds of success, so the logit is often called the log odds. The logit function is a sigmoid function that is symmetric around 0. The logistic link is rather close to the probit link, and since it is easier to handle mathematically, some authors prefer it to the probit link. The logit link is the canonical link for the binomial distribution so it is often the natural choice of link for binary and binomial data. The logit link corresponds to a tolerance distribution that is called the logistic distribution. This distribution has density

$$f\,(y) = \frac{\beta e^{\alpha+\beta y}}{[1+e^{\alpha+\beta y}]^2}.$$

5.1.3 The complementary log-log link

The complementary log-log link is based on arguments originating from a method called dilution assay. This is a method for estimating the number of

active organisms in a solution. The method works as follows. The solution containing the organisms is progressively diluted. Samples from each dilution are applied to plates that contain some growth medium. After some time it is possible to record, for each plate, whether it has been infected by the organism or not. Suppose that the original solution contained N individuals per unit volume. This means that dilution by a factor of two gives a solution with $\frac{N}{2}$ individuals per unit volume. After i dilutions the concentration is $\frac{N}{2^i}$. If the organisms are randomly distributed one would expect the number of individuals per unit volume to follow a Poisson distribution with mean μ_i. Thus, $\mu_i = \frac{N}{2^i}$ or, by taking logarithms, $\log \mu_i = \log N - i \log 2$. The probability that a plate will contain no organisms, assuming a Poisson distribution, is $e^{-\mu_i}$. Thus, if p_i is the probability that growth occurs under dilution i, then $p_i = 1 - e^{-\mu_i}$. Therefore, $\mu_i = -\log(1 - p_i)$ which gives

$$\log \mu_i = \log\left[-\log(1 - p_i)\right]. \tag{5.3}$$

This is the complementary log-log link: $\log\left[-\log(1 - p_i)\right]$. As opposed to the probit and logit links, this function is asymmetric around 0. The tolerance distribution that corresponds to the complementary log-log link is called the extreme value distribution, or a Gumbel distribution, and has density

$$f(y) = \beta \exp\left[(\alpha + \beta y) - e^{(\alpha+\beta y)}\right].$$

The probit, logit and complementary log-log links are compared in Figure 5.2.

5.2 Distributions for binary and binomial data

5.2.1 The Bernoulli distribution

A binary random variable that takes on the values 1 and 0 with probabilities p and $1 - p$, respectively, is said to follow a Bernoulli distribution. The probability function of a Bernoulli random variable y is

$$f(y) = p^y (1 - p)^{1-y} = \begin{cases} 1 - p & \text{if } y = 0 \\ p & \text{if } y = 1 \end{cases}. \tag{5.4}$$

The Bernoulli distribution has mean value $E(y) = p$ and variance $Var(y) = p(1 - p)$.

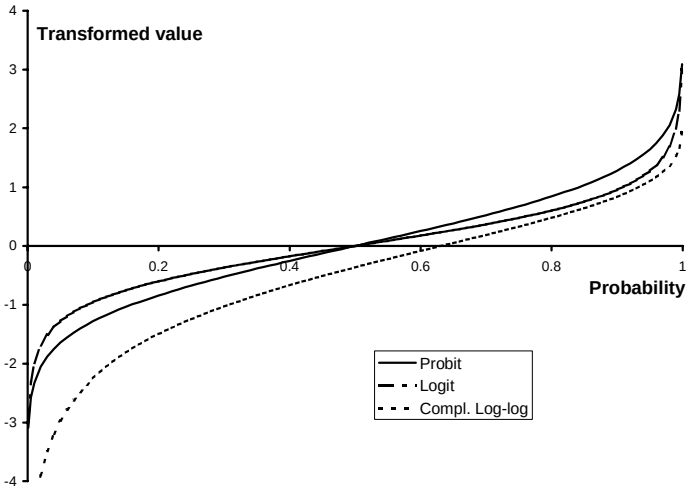


Figure 5.2: *The probit, logit and complementary log-log links*

5.2.2 The Binomial distribution

If a Bernoulli trial is repeated n times such that the trials are independent, then $y =$ the number of successes (1:s) among the n trials follows a binomial distribution with parameters n and p. The probability function of the binomial distribution is

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}. \tag{5.5}$$

The binomial distribution has mean

$$E(y) = np$$

and variance

$$Var(y) = np(1-p).$$

The proportion of successes, $\widehat{p} = \frac{y}{n}$, follows the same distribution, except for a scale factor: $f(y) = f(\widehat{p})$. It holds that $E(\widehat{p}) = p$ and $Var(\widehat{p}) = \frac{p(1-p)}{n}$. As was demonstrated in formula (2.6) on page 37, the binomial distribution is a member of the exponential family. Since the Bernoulli distribution is a special case of the binomial distribution with $n = 1$, even the Bernoulli distribution is an exponential family distribution.

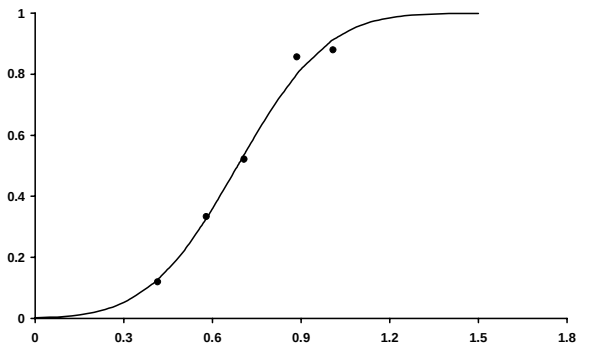
When the binomial distribution is applied for modeling real data, the crucial assumption is the assumption of independence. If independence does not hold, this can often be diagnosed as over-dispersion.

5.3 Probit analysis

Example 5.1 Finney (1947) reported on an experiment on the effect of Rotenone, in different concentrations, when sprayed on the insect *Macrosiphoniella sanborni*, in batches of about fifty. The results were:

Conc	Log(Conc)	No. of insects	No. affected	% affected
10.2	1.01	50	44	88
7.7	0.89	49	42	86
5.1	0.71	46	24	52
3.8	0.58	48	16	33
2.6	0.41	50	6	12

A plot of the relation between the proportion of affected insects and Log(Conc) is given below. A fitted distribution is also included.



Relation between log(Conc) and proportion affected

This situation is an example of a "probit analysis" setting. The dependent variable is a proportion. The probit analysis approach is to assume a linear relation between $\Phi^{-1}(p)$ and log(dose), where Φ^{-1} is the inverse of the cumulative Normal distribution (the so called probit), and p is the proportion affected in the population. This can be achieved as a generalized linear model by specifying a binomial distribution for the response and using a probit link.

The following SAS program was used to analyze these data using Proc Genmod:

```
DATA probit;
  INPUT conc n x;
  logconc=log10(conc);
CARDS;
10.2 50 44
 7.7 49 42
 5.1 46 24
 3.8 48 16
 2.6 50  6
;
PROC GENMOD DATA=probit;
 MODEL x/n = logconc /
    LINK = probit
    DIST = bin
;
```

Part of the output is as follows:

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.PROBIT
Distribution	BINOMIAL
Link Function	PROBIT
Dependent Variable	X
Dependent Variable	N
Observations Used	5
Number Of Events	132
Number Of Trials	243

This part simply gives us confirmation that we are using a Binomial distribution and a Probit link.

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	3	1.7390	0.5797
Scaled Deviance	3	1.7390	0.5797
Pearson Chi-Square	3	1.7289	0.5763
Scaled Pearson X2	3	1.7289	0.5763
Log Likelihood	.	-120.0516	.

This section gives information about the fit of the model to the data. The deviance can be interpreted as a χ^2 variate on 3 degrees of freedom, if the sample is large. In this case, the value is 1.74 which is clearly non-significant, indicating a good fit. Collett (1991) states that "a useful rule of thumb is

that when the deviance on fitting a linear logistic model is approximately equal to its degrees of freedom, the model is satisfactory" (p. 66).

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-2.8875	0.3501	68.0085	0.0001
LOGCONC	1	4.2132	0.4783	77.5919	0.0001
SCALE	0	1.0000	0.0000	.	.

The output finally contains an Analysis of Parameter estimates. This gives estimates of model parameters, their standard errors, and a Wald test of each parameter in the form of a χ^2 test. In this case, the estimated model is

$$\Phi^{-1}(p) = -2.8875 + 4.2132 \cdot \log(\text{conc}).$$

The dose that affects 50% of the animals (ED_{50}) can be calculated: if $p = 0.5$ then $\Phi^{-1}(p) = 0$ from which

$$\begin{aligned} \log(\text{conc}) &= -\frac{\widehat{\beta}_0}{\widehat{\beta}_1} = \frac{2.8875}{4.2132} = 0.68535 \text{ giving} \\ \text{conc} &= 10^{0.68535} = 4.8456. \end{aligned}$$

□

5.4 Logit (logistic) regression

Example 5.2 Since the logit and probit links are very similar, we can alternatively analyze the data in Table 5.1 using a binomial distribution with a logit link function. The program and part of the output are similar to the probit analysis. The fit of the model is excellent, as for the probit analysis case:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	3	1.4241	0.4747
Scaled Deviance	3	1.4241	0.4747
Pearson Chi-Square	3	1.4218	0.4739
Scaled Pearson X2	3	1.4218	0.4739
Log Likelihood	.	-119.8942	.

The parameter estimates are given by the last part of the output:

```
                Analysis Of Parameter Estimates

Parameter     DF    Estimate     Std Err   ChiSquare  Pr>Chi

INTERCEPT      1     -4.8869      0.6429     57.7757  0.0001
LOGCONC        1      7.1462      0.8928     64.0744  0.0001
SCALE          0      1.0000      0.0000            .       .
```

The resulting estimated model is

$$\text{logit}\left(\widehat{p}\right) = \log \frac{\widehat{p}}{1-\widehat{p}} = -4.8869 + 7.1462 \log\left(\text{conc}\right).$$

The estimated model parameters permits us to estimate e.g. the dose that gives a 50% effect (ED_{50}) as the value of $\log\left(\text{conc}\right)$ for which $p = 0.5$. Since $\log \frac{0.5}{1-0.5} = 0$, this value is

$$ED_{50} = -\frac{\widehat{\beta}_0}{\widehat{\beta}_1} = -\frac{-4.8869}{7.1462} = 0.6839$$

which, on the dose scale, is $10^{0.6839} = 4.83$. This is similar to the estimate provided by the probit analysis. Note that the estimated proportion affected at a given concentration can be obtained from

$$\widehat{p} = \frac{\exp\left(-4.8869 + 7.1462 \cdot \log\left(\text{conc}\right)\right)}{1 + \exp\left(-4.8869 + 7.1462 \cdot \log\left(\text{conc}\right)\right)}.$$

□

It can be mentioned that data in the form of proportions have previously often been analyzed as general linear models by using the so called Arc sine transformation $y = \arcsin\left(\sqrt{\widehat{p}}\right)$ (see e.g. Snedecor and Cochran, 1980).

5.5 Multiple logistic regression

5.5.1 Model building

Model building in multiple logistic regression models can be done in essentially the same way as in standard multiple regression.

Example 5.3 The data in Table 5.1, taken from Collett (1991), were collected to explore whether it was possible to diagnose nodal involvement in

prostatic cancer based on non-invasive methods. The variables are:

Age	Age of the patient
Acid	Level of serum acid phosphate
X-ray	Result of x-ray examination (0=negative, 1=positive)
Size	Tumour size (0=small, 1=large)
Grade	Tumour grade (0=less serious, 1=more serious)
Involvement	Nodal involvement (0=no, 1=yes)

The data analytic task is to explore whether the independent variables can be used to predict the probability of nodal involvement. We have two continuous covariates and three covariates coded as dummy variables. Initial analysis of the data suggests that the value of Acid should be log-transformed prior to the analysis.

There are 32 possible linear logistic models, excluding interactions. As a first step in the analysis, all these models were fitted to the data. A summary of the results is given in Table 5.2.

A useful rule-of-thumb in model building is to keep in the model all terms that are significant at, say, the 20% level. In this case, a kind of backward elimination process would start with the full model. We would then delete Grade from the model ($p = 0.29$). In the model with Age, log(acid), x-ray and size, age is not significant ($p = 0.26$). This suggests a model that includes log(acid), x-ray and size; in this model, all terms are significant ($p < 0.05$).

There are no indications of non-linear relations between log(acid) and the probability of nodal involvement. It remains to investigate whether any interactions between the terms in the model would improve the fit. To check this, interaction terms were added to the full model. Since there are five variables, the model was tested with all 10 possible pairwise interactions. The interactions size*grade ($p = 0.01$) and logacid*grade ($p = 0.10$) were judged to be large enough for further consideration. Note that grade was not suggested by the analysis until the interactions were included. We then tried a model with both these interactions. Age could be deleted. The resulting model includes logacid ($p = 0.06$), x-ray ($p = 0.03$), size ($p = 0.21$), grade ($p = 0.19$), logacid*grade ($p = 0.11$), and size*grade ($p = 0.02$). Part of the output is:

```
        Criteria For Assessing Goodness Of Fit

Criterion              DF          Value      Value/DF

Deviance               46        36.2871        0.7889
Scaled Deviance        46        36.2871        0.7889
Pearson Chi-Square     46        42.7826        0.9301
Scaled Pearson X2      46        42.7826        0.9301
Log Likelihood          .       -18.1436             .
```

Age	Acid	Xray	Size	Grade	Involv	Age	Acid	Xray	Size	Grade	Involv
66	.48	0	0	0	0	64	.40	0	1	1	0
68	.56	0	0	0	0	61	.50	0	1	0	0
66	.50	0	0	0	0	64	.50	0	1	1	0
56	.52	0	0	0	0	63	.40	0	1	0	0
58	.50	0	0	0	0	52	.55	0	1	1	0
60	.49	0	0	0	0	66	.59	0	1	1	0
65	.46	1	0	0	0	58	.48	1	1	0	1
60	.62	1	0	0	0	57	.51	1	1	1	1
50	.56	0	0	1	1	65	.49	0	1	0	1
49	.55	1	0	0	0	65	.48	0	1	1	0
61	.62	0	0	0	0	59	.63	1	1	1	0
58	.71	0	0	0	0	61	1.02	0	1	0	0
51	.65	0	0	0	0	53	.76	0	1	0	0
67	.67	1	0	1	1	67	.95	0	1	0	0
67	.47	0	0	1	0	53	.66	0	1	1	0
51	.49	0	0	0	0	65	.84	1	1	1	1
56	.50	0	0	1	0	50	.81	1	1	1	1
60	.78	0	0	0	0	60	.76	1	1	1	1
52	.83	0	0	0	0	45	.70	0	1	1	1
56	.98	0	0	0	0	56	.78	1	1	1	1
67	.52	0	0	0	0	46	.70	0	1	0	1
63	.75	0	0	0	0	67	.67	0	1	0	1
59	.99	0	0	1	1	63	.82	0	1	0	1
64	1.87	0	0	0	0	57	.67	0	1	1	1
61	1.36	1	0	0	1	51	.72	1	1	0	1
56	.82	0	0	0	1	64	.89	1	1	0	1
						68	1.26	1	1	1	1

Table 5.1: *Predictors of nodal involvement on prostate cancer patients*

Terms	Deviance	*df*
(Intercept only)	70.25	52
Age	69.16	51
log(acid)	64.81	51
Xray	59.00	51
Size	62.55	51
Grade	66.20	51
Age, log(acid)	63.65	50
Age, x-ray	57.66	50
Age, size	61.43	50
Age, grade	65.24	50
log(acid), x-ray	55.27	50
log(acid), size	56.48	50
log(acid), grade	59.55	50
x-ray, size	53.35	50
x-ray, grade	56.70	50
size, grade	61.30	50
Age, log(acid), x-ray	53.78	49
Age, log(acid), size	55.22	49
Age, log(acid), grade	58.52	49
Age, x-ray, size	52.09	49
Age, x-ray, grade	55.49	49
Age, size, grade	60.28	49
log(acid), x-ray, size	48.99	49
log(acid), x-ray, grade	52.03	49
log(acid), size, grade	54.51	49
x-ray, size, grade	52.78	49
age, log(acid), x-ray, size	47.68	48
age, log(acid), x-ray, grade	50.79	48
age, log(acid), size, grade	53.38	48
log(acid), x-ray, size, grade	47.78	48
age, x-ray, size, grade	51.57	48
age, log(acid), x-ray, size, grade	46.56	47

Table 5.2: *Deviances for the nodal involvement data*

Analysis Of Parameter Estimates

Parameter			DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT			1	7.2391	3.4133	4.4980	0.0339
LOGACID			1	12.1345	6.5154	3.4686	0.0625
XRAY	0		1	-2.3404	1.0845	4.6571	0.0309
XRAY	1		0	0.0000	0.0000	.	.
SIZE	0		1	2.5098	2.0218	1.5410	0.2145
SIZE	1		0	0.0000	0.0000	.	.
GRADE	0		1	-4.3134	3.2696	1.7404	0.1871
GRADE	1		0	0.0000	0.0000	.	.
LOGACID*GRADE	0		1	-10.4260	6.6403	2.4652	0.1164
LOGACID*GRADE	1		0	0.0000	0.0000	.	.
SIZE*GRADE	0	0	1	-5.6477	2.4346	5.3814	0.0204
SIZE*GRADE	0	1	0	0.0000	0.0000	.	.
SIZE*GRADE	1	0	0	0.0000	0.0000	.	.
SIZE*GRADE	1	1	0	0.0000	0.0000	.	.
SCALE			0	1.0000	0.0000	.	.

The model fits well, with Deviance/df=0.79. Since the size*grade interaction is included in the model, the main effects of size and of grade should also be included. The output suggest the following models for grade 0 and 1, respectively:

Grade 0: $\text{logit}(\widehat{p}) = 2.93 + 1.71 \cdot \log(\text{acid}) - 2.34 \cdot \text{x-ray} - 3.14 \cdot \text{size}$

Grade 1: $\text{logit}(\widehat{p}) = 7.24 + 12.13 \cdot \log(\text{acid}) - 2.34 \cdot \text{x-ray} + 2.51 \cdot \text{size}$

The probability of nodal involvement increases with increasing acid level. The increase is higher for patients with serious (grade 1) tumors. □

5.5.2 Model building tools

A set of tools for model building in logistic regression has been developed. These tools are similar to the tools used in multiple regression analysis. The Logistic procedure in the SAS package includes the following variable selection methods:

Forward selection: Starting with an empty model, the procedure adds, at each step, the variable that would give the lowest p-value of the remaining variables. The procedure stops when all variables have been added, or when no variables meet the pre-specified limit for the p-value.

Backward selection: Starting with a model containing all variables, variables are step by step deleted from the model until all variables remaining in the model meet a specified limit for their p-values. At each step, the variable with the largest p-value is deleted.

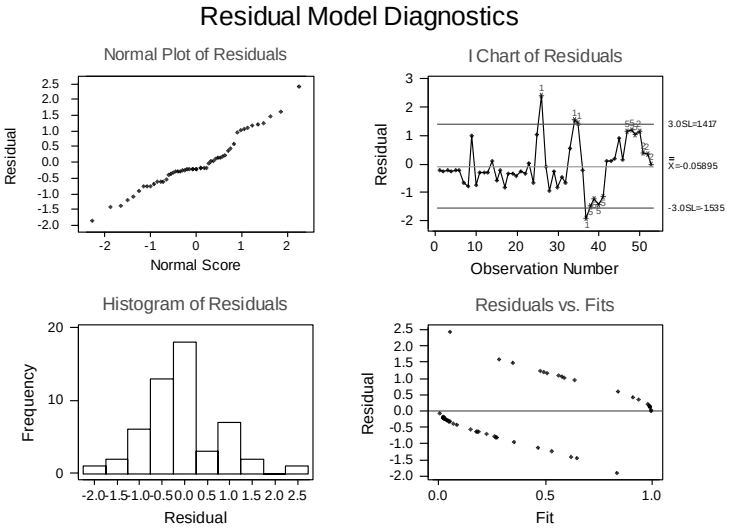


Figure 5.3: *Residual plots for nodal involvement data*

Stepwise selection: This is a modification of the forward selection model. Variables are added to the model step by step. In each step, the procedure also examines whether variables already in the model can be deleted.

Best subset selection: For $k = 1, 2, ...$ up to a user-specified limit, the method identifies a specified number of best models containing k variables. Tests for this method are based on score statistics (see Chapter 2).

Although automatic variable selection methods may sometimes be useful for "quick and dirty" model building, they should be handled with caution. There is no guarantee that an automatic procedure will always come up with the correct answer; see Agresti (1990) for a further discussion.

5.5.3 Model diagnostics

As an illustration of model diagnostics for logistic regression models, the predicted values and the residuals were stored as new variables for the multiple logistic regression data (Table 5.1). Based on these, a set of standard diagnostic plots was prepared using Minitab. These plots are reproduced in Figure 5.3.

It appears that the distribution of the (deviance) residuals is reasonably Normal. The "runs" that are shown in the I chart appear because of the way the data set was sorted. Note that the Residuals vs. Fits plot is not very informative for binary data. This is because the points scatter in two groups: one for observations with $y = 1$ and another group for observations with $y = 0$.

5.6 Odds ratios

If an event occurs with probability p, then the odds in favor of the event is

$$\text{Odds} = \frac{p}{1-p}. \tag{5.6}$$

For example, if an event occurs with probability $p = 0.75$, then the odds in favor of that event is $0.75/(1-0.75) = 3$. This means that a "success" is three times as likely as a "failure". If the odds are known, the probability p can be calculated as $p = \frac{\text{Odds}}{\text{Odds}+1}$.

A comparison between two events, or a comparison between e.g. two groups of individuals with respect to some event, can be made by computing the odds ratio

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}. \tag{5.7}$$

If $p_1 = p_2$ then $OR = 1$. An odds ratio larger than 1 is an indication that the event is more likely in the first group than in the second group. The odds ratio can be estimated from sample data as long as the relevant probabilities can be estimated. Estimated odds ratios are, of course, subject to random variation.

If the probabilities are small, the odds ratio can be used as an approximation to the relative risk, which is defined as

$$RR = \frac{p_1}{p_2}. \tag{5.8}$$

In some sampling schemes, it is not possible to estimate the relative risk but it may be possible to estimate the odds ratio. One example of this is in so called case-control studies. In such studies a number of patients with a certain disease are studied. One (or more) healthy patient is selected as a control for each patient in the study. The presence or absence of certain risk factors is assessed both for the patients and for the controls. Because of the way the sample was selected, the question whether the risk factor is related to disease occurrence cannot be answered by computing a risk ratio, but it may be possible to estimate the odds ratio.

Example 5.4 Freeman (1989) reports on a study designed to assess the relation between smoking and survival of newborn babies. 4915 babies to young mothers were followed during their first year. For each baby it was recorded whether the mother smoked and whether the baby survived the first year. Data are as follows:

	Survived	
Smoker	Yes	No
Yes	499	15
No	4327	74

The probability of dying for babies to smoking mothers is estimated as $15/(499+15) = 0.02918$ and for non-smoking mothers it is $74/(4327+74) = 0.01681$. The odds ratio is $\frac{0.02918/(1-0.02918)}{0.01681/(1-0.01681)} = 1.758$. The odds of death for the baby is higher for smoking mothers. □

Odds ratios can be estimated using logistic regression. Note that in logistic regression we use the model $\log \frac{p}{1-p} = \alpha + \beta x$ where, in this case, x is a dummy variable with value 1 for the smokers and 0 for the nonsmokers. This is the log of the odds, so the odds is $\exp(\alpha + \beta x)$. Using $x = 1$ in the numerator and $x = 0$ in the denominator gives the odds ratio as

$$OR = \frac{\exp(\alpha + \beta)}{\exp(\alpha)} = e^{\beta}$$

Thus, the odds ratio can be obtained by exponentiating the regression coefficient β in a logistic regression. We will use the same data as in the previous example to illustrate this. A SAS program for this analysis can be written as follows:

```
DATA babies;
INPUT smoking $ survival $ n;
CARDS;
Yes Yes  499
Yes No    15
No  Yes 4327
No  No    74
;
PROC GENMOD DATA=babies order=data;
CLASS smoking;
FREQ n;
MODEL survival = smoking/
      dist=bin link=logit ;
RUN;
```

Part of the output is as follow:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	4913	886.9891	0.1805
Scaled Deviance	4913	886.9891	0.1805
Pearson Chi-Square	4913	4914.9964	1.0004
Scaled Pearson X2	4913	4914.9964	1.0004
Log Likelihood		-443.4945	

The model fit, as judged by Deviance/df, is excellent.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-4.0686	0.1172	-4.2983	-3.8388	1204.34	<.0001
smoking	Yes	1	0.5640	0.2871	0.0013	1.1267	3.86	0.0495
smoking	No	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

There is a significant relationship between smoking and the risk of dying for the babies ($p = 0.0495$). The odds ratio can be calculated as $e^{0.5640} = 1.7577$ which is the same result as we got above by hand calculation. But the Genmod procedure also gives a test and a confidence interval for the parameter.

5.7 Overdispersion in binary/binomial models

Overdispersion means that the variance of the response is larger than would be expected for the chosen model. For binomial models, the variance of $y =$"number of successes" is $np\,(1-p)$, and the variance of $\widehat{p} = \frac{y}{n}$ is $\frac{p(1-p)}{n}$. A simple way to illustrate over-dispersion is to consider a simple dose-response experiment where the same dose has been used on two batches of animals. Suppose that the chosen dose has effect on 10 out of 50 animals in one of the replications, and on 20 out of 50 animals in the other replication. This means that there is actually a significant difference between the two replications ($p = 0.029$). In other less extreme cases, there may be a tendency for the responses to differ, even if the results are not significantly different at any given dose. Still, when all replications are considered together, a value of the Deviance/df statistic appreciably above unity may indicate that overdispersion is present in the data.

A common source of over-dispersion is that the data display some form of clustering. This means that the observations are not independent. For example, different batches of animals may come from different parents, and thus be genetically different. One way to model such over-dispersion is to assume that the mean value is still $E(y) = np$ but that the variance takes the form $Var(y) = np(1-p)\sigma^2$, where in the clustered case it can be assumed that $\sigma^2 = 1 + (k-1)\tau^2$. Here, k is the cluster size.

5.7.1 Estimation of the dispersion parameter

One way to account for over-dispersion is to estimate the over-dispersion parameter from the data. If the data have a known cluster structure, this can be done via the between-cluster variance, that can be estimated from

$$\widehat{\sigma}^2 = \frac{1}{r-1}\sum_{j=1}^{r}\frac{(y_j - n_j\widehat{p})}{n_j\widehat{p}(1-\widehat{p})} \tag{5.9}$$

where r is the number of clusters.

If the structure of the clustering is unknown, an alternative way of estimating the dispersion parameter is to use the observed value of Deviance/*df* (or Pearson χ^2/df) as an estimate, and to re-run the analysis using this value. A useful recommendation is to run a "maximal model", that contains all relevant factors, even if they are not significant. The dispersion parameter is estimated from this model. This value of the dispersion parameter is then kept constant in all later analyses of the data.

For simple models, for example models in designed experiments, an alternative is to ask the software to use a Maximum Likelihood estimate of the dispersion parameter. This option is present in Proc Genmod.

5.7.2 Modeling as a beta-binomial distribution

If one can suspect some form of clustering, another approach to the modeling is to assume that y follows a binomial distribution within clusters but that the parameter p follows some random distribution over clusters. If the distribution of p is known, the distribution of y will be a so called compound distribution which can be derived. A rather simple case is obtained when the distribution of p is a Beta distribution. Then, the distribution of y will follow a distribution called the Beta-binomial distribution. However, this distribution is not at present available in the Genmod procedure.

Estimation using Quasi-likelihood methods is an alternative approach to modeling overdispersion. This is discussed in Chapter 8.

5.7.3 An example of over-dispersed data

Example 5.5 *Orobanche* is a parasital plant that grows on the roots of other plants. A number of batches of *Orobanche* seeds of two varieties were grown on extract from Bean or Cucumber roots, and the number of seeds germinating was recorded. The data taken from Collett (1991), are:

O. aegyptiaca 75				O. aegyptiaca 73			
Bean		Cucumber		Bean		Cucumber	
y	n	y	n	y	n	y	n
10	39	5	6	8	16	3	12
23	62	53	74	10	30	22	41
23	81	55	72	8	28	15	30
26	51	32	51	23	45	32	51
17	39	46	79	0	4	3	7
		10	13				

It was of interest to compare the two varieties, and also to compare the two types of host plants. An analysis of these data using a binomial distribution with a logit link revealed that an interaction term was needed. Part of the output for a model containing Variety, Host and Variety*Host, is given below. The model does not fit well (Deviance=33.2778 on 17 df, $p = 0.01$). The ratio Deviance/df is nearly 2, indicating that overdispersion may be present.

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	17	33.2778	1.9575
Scaled Deviance	17	33.2778	1.9575
Pearson Chi-Square	17	31.6511	1.8618
Scaled Pearson X2	17	31.6511	1.8618
Log Likelihood		-543.1106	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter			DF	Estimate	Standard Error
Intercept			1	0.7600	0.1250
variety	73		1	-0.6322	0.2100
variety	75		0	0.0000	0.0000
host	Bean		1	-1.3182	0.1775
host	Cucumber		0	0.0000	0.0000
variety*host	73	Bean	1	0.7781	0.3064
variety*host	73	Cucumber	0	0.0000	0.0000
variety*host	75	Bean	0	0.0000	0.0000
variety*host	75	Cucumber	0	0.0000	0.0000
Scale			0	1.0000	0.0000

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
variety	1	2.53	0.1121
host	1	37.48	<.0001
variety*host	1	6.41	0.0114

As a second analysis, the data were analyzed using the automatic feature in Genmod to estimate the scale parameter from the data using the Maximum Likelihood method. Part of the output was as follows:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	17	33.2778	1.9575
Scaled Deviance	17	17.0000	1.0000
Pearson Chi-Square	17	31.6511	1.8618
Scaled Pearson X2	17	16.1690	0.9511
Log Likelihood		-277.4487	

The procedure now uses a scaled deviance of 1.00. The parameter estimates are identical to those of the previous analysis, but the estimated standard errors are larger when we include a scale parameter. This has the effect that the Variety*Host interaction is no longer significant.

Analysis Of Parameter Estimates

Parameter			DF	Estimate	Standard Error
Intercept			1	0.7600	0.1748
variety	73		1	-0.6322	0.2938
variety	75		0	0.0000	0.0000
host	Bean		1	-1.3182	0.2483
host	Cucumber		0	0.0000	0.0000
variety*host	73	Bean	1	0.7781	0.4287
variety*host	73	Cucumber	0	0.0000	0.0000
variety*host	75	Bean	0	0.0000	0.0000
variety*host	75	Cucumber	0	0.0000	0.0000
Scale			0	1.3991	0.0000

LR Statistics For Type 3 Analysis

Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
variety	1	17	1.29	0.2718	1.29	0.2561
host	1	17	19.15	0.0004	19.15	<.0001
variety*host	1	17	3.27	0.0881	3.27	0.070

□

5.8 Exercises

Exercise 5.1

Species	Expo-sure	Rel. Hum	Temp	Deaths	N	Species	Expo-sure	Rel. Hum	Temp	Deaths	N
A	1	60.0	10	0	20	B	1	60.0	10	0	20
A	1	60.0	15	0	20	B	1	60.0	15	0	20
A	1	60.0	20	0	20	B	1	60.0	20	0	20
A	1	65.8	10	0	20	B	1	65.8	10	0	20
A	1	65.8	15	0	20	B	1	65.8	15	0	20
A	1	65.8	20	0	20	B	1	65.8	20	0	20
A	1	70.5	10	0	20	B	1	70.5	10	0	20
A	1	70.5	15	0	20	B	1	70.5	15	0	20
A	1	70.5	20	0	20	B	1	70.5	20	0	20
A	1	75.8	10	0	20	B	1	75.8	10	0	20
A	1	75.8	15	0	20	B	1	75.8	15	0	20
A	1	75.8	20	0	20	B	1	75.8	20	0	20
A	2	60.0	10	0	20	B	2	60.0	10	0	20
A	2	60.0	15	1	20	B	2	60.0	15	3	20
A	2	60.0	20	1	20	B	2	60.0	20	2	20
A	2	65.8	10	0	20	B	2	65.8	10	0	20
A	2	65.8	15	1	20	B	2	65.8	15	2	20
A	2	65.8	20	0	20	B	2	65.8	20	1	20
A	2	70.5	10	0	20	B	2	70.5	10	0	20
A	2	70.5	15	0	20	B	2	70.5	15	0	20
A	2	70.5	20	0	20	B	2	70.5	20	1	20
A	2	75.8	10	0	20	B	2	75.8	10	1	20
A	2	75.8	15	0	20	B	2	75.8	15	0	20
A	2	75.8	20	0	20	B	2	75.8	20	1	20
A	3	60.0	10	1	20	B	3	60.0	10	7	20
A	3	60.0	15	4	20	B	3	60.0	15	11	20
A	3	60.0	20	5	20	B	3	60.0	20	11	20
A	3	65.8	10	0	20	B	3	65.8	10	4	20
A	3	65.8	15	2	20	B	3	65.8	15	5	20
A	3	65.8	20	4	20	B	3	65.8	20	9	20
A	3	70.5	10	0	20	B	3	70.5	10	2	20
A	3	70.5	15	2	20	B	3	70.5	15	4	20
A	3	70.5	20	3	20	B	3	70.5	20	6	20
A	3	75.8	10	0	20	B	3	75.8	10	2	20
A	3	75.8	15	1	20	B	3	75.8	15	3	20
A	3	75.8	20	2	20	B	3	75.8	20	5	20
A	4	60.0	10	7	20	B	4	60.0	10	12	20
A	4	60.0	15	7	20	B	4	60.0	15	14	20
A	4	60.0	20	7	20	B	4	60.0	20	16	20
A	4	65.8	10	4	20	B	4	65.8	10	10	20
A	4	65.8	15	4	20	B	4	65.8	15	12	20
A	4	65.8	20	7	20	B	4	65.8	20	12	20
A	4	70.5	10	3	20	B	4	70.5	10	5	20
A	4	70.5	15	3	20	B	4	70.5	15	7	20
A	4	70.5	20	5	20	B	4	70.5	20	9	20
A	4	75.8	10	2	20	B	4	75.8	10	4	20
A	4	75.8	15	3	20	B	4	75.8	15	5	20
A	4	75.8	20	3	20	B	4	75.8	20	7	20

The data set given above contains data from an experiment studying the survival of snails. Groups of 20 snails were held for periods of 1, 2, 3 or 4 weeks under controlled conditions, where temperature and humidity were kept at assigned levels. The snails were of two species (A or B). The experiment was

a completely randomized design. The variables are as follows:

Species	Snail species A or B
Exposure	Exposure in weeks (1, 2, 3 or 4)
Humidity	Relative humidity (four levels)
Temp	Temperature in degrees Celsius (3 levels)
Deaths	Number of deaths
N	Number of snails exposed

Analyze these data to find whether Exposure, Humidity, Temp, or interactions between these have any effects on survival probability. Also, make residual diagnostics and leverage diagnostics.

Exercise 5.2 The file Ex5_2.dat gives the following information about passengers travelling on the Titanic when it sank in 1912. Background material for the data can be found on http://www.encyclopedia-titanica.org.

Name	Name of the person
PClass	Passenger class: 1st, 2nd or 3rd
Age	Age of the person
Sex	male or female
Survived	1=survived, 0=died

Find a model that can predict probability of survival as functions of the given covariates, and possible interactions. Note that some age data are missing.

Exercise 5.3 Finney (1947) reported some data on the relative potencies of Rotenone, Deguelin, and a mixture of these. Batches of insects were subjected to these treatments, in different concentrations, and the number of dead insects was recorded. The raw data are:

Treatment	ln(dose)	n	x
Rotenone	1.01	50	44
	0.89	49	42
	0.71	46	24
	0.58	48	16
	0.41	50	6
Deguelin	1.70	48	48
	1.61	50	47
	1.48	49	47
	1.31	48	34
	1.00	48	18
	0.71	49	16
Mixture	1.40	50	48
	1.31	46	43
	1.18	48	38
	1.00	46	27
	0.71	46	22
	0.40	47	7

Analyze these data. In particular, examine whether the regression lines can be assumed to be parallel.

Exercise 5.4 Fahrmeir & Tutz (2001) report some data on the risk of infection from births by Caesarian section. The response variable of interest is the occurrence of infections following the operation. Three dichotomous covariates that might affect the risk of infection were studied:

planned	Was the Caesarian section planned (=1) or not (=0)
risk	Were risk factors such as diabetes, excessive weight or others present (=1) or absent (=0)
antibio	Were antibiotics given as a prophylactic (=1) or not (=0)

The data are included in the following Sas program that also gives the value of the variable infection (1=infection, 0=no infection). The variable wt is the number of observations with a given combination of the other variables. Thus, for example, there were 17 un-infected cases (infection=0) with planned=1. risk=1, and antibio=1.

```
data cesarian;
INPUT planned antibio risk infection wt;
CARDS;
1 1 1 1 1
1 1 1 0 17
1 1 0 1 0
1 1 0 0 2
1 0 1 1 28
1 0 1 0 30
1 0 0 1 8
1 0 0 0 32
0 1 1 1 11
0 1 1 0 87
0 1 0 1 0
0 1 0 0 0
0 0 1 1 23
0 0 1 0 3
0 0 0 1 0
0 0 0 0 9
;
```

The following analyses were run on these data:

1. A binomial Glim model with a logit link, with only the main effects
2. Model 1 plus an interaction planned*antibio
3. Model 1 plus an interaction planned*risk
4. The same as model 3 but with some extra features, discussed below.

Some results:

Model 1

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	8	226.5177	28.3147
Scaled Deviance	8	226.5177	28.3147
Pearson Chi-Square	8	257.2508	32.1563
Scaled Pearson X2	8	257.2508	32.1563
Log Likelihood		-113.2588	

Model 2

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	7	226.4393	32.3485
Scaled Deviance	7	226.4393	32.3485
Pearson Chi-Square	7	254.7440	36.3920
Scaled Pearson X2	7	254.7440	36.3920
Log Likelihood		-113.2196	

Model 3

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	7	216.4759	30.9251
Scaled Deviance	7	216.4759	30.9251
Pearson Chi-Square	7	261.4010	37.3430
Scaled Pearson X2	7	261.4010	37.3430
Log Likelihood		-108.2380	

One problem with Model 3 is that no standard error, and no test, of the parameter for the planned*risk interaction is given by Sas. This is because the likelihood is rather flat which, in turn, depends on cells with observed count = 0. Therefore, Model 4 used the same model as Model 3, but with all values where Wt=0 replaced by Wt=0.5.

Model 4

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	11	231.5512	21.0501
Scaled Deviance	11	231.5512	21.0501
Pearson Chi-Square	11	446.4616	40.5874
Scaled Pearson X2	11	446.4616	40.5874
Log Likelihood		-115.7756	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	2.1440	1.0568	0.0728	4.2152	4.12	0.0425
planned	1	-0.8311	1.1251	-3.0363	1.3742	0.55	0.4601
antibio	1	3.4991	0.5536	2.4141	4.5840	39.95	<.0001
risk	1	-3.7172	1.1637	-5.9980	-1.4364	10.20	0.0014
planned*risk	1	2.4394	1.2477	-0.0060	4.8848	3.82	0.0506
Scale	0	1.0000	0.0000	1.0000	1.0000		

Questions: Use these data to answer the following questions:

A. Compare models 1 and 2 to test whether the planned*antibio interaction is significantly different from zero.

B. Compare models 1 and 3 to test whether the planned*risk interaction is significantly different from zero.

C. Explain why the Deviance in Model 4 has more degrees of freedom than in Model 3.

D. Based on the results for Model 4, estimate the odds ratios for infection for the factors in the model. Note that the program has modeled the probability of not being infected. Calculate the odds ratios for not being infected and, from these, the odds ratios of being infected.

E. Calculate predicted values and raw residuals for the first four observations in the data.

Exercise 5.5 An experiment has been designed in the following way: Two groups of patients (A1 and A2) were used. The groups differed regarding the type of diagnosis for a certain disease. Each group consisted of nine patients.

The patients in the two groups were randomly assigned to three different treatments: B1, B2 and B3, with three patients for each treatment in each group.

The blood pressure (Z) was measured at the beginning of the experiment on each patient.

A binary response variable (Y) was measured on each patient at the end of the experiment. It was modeled as $g(\mu) = \mathbf{XB}$, using some computer package. The final model included the main effects of A and B and their interaction, and the effect of the covariate Z. The slope for Z was different for different treatments, but not for the different patient groups or for the A*B interaction.

A. Write down the complete design matrix $\mathbf{X}$. You should include all dummy variables, even those that are redundant. Covariate values should be represented by some symbol. Also write down the corresponding parameter vector $\mathbf{B}$.

B. The link function used to analyze these data was the logit link $g(p) = \log \frac{p}{1-p}$. What is the inverse g^{-1} of this link function?

6. Response variables as counts

6.1 Log-linear models: introductory example

Count data can be summarized in the form of frequency tables or as contingency tables. The data are then given as the number of observations with each combination of values of some categorical variables. We will first look at a simple example with a contingency table of dimension 2×2.

Example 6.1 Norton and Dunn (1985) studied possible relations between snoring and heart problems. For 2484 persons it was recorded whether the person had any heart problems and whether the person was a snorer. An interesting question is then whether there is any relation between snoring and heart problems. The data are as follows:

	Snores		
Heart problems	Seldom	Often	Total
Yes	59	51	110
No	1958	416	2374
	2017	467	2484

We assume that the persons in the sample constitute a random sample from some population. Denote with p_{ij} the probability that a randomly selected person belongs to row category i and column category j of the table. This can be summarized as follows:

	Snores		
Heart problems	Seldom	Often	Total
Yes	p_{11}	p_{12}	$p_{1\cdot}$
No	p_{21}	p_{22}	$p_{2\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	1

A dot in the subscript indicates a marginal probability. For example, $p_{\cdot 1}$ denotes the probability that a person snores seldom, i.e. $p_{\cdot 1} = p_{11} + p_{21}$. □

6.1.1 A log-linear model for independence

If snoring and heart problems were statistically independent, it would hold that $p_{ij} = p_{i\cdot}p_{\cdot j}$ for all i and j. This is a model that we would like to compare with the more general model that snoring and heart problems are dependent. Instead of modeling the probabilities, we can state the models in terms of expected frequencies $\mu_{ij} = np_{ij}$, where n is the total sample size and μ_{ij} is the expected number in cell (i, j). Thus, the independence model states that

$$\mu_{ij} = np_{i\cdot}p_{\cdot j}.$$

This is a multiplicative model. By taking the logs of both sides we get an additive model assuming independence:

$$\begin{aligned} \log\left(\mu_{ij}\right) &= \log(n) + \log(p_{i\cdot}) + \log(p_{\cdot j}) \\ &= \mu + \alpha_i + \beta_j. \end{aligned} \tag{6.1}$$

In (6.1), α_i denotes the row effect (i.e. the effect of variable A), and β_j denotes the column effect (i.e. the effect of variable B). In log-linear model literature, effects are often denoted with symbols like λ_i^X, but we keep a notation that is in line with the notation of previous chapters. We can see that this model is a linear model (a linear predictor), and that the link function is log. Models of type (6.1) are called log-linear models.

Note that a model for a crosstable of dimension $r \times c$ can include at most $(r-1)$ parameters for the row effects and $(c-1)$ parameters for the column effect. This is analogous to ANOVA models. One way to constrain the parameters is to set the last parameter of each kind equal to zero. In our example, $r = c = 2$ so we need only one parameter α_i and one β_j, for example α_1 and β_1. In GLIM terms, the model for our example data can then be written as

$$\begin{pmatrix} \log(\mu_{11}) \\ \log(\mu_{12}) \\ \log(\mu_{21}) \\ \log(\mu_{22}) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \end{pmatrix}. \tag{6.2}$$

6.1.2 When independence does not hold

If independence does not hold we need to include in the model terms of type $(\alpha\beta)_{ij}$ that account for the dependence. The terms $(\alpha\beta)_{ij}$ represent interaction between the factors A and B, i.e. the effect of one variable depends on the level of the other variable. Then the model becomes

$$\log\left(\mu_{ij}\right) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}\,. \tag{6.3}$$

Any two-dimensional cross-table can be perfectly represented by the model (6.3); this model is called the saturated model. We can test the restrictions imposed by removing the parameters $(\alpha\beta)_{ij}$ by comparing the deviances: the saturated model will have deviance 0 on 0 degrees of freedom, so the deviance from fitting the model (6.1) can be used directly to test the hypothesis of independence.

6.2 Distributions for count data

So far, we have seen that a model for the expected frequencies in a crosstable can be formulated as a log-linear model. This model has the following properties:

The predictor is a linear predictor of the same type as in ANOVA.

The link function is a log function.

It remains to discuss what distributional assumptions to use.

6.2.1 The multinomial distribution

Suppose that a nominal variable Y has k distinct values $y_1, y_2, ..., y_k$ such that no implicit ordering is imposed on the values. In fact, the values might be observations on a single nominal variable, or they might be observations on cell counts in a multidimensional contingency table. The probabilities associated with the different values of Y are $p_1, p_2, ..., p_k$. We make n observations on Y. If the observations are independent, the probability to get n_1 observations with $Y = y_1$, n_2 observations with $Y = y_2$, and so on, is

$$P(n_1, n_2, \ldots, n_k|n) = \frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_k!} p_1^{n_1} \cdot p_2^{n_2} \cdot \ldots \cdot p_k^{n_k} \qquad (6.4)$$

Note that in (6.4), the total sample size $n = n_1 + n_2 + \ldots + n_k$ is regarded as fixed. Also note that the expression simplifies to the binomial distribution for the case $k = 2$. The distribution given in (6.4) is called the multinomial distribution. The multinomial distribution is a multivariate distribution since it describes the joint distribution of $y_1, y_2, ..., y_k$. It can be seen as a multivariate generalization of an exponential family distribution (see e.g. Agresti, 1990).

6.2.2 The product multinomial distribution

A contingency table may have some of its totals fixed by the design of the data collection. For example, 500 males and 500 females might have been interviewed in a survey. In such cases it is not meaningful to talk about the random distribution of the "gender" variable. For such data each "slice" of the table subdivided by gender may be seen as one realization of a multinomial distribution. The joint distribution of all cells of the table is then the product of several multinomial distributions, one for each slice. This joint distribution is called the product multinomial distribution.

6.2.3 The Poisson distribution

Suppose, again, that a nominal variable Y has k distinct values $y_1, y_2, ..., y_k$. We observe counts $n_1, n_2, \ldots, n_k$. The expected number of observations in cell i is μ_i. If the observations arrive randomly, the probability to observe n_i observations in cell i is

$$p(n_i) = \frac{e^{-\mu_i}\mu_i^{n_i}}{n_i!} \tag{6.5}$$

which is the probability function of a Poisson distribution. Note that in this case, the total sample size n is not regarded as fixed. This is the main difference between this sampling scheme and the multinomial case. Since sums of Poisson variables follow a Poisson distribution, n in itself follows a Poisson distribution with mean value $\sum_{i=1}^{k} \mu_i$.

6.2.4 Relation to contingency tables

Contingency tables can be of many different types. In some cases, the total sample size is fixed; an example is when it has been decided that $n = 1000$ individuals will be interviewed about some political question. In some cases even some of the margins of the table may be fixed. An example is when 500 males and 500 females will participate in a survey. A table with a fixed total sample size would suggest a multinomial distribution; if in addition one or more of the margins are fixed we would assume a product multinomial distribution. However, as noted by Agresti (1996), "For most analyses, one need not worry about which sampling model makes the most sense. For the primary inferential methods in this text, the same results occur for the Poisson, multinomial and independent binomial/multinomial sampling models" (p. 19).

Suppose that we observe a contingency table of size $i \times j$. The probability that an observation will fall into cell $(i,\ j)$ is p_{ij}. If the observations are independent and arrive randomly, the number of observations falling into cell $(i,\ j)$ follows a Poisson distribution with mean value μ_{ij}, if the total sample size n is random. If the cell counts n_{ij} follow a Poisson distribution then the conditional distribution of $n_{ij}|n$ is multinomial. The Poisson distribution is often used to model count data since it is rather easy to handle.

Note, however, there is no guarantee that a given set of data will adhere to this assumption. Sometimes the data may show a tendency to "cluster" such that arrival of one observation in a specific cell may increase the probability that the next observation falls into the same cell. This would lead to overdispersion. We will discuss overdispersion for Poisson models in a later section; a distribution called the negative binomial distribution may be used in some such cases. For the moment, however, we will see what happens if we tentatively accept the Poisson assumption for the data on snoring and heart problems.

6.3 Analysis of the example data

Example 6.2 We analyzed the data on page 111 using the Genmod procedure with Poisson distribution and a log link. The program was:

```
DATA snoring;
INPUT snore heart count;
CARDS;
1 1   51
1 0  416
0 1   59
0 0 1958
;
PROC GENMOD DATA=snoring;
CLASS snore heart;
MODEL count = snore heart /
            LINK = log
            DIST = poisson
;
RUN;
```

The output contains the following information:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	1	45.7191	45.7191
Scaled Deviance	1	45.7191	45.7191
Pearson Chi-Square	1	57.2805	57.2805
Scaled Pearson X2	1	57.2805	57.2805
Log Likelihood	.	15284.0145	.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	3.0292	0.1041	847.2987	0.0001
SNORE	0	1	1.4630	0.0514	811.6746	0.0001
SNORE	1	0	0.0000	0.0000	.	.
HEART	0	1	3.0719	0.0975	992.0240	0.0001
HEART	1	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

A similar analysis that includes an interaction term would produce a deviance of 0 on 0 *df*. Thus, the difference between our model and the saturated model can be tested; the difference in deviance is 45.7 on 1 degree of freedom which is highly significant when compared with the corresponding χ^2 limit with 1 *df*. We conclude that snoring and heart problems do not seem to be independent. Note that the Pearson chi-square of 57.28 on 1 *df* presented in the output is based on the textbook formula

$$\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - \widehat{\mu}_{ij}\right)^2}{\widehat{\mu}_{ij}}.$$

The conclusion is the same, in this case, but the tests are not identical.

The output also gives us estimates of the three parameters of the model: $\widehat{\mu} = 3.0292$, $\widehat{\alpha}_1 = 1.4630$ and $\widehat{\beta}_1 = 3.0719$. An analysis of the saturated model would give an estimate of the interaction parameter as $\widehat{(\alpha\beta)}_{11} = 1.4033$. From this we can calculate the odds ratio OR as

$$OR = \exp(1.4033) = 4.07.$$

Patients who snore have a four times larger odds of having heart problems. Odds ratios in log-linear models is further discussed in a later section. □

6.4 Testing independence in an r×c crosstable

The methods discussed so far can be extended to the analysis of cross-tables of dimension $r \times c$.

Example 6.3 Sokal and Rohlf (1973) presented data on the color of the Tiger beetle (*Cicindela fulgida*) for beetles collected during different seasons. The results are given as:

Season	Red	Other	Total
Early spring	29	11	40
Late spring	273	191	464
Early summer	8	31	39
Late summer	64	64	128
Total	374	297	671

A standard analysis of these data would be to test whether there is independence between season and color through a χ^2 test. The corresponding GLIM approach is to model the expected number of beetles as a function of season and color. The observed numbers in each cell are assumed to be generated from an underlying Poisson distribution. The canonical link for the Poisson distribution is the log link. Thus, a Genmod program for these data is

```
DATA chisq;
INPUT season $ color $ no;
CARDS;
Early_spring red    29
Early_spring other  11
Late_spring  red   273
Late_spring  other 191
Early_summer red     8
Early_summer other  31
Late_summer  red    64
Late_summer  other  64
;
PROC GENMOD DATA=chisq;
CLASS season color;
MODEL no = season color /
           DIST=poisson
           LINK=log ;
run;
```

Part of the output is

```
        Criteria For Assessing Goodness Of Fit

Criterion              DF         Value      Value/DF

Deviance                3       28.5964        9.5321
Scaled Deviance         3       28.5964        9.5321
Pearson Chi-Square      3       27.6840        9.2280
Scaled Pearson X2       3       27.6840        9.2280
Log Likelihood          .     2628.7264             .
```

The deviance is 28.6 on 3 *df* which is highly significant. The Pearson chi-square is again the same value as would be obtained from a standard chi-square test; it is also highly significant. Formally, independence is tested by comparing the deviance of this model with the deviance that would be obtained if the Season*Color interaction was included in the model. This saturated model has deviance 0.00 on 0 *df*. Thus, the deviance 28.6 is a large-sample test of independence between color and season. □

6.5 Higher-order tables

6.5.1 A three-way table

The arguments used above for the analysis of two-dimensional contingency tables can be generalized to tables of higher order. A general (saturated) model for a three-way table can be written as

$$\log\left(\mu_{ijk}\right) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} \quad (6.6)$$

An important part of the analysis is to decide which terms to include in the model.

Example 6.4 The table below contains data from a survey from Wright State University in 1992[1]. 2276 high school seniors were asked whether they had ever used Alcohol (A), Cigarettes (C) and/or Marijuana (M). This is a three-way contingency table of dimension $2 \times 2 \times 2$.

Alcohol use	Cigarette use	Marijuana use Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

□

[1] Data quoted from Agresti (1996) who credited the data to Professor Harry Khamis.

6.5.2 Types of independence

Models for data of the type given in the last example can include the main effects of A, C and M and different interactions containing these. The presence of an interaction, for example `A*C`, means that students who use alcohol have a higher (or lower) probability of also using cigarettes. One way of interpreting interactions is to calculate odds ratios; we will return to this topic soon.

A model of type `A C M A*C A*M` would permit interaction between A and C, and between A and M, but not between C and M. C and M are then said to be conditionally independent, controlling for A.

A model that only contains the main effects, i.e. the model `A C M` is called a mutual independence model. In this example this would mean that use of one drug does not change the risk of using any other drug.

A model that contains all interactions up to a certain level, but no higher-order interactions, is called a homogenous association model.

6.5.3 Genmod analysis of the drug use data

The saturated model that contains all main effects and all two- and threeway interactions was fitted to the data as a baseline. The three-way interaction `A*C*M` was not significant ($p = 0.53$). The output for the homogenous association model containing all two-way interactions was as follows:

```
        Criteria For Assessing Goodness Of Fit

Criterion              DF         Value      Value/DF

Deviance                1        0.3740        0.3740
Scaled Deviance         1        0.3740        0.3740
Pearson Chi-Square      1        0.4011        0.4011
Scaled Pearson X2       1        0.4011        0.4011
Log Likelihood          .    12010.6124             .
```

The fit of this model is good; a simple rule of thumb is that Value/df should not be too much larger than 1. The parameter estimates for this model are as follows:

Analysis Of Parameter Estimates

Parameter			DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT			1	6.8139	0.0331	42312.0532	0.0001
A	No		1	-5.5283	0.4522	149.4518	0.0001
A	Yes		0	0.0000	0.0000	.	.
C	No		1	-3.0158	0.1516	395.6463	0.0001
C	Yes		0	0.0000	0.0000	.	.
M	No		1	-0.5249	0.0543	93.4854	0.0001
M	Yes		0	0.0000	0.0000	.	.
A*C	No	No	1	2.0545	0.1741	139.3180	0.0001
A*C	No	Yes	0	0.0000	0.0000	.	.
A*C	Yes	No	0	0.0000	0.0000	.	.
A*C	Yes	Yes	0	0.0000	0.0000	.	.
A*M	No	No	1	2.9860	0.4647	41.2933	0.0001
A*M	No	Yes	0	0.0000	0.0000	.	.
A*M	Yes	No	0	0.0000	0.0000	.	.
A*M	Yes	Yes	0	0.0000	0.0000	.	.
C*M	No	No	1	2.8479	0.1638	302.1409	0.0001
C*M	No	Yes	0	0.0000	0.0000	.	.
C*M	Yes	No	0	0.0000	0.0000	.	.
C*M	Yes	Yes	0	0.0000	0.0000	.	.
SCALE			0	1.0000	0.0000	.	.

All remaining interactions in the model are highly significant which means that no further simplification of the model is suggested by the data.

6.5.4 Interpretation through Odds ratios

Consider, for the moment, a $2 \times 2 \times k$ cross-table of variables X, Y and Z. Within a fixed level j of Z, the conditional odds ratio for describing the relationship between X and Y is

$$\theta_{XY(j)} = \frac{\mu_{11j}\mu_{22j}}{\mu_{12j}\mu_{21j}} \tag{6.7}$$

where μ denotes expected values. In contrast, in the marginal odds ratio the value of the variable Z is ignored and we calculate the odds ratio as

$$\theta_{XY} = \frac{\mu_{11\cdot}\mu_{22\cdot}}{\mu_{12\cdot}\mu_{21\cdot}} \tag{6.8}$$

where the dot indicates summation over all levels of Z. The odds ratios can be estimated from the parameter estimates; it holds that, for example,

$$\widehat{\theta}_{XY} = \exp\left[\widehat{(\alpha\beta)}_{11} + \widehat{(\alpha\beta)}_{22} - \widehat{(\alpha\beta)}_{12} - \widehat{(\alpha\beta)}_{21}\right] \tag{6.9}$$

In our drug use example, the chosen model does not contain any three-way interaction, and only one parameter is estimable for each interaction. Thus, the partial odds ratios for the two-way interactions can be estimated as:

A*C: $\exp(2.0545) = 7.80$

A*M: $\exp(2.9860) = 19.81$

C*M: $\exp(2.8479) = 17.25$

As an example of an interpretation, a student who has tried alcohol has an odds of also having tried marijuana of 19.81, regardless of reported cigarette use.

6.6 Relation to logistic regression

6.6.1 Binary response

If one binary variable in a contingency table can be regarded as the response, an alternative to the log-linear model would be to model the probability of response as a function of the other variables in the table. This can be done using logistic regression methods as outlined in Chapter 5. As a comparison, we will analyze the data on page 111 as a logistic regression. A Genmod program for this analysis is:

```
DATA snoring;
INPUT x n snoring $;
CARDS;
51  467 Yes
59 2017  No
RUN;
PROC GENMOD DATA=snoring ORDER=data;
CLASS snoring;
MODEL x/n = snoring /
  DIST=Binomial LINK=logit;
RUN;
```

The corresponding output is

```
Analysis Of Parameter Estimates

                              Standard   Wald 95% Confidence      Chi-
Parameter       DF   Estimate     Error         Limits          Square   Pr > ChiSq

Intercept        1    -3.5021    0.1321    -3.7611    -3.2432    702.47       <.0001
snoring   Yes    1     1.4033    0.1987     1.0139     1.7927     49.89       <.0001
snoring   No     0     0.0000    0.0000     0.0000     0.0000       .          .
Scale            0     1.0000    0.0000     1.0000     1.0000
```

We note that the parameter estimate for snoring is 1.4033. This is the same as the estimate of the interaction parameter for the saturated log-linear model.

The odds ratio is

$$OR = \exp(1.4033) = 4.07$$

which is also the same as in the log-linear model. This suggests that contingency tables where one of the variables may be regarded as a binary response can be analyzed either as a log-linear model or using logistic regression. Note, however, that the models are written in different ways. The saturated log-linear model regards the counts as functions of the row and column variables and their interaction: `count = snoring heart snoring*heart`. The saturated logistic model regards the proportion of persons with heart problems as a function of snoring status: `x/n = snoring`. Although the models are written in different ways, the results and the interpretations are identical.

6.6.2 Nominal logistic regression

In log-linear models there is no variable that is regarded as the "dependent" variable. The treatment of the row and column variables is symmetric. In some cases it may be preferable to regard one nominal variable as the response. In such cases data can be modeled using nominal logistic regression. The idea is as follows:

One category of the nominal response variable is selected as a baseline, or reference, category. If category 1 is the baseline, the logits for the other categories, compared with the first, are

$$\text{logit}(p_j) = \log\left(\frac{p_j}{p_1}\right) = \mathbf{X}\boldsymbol{\beta}.$$

Thus we write $(j-1)$ logit equations, one for each category except for the baseline. These logit equations should be estimated simultaneously, which makes the problem multivatiate. At present, nominal logit models are not available in Proc Genmod, except for the case of ordinal response which is discussed in Chapter 7.

6.7 Capture-recapture data

Capture-recapture data provide an interesting application of log-linear models. Suppose that there are M individuals in a population; M is unknown and we want to estimate M. We capture and mark n_1 of the individuals. After some time we capture another n_2 individuals. It turns out that s of

these were marked. It is now relatively straightforward to estimate M as

$$\widehat{M} = \frac{n_1}{\hat{p}} = \frac{n_2 \cdot n_1}{s} \tag{6.10}$$

If the individuals are captured on three occasions, the data can be written as a three-way contingency table. There are eight different "capture patterns":

Notation	Captured at occasion
n_{123}	1, 2 and 3
$n_{\bar{1}23}$	2 and 3
$n_{1\bar{2}3}$	1 and 3
$n_{\bar{1}\bar{2}3}$	3
$n_{12\bar{3}}$	1 and 2
$n_{\bar{1}2\bar{3}}$	2
$n_{1\bar{2}\bar{3}}$	1
$n_{\bar{1}\bar{2}\bar{3}}$	None

If we assume independence between occasions, the probability that an individual is never captured is

$$\hat{p}_{\bar{1}\bar{2}\bar{3}} = (1 - \frac{n_1}{M})(1 - \frac{n_2}{M})(1 - \frac{n_3}{M}) \tag{6.11}$$

Thus, an estimator of the number of individuals that have never been captured is

$$\hat{n}_{\bar{1}\bar{2}\bar{3}} = M \cdot \hat{p}_{\bar{1}\bar{2}\bar{3}} = M \cdot (1 - \frac{n_1}{M})(1 - \frac{n_2}{M})(1 - \frac{n_3}{M}) \tag{6.12}$$

and an estimate of the unknown population size can be obtained by solving

$$M = n + M(1 - \frac{n_1}{M})(1 - \frac{n_2}{M})(1 - \frac{n_3}{M}) \tag{6.13}$$

for M, where n is the number of individuals that have been captured at least once.

There are some drawbacks with the method outlined so far. We have to assume independence between occasions, and we only use the information in the margins of the table. A more flexible analysis of this kind of data can be obtained by using log-linear models.

If the occasions are independent, it would hold that, for example, $p_{123} = p_1 p_2 p_3$. The expected number of individuals in this cell would then be

$$\mu_{123} = M p_1 p_2 p_3 \tag{6.14}$$

Taking logarithms,

$$\ln(\mu_{123}) = \ln M + \ln p_1 + \ln p_2 + \ln p_3 = \mu + \alpha_1 + \beta_1 + \gamma_1 \qquad (6.15)$$

where the occasions correspond to α, β and γ, respectively.

In a similar way, we could write the expected numbers in all cells as linear functions of parameters. This is a log-linear model. If the occasions are not independent, we can include parameters like $(\alpha\beta)_{ij}$ that account for the dependence. Thus, a general log-linear model can be written as

$$\ln(\mu_{ijk}) = \mu + \alpha_1 + \beta_1 + \gamma_1 + (\alpha\beta)_{11} + (\alpha\gamma)_{11} + (\beta\gamma)_{11} + (\alpha\beta\gamma)_{111} \qquad (6.16)$$

Log-linear models are now often used to model capture-recapture data; see Olsson (2000). The model specification includes a Poisson distribution for the numbers in each cell of the table, a log link and a feature to account for the fact that it is impossible to observe $n_{\bar{1}\bar{2}\bar{3}}$; the number of individuals that have never been captured.

Example 6.5 Table 6.1 summarizes information about persons who were heavy misusers of drugs in Sweden in 1979. The individuals could appear in registers within the health care system, social authorities, penal system, police or customs, or others. These correspond to the "captures", in a capture-recapture sense. It is reasonable to assume that some of these sources of information are related. Thus, for example, an individual who has been taken in charge by police is quite likely to appear also in the penal system at some stage. Thus, interactions between some or all of these sources are likely.

The SAS programs that were used for analysis had the structure

```
proc genmod;
  class x1 x2 x3 x4 x5;
  model count=x1 x2 x3 x4 x5 /
  dist=Poisson obstats residuals;
  weight w;
run;
```

In this program, x1 to x5 refer to the following sources of information:

Code	Source of information
x1	Health care
x2	Social authorities
x3	Penal system
x4	Police, customs
x5	Others

Table 6.1: *Swedish drug addicts with different capture patterns in 1979.*

Hospital care	Social authorities	Penal system	Police, customs	Others	Count
0	0	0	0	0	0
0	0	0	0	1	45
0	0	0	1	0	2080
0	0	0	1	1	11
0	0	1	0	0	1056
0	0	1	0	1	11
0	0	1	1	0	942
0	0	1	1	1	9
0	1	0	0	0	1011
0	1	0	0	1	59
0	1	0	1	0	381
0	1	0	1	1	15
0	1	1	0	0	245
0	1	1	0	1	18
0	1	1	1	0	345
0	1	1	1	1	13
1	0	0	0	0	828
1	0	0	0	1	7
1	0	0	1	0	179
1	0	0	1	1	1
1	0	1	0	0	191
1	0	1	0	1	3
1	0	1	1	0	137
1	0	1	1	1	1
1	1	0	0	0	264
1	1	0	0	1	18
1	1	0	1	0	132
1	1	0	1	1	9
1	1	1	0	0	144
1	1	1	0	1	16
1	1	1	1	0	133
1	1	1	1	1	15

The weight w has been set to 1 for all combinations except the combination where all x1,...,x5 are 0. This combination cannot be observed and is a structural zero.

In the program example it is implicitly assumed that the different sources of information are independent. However, interactions can be included in the model as e.g. x1*x2. In this rather large data set all two-way interactions were significant. In addition, the interaction x1*x3*x4 was also significant. Thus, the final model was

```
count =x1|x2|x3|x4|x5 @2 x1*x3*x4;
```

This model had a good fit to the data ($\chi^2 = 17.5$ on 14 df). The model estimates the number of uncaptured individuals as 8878 individuals with confidence interval $7640 - 10317$ individuals. This would mean that the number of drug addicts in Sweden in 1979 was $8319 + 8878 = 17197$ individuals. This is more than 5000 individuals higher than the published result, which was 12000 (Socialdepartementet 1980). The published result was obtained through capture-recapture methods but assuming that the different sources of information are independent. □

6.8 Poisson regression models

We have seen that log-linear models for cross-tabulations can be handled as generalized linear models. The linear predictor then consists of a design matrix that contains dummy variables for the different margins of the table. It is quite possible to introduce quantitative variables into the model, in a similar way as for regression models.

Example 6.6 The table below, taken from Haberman (1978), shows the distribution of stressful events reported by 147 subjects who have experienced exactly one stressful event. The table gives the number of persons reporting a stressful event 1, 2, ..., 18 months prior to the interview. We want to model the occurrence of stressful events as a function of time.

Months	1	2	3	4	5	6	7	8	9
Number	15	11	14	17	5	11	10	4	8
Months	10	11	12	13	14	15	16	17	18
Number	10	7	9	11	3	6	1	1	4

One approach to modelling the occurrence of stressful events as a function of X=months is to assume that the number of persons responding for any month is a Poisson variate. The canonical link for the Poisson distribution

is log, so a first attempt to modelling these data is to assume that

$$\log(\mu) = \beta_0 + \beta_1 x \tag{6.17}$$

This is a generalized linear model with a Poisson distribution, a log link and a simple linear predictor. A SAS program for this model can be written as

```
DATA stress;
INPUT months number @@;
CARDS;
 1 15  2 11  3 14  4 17  5  5  6 11  7 10  8  4  9  8 10 10
11  7 12  9 13 11 14  3 15  6 16  1 17  1 18  4
;
PROC GENMOD DATA=stress;
MODEL number = months / DIST=poisson LINK=log OBSTATS RESIDUALS;
MAKE 'obstats' out=ut;
RUN;
```

Part of the output is

```
Criteria For Assessing Goodness Of Fit

 Criterion              DF         Value      Value/DF

 Deviance               16       24.5704        1.5356
 Scaled Deviance        16       24.5704        1.5356
 Pearson Chi-Square     16       22.7145        1.4197
 Scaled Pearson X2      16       22.7145        1.4197
 Log Likelihood          .      174.8451             .
```

The data have a less than perfect fit to the model, with Value/df=1.53; the p-value is 0.078.

```
                Analysis Of Parameter Estimates

Parameter    DF    Estimate     Std Err   ChiSquare  Pr>Chi

INTERCEPT     1      2.8032      0.1482    357.9476  0.0001
MONTHS        1     -0.0838      0.0168     24.8639  0.0001
SCALE         0      1.0000      0.0000           .       .
```

We find that the memory of stressful events fades away as $\log(\mu) = 2.80 - 0.084x$. A plot of the data, along with the fitted regression line, is given as Figure 6.1. Figure 6.2 shows the data and regression line with a log scale for the y-axis. □

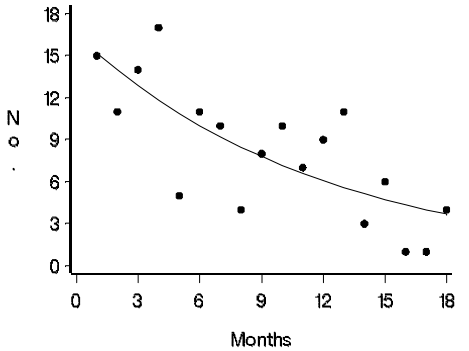


Figure 6.1: *Distribution of persons remembering stressful events*

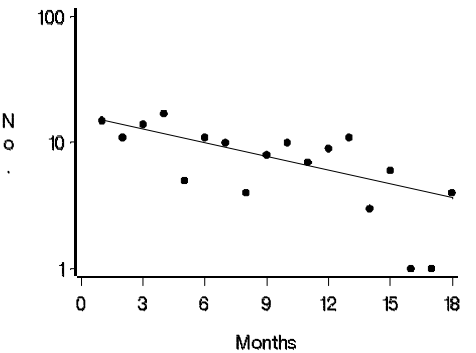


Figure 6.2: *Distribution of persons remembering stressful events; log scale*

6.9 A designed experiment with a Poisson distribution

Example 6.7 The number of wireworms counted in the plots of a Latin square experiment following soil fumigations in the previous year is given in the following table[2].

	1	2	3	4	5
1	P 3	O 2	N 5	K 1	M 4
2	M 6	K 0	O 6	N 4	P 4
3	O 4	M 9	K 1	P 6	N 5
4	N 17	P 8	M 8	O 9	K 0
5	K 4	N 4	P 2	M 4	O 8

We may model the number of wireworms in a certain plot as a Poisson distribution. The design includes a Row effect, a Column effect and a Treatment effect. Thus, an "ANOVA-like" model for these data can be written as

$$g(\mu) = \beta_0 + \alpha_i + \beta_j + \tau_k \tag{6.18}$$

where β_0 is a general mean, α_i is a row effect, β_j is a column effect and τ_k is the effect of treatment k.

A SAS program for analysis of these data using Proc Genmod is:

```
DATA Poisson;
INPUT Row Col Treat $ Count;
CARDS;
1 1 P 3
1 2 O 2
1 3 N 5
... More data lines ...
5 4 M 4
5 5 O 8
;
PROC GENMOD DATA=Poisson;
CLASS row col treat;
MODEL Count = row col treat /
      Dist=Poisson
      Link=Log
      Type3;
RUN;
```

[2]Data from Snedecor and Cochran (1980). The original analysis was an Anova on data transformed as $y = \sqrt{x+1}$

The output is:

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.POISSON
Distribution	POISSON
Link Function	LOG
Dependent Variable	COUNT
Observations Used	25

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	12	19.5080	1.6257
Scaled Deviance	12	19.5080	1.6257
Pearson Chi-Square	12	18.0096	1.5008
Scaled Pearson X2	12	18.0096	1.5008
Log Likelihood	.	97.0980	.

The fit of the model is reasonable but not perfect; the p value is 0.077. Ideally, Value/df should be closer to 1.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	1.4708	0.3519	17.4670	0.0001
ROW	1	1	-0.4419	0.3404	1.6851	0.1942
ROW	2	1	-0.1751	0.3175	0.3041	0.5813
ROW	3	1	0.0451	0.2980	0.0229	0.8796
ROW	4	1	0.5699	0.2729	4.3618	0.0368
ROW	5	0	0.0000	0.0000	.	.
COL	1	1	0.3045	0.2892	1.1087	0.2924
COL	2	1	-0.0506	0.3099	0.0267	0.8703
COL	3	1	-0.0936	0.3207	0.0852	0.7704
COL	4	1	-0.0636	0.3093	0.0423	0.8370
COL	5	0	0.0000	0.0000	.	.
TREAT	K	1	-1.3797	0.4627	8.8906	0.0029
TREAT	M	1	0.2910	0.2789	1.0888	0.2967
TREAT	N	1	0.3324	0.2760	1.4502	0.2285
TREAT	O	1	0.2003	0.2854	0.4928	0.4827
TREAT	P	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

```
 LR Statistics For Type 3 Analysis

Source       DF   ChiSquare  Pr>Chi

ROW           4     14.3595  0.0062
COL           4      2.8225  0.5880
TREAT         4     25.1934  0.0001
```

We find a significant Row effect and a highly significant Treatment effect. It is interesting to note that the GLM analysis of square root transformed data, as suggested by Snedecor and Cochran (1980, results in a significant treatment effect ($p = 0.02$) but no significant row or column effects. This may be related to the fact that the model fit is not perfect. We will return to these data later. □

6.10 Rate data

Events that may be assumed to be essentially Poisson are sometimes recorded on units of different size. For example, the number of crimes recorded in a number of cities depends on the size of the city, such that "crimes per 1000 inhabitants" is a meaningful measure of crime rate. Data of this type are called *rate data*.

If we denote the measure of size with t, we can model this type of data as

$$\log\left(\frac{\mu}{t}\right) = \mathbf{X}\boldsymbol{\beta} \tag{6.19}$$

which means that

$$\log(\mu) = \log(t) + \mathbf{X}\boldsymbol{\beta} \tag{6.20}$$

The adjustment term $\log(t)$ is called an *offset*. The offset can easily be included in models analyzed with e.g. Proc Genmod.

Example 6.8 The data below, quoted from Agresti (1996), are accident rates for elderly drivers, subdivided by sex. For each sex the number of person years (in thousands) is also given. The data refer to 16262 Medicaid enrollees.

	Females	Males
No. of accidents	175	320
No. of person years ('000)	17.30	21.40

From the raw data we can calculate accident rates as $175/17.30 = 10.1$ per 1000 person years for females and $320/21.40 = 15.0$ per 1000 person years for

males. A simple way to model these data is to use a generalized linear model with a Poisson distribution, a log link, and to use the number of person years as an offset. This is done with the following program:

```
DATA accident;
  INPUT sex $ accident persyear;
  logpy=log(persyear);
CARDS;
Male    320 21.400
Female  175 17.300
;
PROC GENMOD DATA=accident;
CLASS sex;
MODEL accident = sex /
                 LINK   = log
                 DIST   = poisson
                 OFFSET = logpy
;
RUN;
```

The output is

Criterion	DF	Value	Value/DF
Deviance	0	0.0000	.
Scaled Deviance	0	0.0000	.
Pearson Chi-Square	0	0.0000	.
Scaled Pearson X2	0	0.0000	.
Log Likelihood	.	2254.7003	.

The model is a saturated model so we can't assess the over-all fit of the model by using the deviance.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	2.7049	0.0559	2341.3269	0.0001
SEX	Female	1	-0.3909	0.0940	17.2824	0.0001
SEX	Male	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

The parameter estimate for females is -0.39. The model can be written as

$$\log(\mu) = \log(t) + \beta_0 + \beta_1 x$$

where x is a dummy variable taking the value 1 for females and 0 for males. Thus the estimate can be interpreted such that the odds ratio is $e^{-0.3909} = 0.676$. The risk of having an accident for a female is 68% of the risk for men. This difference is significant; however, other factors that may affect the risk of accident, for example differences in driving distance, are not included in this model. □

6.11 Overdispersion in Poisson models

Overdispersion means that the variance of the response variable is larger than would be expected for the chosen distribution. For Poisson data we would expect the variance to be equal to the mean.

As noted earlier, the presence of overdispersion may be related to mistakes in the formulation of the generalized linear model: the distribution, the link function and/or the linear predictor. The effects of overdispersion is that p-values for tests are deflated: it becomes "too easy" to get significant results.

6.11.1 Modeling the scale parameter

If the model is correct, overdispersion may be caused by heterogeneity among the observations. One way to account for such heterogeneity is to introduce a scale parameter ϕ into the variance function. Thus, we would assume that $Var(Y) = \phi\sigma^2$, where the parameter ϕ can be estimated as (Deviance) $/df$ or as χ^2/df, where χ^2 is the Pearson Chi-square.

Example 6.9 In the analysis of data on number of wireworms in a Latin square experiment (page 129), there were some indications of overdispersion. The ratio Deviance/df was 1.63. We re-analyze these data but ask the Genmod procedure to use 1.63 as an estimate of the dispersion parameter ϕ. The program is:

```
PROC GENMOD data=poisson;
CLASS row col treat;
MODEL count = row col treat /
              dist=Poisson Link=log
              type3
              scale=1.63 ;
RUN;
```

Output:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	12	19.5080	1.6257
Scaled Deviance	12	7.3424	0.6119
Pearson Chi-Square	12	18.0096	1.5008
Scaled Pearson X2	12	6.7784	0.5649
Log Likelihood	.	36.5456	.

```
LR Statistics For Type 3 Analysis

Source       DF   ChiSquare  Pr>Chi

ROW           4      5.4046  0.2482
COL           4      1.0623  0.9002
TREAT         4      9.4823  0.0501
```

Fixing the scale parameter to 1.63 has a rather dramatic effect on the result. In our previous analysis of these data, the treatment effect was highly significant ($p = 0.0001$), and the row effect was significant ($p = 0.0062$). In our new analysis even the treatment effect is above the 0.05 limit. In the original analysis of these data (Snedecor and Cochran, 1980), only the treatment effect was significant ($p = 0.021$). Note that the Genmod procedure has an automatic feature to base the analysis on a scale parameter estimated by the Maximum Likelihood method; see the SAS manual for details. □

6.11.2 Modeling as a Negative binomial distribution

If a Poisson model shows signs of overdispersion, an alternative approach is to replace the Poisson distribution with a Negative binomial distribution. This idea can be traced back to Student (1907), who studied counts of red blood cells. This distribution can be derived in two ways.

For a series of Bernouilli trials, suppose that we are studying the number of trials (y) until we have recorded r successes. The probability of success is p. The distribution for y is

$$\begin{aligned} P(y) &= \binom{y-1}{r-1} p^r (1-p)^{y-r} \\ \text{for } y &= r,\ (r+1), ... \end{aligned} \tag{6.21}$$

This is the binomial waiting time distribution. If $r = 1$, it is called a geometric distribution. Using the Gamma function, the distribution can be defined even for non-integer values of r. When r is an integer, it is called the Pascal distribution. The distribution has mean value $E(y) = \frac{r}{p}$ and variance $Var(y) = \frac{r(1-p)}{p^2}$.

A second way to derive the negative binomial distribution is as a so called compund distribution. Suppose that the response for individual i can be modeled as a Poisson distribution with mean value μ_i. Suppose further that the distribution of the mean values μ_i over individuals follows a Gamma distribution. It can be shown that the resulting compound distribution for y is a negative binomial distribution.

The negative binomial distribution has a higher probability for the zero count, and a longer tail, than a Poisson distribution with the same mean value.

Because of this, and because of the relation to compound distributions, it is often used as an alternative to the Poisson distribution when over-dispersion can be suspected. The negative binomial distribution is available in Proc Genmod in SAS, version 8.

6.12 Diagnostics

Model diagnostics for Poisson models follows the same lines as for other generalized linear models.

Example 6.10 We can illustrate some diagnostic plots using data from the Wireworm example on page 129. The residuals and predicted values were stored in a file. The standardized deviance residuals were plotted against the predicted values, and a normal probability plot of the residuals was prepared. The results are given in Figure 6.3 and Figure 6.4. Both plots indicate a reasonable behavior of the residuals. We cannot see any irregularities in the plot of residuals against fitted values, and the normal plot is rather linear. □

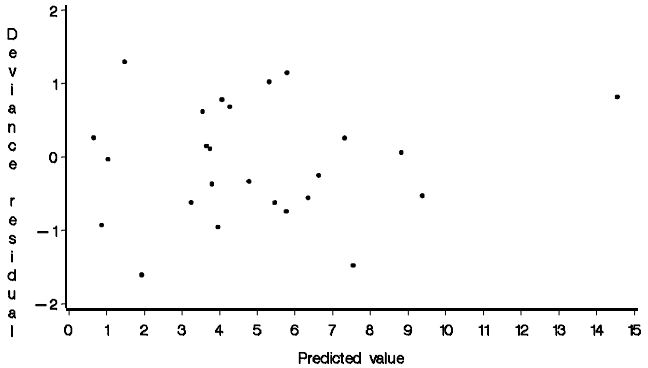


Figure 6.3: *Plot of standardized deviance residuals against fitted values for the Wireworm example.*

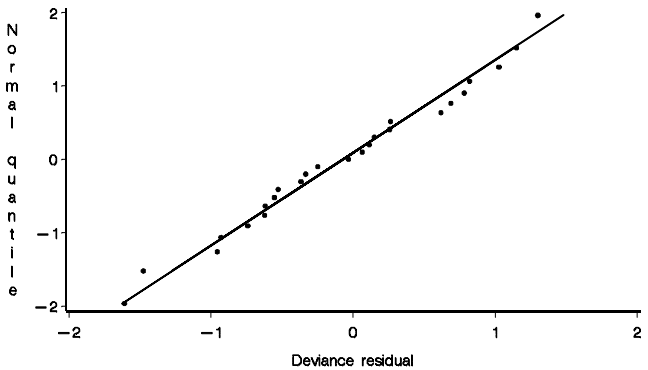


Figure 6.4: *Normal probability plot for the wireworm data.*

6.13 Exercises

Exercise 6.1 The data in Exercise 1.4 are of a kind that can often be approximated by a Poisson distribution. Re-analyze the data using Poisson regression. Prepare a graph of the relation and compare the results with the results from Exercise 1.4. The data are repeated here for your convenience:

Gowen and Price counted the number of lesions of Aucuba mosaic virus after exposure to X-rays for various times. The results were:

Minutes exposure	Count
0	271
15	108
30	59
45	29
60	12

Exercise 6.2 The following data consist of failures of pieces of electronic equipment operating in two modes. For each observation, Mode1 is the time spent in one mode and Mode2 is the time spent in the other. The total number of failures recorded in each period is also recorded.

Mode1	Mode2	Failures
33.3	25.3	15
52.2	14.4	9
64.7	32.5	14
137.0	20.5	24
125.9	97.6	27
116.3	53.6	27
131.7	56.6	23
85	87.3	18
91.9	47.8	22

Fit a Poisson regression model to these data, using Failures as a dependent variable and Mode1 and Mode2 as predictors. In the original analysis (Jørgensen 1961) an Identity link was used. Try this, but also try a log link. Which model seems to fit best?

Exercise 6.3 The following data was taken from the Statlib database (Internet address http://lib.stat.edu/datasets). They have also been used by McCullagh and Nelder (1989). The source of the data is the Lloyd Register of Shipping. The purpose of the analysis is to find variables that are related to the number of damage incidents for ships. The following variables

are available:

type	Ship type A, B, C, D or E
yr_constr	Year of construction in 5-year intervals
per_op	Period of operation: 1960-74, 1975-79
mon_serv	Aggregate months service for ships in this cell
incident	Number of damage incidents

Type	yr_constr	per_op	mon_serv	incident
A	60	60	127	0
A	60	75	63	0
A	65	60	1095	3
A	65	75	1095	4
A	70	60	1512	6
A	70	75	3353	18
A	75	60	0	*
A	75	75	2244	11
B	60	60	44882	39
B	60	75	17176	29
B	65	60	28609	58
B	65	75	20370	53
B	70	60	7064	12
B	70	75	13099	44
B	75	60	0	*
B	75	75	7117	18
C	60	60	1179	1
C	60	75	552	1
C	65	60	781	0
C	65	75	676	1
C	70	60	783	6
C	70	75	1948	2
C	75	60	0	*
C	75	75	274	1
D	60	60	251	0
D	60	75	105	0
D	65	60	288	0
D	65	75	192	0
D	70	60	349	2
D	70	75	1208	11
D	75	60	0	*
D	75	75	2051	4
E	60	60	45	0
E	60	75	0	*
E	65	60	789	7
E	65	75	437	7
E	70	60	1157	5
E	70	75	2161	12
E	75	60	0	*
E	75	75	542	1

The number of damage incidents is reported as * if it is a "structural zero". Fit a model predicting the number of damage incidents based on the other variables. Use the following instructions:

- Use a Poisson model with a log link.
- Use log (Aggregate months service) as an offset.
- Use all predictors as class variables. Include any necessary interactions.

- If necessary, try to model any overdispersion in the data.

Note: some of the observations are "structural zeros". For example, a ship constructed in 1975-79 cannot operate during the period 1960-74.

Exercise 6.4 The data in table 6.8 (page 131) are rather simple. It is quite possible to calculate the parameter estimates by hand. Make that calculation.

Exercise 6.5 An experiment analyzes imperfection rates for two processes used to fabricate silicon wafers for computer chips. For treatment A applied to 10 wafers the number of imperfections are 8, 7, 6, 6, 3, 4, 7, 2, 3, 4. Treatment B applied to 10 wafers has 9, 9, 8, 14, 8, 13, 11, 5, 7, 6 imperfections. The counts were treated as independent Poisson variates in a generalized linear model with a log link. Parts of the results were:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	18	16.2676	0.9038
Scaled Deviance	18	16.2676	0.9038
Pearson Chi-Square	18	16.0444	0.8914
Scaled Pearson X2	18	16.0444	0.8914
Log Likelihood		138.2221	

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	1.6094	0.1414			
treat	1	0.5878	0.1764			
Scale	0	1.0000	0.0000			

A model with only an intercept term gave a deviance of 27.857 on 19 degrees of freedom.

A. Test the hypothesis $H_0 : \mu_A = \mu_B$ using (i) a Likelihood ratio test; (ii) a Wald test.

B. Construct a 95% confidence interval for μ_B/μ_A. Hint: What is the relationship between the parameter β and μ_B/μ_A?

Exercise 6.6 The following table (from Agresti 1996) gives the number of train miles (in millions) and the number of collisions involving British Rail passenger trains between 1970 and 1984. Is it plausible that the collision counts are independent Poisson variates? Respond by testing a model with only an intercept term. Also, examine whether inclusion of log(miles) as an offset would improve the fit.

Year	Collisions	Miles	Year	Collisions	Miles
1970	3	281	1977	4	264
1971	6	276	1978	1	267
1972	4	268	1979	7	265
1973	7	269	1980	3	267
1974	6	281	1981	5	260
1975	2	271	1982	6	231
1976	2	265	1983	1	249

Exercise 6.7 Rosenberg et al (1988) studied the relationship between coffee drinking, smoking and the risk for myocardial infarction in a case-control study for men under 55 years of age. The data are as follows:

	Cigarettes per day							
Coffee	0		1-24		25-34		35-	
per day	Case	Control	Case	Control	Case	Control	Case	Control
0	66	123	30	52	15	12	36	13
1-2	141	179	59	45	53	22	69	25
3-4	113	106	63	65	55	16	119	30
5-	129	80	102	58	118	44	373	85

A. Analyze these data using smoking and coffee drinking as qualitative variables.

B. Assign scores to smoking and coffee drinking and re-analyze the data using these scores as quantitative variables.

C. Compare the analyses in A. and B. in terms of fit. Perform residual analyses.

Exercise 6.8 Even before the space shuttle Challenger exploded on January 20, 1986, NASA had collected data from 23 earlier launches. One part of these data was the number of O-rings that had been damaged at each launch. O-rings are a kind of gaskets that will prevent hot gas from leaking during takeoff. In total there were six such O-rings at the Challenger. The data included the number of damaged O-rings, and the temperature (in Fahrenheit) at the time of the launch. On the fateful day when the Challenger exploded, the temperature was 31°F.

One might ask whether the probability that an O-ring is damaged is related to the temperature. The following data are available:

No. of Defective O-rings	Temperature ^{o}F	No. of Defective O-rings	Temperature ^{o}F
2	53	0	70
1	57	1	70
1	58	1	70
1	63	0	72
0	66	0	73
0	67	0	75
0	67	2	75
0	67	0	76
0	68	0	76
0	69	0	78
0	70	0	79
		0	81

A statistician fitted two alternative Generalized linear models to these data: one model with a Poisson distribution and a log link, and another model with a binomial distribution and a logit link. Part of the output from these two analyses are presented below. Deviances for "null models" that only include an intercept were 22.434 (22 *df*, Poisson model) and 24.2304 (22 *df*, binomial model).

Poisson model:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	21	16.8337	0.8016
Scaled Deviance	21	16.8337	0.8016
Pearson Chi-Square	21	28.1745	1.3416
Scaled Pearson X2	21	28.1745	1.3416
Log Likelihood		-14.6442	

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	5.9691	2.7628			
Temp	1	-0.1034	0.0430			
Scale	0	1.0000	0.0000			

Binomial model: Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	21	18.0863	0.8613
Scaled Deviance	21	18.0863	0.8613
Pearson Chi-Square	21	29.9802	1.4276
Scaled Pearson X2	21	29.9802	1.4276
Log Likelihood		-30.1982	

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	5.0850	3.0525				
Temp	1	-0.1156	0.0470				
Scale	0	1.0000	0.0000				

A. Test whether temperature has any significant effect on the failure of O-rings using

i) the Poisson model

ii) the binomial model

B. Predict the outcome of the response variable if the temperature is 31°F

i) for the Poisson model

ii) for the binomial model

C. Which of the two models do you prefer? Explain why!

D. Using your preferred model, calculate the probability that three or more of the O-rings fail if the temperature is 31°F.

Exercise 6.9 Agresti (1996) discusses analysis of a set of accident data from Maine. Passengers in all traffic accidents during 1991 were classified by:

Gender	Gender of the person (F or M)
Location	Place of the accident: Urban or Rural
Belt	Whether the person used seat belt (Y or N)
Injury	Whether the person was injured in the accident (Y or N)

A total of 68694 passengers were included in the data.

A log-linear model was fitted to these data using Proc Genmod. All main effects and two-way interactions were included. A model with a three-way interaction fitted slightly better but is not discussed here. Part of the results were:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	5	23.3510	4.6702
Scaled Deviance	5	23.3510	4.6702
Pearson Chi-Square	5	23.3752	4.6750
Scaled Pearson X2	5	23.3752	4.6750
Log Likelihood		536762.6081	

Analysis Of Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% Limits		Chi-Square	Pr > ChiSq
Intercept			1	5.9599	0.0314	5.8984	6.0213	36133.0	<.0001
gender	F		1	0.6212	0.0288	0.5647	0.6777	463.89	<.0001
gender	M		0	0.0000	0.0000	0.0000	0.0000	.	.
location	R		1	0.2906	0.0290	0.2337	0.3475	100.16	<.0001
location	U		0	0.0000	0.0000	0.0000	0.0000	.	.
belt	N		1	0.7796	0.0291	0.7225	0.8367	716.17	<.0001
belt	Y		0	0.0000	0.0000	0.0000	0.0000	.	.
injury	N		1	3.3309	0.0310	3.2702	3.3916	11563.8	<.0001
injury	Y		0	0.0000	0.0000	0.0000	0.0000	.	.
gender*location	F	R	1	-0.2099	0.0161	-0.2415	-0.1783	169.50	<.0001
gender*location	F	U	0	0.0000	0.0000	0.0000	0.0000	.	.
gender*location	M	R	0	0.0000	0.0000	0.0000	0.0000	.	.
gender*location	M	U	0	0.0000	0.0000	0.0000	0.0000	.	.
gender*belt	F	N	1	-0.4599	0.0157	-0.4907	-0.4292	860.14	<.0001
gender*belt	F	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
gender*belt	M	N	0	0.0000	0.0000	0.0000	0.0000	.	.
gender*belt	M	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
gender*injury	F	N	1	-0.5405	0.0272	-0.5939	-0.4872	394.36	<.0001
gender*injury	F	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
gender*injury	M	N	0	0.0000	0.0000	0.0000	0.0000	.	.
gender*injury	M	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
location*belt	R	N	1	-0.0849	0.0162	-0.1167	-0.0532	27.50	<.0001
location*belt	R	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
location*belt	U	N	0	0.0000	0.0000	0.0000	0.0000	.	.
location*belt	U	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
location*injury	R	N	1	-0.7550	0.0269	-0.8078	-0.7022	784.94	<.0001
location*injury	R	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
location*injury	U	N	0	0.0000	0.0000	0.0000	0.0000	.	.
location*injury	U	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
belt*injury	N	N	1	-0.8140	0.0276	-0.8681	-0.7599	868.65	<.0001
belt*injury	N	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
belt*injury	Y	N	0	0.0000	0.0000	0.0000	0.0000	.	.
belt*injury	Y	Y	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale			0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Calculate and interpret estimated odds ratios for the different factors.

7. Ordinal response

Response variables in the form of judgements or other ordered classifications are called ordinal response variables. Examples of such variables are diagnostics of patients (improved, no change, worse); classification of potatoes (ordinary, high quality, extra high quality); answers to opinion items (agree completely; agree; undecided; disagree; disagree completely); and school marks.

Ordinal response variables can be analyzed as nominal response using the methods outlined in Chapter 6. However, this kind of analysis would disregard an important part of the information in the data, namely the fact that the categories are ordered. Alternatively, ordinal data are sometimes analyzed as if the data had been numeric, using some scoring of the response. This approach is often unsatisfactory since the data are then assumed to be "better" than they actually are. Several suggestions on the modeling of ordinal response data have been put forward in the literature. We will briefly review some of these approaches from the point of view of generalized linear models.

7.1 Arbitrary scoring

Example 7.1 Norton and Dunn (1985) studied the relation between snoring and heart problems for a sample of 2484 patients. The data were obtained through interviews with the patients. The amount of snoring was assessed on a scale ranging from "Never" to "Always", which is an ordinal variable. An interesting question is whether there is any relation between snoring and heart problems. The data are given in the following table:

			Snoring		
Heart problems	Never	Some-times	Often	Always	Total
Yes	24	35	21	30	110
No	1355	603	192	224	2374
Total	1379	638	213	254	2484

□

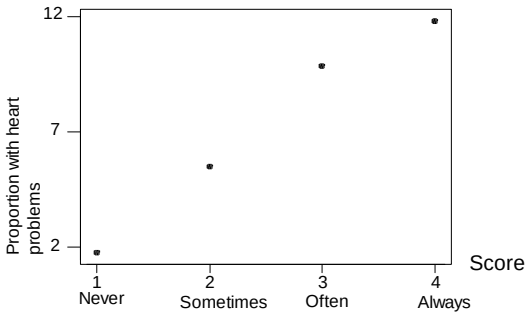


Figure 7.1: *Relation between heart problems and snoring*

The main interest lies in studying a possible dependence between snoring and heart problems. A simple approach to analyzing these data is to ignore the ordinal nature of the data and use a simple χ^2 test of independence or, in this context, the corresponding log-linear model

$$\log\left(\mu_{ij}\right) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \tag{7.1}$$

This is a saturated model. The test of the hypothesis of independence corresponds to testing the hypothesis that the parameter $(\alpha\beta)_{ij}$ is zero. For the data on page 145, this gives a deviance of 21.97 on 3 *df* ($p < 0.0001$) which is, of course, highly significant. This analysis, however, does not use the fact that the snoring variable is ordinal.

A plot (Figure 7.1) of the percentage of persons with heart problems in each snoring category suggests that this percentage increases nearly linearly, if we choose the arbitrary scores (1, 2, 3, 4) for the snoring categories.

This suggests a simple way of accounting for the ordinal nature of the data. Instead of entering the dependence between the variables as the interaction term $(\alpha\beta)_{ij}$, as in the saturated model (7.1), we write the model as

$$\log\left(\mu_{ij}\right) = \mu + \alpha_i + \beta_j + \gamma \cdot u_i \cdot v_j \tag{7.2}$$

where u_i are arbitrary scores for the row variable and v_j are scores for the column variable. In this model, the term $\gamma \cdot u_i \cdot v_j$ captures the linear part of the dependence between the scores. This model is called a linear by linear association model (LL model; see Agresti, 1996).

In a Genmod analysis according to this model we need to arrange the data according to the following data step:

```
DATA snoring;
INPUT heart $ snore $ freq u v;
CARDS;
Yes Never        24 1  1
Yes Sometimes    35 1  2
Yes Often        21 1  3
Yes Always       30 1  4
No  Never      1355 0  1
No  Sometimes   603 0  2
No  Often       192 0  3
No  Always      224 0  4
;
```

The model request in Genmod can be written as

```
PROC GENMOD DATA=snoring;
  CLASS heart snore;
  MODEL freq = heart snore u*v/
             DIST=poisson
             LINK=log ;
RUN;
```

Parts of the output is

```
                Criteria For Assessing Goodness Of Fit

          Criterion             DF         Value      Value/DF

          Deviance               2        6.2398        3.1199
          Scaled Deviance        2        6.2398        3.1199
          Pearson Chi-Square     2        6.3640        3.1820
          Scaled Pearson X2      2        6.3640        3.1820
          Log Likelihood         .    13733.2247             .


                    Analysis Of Parameter Estimates

   Parameter              DF    Estimate     Std Err   ChiSquare  Pr>Chi

   INTERCEPT               1      1.9833      0.2258     77.1306  0.0001
   HEART      No           1      4.4319      0.2256    386.0188  0.0001
   HEART      Yes          0      0.0000      0.0000           .       .
   SNORE      Always       1     -1.0289      0.0773    177.1304  0.0001
   SNORE      Never        1      0.7912      0.0479    272.4822  0.0001
   SNORE      Often        1     -1.1353      0.0794    204.5545  0.0001
   SNORE      Sometime     0      0.0000      0.0000           .       .
   U*V                     1      0.6545      0.0825     62.8977  0.0001
   SCALE                   0      1.0000      0.0000           .       .

NOTE:  The scale parameter was held fixed.
```

Earlier we found that the independence model gives a deviance of 21.97 on 3 *df*. If we include the single parameter γ for the linear by linear association

model we get a model with 2 *df* and a deviance of 6.24. The difference is 15.73 on 1 *df* which is highly significant. The Wald test of the parameter for the linear by linear association is also highly significant ($\chi^2 = 62.9$ on 1 *df*; $p < 0.0001$). This indicates that most of the dependence between snoring and heart problems is captured by the linear interaction term. The original analysis using a simple χ^2 test indicated "some form of relationship" between snoring and heart problems. The linear by linear association model suggests that snoring and heart problems may have a positive relationship.

7.2 RC models

The method of arbitrary scoring is often useful, but it is subjective in the sense that different choices of scores for the ordinal variables may result in different conclusions. An approach that has been suggested (see e.g. Andersen, 1980) is to include the row and column scores as parameters of the model. Thus, the model can be written as

$$\log\left(\mu_{ij}\right) = \mu + \alpha_i + \beta_j + \gamma \cdot \mu_i \cdot v_j \tag{7.3}$$

where μ_i and v_j are now parameters to be estimated from the data. This model, called an RC model, is nonlinear since it includes a product term in the row and column scores. Thus, the model is not formally a generalized linear model. However, Agresti (1985) suggested methods for fitting this model using standard software. This method is iterative. The row scores are kept fixed and the model is fitted for the column scores. These column scores are then kept fixed and the row scores are estimated. These two steps are continued until convergence. The method seems to converge in most cases.

7.3 Proportional odds

The proportional odds model for an ordinal response variable is a model for cumulative probabilities of type $P\left(Y \leq j\right) = p_1 + p_2 + \ldots + p_j$, where for simplicity we index the categories of the response variable with integers. The cumulative logits are defined as

$$\text{logit}\left(P\left(Y \leq j\right)\right) = \log \frac{P\left(Y \leq j\right)}{1 - P\left(Y \leq j\right)} \tag{7.4}$$

The cumulative logits are defined for each of the categories of the response except the first one. Thus, for a response variable with 5 categories we would get $5 - 1 = 4$ different cumulative logits.

The proportional odds model for ordinal response suggests that all these cumulative logit functions can be modeled as

$$\text{logit}\left(P\left(Y \leq j\right)\right) = \alpha_j + \beta x \tag{7.5}$$

i.e. the functions have different intercepts α_i but a common slope β. This means that the odds ratio, for two different values x_1 and x_2 of the predictor x, has the form

$$\frac{P\left(Y \leq j|x_2\right)/P\left(Y > j|x_2\right)}{P\left(Y \leq j|x_1\right)/P\left(Y > j|x_1\right)}. \tag{7.6}$$

The log of this odds ratio equals $\beta\left(x_2 - x_1\right)$, i.e. the log odds is proportional to the difference between x_2 and x_1. This is why the model is called the proportional odds model.

The proportional odds model is not formally a (univariate) generalized linear model, although it can be seen as a kind of multivariate Glim. The model states that the different cumulative logits, for the different ordinal values of the response, are all parallel but with different intercepts. Thus, the model gives, in a sense, a set of $k-1$ related models if the response has k scale steps. The Genmod procedure in SAS version 8 (SAS 2000b), as well as the Logistic procedure in SAS, can handle this type of models.

Example 7.2 We continue with the analysis of the data on page 145. For the sake of illustration we use the ordinal snoring variable as the response, and analyze the data to explore whether the risk of snoring depends on whether the patient has heart problems. A simple way of analyzing this type of data is to use the Logistic procedure of the SAS package:

```
PROC LOGISTIC DATA=snoring;
  FREQ freq;
  MODEL v = u;
RUN;
```

The following output is obtained:

```
             Score Test for the Proportional Odds Assumption

                Chi-Square = 1.1127 with 2 DF (p=0.5733)


   Model Fitting Information and Testing Global Null Hypothesis BETA=0

                          Intercept
                Intercept     and
Criterion         Only    Covariates    Chi-Square for Covariates

AIC             5568.351    5505.632          .
SC              5585.804    5528.903          .
-2 LOG L        5562.351    5497.632      64.719 with 1 DF (p=0.0001)
Score                .           .      68.217 with 1 DF (p=0.0001)
```

The proportional odds assumption (i.e. the assumption of a common slope) cannot be rejected ($p = 0.57$). The hypothesis that β is zero is rejected ($p < 0.0001$). The estimates of the parameters α_1, α_2 and α_3 are given in the next part of the output, along with an estimate of the common slope β. The intercept for the last category is set equal to zero.

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCP1	1	0.2824	0.0414	46.5871	0.0001	.	.
INTERCP2	1	1.5545	0.0534	846.5227	0.0001	.	.
INTERCP3	1	2.2792	0.0687	1101.5470	0.0001	.	.
U	1	-1.4209	0.1774	64.1619	0.0001	-0.161188	0.242

A similar analysis can be done using Proc Genmod in SAS version 8 or later. The program can be written as

```
PROC GENMOD data=snoring order=data;
FREQ freq;
CLASS heart;
MODEL v = u
  /dist=multinomial
   link=cumlogit;
RUN;
```

The information is essentially the same as in the Logistic procedure but the standard error estimates are slightly different. Also, Proc Genmod does not automatically test the common slope assumption.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept1	1	0.2824	0.0414	0.2012	0.3635	46.53	<.0001
Intercept2	1	1.5545	0.0535	1.4497	1.6594	844.88	<.0001
Intercept3	1	2.2792	0.0686	2.1446	2.4137	1102.40	<.0001
u	1	-1.4208	0.1742	-1.7624	-1.0793	66.49	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

□

7.4 Latent variables

Another point of view when analyzing ordinal response variables is to assume that the observed ordinal variable Y is related to some underlying, latent,

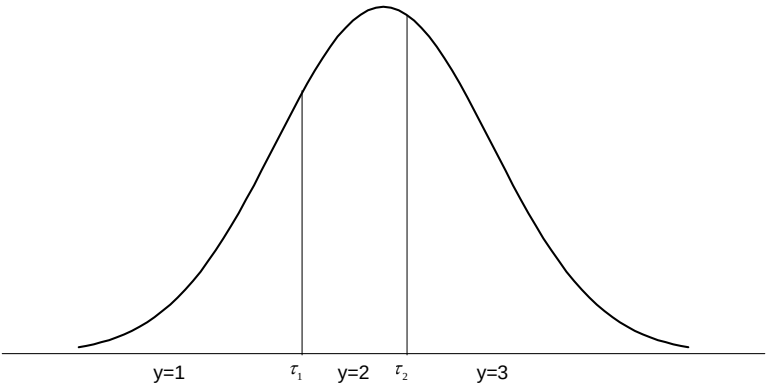


Figure 7.2: *Ordinal variable with three scale steps generated by cutting a continuous variable at two thresholds*

variable η through a relation of type

$$\begin{array}{lll} y = 1 & \text{if} & \eta < \tau_1 \\ y = 2 & \text{if} & \tau_1 \leq \eta < \tau_2 \\ \quad \vdots & & \\ y = s & \text{if} & \tau_{s-1} \leq \eta \end{array} \tag{7.7}$$

An example of this point of view is illustrated in Figure 7.2, where the latent variable is assumed to have a symmetric distribution, for example a logistic or a Normal distribution.

Although (7.7) can be formally seen as a kind of link function, modelling the data by assuming a latent variable underlying the ordinal response is not formally a generalized linear model.

However, it can be shown (see e.g. McCullagh and Nelder, 1989) that the latent variable approach gives a model that is identical to the proportional odds model with a logit link, for the case where the latent variable has a logistic distribution. The estimated intercepts would be the estimated thresholds for the latent variables.

In a similar way, a proportional odds model using a complementary log-log link corresponds to a latent variable having a so called extreme value distribution. This is the well-known proportional hazards model used in survival analysis (Cox 1972).

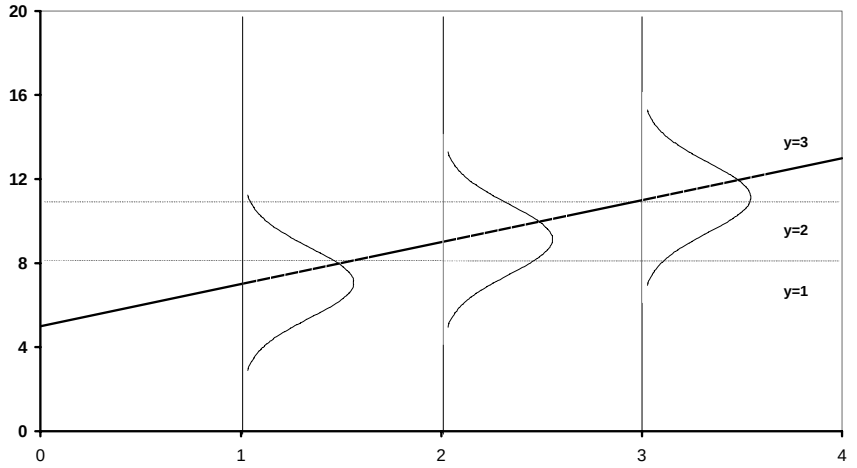


Figure 7.3: *An ordinal regression model*

A similar approach can also be used for the case where the latent variable is assumed to follow a Normal distribution. In the Genmod or Logistic procedures in SAS it is possible to specify the form of the link function to be logistic, complementary log-log, or Normal. This leads to a class of models called ordinal regression models, for example ordinal logit regression or ordinal probit regression. The concept of an ordinal regression model can be illustrated as in Figure 7.3. We observe the ordinal variable y that has values 1, 2 or 3. $y = 1$ is observed if the latent variable η is smaller than the lowest threshold which has a value close to 8. We observe $y = 2$ if, approximately, $8 \leq \eta < 11.5$ and we observe $y = 3$ if $\eta > 11.5$. In practice the scale of η cannot be determined. The scale of η can be chosen arbitrarily, for example such that the distribution of η is standard Normal for one of the values of x.

Note that probit models and logistic regression models can also be derived as models with latent variables. In these cases it is assumed that the observations are generated by a latent variable: if this latent variable is smaller than a threshold τ we observe $Y = 1$, else $Y = 0$; see Figure 5.1 on page 86.

As a comparison with the results given on page 149 we have analyzed the data on page 145 using the Logistic procedure and a Normal link. The following results were obtained:

```
Score Test for the Equal Slopes Assumption

 Chi-Square = 2.6895 with 2 DF (p=0.2606)

Model Fitting Information and Testing Global Null Hypothesis BETA=0

                          Intercept
                Intercept       and
Criterion         Only      Covariates    Chi-Square for Covariates

AIC             5568.351      5507.436         .
SC              5585.804      5530.706         .
-2 LOG L        5562.351      5499.436       62.916 with 1 DF (p=0.0001)
Score                .             .        70.693 with 1 DF (p=0.0001)


Analysis of Maximum Likelihood Estimates

                Parameter  Standard     Wald        Pr >     Standardized
 Variable  DF    Estimate    Error   Chi-Square  Chi-Square    Estimate

 INTERCP1  1      0.1749    0.0258     46.0795      0.0001            .
 INTERCP2  1      0.9355    0.0300    974.8408      0.0001            .
 INTERCP3  1      1.3266    0.0352   1420.1155      0.0001            .
 U         1     -0.8415    0.1071     61.6773      0.0001     -0.173150
```

The fit of the model, and the conclusions, are similar to the logistic model. The three thresholds are estimated to be 0.17; 0.94; and 1.33. For $x = 0$ this would give the probabilities as 0.5694, 0.2558, 0.0825 and 0.0923. For $x = 1$ the mean value of η is -0.8415 so the probabilities are 0.8463, 0.1159, 0.0227 and 0.0151.

7.5 A Genmod example

Example 7.3 Koch and Edwards (1988) considered analysis of data from a clinical trial on the response to treatment for arthritis pain. The data are as follows:

		Response		
Gender	Treatment	Marked	Some	None
Female	Active	16	5	6
Female	Placebo	6	7	19
Male	Active	5	2	7
Male	Placebo	1	0	10

The object is to model the response as a function of gender and treatment. We will attempt a proportional odds model for the cumulative logits and the

cumulative probits, using the Genmod procedure of SAS (2000b). The data were input in a form where the data lines had the form

```
F A 3 16
F A 2  5
...
```

The program was written as follows:

```
PROC GENMOD data=Koch order=formatted;
  CLASS gender treat;
  FREQ count;
  MODEL response = gender treat gender*treat/
       LINK=cumlogit aggregate=response TYPE3;
RUN;
```

Part of the output was:

Analysis Of Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square
Intercept1			1	3.6746	1.0125	1.6901	5.6591	13.17
Intercept2			1	4.5251	1.0341	2.4983	6.5519	19.15
gender	F		1	-3.2358	1.0710	-5.3350	-1.1366	9.13
gender	M		0	0.0000	0.0000	0.0000	0.0000	.
treat	A		1	-3.7826	1.1390	-6.0150	-1.5503	11.03
treat	P		0	0.0000	0.0000	0.0000	0.0000	.
gender*treat	F	A	1	2.1110	1.2461	-0.3312	4.5533	2.87
gender*treat	F	P	0	0.0000	0.0000	0.0000	0.0000	.
gender*treat	M	A	0	0.0000	0.0000	0.0000	0.0000	.
gender*treat	M	P	0	0.0000	0.0000	0.0000	0.0000	.
Scale			0	1.0000	0.0000	1.0000	1.0000	

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
gender	1	18.01	<.0001
treat	1	28.15	<.0001
gender*treat	1	3.60	0.0579

We can note that there is a slight (but not significant) interaction; that there are significant gender differences and that the treatment has a significant effect. The signs of the parameters indicate that patients on active treatment experienced a higher degree of pain relief and that the females experienced better pain relief than the males. The cumulative probit model gave similar results except that the interaction term was further from being significant ($p = 0.11$). □

7.6 Exercises

Exercise 7.1 Ezdinli et al (1976) studied two treatments against lymphocytic lymphoma. After the experiment the tumour of each patient was graded on an ordinal scale from "Complete response" to "Progression". Examine whether the treatments differ in their efficiacy by fitting an appropriate ordinal regression model. You are also free to analyze the data using other methods that you may have met during your training.

	Treatment		Total
	BP	CP	
Complete response	26	31	57
Partial response	51	59	110
No change	21	11	32
Progression	40	34	74
Total	138	135	273

Exercise 7.2 The following data, from Hosmer and Lemeshow, (1989), come from a survey on women's attitudes towards mammography. The women were asked the question "How likely is it that mammography could find a new case of breast cancer". They were also asked about recent experience of mammography. Results:

Mammography	Detection of breast cancer		
experience	Not likely	Somewhat likely	Very likely
Never	13	77	144
Over 1 year ago	4	16	54
Within the past year	1	12	91

Analyze these data.

8. Additional topics

8.1 Variance heterogeneity

In general linear models, it is not uncommon that diagnostic tools indicate that the variance is not constant. This might indicate that the choice of distribution is wrong, such that some distribution where the variance depends on the mean should be chosen instead of the Normal distribution.

An alternative approach, suggested by Aitkin (1987) works as follows. The response for observation i is modeled using the linear predictor

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + e_i \tag{8.1}$$

where we assume that $e_i \sim N\left(0, \sigma_i^2\right)$. The variance σ_i^2 is modeled as

$$\sigma_i^2 = \exp\left(\lambda \mathbf{z}_i\right). \tag{8.2}$$

Here, z is a vector that contains some or all of the predictors $\mathbf{x}$, and $\boldsymbol{\lambda}$ is a vector of parameters to be estimated. Thus, the problem is to estimate the parameters of the linear predictor, as well as the parameters in the model for the variance. The estimation procedure suggested by Aitkin (1987) to estimate the parameter vector $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is to iterate between two generalized linear models. One model is a model with a Normal distribution and an identity link, (8.1), and the other model fits the squared residuals from this model to a Gamma distribution using a log link, corresponding to (8.2). Aitkin showed that this process produces the ML estimates, on convergence. A SAS macro for this process is given in the SAS (1997) manual.

Example 8.1 In our analysis of the data on page 16 we found that the data strongly suggested variance heterogeneity, but that the distribution, for each contrast medium, was rather symmetric. This may indicate that the variance heterogeneity can be modeled using Aitkin's method. The procedure produces one set of estimates for the mean model and a second set of estimates for the variance model. For these data we obtained the following results for the mean model:

Mean model

OBS	PARM	LEVEL1	DF	ESTIMATE	STDERR	CHISQ	PVAL
1	INTERCEPT		1	9.9075	0.3536	785.2685	0.0001
2	MEDIUM	Diatrizo	1	11.4845	0.5701	405.8269	0.0001
3	MEDIUM	Hexabrix	1	-5.9475	0.4859	149.8140	0.0001
4	MEDIUM	Isovist	1	-8.2431	0.4859	287.7796	0.0001
5	MEDIUM	Mannitol	1	-2.2197	0.4859	20.8680	0.0001
6	MEDIUM	Omnipaqu	1	-1.5175	0.4743	10.2347	0.0014
7	MEDIUM	Ringer	1	-9.6975	0.5175	351.0883	0.0001
8	MEDIUM	Ultravis	0	0.0000	0.0000	.	.
9	SCALE		0	1.0000	0.0000	.	.

The estimates for the variance model were as follows:

Variance model

OBS	PARM	LEVEL1	DF	ESTIMATE	STDERR	CHISQ	PVAL
1	INTERCEPT		1	2.9560	0.5000	34.9524	0.0001
2	MEDIUM	Diatrizo	1	1.3196	0.8062	2.6788	0.1017
3	MEDIUM	Hexabrix	1	-1.4736	0.6872	4.5982	0.0320
4	MEDIUM	Isovist	1	-2.1732	0.6872	10.0016	0.0016
5	MEDIUM	Mannitol	1	-0.3517	0.6872	0.2620	0.6087
6	MEDIUM	Omnipaqu	1	0.0905	0.6708	0.0182	0.8927
7	MEDIUM	Ringer	1	-6.2914	0.7319	73.8867	0.0001
8	MEDIUM	Ultravis	0	0.0000	0.0000	.	.
9	SCALE		0	0.5000	0.0000	.	.

The procedure converged after two iterations producing an overall deviance of 257.73 on 43 *df*. □

8.2 Survival models

Survival data are data for which the response is the time a subject has survived a certain treatment or condition. Survival models are used in epidemiology, as well as in lifetime testing in industry. Censoring is a special feature of survival data. Censoring means that the survival time is not known for all individuals when the study is finished. For right censored observations we only know that the survival time is at least the time at which censoring occurred. Left censoring, i.e. observations for which we do not know e.g. the duration of disease when the study started, is also possible.

Denote the density function for the survival time with $f(t)$, and let the corresponding distribution function be $F(t) = \int_{-\infty}^{t} f(s)\,ds$. The survival function is defined as

$$S(t) = 1 - F(t) \tag{8.3}$$

and the hazard function is defined as

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d\log(S(t))}{dt} \tag{8.4}$$

The hazard function measures the instantaneous risk of dying, i.e. the probability of dying in the next small time interval of duration dt. The cumulative hazard function is

$$H(t) = \int_{-\infty}^{t} h(s)\, ds \tag{8.5}$$

Modelling of survival data includes choosing a suitable distribution for the survival times or, which is equivalent, choosing a hazard function. This can be done in different ways:

1. In nonparametric modelling, the survival function is not specified, but is estimated nonparametrically through the observed survival distribution. This is the basis for the so called Kaplan-Meier estimates of the survival function.

2. In parametric models, the distribution of survival times is assumed to have some specified parametric form. The exponential distribution, Weibull distribution or extreme value distribution are often used to model survival times.

3. A semiparametric approach is to leave the distribution unspecified but to assume that the hazard function changes in steps which occur at the observed events.

We will here only give examples of the parametric approach. For a more thorough description of analysis of survival data, reference is made to standard textbooks such as Klein and Moeschberger (1997).

8.2.1 An example

Although survival models are often discussed in texts on generalized linear models, the treatment of censoring makes it more convenient to analyze general survival data using special programs. However, data where there is no censoring can be analyzed using standard GLIM software, as long as the desired survival distribution belongs to the exponential family.

Example 8.2 Feigl and Zelen (1965) analyzed the survival times for leukemia patients classified as AG positive or AG negative. The white cell count

Table 8.1: *Survival of leukemia patients*

AG +		AG −	
WBC	Surv.	WBC	Surv.
2300	65	4400	56
750	156	3000	65
4300	100	4000	17
2600	134	1500	7
6000	16	9000	16
10500	108	5300	22
10000	121	10000	3
17000	4	19000	4
5400	39	27000	2
7000	143	28000	3
9400	56	31000	8
32000	26	26000	4
35000	22	21000	3
100000	1	79000	30
100000	1	100000	4
52000	5	100000	43
100000	65		

(WBC) for each patient is also given. The data are reproduced in Table 8.1.

As a first attempt, we model the data using a Gamma distribution. The log of the WBC was used. The interaction ag*logwbc was not significant. The program is

```
PROC GENMOD data=feigl;
  CLASS ag;
  MODEL survival = ag logwbc  /
                   DIST=gamma obstats residuals;
  MAKE obstats out=ut;
RUN;
```

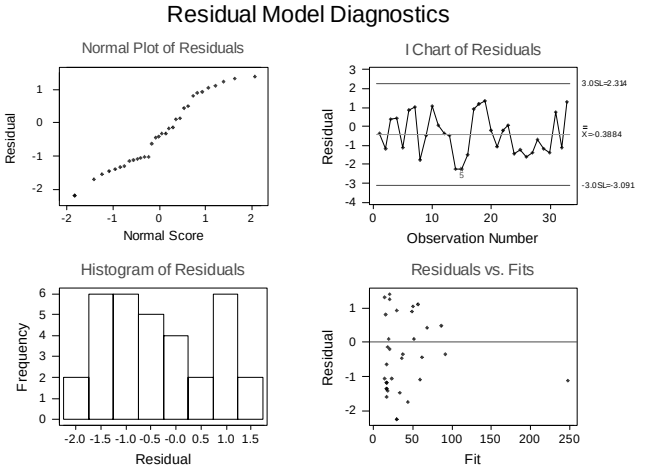


Figure 8.1: *Residual plots for Leukemia data*

Parts of the output is

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	30	40.0440	1.3348
Scaled Deviance	30	38.2985	1.2766
Pearson Chi-Square	30	29.6222	0.9874
Scaled Pearson X2	30	28.3310	0.9444
Log Likelihood	.	-146.3814	.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-0.0020	0.0262	0.0056	0.9403
AG	+	1	-0.0344	0.0150	5.2495	0.0220
AG	-	0	0.0000	0.0000	.	.
LOGWBC		1	0.0061	0.0024	6.5899	0.0103
SCALE		1	0.9564	0.2065	.	.

The fit of the model is reasonable with a scaled deviance of 38.3 on 30 df, Deviance/df=1.28. The effects of WBC and AG are both significant. We asked the procedure to output predicted values and standardized Deviance residuals to a file. A residual plot based on these data is given in Figure 8.1. The distribution seems reasonable; one very large fitted value stands out. □

8.3 Quasi-likelihood

In general linear models, the assumption that the observation come from a Normal distribution is not crucial. Estimation of parameters in GLM:s is often done using some variation of Least squares, for which certain optimality properties are valid even under non-normality. Thus, we can estimate parameters in, for example, regression models, without being too much worried by non-normality.

Quasi-likelihood can give a similar peace of mind to users of generalized linear models. In principle, we need to specify a distribution (Poisson, binomial, Normal etc.) when we fit generalized linear models. However, Wedderburn (1974) noted the following property of generalized linear models.

The score equations for the regression coefficients $\boldsymbol{\beta}$ have the form

$$\sum_i \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} v_i^{-1} \left(y_i - \mu_i \left(\boldsymbol{\beta}\right)\right) = 0 \tag{8.6}$$

Note that this expression only contains the first two moments, the mean μ_i and the variance v_i. Wedderburn (1974) suggested that this can be used to define a class of estimators that do not require explicit expressions for the distributions. A type of generalized linear models can be constructed by specifying the linear predictor η and the way the variance v depends on μ.

The integral of (8.6) can be seen as a kind of likelihood function. This integral is

$$Q\left(y_i, \mu_i\right) = \int_{-\infty}^{\mu_i} \frac{y_i - t}{v_i} dt + f\left(y_i\right) \tag{8.7}$$

where $f\left(y_i\right)$ is some arbitrary function of y_i. $Q\left(y_i, \mu_i\right)$ is called a quasi-likelihood. Maximizing (8.7) with respect to the parameters of the model yields quasi-likelihood (QL) estimators. QL estimators can be shown to have nice asymptotic properties. First, they are consistent, regardless of whether the variance assumption $v_i = V\left(\mu_i\right)$ is true, as long as the linear predictor is correctly specified. Secondly, QL estimators are asymptotically unbiased and efficient among the class of estimating equations which are linear functions of the data (McCullagh, 1983).

Estimators of the variances of QL estimators can be obtained in different ways. The matrix $\mathbf{I}_\theta$ of second order derivatives of (8.7) gives the QL equivalent of the Fisher information matrix. The inverse $\mathbf{I}_\theta^{-1}$ is an estimator of the covariance matrix of the parameter estimates. This is called the model-based estimator of $Cov\left(\widehat{\beta}\right)$. An alternative approach is to use the so called empirical, or robust, estimator, which is less sensitive to assumptions regarding

variances and covariances. This is also called the sandwich estimator. It has general form

$$\widehat{Cov}\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{I}_\theta^{-1}\mathbf{I}_1\mathbf{I}_\theta^{-1}$$

where $\mathbf{I}_\theta$ is the information matrix and

$$\mathbf{I}_1 = \sum_{i=1}^{k} \frac{\partial \boldsymbol{\mu}_i^{'}}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i^{'}}{\partial \boldsymbol{\beta}}.$$

Software supporting quasi-likelihood may have options to choose between model-based and robust variance estimators.

The quasi-likelihood approach can be used in over-dispersed models. It is also used in the GEE method for analysis of repeated measures data, and in the method for analysis of mixed generalized linear models discussed below.

8.4 Quasi-likelihood for modeling overdispersion

The quasi-likelihood approach is sometimes useful when the data show signs of over-dispersion. Since the emprical variance estimates obtained in QL estimation are rather robust against the variance assumption, QL estimation is a viable alternative to the methods for modeling over-dispersion presented in earlier chapters, at least if the sample is reasonably large. We will illustrate this idea based on a set of data from Liang and Hanfelt (1994).

Example 8.3 Two groups of rats, each consisting of sixteen pregnant females, were fed different diets during pregnancy and lactation. The control diet was a standard food whereas the treatment diet contained a certain chemical agent. After three weeks it was recorded how many of the live born pups that still were alive. The data are given as x/n where x is the number of surviving pups and n is the total litter size.

Control	13/13	12/12	9/9	9/9	8/8	8/8	12/13	11/12
	9/10	9/10	8/9	11/13	4/5	5/7	7/10	7/10
Treated	12/12	11/11	10/10	9/9	10/11	9/10	9/10	8/9
	8/9	4/5	7/9	4/7	5/10	3/6	3/10	0/7

A standard logistic model has a rather bad fit with a deviance of 86.19 on 30 df, $p < 0.0001$. In this model the treatment effect is significant, both when we use a Wald test ($p = 0.0036$) and when we use a likelihood ratio test ($p = 0.0027$).

The bad fit may be caused by heterogeneity among the females: different females may have different ability to take care of their pups. If it can be assumed that the dispersion parameter is the same in both groups, this can be modeled by including a dispersion parameter in the model, as discussed in Chapter 5. Such a model gives a non-significant treatment effect (Wald test: $p = 0.0855$; LR test: $p = 0.0765$.)

The quasi-likelihood estimates can be obtained in Proc Genmod by using the following trick. Proc Genmod can use QL, but only in repeated-measures models. We can then request a repeated-measures analysis but with only one measurement per female. The program can be written as

```
PROC GENMOD data=tera;
CLASS treat litter;
MODEL x/n=treat /
      DIST=bin LINK=logit type3;
      REPEATED subject=litter;
RUN;
```

The output is given in two parts. The first part uses the model-based estimates of variances and these results are identical to the first output. The second part presents the QL results:

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		1.2220	0.3813	0.4747	1.9693	3.20	0.0014
treat	c	0.9612	0.4751	0.0300	1.8925	2.02	0.0431
treat	t	0.0000	0.0000	0.0000	0.0000	.	.

Score Statistics For Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
treat	1	2.89	0.0890

In these results the Wald test is significant ($p = 0.043$) but the Score test is not ($p = 0.0890$). □

In the paper by Liang and Hanfelt (1994), a simulation study compared the performances of different methods for allowing for overdispersion in this type of data. The methods included modeling as a beta-binomial distribution, and two QL approaches. In the simulations, the overdispersion parameter was

different for different treatments. Nevertheless, the QL approach assuming constant overdispersion performed surprisingly well. It was also concluded that results based on the beta-binomial distribution "can lead to severe bias in the estimation of the dose-response relationship" (p. 878.) Thus, the QL approach seems to be a useful and rather robust tool for modeling overdispersed data.

8.5 Repeated measures: the GEE approach

Suppose that measurements have been made on the same individuals on k occasions. The responses can then be represented as Y_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, k$. Subject i has measurements on n_i occasions, and we have a total of $\sum_{i=1}^{k} n_i$ measurements. This type of data is called repeated measures data.

The main problem with repeated measures data is that observations within one individual are correlated. There are several ways to model this correlation. We will here only consider the Generalized estimating equations approach of Liang and Zeger (1986); see also Diggle, Liang and Zeger (1994). This approach is available in the Genmod procedure in SAS (2000b). The GEE approach can be seen as an extension of the quasi-likelihood approach to a multivariate mean vector.

Models for repeated measures data have the same basic components as other generalized linear models. We need to specify a link function, a distribution and a linear predictor. But in addition we need to consider how observations within individuals are correlated.

Suppose that we store all data for individual i in the vector $\mathbf{Y}_i$ that has elements $\mathbf{Y}_i = [Y_{i1}, ..., Y_{in_i}]'$. The corresponding vector of mean values is $\boldsymbol{\mu}_i = \left[\mu_{i1}, ..., \mu_{in_i}\right]'$. Let $\mathbf{V}_i$ be the covariance matrix of $\mathbf{Y}_i$. The values of the independent variables for individual i at measurement (occasion) j are collected in the vector $\mathbf{x}_{ij}^{'} = [x_{ij1}, ..., x_{ijp}]'$.

The vector $\boldsymbol{\beta}$ contains the parameters to be estimated. The GEE approach means that we estimate the parameters by solving the GEE equation

$$\sum_{i=1}^{k} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} \left(\mathbf{Y}_i - \boldsymbol{\mu}_i \left(\boldsymbol{\beta}\right)\right) = 0 \tag{8.8}$$

This is similar to the quasi-likelihood equation (8.6), but it is here written in matrix form. It can be shown that the multivariate quasi-likelihood approach provides consistent estimates of the parameters even if the covariance

structure is incorrectly specified. Estimates of the variances and covariances of $\widehat{\boldsymbol{\beta}}$ can be obtained in two ways. The model-based approach assumes that the model is correctly specified. The robust approach provides consistent estimates of variances and covariances even if $\mathbf{V}_i$ is incorrectly specified. Both approaches are available in Proc Genmod.

The correlations between measurement occasions are modeled by a vector of parameters $\boldsymbol{\alpha}$. The following correlation structures are available in Proc Genmod:

Fixed (user-specified): $Corr\left(Y_{ij}, Y_{ik}\right) = r_{jk}$.

m-dependent: $Corr\left(Y_{ij}, Y_{ik}\right) = \begin{cases} 1 \text{ if } t = 0 \\ \alpha_t \text{ if } t = 1, ..., m \\ 0 \text{ if } t > m \end{cases}$ where t is the time span between the observations. The correlation is 0 for occasions more than m time units apart.

Exchangeable: $Corr\left(Y_{ij}, Y_{ik}\right) = \begin{cases} 1 \text{ if } j = k \\ \alpha \text{ if } j \neq k \end{cases}$. All correlations are equal.

Unstructured: $Corr\left(Y_{ij}, Y_{ik}\right) = \begin{cases} 1 \text{ if } j = k \\ \alpha_{jk} \text{ if } j \neq k \end{cases}$

Autoregressive, AR(1): $Corr\left(Y_{ij}, Y_{ik}\right) = \alpha^t$ for $t = 0, 1, ..., n_i - j$

As usual, the choice of model for the covariance structure is a compromise between realism and parsimony. A model with more parameters is often more realistic, but may be more difficult to interpret and may give convergence problems. The fixed structure means that the user enters all correlations, so there are no parameters to estimate. The exchangeable structure includes only one parameter. The AR(1) structure also has only one parameter but it is often intuitively appealing since the correlations decrease with increasing distance. If we assume unstructured correlations we need to estimate $k\left(k-1\right)/2$ correlations, while the m-dependent correlation structure includes fewer correlations.

Example 8.4 Sixteen children (taken from the data of Lipsitz et al, 1994) were followed from the age of 9 to the age of 12. The children were from two different cities. The binary response variable was the wheezing status of the child. The explanatory variables were city; age; and maternal smoking status. The structure of the data is given in Table 8.2; the complete data set is available from the publishers home page as the file Wheezing.dat.

Table 8.2: *Structure of the data on wheezing status*

Child	City	Age	Smoke	Wheeze
1	Portage	9	0	1
1	Portage	10	0	1
1	Portage	11	0	1
1	Portage	12	0	0
2	Kingston	9	1	1
2	Kingston	10	2	1
2	Kingston	11	2	0

A Genmod program for analysis of these data can be written as

```
PROC GENMOD DATA=wheezing;
  CLASS child city;
  MODEL wheeze = city age smoke
             /dist=bin link=logit;
REPEATED subject=child / type = exch covb corrw;
RUN;
```

This program models the probability of wheezing as a function of city, age and maternal smoking. The effects of age and smoking are assumed to be linear. A binomial distribution with a logit link is used.

The Repeated statement indicates that there are several measurements for each child. These are correlated, with an exchangeable correlation structure as described above. Additional output is also requested.

The following output is obtained:

```
          Criteria For Assessing Goodness Of Fit

Criterion              DF          Value      Value/DF

Deviance               60        76.9380        1.2823
Scaled Deviance        60        76.9380        1.2823
Pearson Chi-Square     60        63.9651        1.0661
Scaled Pearson X2      60        63.9651        1.0661
Log Likelihood          .       -38.4690             .
```

The model fits the data reasonably well.

Covariance Matrix (Model-Based)
Covariances are Above the Diagonal and Correlations are Below

Parameter Number	PRM1	PRM2	PRM4	PRM5
PRM1	5.71511	-0.22386	-0.53133	0.01658
PRM2	-0.13847	0.45733	-0.002411	0.01877
PRM4	-0.96838	-0.01553	0.05268	-0.01658
PRM5	0.01587	0.06353	-0.16530	0.19088

Covariance Matrix (Empirical)
Covariances are Above the Diagonal and Correlations are Below

Parameter Number	PRM1	PRM2	PRM4	PRM5
PRM1	9.33891	-0.85121	-0.83232	-0.16667
PRM2	-0.40467	0.47378	0.05737	0.04007
PRM4	-0.97676	0.29893	0.07775	-0.002201
PRM5	-0.15108	0.16125	-0.02187	0.13032

The covariance matrix of $\widehat{\beta}$ is estimated in two ways: assuming that the model for V is correct, and using the "robust" method.

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter		Estimate	Empirical Std Err	95% Confidence Limits Lower	Upper	Z	Pr>|Z
INTERCEPT		1.2754	3.0560	-4.7141	7.2650	0.4174	0.6764
CITY	Kingston	0.1219	0.6883	-1.2272	1.4709	0.1771	0.8595
CITY	Portage	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AGE		-0.2036	0.2788	-0.7501	0.3429	-.7302	0.4652
SMOKE		-0.0928	0.3610	-0.8003	0.6147	-.2571	0.7971
Scale		0.9991	.	.	.	.	

Estimates of the parameters of the model are given, along with their empirical standard error estimates. None of the parameters are significantly different from 0. This, of course, may be related to the rather small sample size.

□

8.6 Mixed Generalized Linear Models

Mixed models are models where some of the independent variables are assumed to be fixed, i.e. chosen beforehand, while others are seen as randomly sampled from some population or distribution. Mixed models have proven to be very useful in modeling different phenomena. An example of an application of mixed models is when several measurements have been taken on the

same individual. In such cases the effect of individual can often be included in the model as a random effect.

A mixed linear model for a continuous response variable y can be written, for each individual i, as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \mathbf{e}_i \tag{8.9}$$

In (8.9), $\mathbf{y}_i$ is the $n_i \times 1$ response vector for individual i, $\mathbf{X}_i$ is a $n_i \times p$ design matrix that contains values for the fixed effect variables, $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector for the fixed effects, $\mathbf{Z}_i$ is a $n_i \times q$ matrix that contains the random effects variables, and $\mathbf{u}_i$ is a $q \times 1$ vector of random effects. In mixed models based on Normal theory it is often assumed that $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{D})$ and that $\mathbf{e}_i \sim N(\mathbf{0}, \Sigma_i)$. Σ_i is often chosen to be equal to $\sigma^2\mathbf{I}_{n_i}$, where $\mathbf{I}_{n_i}$ is the identity matrix of dimension n_i. $\mathbf{D}$ is a general covariance matrix of dimension $q \times q$.

In general, the actual effects of the random factors is not of primary concern. The parameters of interest in a model such as (8.9) are often the regression parameters $\boldsymbol{\beta}$; and estimates of the variance components. A special SAS procedure, Proc Mixed, can be used for fitting mixed linear models in cases where the response variable is continuous and approximately normally distributed.

There are situations when the response is of a type amenable for GLIM estimation but where there would be a need to assume that some of the independent variables are random. Breslow and Clayton (1993), and Wolfinger and O'Connell (1993) have explored a pseudo-likelihood approach to fitting models such as (8.9) but where the distributions are free to be any member of the exponential family, and where a link function is used to model the expected response as a function of the linear predictor. A SAS macro Glimmix has been written to do the estimation. Essentially, the macro iterates between Proc Mixed and Proc Genmod. The method and the macro are described in Littell et al (1996).

Example 8.5 Thirty-three children between the ages of 6 and 16 years, all suffering from monosymptomatic nocturnal enuresis, were enrolled in a study. The study was carried out with a double-blind randomized three-period cross-over design. The children received 0.4 mg. Desmopressin, 0.8 mg. Desmopressin, or placebo tablets at bedtime for five consecutive nights with each dosage. A wash-out period of at least 48 hours without any medication was interspersed between treatment periods. Wet and dry nights were documented; for more details about the study and its analysis see Neveus et al (1999), and Olsson and Neveus (2000). The data consisted of nightly recordings, where a dry night was recorded as 1 and a wet night as 0. The nights were grouped into sets of five nights where the same treatment had been given. The structure of the data is given in Table 8.3. Only one patient is listed; the original data set contained 33 patients.

Table 8.3: *Raw data for one patient in the enuresis study*

Patient	Period	Dose	Night	Dry
1	1	1	1	1
1	1	1	2	1
1	1	1	3	1
1	1	1	4	1
1	1	1	5	1
1	2	0	1	0
1	2	0	2	0
1	2	0	3	1
1	2	0	4	0
1	2	0	5	0
1	3	2	1	1
1	3	2	2	1
1	3	2	3	1
1	3	2	4	1
1	3	2	5	1

Following Jones and Kenward (1989), the linear predictor part of a general model for our data may be written as

$$\eta_{ijk} = \mu + s_{ik} + \pi_j + \tau_{d[i,j]} + \lambda_{d[i,j-1]} \tag{8.10}$$

In (8.10), μ is a general mean; s_{ik} is the random effect of patient k in sequence i, π_j is the effect of period j; $\tau_{d[i,j]}$ is the direct effect of the treatment administered in period j of group i; and $\lambda_{d[i,j-1]}$ is the carry-over effect of the treatment administered in period $j-1$ of group i.

The model further includes a logistic link function:

$$\eta_{ijk} = \log \frac{\mu_{ijk}}{1-\mu_{ijk}} \tag{8.11}$$

Finally, the model assumes a binomial distribution of the observations.

Models containing different combinations of model parameters were tested. The results are summarized in the following table. The numbers in the table are p-values to assess the significance of the different factors. Patient was included as a random factor in all models.

Effects included	Dose	Period	Sequence	After effect
Dose	.0001			
Dose, Seq	.0001		.7442	
Dose, Period	.0001	.0938		
Dose, After eff.	.0001			.6272
Dose, After eff., Period	.0001	.0759		.8713
Dose, After eff., Seq.	.0001		.7762	.6573
Dose, Period, Seq	.0001	.0898	.7577	

Based on these results, it was concluded that a model containing a random Patient effect, and fixed effects of Dose and Period, provided an appropriate description of the data. Neither the sequence effect nor the after effect was anywhere close to being significant in any of the analyses. Further analyses using pairwise comparisons revealed that there were no significant differences between doses but that the drug had a significant effect at both doses. □

8.7 Exercises

Exercise 8.1 Survival times in weeks were recorded for patients with acute leukaemia. For each patient the white cell count (wbc, in thousands) and the AG factor was also recorded. Patients with positive AG factor had Auer rods and/or granulate of the leukemia cells in the bone marrow at diagnosis while the AG negative patients had not.

Time	wbc	AG	Time	wbc	AG
65	2.3	1	65	3.0	0
108	10.5	1	3	10.0	0
56	9.4	1	4	26.0	0
5	52.0	1	22	5.3	0
143	7.0	1	17	4.0	0
156	0.8	1	4	19.0	0
121	10.0	1	3	21.0	0
26	32.0	1	8	31.0	0
65	100.0	1	7	1.5	0
1	100.0	1	2	27.0	0
100	4.3	1	30	79.0	0
4	17.0	1	43	100.0	0
22	35.0	1	16	9.0	0
56	4.4	1	3	28.0	0
134	2.6	1	4	100.0	0
39	5.4	1			
1	100.0	1			
16	6.0	1			

The four largest wbc values are actually larger than 100. Construct a model that can predict survival time based on wbc and ag. Note that the wbc value may need to be transformed. Also note that there are no censored observations so a straight-forward application of generalized linear models is possible. Try a survival distribution based on the Gamma distribution.

Exercise 8.2 The data in Exercise 1.2 show some signs of being heteroscedastic. Re-analyze these data using the method discussed in Section 8. The data are as follows:

The level of cortisol has been measured for three groups of patients with different syndromes: a) adenoma b) bilateral hyperplasia c) cardinoma. The

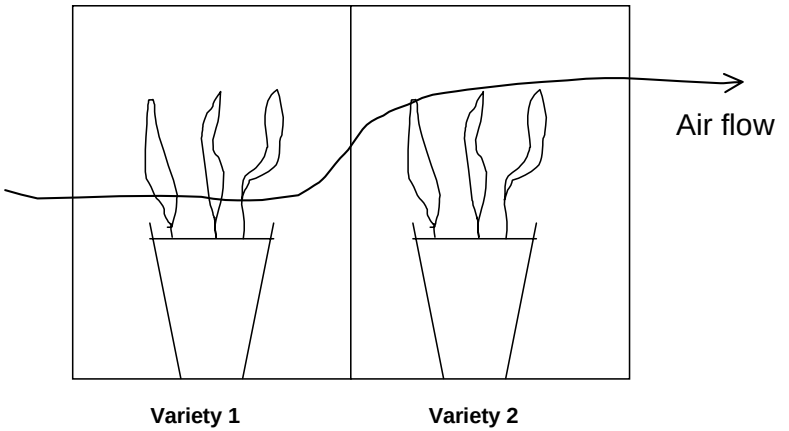


Figure 8.2: *Experimental layout for lice experiment*

results are summarized in the following table:

a	b	c
3.1	8.3	10.2
3.0	3.8	9.2
1.9	3.9	9.6
3.8	7.8	53.8
4.1	9.1	15.8
1.9	15.4	
	7.7	
	6.5	
	5.7	
	13.6	

Exercise 8.3 An experiment on lice preferences for different varieties of plants (Nincovic et al, 2002) was preformed in the following way: Plants of one variety (Variety 1) were placed in a box. An adjacent box contained plants of some other variety (Variety 2); see Figure 8.2. Air was allowed to flow through the boxes from Variety 1 to Variety 2. Tubes were placed on four leaves of the plant of Variety 2. 10 lice were placed in each tube. After about two hours it was recorded how many of the 10 lice that were eating from the plant.

The experiment was designed to answer the following types of questions: Are the eating preferences of the lice different for different varieties? Are the

eating preferences affected by the smell from Variety 1?

The structure of the raw data was as follows; only a few observations are listed. The complete dataset contains 320 observations and is listed at the end of the exercise.

Pot	Tube	x2	n2	Var1	Var2
1	1	9	10	F	K
1	2	7	10	F	K
1	3	7	10	F	K
1	4	8	10	F	K
2	5	9	10	F	K
2	6	10	10	F	K
2	7	10	10	F	K
2	8	10	10	F	K
3	9	10	10	F	K
3	10	10	10	F	K
3	11	9	10	F	K

Pot indicates pot number and Tube indicates tube number. n2 is the number of lice in the tube, and x2 is the number of lice eating after two hours. Var1 and Var2 are codes for Variety 1 and Variety 2, respectively.

Formulate a model for these data that can answer the question whether the eating preferences of the lice depends on Variety 1, Variety 2 or a combination of these. Hint: Since repeated observations are made on the same plants it may be reasonable to include Pot as a random factor in the model.

Pot	Tube	x2	n2	Var1	Var2
1	1	9	10	F	K
1	2	7	10	F	K
1	3	7	10	F	K
1	4	8	10	F	K
2	5	9	10	F	K
2	6	10	10	F	K
2	7	10	10	F	K
2	8	10	10	F	K
3	9	10	10	F	K
3	10	10	10	F	K
3	11	9	10	F	K
3	12	7	10	F	K
4	13	8	10	F	K
4	14	9	10	F	K
4	15	8	10	F	K
4	16	9	10	F	K
5	17	9	10	F	K
5	18	9	10	F	K
5	19	10	10	F	K
5	20	10	10	F	K
1	1	8	10	K	F
1	2	7	10	K	F
1	3	10	10	K	F
1	4	7	10	K	F
2	5	9	10	K	F
2	6	8	10	K	F
2	7	10	10	K	F
2	8	8	10	K	F
3	9	10	10	K	F
3	10	10	10	K	F
3	11	10	10	K	F
3	12	10	10	K	F
4	13	7	10	K	F
4	14	10	10	K	F
4	15	7	10	K	F
4	16	10	10	K	F
5	17	8	10	K	F
5	18	7	10	K	F
5	19	10	10	K	F
5	20	7	10	K	F
1	1	9	10	H	K
1	2	7	10	H	K
1	3	7	10	H	K
1	4	8	10	H	K
2	5	7	10	H	K
2	6	10	10	H	K
2	7	9	10	H	K
2	8	8	10	H	K
3	9	8	10	H	K
3	10	9	10	H	K
3	11	7	10	H	K
3	12	9	10	H	K
4	13	10	10	H	K
4	14	8	10	H	K
4	15	6	10	H	K
4	16	9	10	H	K
5	17	9	10	H	K
5	18	8	10	H	K
5	19	8	10	H	K
5	20	9	10	H	K
1	1	8	10	K	H
1	2	8	10	K	H
1	3	8	10	K	H
1	4	8	10	K	H
2	5	7	10	K	H
2	6	8	10	K	H
2	7	8	10	K	H
2	8	7	10	K	H
3	9	9	10	K	H
3	10	8	10	K	H
3	11	9	10	K	H
3	12	6	10	K	H
4	13	7	10	K	H
4	14	8	10	K	H
4	15	8	10	K	H
4	16	9	10	K	H
5	17	6	10	K	H
5	18	9	10	K	H
5	19	7	10	K	H
5	20	8	10	K	H
1	1	9	10	F	H
1	2	7	10	F	H
1	3	9	10	F	H
1	4	8	10	F	H
2	5	7	10	F	H
2	6	9	10	F	H
2	7	8	10	F	H
2	8	8	10	F	H
3	9	7	10	F	H
3	10	7	10	F	H
3	11	7	10	F	H
3	12	6	10	F	H
4	13	8	10	F	H
4	14	8	10	F	H
4	15	8	10	F	H
4	16	7	10	F	H
5	17	7	10	F	H
5	18	7	10	F	H
5	19	2	10	F	H
5	20	7	10	F	H
1	1	9	10	H	F
1	2	10	10	H	F
1	3	9	10	H	F
1	4	9	10	H	F
2	5	7	10	H	F
2	6	9	10	H	F
2	7	8	10	H	F
2	8	7	10	H	F
3	9	7	10	H	F
3	10	7	10	H	F
3	11	10	10	H	F
3	12	7	10	H	F
4	13	8	10	H	F
4	14	9	10	H	F
4	15	10	10	H	F
4	16	8	10	H	F
5	17	10	10	H	F
5	18	9	10	H	F
5	19	8	10	H	F
5	20	7	10	H	F
1	1	10	10	A	K
1	2	8	10	A	K
1	3	9	10	A	K
1	4	7	10	A	K
2	5	8	10	A	K
2	6	6	10	A	K
2	7	8	10	A	K
2	8	7	10	A	K
3	9	8	10	A	K
3	10	8	10	A	K
3	11	10	10	A	K
3	12	9	10	A	K
4	13	7	10	A	K
4	14	8	10	A	K
4	15	10	10	A	K
4	16	8	10	A	K
5	17	7	10	A	K
5	18	10	10	A	K
5	19	8	10	A	K
5	20	7	10	A	K
1	1	7	10	K	A
1	2	9	10	K	A
1	3	10	10	K	A
1	4	8	10	K	A
2	5	8	10	K	A
2	6	8	10	K	A
2	7	8	10	K	A
2	8	7	10	K	A

3	9	6	10	K	A	1	3	9	10	H	A
3	10	3	10	K	A	1	4	8	10	H	A
3	11	7	10	K	A	2	5	9	10	H	A
3	12	8	10	K	A	2	6	10	10	H	A
4	13	4	10	K	A	2	7	8	10	H	A
4	14	8	10	K	A	2	8	5	10	H	A
4	15	9	10	K	A	3	9	5	10	H	A
4	16	8	10	K	A	3	10	9	10	H	A
5	17	6	10	K	A	3	11	8	10	H	A
5	18	5	10	K	A	3	12	8	10	H	A
5	19	10	10	K	A	4	13	10	10	H	A
5	20	7	10	K	A	4	14	9	10	H	A
1	1	8	10	A	F	4	15	9	10	H	A
1	2	5	10	A	F	4	16	9	10	H	A
1	3	5	10	A	F	5	17	8	10	H	A
1	4	6	10	A	F	5	18	7	10	H	A
2	5	8	10	A	F	5	19	9	10	H	A
2	6	10	10	A	F	5	20	5	10	H	A
2	7	3	10	A	F	1	1	9	10	K	K
2	8	7	10	A	F	1	2	9	10	K	K
3	9	6	10	A	F	1	3	9	10	K	K
3	10	9	10	A	F	1	4	5	8	K	K
3	11	6	10	A	F	2	5	10	10	K	K
3	12	8	10	A	F	2	6	8	10	K	K
4	13	8	10	A	F	2	7	9	10	K	K
4	14	5	10	A	F	2	8	7	10	K	K
4	15	8	10	A	F	3	9	8	10	K	K
4	16	6	10	A	F	3	10	10	10	K	K
5	17	5	10	A	F	3	11	7	10	K	K
5	18	6	10	A	F	3	12	10	10	K	K
5	19	9	10	A	F	4	13	9	10	K	K
5	20	7	10	A	F	4	14	7	10	K	K
1	1	8	10	F	A	4	15	8	10	K	K
1	2	10	10	F	A	4	16	8	10	K	K
1	3	9	10	F	A	5	17	8	10	K	K
1	4	9	10	F	A	5	18	7	10	K	K
2	5	8	10	F	A	5	19	10	10	K	K
2	6	8	10	F	A	5	20	9	10	K	K
2	7	7	10	F	A	1	1	9	10	A	A
2	8	7	10	F	A	1	2	10	10	A	A
3	9	7	10	F	A	1	3	10	10	A	A
3	10	8	10	F	A	1	4	10	10	A	A
3	11	10	10	F	A	2	5	10	10	A	A
3	12	6	10	F	A	2	6	6	10	A	A
4	13	10	10	F	A	2	7	9	10	A	A
4	14	9	10	F	A	2	8	8	10	A	A
4	15	8	10	F	A	3	9	8	10	A	A
4	16	9	10	F	A	3	10	10	10	A	A
5	17	9	10	F	A	3	11	10	10	A	A
5	18	8	10	F	A	3	12	8	10	A	A
5	19	8	10	F	A	4	13	4	10	A	A
5	20	9	10	F	A	4	14	8	10	A	A
1	1	7	10	A	H	4	15	8	10	A	A
1	2	7	10	A	H	4	16	8	10	A	A
1	3	6	10	A	H	5	17	7	10	A	A
1	4	4	10	A	H	5	18	9	10	A	A
2	5	8	10	A	H	5	19	9	10	A	A
2	6	7	10	A	H	5	20	8	10	A	A
2	7	5	10	A	H	1	1	9	10	F	F
2	8	7	10	A	H	1	2	8	10	F	F
3	9	7	10	A	H	1	3	9	10	F	F
3	10	5	10	A	H	1	4	8	10	F	F
3	11	6	10	A	H	2	5	9	10	F	F
3	12	6	10	A	H	2	6	9	10	F	F
4	13	10	10	A	H	2	7	8	10	F	F
4	14	9	10	A	H	2	8	9	10	F	F
4	15	8	10	A	H	3	9	8	10	F	F
4	16	8	10	A	H	3	10	8	10	F	F
5	17	6	10	A	H	3	11	8	10	F	F
5	18	7	10	A	H	3	12	7	10	F	F
5	19	5	10	A	H	4	13	8	10	F	F
5	20	8	10	A	H	4	14	8	10	F	F
1	1	7	10	H	A	4	15	8	10	F	F
1	2	6	10	H	A	4	16	9	10	F	F

5	17	10	10	F	F
5	18	9	10	F	F
5	19	10	10	F	F
5	20	8	10	F	F
1	1	9	10	H	H
1	2	7	10	H	H
1	3	8	10	H	H
1	4	7	10	H	H
2	5	8	10	H	H
2	6	6	10	H	H
2	7	8	10	H	H
2	8	7	10	H	H
3	9	10	10	H	H
3	10	7	10	H	H
3	11	9	10	H	H
3	12	8	10	H	H
4	13	8	10	H	H
4	14	10	10	H	H
4	15	8	10	H	H
4	16	9	10	H	H
5	17	10	10	H	H
5	18	10	10	H	H
5	19	8	10	H	H
5	20	9	10	H	H

Appendix A: Introduction to matrix algebra

Some basic definitions

Definition: A vector is an ordered set of numbers. Each number has a given position.

Example: $x = \begin{pmatrix} 5 \\ 3 \\ 8 \end{pmatrix}$ is a column vector with 3 elements.

Example: $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ is a column vector with n elements.

Definition: A matrix is a two-dimensional (rectangular) ordered set of numbers.

Example: $\mathbf{A} = \begin{pmatrix} 1 & 2 & 4 \\ 1 & 6 & 3 \end{pmatrix}$ is a matrix with two rows and three columns.

Example: $\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & & b_{1c} \\ b_{21} & b_{22} & & \\ & & \ddots & \\ b_{r1} & b_{r2} & & b_{rc} \end{pmatrix}$ is a matrix with r rows and c columns. The general element of the matrix $\mathbf{B}$ is b_{ij}. The first index denotes row, the second index denotes column.

Vectors are often written using lowercase symbols like $\mathbf{x}$, while matrices are often written using uppercase letters like $\mathbf{A}$. Both matrices and vectors are written in **bold**.

The dimension of a matrix

Definition: A matrix that has r rows and c columns is said to have dimension $r \times c$.

Definition: A column matrix with n rows has dimension $n \times 1$.

Definition: A row matrix with m columns has dimension $1 \times m$.

Definition: A scalar, i.e. a number, is a matrix that has dimension 1×1.

The transpose of a matrix

Transposing a matrix means to interchange rows and columns. If $\mathbf{A}$ is a matrix of dimension $r \times c$ then the transpose of $\mathbf{A}$ is a matrix of dimension $c \times r$. The transpose operator is denoted with a prime, $'$, so the transpose of $\mathbf{A}$ is denoted with $\mathbf{A}'$ (Some textbooks indicate a transpose by using the letter T).

For the elements $a_{ij}^{'}$ of $\mathbf{A}'$ it holds that

$$a_{ij}^{'} = a_{ji}$$

If $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ is a column vector, then $\mathbf{x}' = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix}$ is a row vector with n elements.

Example: The transpose of the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 4 \\ 1 & 6 & 3 \end{pmatrix}$$

is

$$\mathbf{A}' = \begin{pmatrix} 1 & 1 \\ 2 & 6 \\ 4 & 3 \end{pmatrix}.$$

Some special types of matrices

Definition: A matrix where the number of rows = number of columns (i.e. $r = c$) is a square matrix.

Definition: A square matrix that is unchanged when transposed is symmetric.

Example: The matrix $\mathbf{A} = \begin{pmatrix} 3 & 0 & -1 \\ 0 & 1 & 2 \\ -1 & 2 & 4 \end{pmatrix}$ is square and symmetric.

Definition: The elements a_{ii} in a square matrix are called the diagonal elements.

Definition: An identity matrix $\mathbf{I}$ is a symmetric matrix where all elements are 0, except that the diagonal elements are 1: $\mathbf{I} = \begin{pmatrix} 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ & & \ddots & \\ 0 & 0 & & 1 \end{pmatrix}$

Definition: A diagonal matrix is a matrix where all elements are 0, except for the diagonal elements: $\mathbf{D}(a_i) = \begin{pmatrix} a_1 & 0 & & 0 \\ 0 & a_2 & & 0 \\ & & \ddots & \\ 0 & 0 & & a_r \end{pmatrix}$.

Definition: A unit vector is a vector where all elements are 1: $\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$.

The transpose is $\mathbf{1}' = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}$.

Calculations on matrices

Addition, subtraction and multiplication can be defined for matrices.

Definition: Equality: Two matrices A and B with the same dimension $r \times c$ are equal if and only if $a_{ij} = b_{ij}$ for all i and j, i.e. if all elements are equal.

Definition: Addition: The sum of two matrices $\mathbf{A}$ and $\mathbf{B}$ that have the same dimension is the matrix that consists of the sum of the elements of $\mathbf{A}$ and $\mathbf{B}$.

Example: If $\mathbf{A} = \begin{pmatrix} 1 & 2 & 4 \\ 1 & 6 & 3 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 3 & 9 & 6 \\ 4 & 2 & 1 \end{pmatrix}$ then

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} 4 & 11 & 10 \\ 5 & 8 & 4 \end{pmatrix}.$$

Example: If $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{1c} \\ a_{21} & a_{22} & \\ & & \\ a_{r1} & a_{r2} & a_{rc} \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & b_{1c} \\ b_{21} & b_{22} & \\ & & \\ b_{r1} & b_{r2} & b_{rc} \end{pmatrix}$ then

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & a_{1c} + b_{1c} \\ a_{21} + b_{21} & a_{22} + b_{22} & \\ & & \\ a_{r1} + b_{r1} & a_{r2} + b_{r2} & a_{rc} + b_{rc} \end{pmatrix}.$$

Definition: Subtraction: Matrix subtraction is defined in an analogous way. It holds that

$$\begin{aligned} \mathbf{A} + \mathbf{B} &= \mathbf{B} + \mathbf{A} \\ \mathbf{A} + (\mathbf{B} + \mathbf{C}) &= (\mathbf{A} + \mathbf{B}) + \mathbf{C} \\ \mathbf{A} - (\mathbf{B} - \mathbf{C}) &= (\mathbf{A} - \mathbf{B}) + \mathbf{C} \end{aligned}$$

For matrices that do not have the same dimensions, addition and subtraction are not defined.

Matrix multiplication

Multiplication by a scalar

To multiply a matrix $\mathbf{A}$ by a scalar (= a number) c means that all elements in $\mathbf{A}$ are multiplied by c.

Example: If $\mathbf{A} = \begin{pmatrix} 1 & 2 & 4 \\ 1 & 6 & 3 \end{pmatrix}$ then $4 \cdot \mathbf{A} = \begin{pmatrix} 4 & 8 & 16 \\ 4 & 24 & 12 \end{pmatrix}$

Example: $k \cdot \mathbf{A} = \begin{pmatrix} k \cdot a_{11} & k \cdot a_{12} & k \cdot a_{1c} \\ k \cdot a_{21} & k \cdot a_{22} & \\ & & \\ k \cdot a_{r1} & k \cdot a_{r2} & k \cdot a_{rc} \end{pmatrix}$.

Multiplication by a matrix

Matrix multiplication of type $\mathbf{C} = \mathbf{A} \cdot \mathbf{B}$ is defined only if the number of columns in $\mathbf{A}$ is equal to the number of rows in $\mathbf{B}$. If $\mathbf{A}$ has dimension $p \times r$ and $\mathbf{B}$ has dimension $r \times q$ then the product $\mathbf{A} \cdot \mathbf{B}$ will have dimension $p \times q$. The elements of $\mathbf{C}$ are calculated as

$$c_{ij} = \sum_{k=1}^{r} a_{ik} b_{kj}.$$

Example: If $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ -1 & 0 & 1 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 6 & 5 & 4 \\ -1 & 1 & -1 \\ 0 & 2 & 0 \end{pmatrix}$ then $\mathbf{AB} =$
$\begin{pmatrix} 1\cdot 6+2\cdot(-1)+3\cdot 0 & 1\cdot 5+2\cdot 1+3\cdot 2 & 1\cdot 4+2\cdot(-1)+3\cdot 0 \\ -1\cdot 6+0\cdot(-1)+1\cdot 0 & -1\cdot 5+0\cdot 1+1\cdot 2 & -1\cdot 4+0\cdot(-1)+1\cdot 0 \end{pmatrix} =$
$\begin{pmatrix} 4 & 13 & 2 \\ -6 & -3 & -4 \end{pmatrix}$.

Calculation rules of multiplication

It holds that

$$\begin{aligned} \mathbf{A}(\mathbf{B}+\mathbf{C}) &= \mathbf{A}\cdot\mathbf{B}+\mathbf{A}\cdot\mathbf{C} \\ \mathbf{A}(\mathbf{B}\cdot\mathbf{C}) &= (\mathbf{A}\cdot\mathbf{B})\cdot\mathbf{C}. \end{aligned}$$

Note that in general, $\mathbf{AB} \neq \mathbf{BA}$. The order has importance for multiplication. In the expression $\mathbf{AB}$ the matrix $\mathbf{A}$ has been post-multiplied with the matrix $\mathbf{B}$. In the expression $\mathbf{BA}$ the matrix $\mathbf{A}$ has been pre-multiplied with the matrix $\mathbf{B}$. Note that $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.

Idempotent matrices

Definition: A matrix $\mathbf{A}$ is idempotent if $\mathbf{A}\cdot\mathbf{A} = \mathbf{A}$.

The inverse of a matrix

Definition: The inverse of a square matrix $\mathbf{A}$ is the unique matrix $\mathbf{A}^{-1}$ for which it holds that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. That is: the matrix multiplied with its inverse results in the unit matrix. (Note that the same rule holds for scalars: $3\cdot 3^{-1} = 3\cdot\frac{1}{3} = \frac{3}{3} = 1$).

Example: The inverse of the matrix $\mathbf{A} = \begin{pmatrix} 5 & 10 \\ 3 & 2 \end{pmatrix}$ is

$$\mathbf{A}^{-1} = \begin{pmatrix} -0.1 & 0.5 \\ 0.15 & -0.25 \end{pmatrix}.$$

To verify this we calculate

$$\begin{aligned}\mathbf{A}\cdot\mathbf{A}^{-1} &= \begin{pmatrix} 5 & 10 \\ 3 & 2 \end{pmatrix}\begin{pmatrix} -0.1 & 0.5 \\ 0.15 & -0.25 \end{pmatrix} \\ &= \begin{pmatrix} 5\cdot(-0.1)+10\cdot 0.15 & 5\cdot 0.5+10\cdot(-0.25) \\ 3\cdot(-0.1)+2\cdot 0.15 & 3\cdot 0.5+2\cdot(-0.25) \end{pmatrix} \\ &= \begin{pmatrix} 1.0 & 0 \\ 0 & 1.0 \end{pmatrix} = \mathbf{I}.\end{aligned}$$

It is possible that the inverse $\mathbf{A}^{-1}$ does not exist. $\mathbf{A}$ is then said to be singular. The following relations hold for inverses:

The inverse of a symmetric matrix is symmetric

$$(\mathbf{A}')^{-1} = \left(\mathbf{A}^{-1}\right)'.$$

The inverse of a product of several matrices is obtained by taking the product of the inverses, in opposite order:

$$(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}.$$

If c is a scalar different from zero, then

$$(c\mathbf{A})^{-1} = \frac{1}{c}\mathbf{A}^{-1}.$$

Generalized inverses

A matrix $\mathbf{B}$ is said to be a generalized inverse of the matrix $\mathbf{A}$ if $\mathbf{ABA} = \mathbf{A}$. The generalized inverse of a matrix $\mathbf{A}$ is denoted with $\mathbf{A}^{-}$. If $\mathbf{A}$ is nonsingular then $\mathbf{A}^{-} = \mathbf{A}^{-1}$. When $\mathbf{A}$ is singular, $\mathbf{A}^{-}$ is not unique. A generalized inverse of a matrix $\mathbf{A}$ can be calculated as

$$\mathbf{A}^{-} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'.$$

The rank of a matrix

Definition: Two vectors are linearly dependent if the elements of one vector are proportional to the elements of the other vector.

Example: If $\mathbf{x}' = \begin{pmatrix} 1 & 0 & 1 \end{pmatrix}$ and $\mathbf{y}' = \begin{pmatrix} 4 & 0 & 4 \end{pmatrix}$ then the vectors $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent.

Definition: A set of vectors are linearly independent if it is impossible to write any one of the vectors as a linear combination of the others.

Example: The vectors $\mathbf{t}' = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$, $\mathbf{u}' = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$ and $\mathbf{v}' = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}$ are linearly independent.

Definition: The degree of linear independence among a set of vectors is called the rank of the matrix that is composed by the vectors.

The following properties hold for the rank of a matrix:

The rank of $\mathbf{A}^{-1}$ is equal to the rank of $\mathbf{A}$.

The rank of $\mathbf{A}'\mathbf{A}$ is equal to the rank of $\mathbf{A}$ (It is also true that the rank of $\mathbf{A}\mathbf{A}'$ is equal to the rank of $\mathbf{A}$).

The rank of a matrix $\mathbf{A}$ does not change if $\mathbf{A}$ is pre- or postmultiplied with a nonsingular matrix.

Determinants

To each square matrix $\mathbf{A}$ belongs a unique scalar that is called the determinant of A. The determinant of A is written as $|\mathbf{A}|$. The determinant of a matrix of dimension n can be calculated as $|\mathbf{A}| = \sum (-1)^{\#(\pi(n))} \prod_{i=1}^{n} a_{\pi_i,i}$. Here, $\pi(n)$ denotes any permutation of the numbers $1, 2, \ldots n$. $\#\pi(n)$ denotes the number of inversions of a permutation $\pi(n)$. This is the number of exchanges of pairs of the numbers in $\pi(n)$ that are needed to bring them back into natural order. Determinants of small matrices can be calculated by hand, but for larger matrices we prefer to leave the work to computers.

If $\mathbf{A}$ is singular, then the determinant $|\mathbf{A}| = 0$.

Eigenvalues and eigenvectors

To each symmetric square matrix $\mathbf{A}$ of dimension $n \times n$ belongs n scalars that are called the eigenvalues of $\mathbf{A}$. These are solutions to the equation

$$|\mathbf{A} - \lambda\mathbf{I}| = \mathbf{0}.$$

The eigenvalues have the following properties:

The product of all eigenvalues of $\mathbf{A}$ is equal to $|\mathbf{A}|$.

The sum of all eigenvalues of $\mathbf{A}$ is equal to $tr(\mathbf{A})$, which is the sum of the diagonal elements of $\mathbf{A}$. The symbol $tr(\mathbf{A})$ can be read as "the trace of $\mathbf{A}$".

Some statistical formulas on matrix form

$$\mathbf{x}'\mathbf{x} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \sum_{i=1}^{n} x_i^2$$

$$\mathbf{x}'\mathbf{y} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \sum_{i=1}^{n} x_i y_i$$

$$\mathbf{1}'\mathbf{y} = \sum_{i=1}^{n} y_i \qquad \mathbf{1}'\mathbf{1} = n \qquad \mathbf{1}'\mathbf{y} n^{-1} = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y} = \overline{y}$$

Further reading

This chapter has only given a very brief and sketchy introduction to matrix algebra. A more complete treatment can be found in textbooks such as Searle (1982).

Appendix B: Inference using likelihood methods

The likelihood function

Suppose that we want to estimate the (single) parameter θ in some distribution. We assume that the distribution has some density function $f(x;\theta)$; we use the term "density function" whether x is continuous or discrete. We take a random sample of size n from the distribution and end up with the observation vector $\mathbf{x}' = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix}$.

The likelihood function of our sample is defined as

$$L = \prod_{i=1}^{n} f(x_i;\theta) \tag{B.1}$$

For discrete distributions, L is the probability of obtaining our sample. For continuous distributions we use the term "likelihood" rather than probability since the probability of obtaining any specified value of x is zero. In either case, L indicates how likely our sample is, given the value of θ.

The Maximum Likelihood estimator of θ is the value $\widehat{\theta}$ which maximizes the likelihood function L. This seems intuitively sensible: we choose as our estimator the value of θ for which our sample of observations is most likely.

In many cases it is more convenient to work with the log of the likelihood function. There are three reasons for this. First, the log function is monotone which means that L and $l = \log(L)$ have their maxima for the same parameter values. Secondly, the behavior of L can often be such that it is difficult numerically to find the maximum. Thirdly, if we take logs, we will replace the product sign with a summation sign which makes derivations somewhat easier. Thus, maximizing the likelihood (B.1) is equivalent to maximizing the log likelihood

$$l = \log(L) = \sum_{i=1}^{n} \log(f(x_i;\theta)) \tag{B.2}$$

with respect to θ. This is done by differentiating (B.2) with respect to θ. This gives the so called score equation

$$\frac{dl}{d\theta} = \frac{d\left(\sum_{i=1}^{n} \log\left(f\left(x_i;\theta\right)\right)\right)}{d\theta} = \sum_{i=1}^{n} \frac{f'\left(x_i;\theta\right)}{f\left(x_i;\theta\right)} = 0. \tag{B.3}$$

The Cramér-Rao inequality

We state without proof the following theorem: The variance of any unbiased estimator of θ must follow the Cramér-Rao inequality

$$Var\left(\widehat{\theta}\right) \geq \mathbf{I}_{\theta}^{-1} \tag{B.4}$$

where

$$\mathbf{I}_{\theta} = E\left[\left(\frac{d\log\left(L\left(\theta;x\right)\right)}{d\theta}\right)^{2}\right] = E\left[\left(\frac{dl}{d\theta}\right)^{2}\right] \tag{B.5}$$

$\mathbf{I}_{\theta}^{-1}$ is called the Cramér-Rao lower bound. $\mathbf{I}_{\theta}$ is called the Fisher information about θ. The connection between variance and information is that an estimator that has small variance gives us more information about the parameter.

Properties of Maximum Likelihood estimators

The following properties of Maximum Likelihood estimators hold under fairly weak regularity conditions:

Maximum Likelihood estimators can be biased or unbiased.

Maximum Likelihood estimators are consistent.

Maximum Likelihood estimators are asymptotically efficient.

Maximum Likelihood estimators are asymptotically normally distributed.

The asymptotic efficiency means that the variance of ML estimators approaches the Cramer-Rao lower bound as n increases. This means that, in large samples, we can regard $\widehat{\theta}$ as normally distributed with mean θ and variance $\mathbf{I}_{\theta}^{-1}$:

$$\widehat{\theta} \sim N\left(\theta, \mathbf{I}_{\theta}^{-1}\right).$$

Distributions with many parameters

So far, we have discussed Maximum Likelihood estimation of a single parameter θ. In the case where the distribution has, say, p parameters, the expressions we have given so far must be written as vectors and matrices. If we have an observation vector $\mathbf{x}$ of dimension $n \cdot 1$ and a parameter vector $\boldsymbol{\theta}$ of dimension $p \cdot 1$ the log likelihood equation can be written as

$$l = \log(L) = \sum_{i=1}^{n} \log\left(f\left(x_i; \boldsymbol{\theta}\right)\right). \tag{B.6}$$

l should be maximized with respect to all elements of $\boldsymbol{\theta}$. The set of p score equations is

$$\frac{\partial l}{\partial \theta_j} = \frac{\partial\left(\sum_{i=1}^{n} \log\left(f\left(x_i; \boldsymbol{\theta}\right)\right)\right)}{\partial \theta_j} = 0 \tag{B.7}$$

The asymptotic covariance matrix of $\boldsymbol{\theta}$ is the inverse of the Fisher information matrix $\mathbf{I}_\theta$ that has as its (j,k):th element

$$I_{j,k} = E\left[\left(\frac{\partial l}{\partial \theta_j}\right)\left(\frac{\partial l}{\partial \theta_k}\right)\right] \tag{B.8}$$

The Maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ of the parameter vector $\boldsymbol{\theta}$ is asymptotically multivariate Normal with mean $\boldsymbol{\theta}$ and covariance matrix $\mathbf{I}_\theta^{-1}$.

Numerical procedures

For complex distributions the score equations may be difficult to solve analytically. Numerical procedures have been developed that mostly, but not always, converge to the solution. Two commonly used procedures are the Newton-Raphson method and Fisher's method of scoring.

The Newton-Raphson method

We wish to maximize the log likelihood $l\left(\boldsymbol{\theta}; \mathbf{x}\right)$. Denote the vector of first derivatives of the log likelihood with respect to the elements of $\boldsymbol{\theta}$ with $\mathbf{g}\left(\boldsymbol{\theta}\right)$, and denote the matrix of second derivatives with $\mathbf{H}\left(\boldsymbol{\theta}\right)$. Thus, the (j,k):th element of $\mathbf{H}$ is $\partial^2 l / \partial\theta_j \partial\theta_k$. The matrix $\mathbf{H}$ is known as the Hessian matrix.

Suppose that we guess an initial estimate $\boldsymbol{\theta}_0$ of $\widehat{\boldsymbol{\theta}}$. The method works by a Taylor series expansion of $\mathbf{g}(\boldsymbol{\theta})$ around $\widehat{\boldsymbol{\theta}}$:

$$\mathbf{g}\left(\widehat{\boldsymbol{\theta}}\right) = \mathbf{g}(\boldsymbol{\theta}_0) + \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\mathbf{H}(\boldsymbol{\theta}_0).$$

Since $\mathbf{g}\left(\widehat{\boldsymbol{\theta}}\right) = 0$, this leads to a new approximation

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 - \mathbf{g}(\boldsymbol{\theta}_0)\mathbf{H}^{-1}(\boldsymbol{\theta}_0). \tag{B.9}$$

We can now substitute $\boldsymbol{\theta}_1$ for $\boldsymbol{\theta}_0$ in (B.9). We get a series of estimates $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, and so on until the process has converged.

Fisher's scoring

Fisher's scoring method is a variation of the Newton-Raphson method. The basic idea is to replace the Hessian matrix $\mathbf{H}$ with its expected value. It holds that $E[\mathbf{H}(\boldsymbol{\theta})] = -\mathbf{I}_\theta$, the Fisher information matrix.

There are two advantages to using the expected Hessian rather than the Hessian itself. First, it can be shown that

$$E\left(\frac{\partial^2 l}{\partial\theta_j\partial\theta_k}\right) = -E\left[\left(\frac{\partial l}{\partial\theta_j}\right)\left(\frac{\partial l}{\partial\theta_k}\right)\right] \tag{B.10}$$

Thus, to calculate the expected Hessian we do not need to evaluate the second order derivatives; it suffices to calculate the first-order derivatives of type $\frac{\partial l}{\partial\theta_j}$. A second advantage is that the expected Hessian is guaranteed to be positive definite so some non-convergence problems with the Newton-Raphson method do not occur. On the other hand, Fisher's scoring method often converges more slowly than the Newton-Raphson method. However, for distributions in the exponential family, the Newton-Raphson method and Fisher's scoring method are equivalent.

Fisher's scoring method can be regarded, at each step, as a kind of weighted least squares procedure. In the generalized linear model context, the method is also called Iteratively reweighted least squares.

Bibliography

[1] Aanes, W. A. (1961): Pingue (*Hymenoxys richardsonii*) poisoning in sheep. *American J. of Veterinary Research*, **22**, 47-52.

[2] Agresti, A. (1984): *Analysis of ordered categorical data*. New York, WIley.

[3] Agresti, A. (1990): *Categorical data analysis*. New York, Wiley.

[4] Agresti, A. (1996): *An introduction to categorical data analysis*. New York, Wiley.

[5] Aitkin, M. (1987): Modelling variance heterogeneity in normal regression using GLIM. *Applied statistics*, **36**, 332-339.

[6] Akaike, H. (1973): Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N. and Csàki, F. (eds): *Second international symposium on inference theory*, Budapest, Akadèmiai Kiadó, pp. 267-281.

[7] Andersen, E. B. (1980): *Discrete statistical models with social science applications*. Amsterdam, North-Holland.

[8] Anscombe, F. J. (1953): Contribution to the discussion of H. Hotelling's paper. *J. Roy. Stat. Soc, B*, **15**, 229-30.

[9] Armitage, P. and Colton, T. (1998): *Encyclopedia of Biostatistics*. Chichester, Wiley.

[10] Ben-Akiva, M. and Lerman, S. R. (1985): *Discrete choice analysis: Theory and application to travel demand*. Cambridge, MIT press.

[11] Box, G. E. P. and Cox, D. R. (1964): An analysis of transformations. *J. Roy. Stat. Soc., A,* **143**, 383-430.

[12] Breslow, N. R. and Clayton, D. G. (1993): Approximate inference in generalized linear mixed models. *JASA*, **88**, 9-25.

[13] Brown, B. W.: (1980): Prediction analysis for binary data. In: *Biostatistics Casebook*, Eds. R. J. Miller, B. Efron, B. Brown and L. E. Moses. New York, Wiley.

[14] Christensen, R. (1996): *Analysis of variance, design and regression.* London, Chapman & Hall.

[15] Cicirelli, M. F., Robinson, K. R. and Smith, L. D. (1983): Internal pH of Xenopus oocytes: a study of the mechanism and role of pH changes during meotic maturation. *Developmental Biology*, **100**, 133-146.

[16] Collett, D. (1991): *Modelling binary data.* London, Chapman and Hall.

[17] Cox, D. R. (1972): Regression models and life tables. *J. Roy. Stat. Soc, B,* **34**, 187-220.

[18] Cox, D. R. and Lewis, P. A. W. (1966): *The statistical analysis of series of events.* London, Chapman & Hall.

[19] Cox, D. R. and Snell, E. J. (1989): *The analysis of binary data*, 2nd ed. London, Chapman and Hall.

[20] Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994): *Analysis of longitudinal data.* Oxford: Clarendon press.

[21] Dobson, A. J. (2002): *An introduction to generalized linear models, second edition.* London: Chapman & Hall/CRC Press.

[22] Draper, N. R. and Smith, H. (1998): *Applied regression analysis, 3rd Ed.* New York, Wiley.

[23] Ezdinli, E., Pocock, S., Berard, C. W. et al (1976): Comparison of intensive versus moderate chemotherapy of lymphocytic lymphomas: a progress report. *Cancer*, **38**, 1060-1068.

[24] Fahrmeir, L. and Tutz, G. (1994; 2001): *Multivariate statistical modeling based on generalized linear models.* New York, Springer.

[25] Feigl, P. and Zelen, M. (1965): Estimation of exponential survival probabilities with concomitant information. *Biometrics*, **21**, 826-838.

[26] Finney, D. J. (1947, 1952): *Probit analysis. A statistical treatment of the sigmoid response curve.* Cambridge, Cambridge University Press.

[27] Freeman, D. H. (1987): *Applied categorical data analysis.* New York, Marcel Dekker.

[28] Gill, J. and Laughton, C. D. (2000): *Generalized linear models: a unified approach.* New York, Sage publications.

[29] Francis, B., Green, M. and Payne, C. (Eds.) (1993): *The GLIM system manual, Release 4*. London, Clarendon press.

[30] Haberman, S. (1978): *Analysis of qualitative data*. Vol. 1: Introductory topics. New York, Academic Press.

[31] Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. and Ostrowski, E. (1994): *A handbook of small data sets*. London, Chapman & Hall.

[32] Horowitz, J. (1982): An evaluation of the usefulness of two standard goodness-of-fit indicators for comparing non-nested random utility models. *Trans. Research Record*, 874, 19-25.

[33] Hosmer, D. W. and Lemeshow, S. (1989): *Applied logistic regression*. New York, Wiley.

[34] Hurn, M. W., Barker, N. W. and Magath, T. D. (1945): The determination of prothrombin time following the administration of dicumarol with specific reference to thromboplastin. *J. Lab. Clin. Med.*, **30**, 432-447.

[35] Hutcheson, G. D. (1999): *Introductory statistics using Generalized Linear Models*. New York, Sage publications.

[36] Jones, B. and Kenward, M. G.: *Design and analysis of cross-over trials*. London, Chapman and Hall.

[37] Jørgensen, B. (1987): Exponential dispersion models. *Journal of the Royal Statistical Society*, **B49**, 127-162.

[38] Klein, J. and Moeschberger, M. (1997): *Survival analysis: techniques for censored and truncated data*. New York, Springer.

[39] Koch, G. G. and Edwards, S. (1988): Clinical efficiacy trials with categorical data. In: *Biopharmaceutical statistics for drug development*, K. E. Peace, ed. New York, Marcel Dekker, pp. 403-451.

[40] Leemis, L. M. (1986): Relationships among common univariate distributions. *American Statistician*, **40**, 134-146.

[41] Liang, K-Y. and Zeger, S. L. (1986): Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

[42] Liang, K-Y and Hanfelt, J. (1994): On the use of the quasi-likelihood method in teratological experiments. Biometrics, **50**, 872-880.

[43] Lindahl, B., Stenlid, J., Olsson, S. and Finlay, R. (1999): Translocation of ^{32}P between interacting mycelia of a wood-decomposing fungus and ectomycorrhizal fungi in microcosm systems. *New Phytol.*, **144**, 183-193.

[44] Lindsey, J. K. (1997): *Applying generalized linear models*. New York, Springer.

[45] Lipsitz, S. H., Fitzmaurice, G. M., Orav, E. J. and Laird, N. M. (1994): Performance of generalized estimating equations in practical situations. *Biometrics*, **50**, 270-278.

[46] Liss, P., Nygren, A., Olsson, U., Ulfendahl, H. R. and Eriksson, U.: Effects of contrast media and Mannitol on renal medullary blood flow and renal blood cell aggregation in the rat kidney. *Kidney International*, 1996, **49**, 1268-1275.

[47] Littell, R. C., Milliken, G. A., Stroup, W. W. and Wolfinger, R. D. (1996): *SAS system for mixed models*. Cary, N. C., SAS Institute Inc.

[48] McCullagh, P. (1983): Quasi-likelihood functions. *Annals of Statistics*, **11**, 59-67.

[49] McCullagh, P. and Nelder, J. A. (1989): *Generalized Linear Models*. London, Chapman and Hall.

[50] Minitab Inc. (1998): *Minitab User's Guide, Release 12*. State College, Minitab Inc.

[51] Montgomery, D. C. (1984): *Design and analysis of experiments*. New York, Wiley.

[52] Nagelkerke, N. J. D. (1991): A note on a general definition of the coefficient of determination. *Biometrika*, **78**, 691-692.

[53] Nelder, J. (1971): Discussion on papers by Wynn, Bloomfield, O´Neill and Wetherill. *JRSS (B)*, 33, 244-246.

[54] Neveus T., Läckgren G., Tuvemo T., Olsson U. and Stenberg A. (1999): Desmopressin-resistant Enuresis: Pathogenetic and Therapeutic Considerations. *Journal of Urology*, 1999, **162**, 2136.

[55] Ninkovic, V., Olsson, U. and Pettersson, J. (2002). Mixing barley cultivars affect aphid host plant acceptance in field experiments. *Entomologia Experimentalis et Applicata*, in press.

[56] Norton, P. G. and Dunn, E. V. (1985): Snoring as a risk factor for disease. *British Medical Journal*, **291**, 630-632.

[57] Olsson, U. (2000): Estimation of the number of drug addicts in Sweden - an application of capture-recapture methodology. Swedish University of Agricultural Sciences, Department of Statistics, Report 55.

[58] Olsson, U. and Neveus, T. (2000): Generalized Linear Mixed Models used for Evaluating Enuresis Therapy. Swedish University of Agricultural Sciences, Department of Statistics, Report 54.

[59] Rea, T. M., Nash, J. F., Zabik, J. E., Born, G. S. and Kessler, W. V. (1984): Effects of Toulene inhalation on brain biogenic amines in the rat. *Toxicology*, **31**, 143-150.

[60] Rosenberg, L., Palmer, J. R., Kelly, J. P., Kaufman, D. W. and Shapiro, S. (1988): Coffee drinking and nonfatal myocaridal infarction in men under 55 years of age. *Am J. Epidemiol.*, **128**, 570-578.

[61] Samuels, M. and Witmer, J. A. (1999): *Statistics for the life sciences.* Upper Saddle River, NJ: Prentice-Hall.

[62] SAS Institute Inc. (1997): *SAS/Stat software: Changes and enhancements through release 6.12.* Cary, NC. SAS Institute Inc.

[63] SAS Institute Inc. (2000a): *JMP software, version 4.* Cary, NC. SAS Institute Inc.

[64] SAS Institute Inc. (2000b): *SAS/Stat user's guide, Version 8.* Cary, NC. SAS Institute Inc.

[65] Searle, S. R.: *Matrix Algebra Useful for Statistics.* New York, Wiley, 1982.

[66] Sen, A. and Srivastava, M. (1990): *Regression analysis. Theory, methods and applications.* New York, Springer.

[67] Snedecor, G. W. and Cochran, W. G. (1980): *Statistical methods*, 7th ed. Ames, Iowa, The Iowa State University Press.

[68] Socialdepartementet: *Tungt narkotikamissbruk - en totalundersökning 1979.* Rapport från utredningen om narkotikamissbrukets omfattning (UNO). Stockholm: Socialdepartementet (Ds S 1980:5). (Heavy drug use - a comprehensive survey; in Swedish).

[69] Sokal, R. R. and Rohlf, F. J. (1973): *Introduction to biostatistics.* San Fransisco, Freeman.

[70] Student (W. S. Gossett) (1907): On error of counting with an haemocytometer. *Biometrika*, **5**, 351-360.

[71] Wedderburn, R. W. M. (1974): Quasi-likelihood function, generalized linear models and the Gauss-Newton method. *Biometrika*, **61**, 439-477.

[72] Williams, D. A. (1982): Extra-binomial variation in linear logistic models. *Applied Statistics*, **31**, 144-148.

[73] Wolfinger, R. and O'Connell, M. (1993): Generalized linear models: a pseudo-likelihood approach. *J. Statist. Comput. Simul.*, **48**, 233-243.

[74] Zagal, E., Bjarnason, S. and Olsson, U. (1993): Carbon and nitrogen in the root-zone of Barley supplied with nitrogen fertilizer at two rates. *Plant and Soil*, **157**, 51-63.

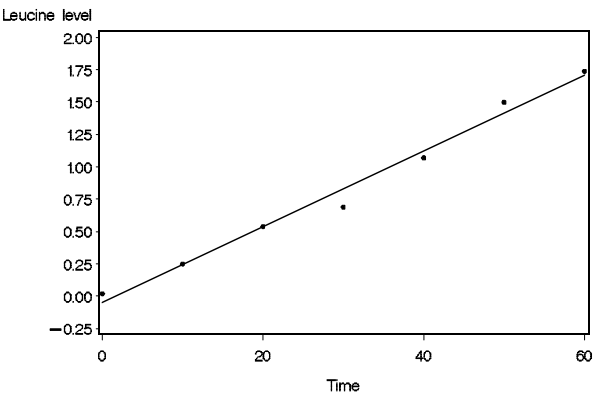
Solutions to the exercises

Exercise 1.1

A. The model can be written as $y_i = \alpha + \beta t_i + e_i$. $e_i \sim N\left(0; \sigma^2\right)$. A regression analysis, using the GLM procedure of the SAS package, gives the following results:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-.0475000000	0.05719172	-0.83	0.4441
time	0.0292500000	0.00158621	18.44	<.0001

B. The estimated regression equation is $\widehat{y} = -0.0475 + 0.02925t$. A plot of the data and the regression line is as follows.



C. The Anova table is

Dependent Variable: leucine Leucine level

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.39557500	2.39557500	340.04	<.0001
Error	5	0.03522500	0.00704500		
Corrected Total	6	2.43080000			

It can be concluded that the leucine level increases with time. The increase is nearly linear, in the studied time range.

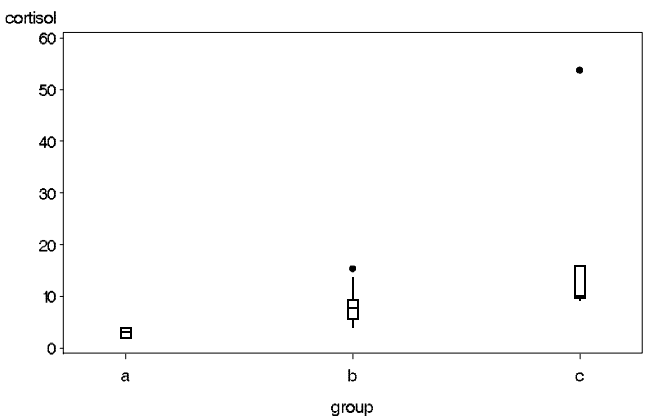
Exercise 1.2

A. This is a one-way Anova model: $y_{ij} = \mu + \alpha_i + e_{ij}$; $e_{ij} \sim N\left(0; \sigma^2\right)$. We wish to test the null hypothesis that there are no group differences, i.e. H_0: $\sum n_i \alpha_i^2 = 0$. The Anova table produced by Proc GLM is as follows:

```
Dependent Variable: cortisol

                                  Sum of
Source                     DF       Squares   Mean Square  F Value  Pr > F

Model                       2    795.692190    397.846095     4.44  0.0271
Error                      18   1614.017333     89.667630
Corrected Total            20   2409.709524


          R-Square     Coeff Var      Root MSE    cortisol Mean

          0.330203      100.3306      9.469299         9.438095


Source                     DF   Type III SS   Mean Square  F Value  Pr > F

group                       2   795.6921905   397.8460952     4.44  0.0271
```

The results suggest that there are significant differences between the groups ($p = 0.0271$). To study these differences we prepare a table of mean values, and a box plot:

```
The GLM Procedure

           Level of            ----------cortisol----------
           group         N             Mean          Std Dev

           a             6        2.9666667        0.9244818
           b            10        8.1800000        3.7891072
           c             5       19.7200000       19.2388149
```



B. The sample standard deviations are rather different in the different groups. Since the model assumes that the population variances are equal, the analysis presented above may not be the optimal one. The box plot suggests that one or two observations may be outliers.

Exercise 1.3

A. We want to compare two competing models:

i) $y_{ijk} = \mu + \alpha_i + \beta x_j + e_{ijk}$. Equal slopes (no interaction)

ii) $y_{ijk} = \mu + \alpha_i + \beta x_j + (\alpha\beta)_{ij} x_j + e_{ijk}$. Different slopes (interaction exists). The corresponding GLM outputs are presented below.

Model i)

```
Dependent Variable: co2

                                 Sum of
Source                    DF       Squares   Mean Square  F Value  Pr > F

Model                      2   2350.424299   1175.212150    44.93  <.0001
Error                     21    549.234956     26.154046
Corrected Total           23   2899.659255


            R-Square     Coeff Var      Root MSE      co2 Mean

            0.810586      19.53056      5.114103      26.18513


Source                    DF   Type III SS   Mean Square  F Value  Pr > F

level                      1    264.809910    264.809910    10.13  0.0045
days                       1   2085.614389   2085.614389    79.74  <.0001
```

Model ii)

```
Dependent Variable: co2

                                 Sum of
Source                    DF       Squares   Mean Square  F Value  Pr > F

Model                      3   2445.093241    815.031080    35.86  <.0001
Error                     20    454.566014     22.728301
Corrected Total           23   2899.659255


            R-Square     Coeff Var      Root MSE      co2 Mean

            0.843235      18.20660      4.767421      26.18513


Source                    DF   Type III SS   Mean Square  F Value  Pr > F

level                      1     47.785031     47.785031     2.10  0.1626
days                       1   2085.614389   2085.614389    91.76  <.0001
days*level                 1     94.668942     94.668942     4.17  0.0547
```

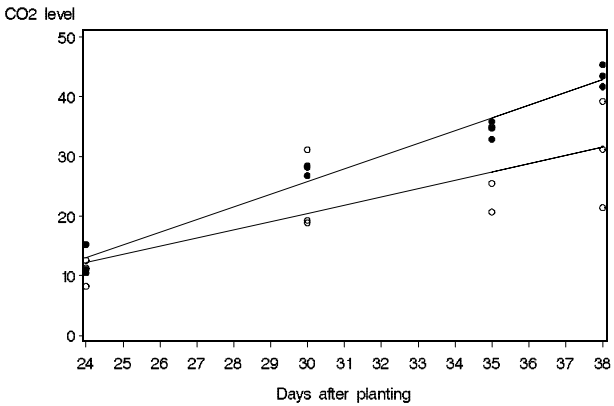
The test of parallelism is not significant ($p = 0.0547$). Still, for model building purposes, I would prefer to retain the interaction term in the model; see the discussion on model building strategy in the text. Thus, I would use Model 2 for interpretation and plotting.

B. Estimates of model parameters for model ii) are as follows:

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		-38.11803843 B	8.34443339	-4.57	0.0002
level	High	17.11095713 B	11.80081087	1.45	0.1626
level	Low	0.00000000 B	.	.	.
days		2.12991722 B	0.25921766	8.22	<.0001
days*level	High	-0.74816925 B	0.36658913	-2.04	0.0547
days*level	Low	0.00000000 B	.	.	.

Thus, the predicted value for High nitrogen level is $-38.1180+17.1110+2.1299 \cdot 35 - 0.7482 \cdot 35 = 27.353$. The predicted value for Low level is similarly $-38.1180 + 2.1299 \cdot 35 = 36.429$.

C. A graph of the model that does not assume parallel regression lines is:



D. There are strongly significant effects of time and of nitrogen level. The interaction, although not formally significant, indicates that the increase of CO2 emission may be somewhat faster for the low nitrogen treatment.

Exercise 1.4

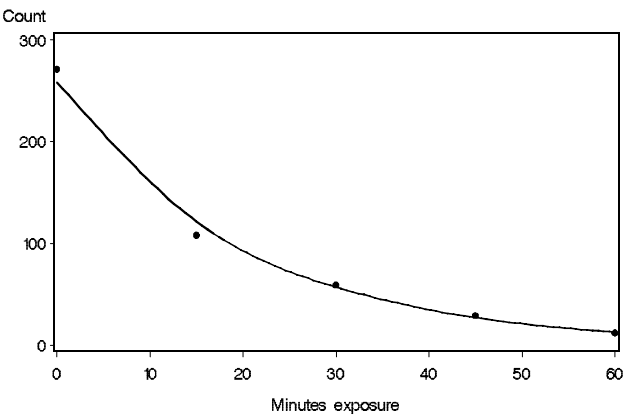
A. After taking logs of the count variable, a regression output is as follows:

```
Dependent Variable: logcount

                                  Sum of
Source                     DF       Squares   Mean Square  F Value  Pr > F

Model                       1    5.69913226    5.69913226   667.04  0.0001
Error                       3    0.02563161    0.00854387
Corrected Total             4    5.72476387


          R-Square     Coeff Var      Root MSE    logcount Mean

          0.995523      2.286363      0.092433          4.042798

Source                     DF   Type III SS   Mean Square  F Value  Pr > F
minutes                     1    5.69913226    5.69913226   667.04  0.0001


                                     Standard
    Parameter         Estimate           Error    t Value    Pr > |t|

    Intercept      5.552649942      0.07159833      77.55      <.0001
    minutes       -0.050328398      0.00194866     -25.83      0.0001
```

B. If we assume that there is a multiplicative residual in the original model, we get: $y = Ae^{-Bx} \cdot \epsilon$. This gives, after taking logs, $\log(y) = \log(A) - Bx + e$ (where $e = \log \epsilon$) which is a linear model.

There is a strong relationship between log(count) and time.

C. We prefer to make the graph on the original scale. Thus, we calculate predicted values and take the anti-logs of these. The corresponding graph is:



Exercise 2.1

We write the density as $f(x) = \lambda e^{-\lambda x} = e^{\log \lambda - \lambda x}$ which is an exponential family distribution. If we use $\theta = -\lambda$, then $b(\theta) = \log(-\theta)$,

$a(\phi) = 1$, and $c(y, \phi) = 0$. For the variance function, we find that $b' = \frac{d}{d\theta}(\log(-\theta)) = \frac{1}{\theta}$ and $b'' = \frac{d}{d\theta}\left(\frac{1}{\theta}\right) = -\frac{1}{\theta^2}$.

Exercise 2.2

A. We write the distribution as

$\frac{e^{-\lambda}\lambda^{y_i}}{(1-e^{-\lambda})y_i!} = \frac{e^{-\lambda}e^{y_i \ln \lambda}}{e^{\ln(1-e^{-\lambda})}e^{\ln(y_i!)}} = e^{\left[-\lambda + y_i \ln \lambda - \ln\left(1-e^{-\lambda}\right) - \ln(y_i!)\right]}$ which is an Exponential family with $\theta = \ln \lambda$, $a(\phi) = 1$, $b(\theta) = -\lambda - \ln\left(1 - e^{-\lambda}\right)$ and $c(y, \phi) = -\ln(y_i!)$.

B. If we insert $\lambda = e^{\theta}$ into the expression for $b(\cdot)$, we get $b(\theta) = -\exp(\theta) - \ln\left(1 - e^{-e^{\theta}}\right)$. The derivatives of b with respect to θ are:

$b' = \frac{d}{d\theta}\left(-\exp(\theta) - \ln\left(1 - e^{-e^{\theta}}\right)\right) = \frac{e^{\theta}}{-1+e^{-e^{\theta}}}$

$b'' = \frac{d}{d\theta}\left(\frac{e^{\theta}}{-1+e^{-e^{\theta}}}\right) = -\frac{e^{\theta}-e^{\theta-e^{\theta}}-e^{2\theta-e^{\theta}}}{\left(-1+e^{-e^{\theta}}\right)^2}$ which is the required variance function.

Exercise 2.3

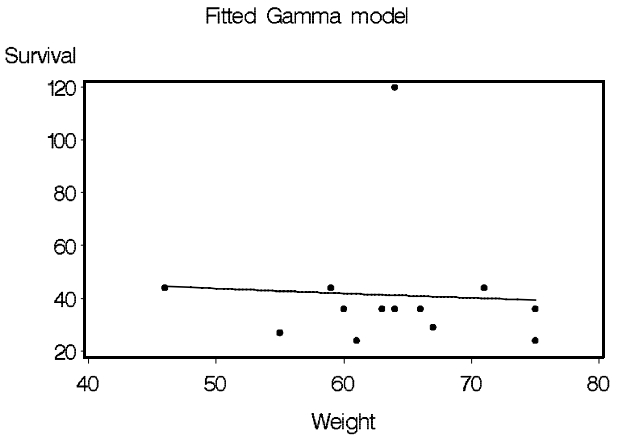
It is not easy to find a well-fitting model for these data. One of the best models is probably the one with a Gamma distribution and an inverse link, but other models might also be considered. However, most models we have tried do not indicate any significant relation between weight and survival:

Criterion	DF	Value	Value/DF
Deviance	11	2.5315	0.2301
Scaled Deviance	11	13.4077	1.2189
Pearson Chi-Square	11	4.3154	0.3923
Scaled Pearson X2	11	22.8557	2.0778
Log Likelihood		-55.0956	

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	0.0178	0.0239	-0.0291	0.0647	0.56	0.4560
weight	1	0.0001	0.0004	-0.0006	0.0008	0.07	0.7878
Scale	1	5.2964	2.0154	2.5123	11.1655		

A graph of the data and the fitted line may explain why: one observation has an unusually long survival time. Since we have no other information about the data, deletion of this observation cannot be justified.

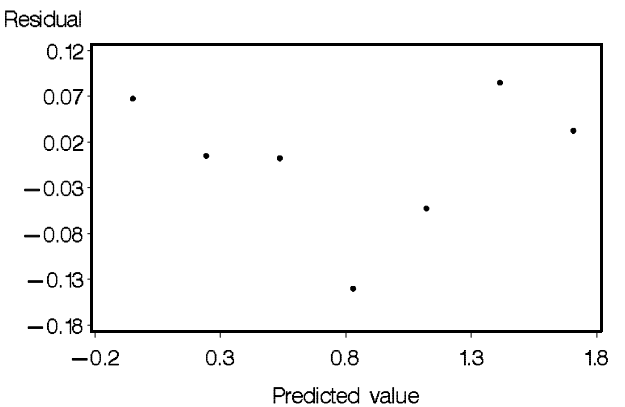


Exercise 3.1

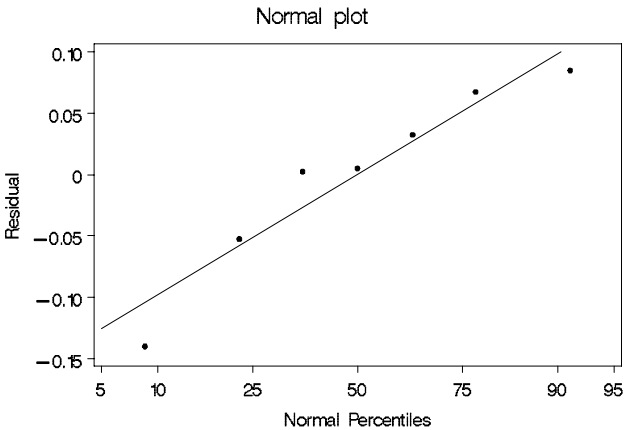
A. Data, predicted values and residuals are:

Obs	time	leucine	pred	res
1	0	0.02	-0.0475	0.0675
2	10	0.25	0.2450	0.0050
3	20	0.54	0.5375	0.0025
4	30	0.69	0.8300	-0.1400
5	40	1.07	1.1225	-0.0525
6	50	1.50	1.4150	0.0850
7	60	1.74	1.7075	0.0325

B. A plot of residuals against fitted values indicates no serious deviations from homoscedasticity, but this is difficult to see in such a small data set.



C. The Normal probability plot was obtained using Proc Univariate in SAS:



D. The influence diagnostics can be obtained from Proc Reg:

Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	------DFBETAS----- Intercept	time
1	0.0675	1.1284	0.4643	1.6783	1.0504	1.0504	-0.8740
2	0.005000	0.0631	0.2857	2.1832	0.0399	0.0391	-0.0282
3	0.002500	0.0294	0.1786	1.9014	0.0137	0.0119	-0.0061
4	-0.1400	-2.7205	0.1429	0.2244	-1.1106	-0.6161	0.0000
5	-0.0525	-0.6490	0.1786	1.5570	-0.3026	-0.0375	-0.1353
6	0.0850	1.2694	0.2857	1.1116	0.8028	-0.1574	0.5677
7	0.0325	0.4870	0.4643	2.5993	0.4534	-0.1744	0.3772

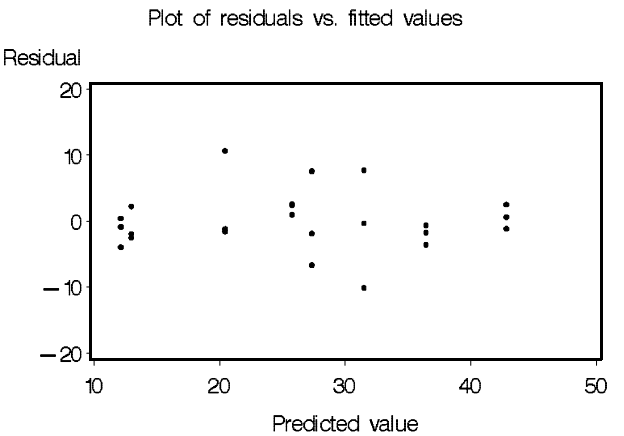
Since there are $n = 7$ observations and $p = 2$ parameters the average leverage is $2/7 = 0.286$. The rule of thumb that an observation is influential if $h > 2 \cdot p/n$ would suggest that observations with $h > 0.571$ are influential. None of the observations have a "Hat Diag" value above this limit.

Exercise 3.2

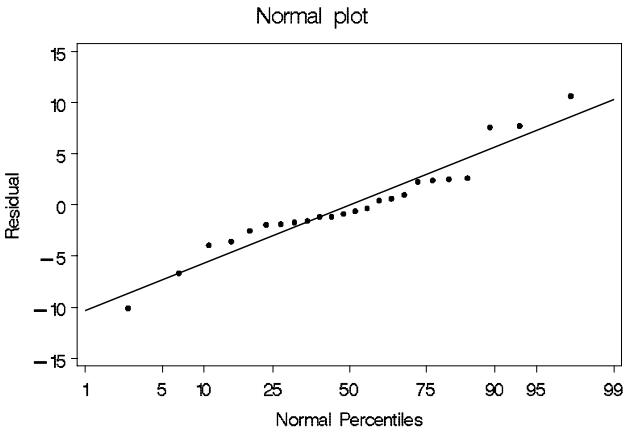
A. and D. Data, predicted values, and leverage values (diagonal elements from the Hat matrix) are as follows:

Obs	LEVEL	days	Co2	pred	res	hat
1	H	24	8.220	12.1549	-3.9349	0.26090
2	H	24	12.594	12.1549	0.4391	0.26090
3	H	24	11.301	12.1549	-0.8539	0.26090
4	L	24	15.255	13.0000	2.2550	0.26090
5	L	24	11.069	13.0000	-1.9310	0.26090
6	L	24	10.481	13.0000	-2.5190	0.26090
7	H	30	19.296	20.4454	-1.1494	0.09239
8	H	30	31.115	20.4454	10.6696	0.09239
9	H	30	18.891	20.4454	-1.5544	0.09239
10	L	30	28.200	25.7795	2.4205	0.09239
11	L	30	26.765	25.7795	0.9855	0.09239
12	L	30	28.414	25.7795	2.6345	0.09239
13	H	35	25.479	27.3541	-1.8751	0.11456
14	H	35	34.951	27.3541	7.5969	0.11456
15	H	35	20.688	27.3541	-6.6661	0.11456
16	L	35	32.862	36.4291	-3.5671	0.11456
17	L	35	34.730	36.4291	-1.6991	0.11456
18	L	35	35.830	36.4291	-0.5991	0.11456
19	H	38	31.186	31.4993	-0.3133	0.19882
20	H	38	39.237	31.4993	7.7377	0.19882
21	H	38	21.403	31.4993	-10.0963	0.19882
22	L	38	41.677	42.8188	-1.1418	0.19882
23	L	38	43.448	42.8188	0.6292	0.19882
24	L	38	45.351	42.8188	2.5322	0.19882

B. The plot of residuals against fitted values shows no large differences in variance:



C. The Normal probability plot has a slight "bend":



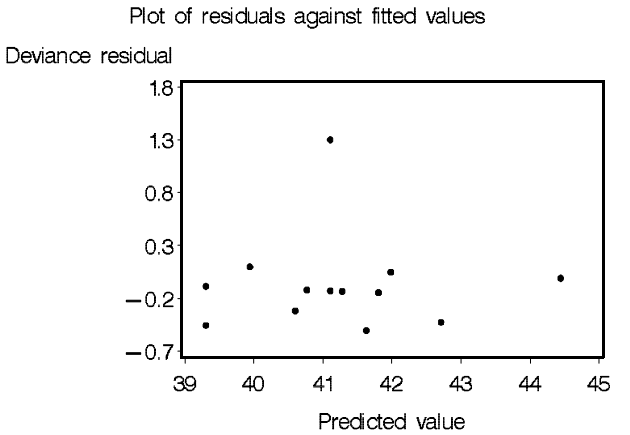
The limit for influential observations is $2 \cdot p/n = 2 \cdot 4/24 = 0.333$. The Hat values of all observations are below this limit.

Exercise 3.3

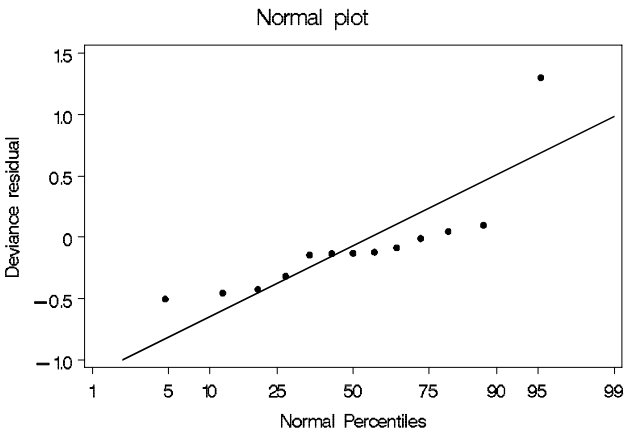
A. Predicted values and deviance residuals are as follows:

Obs	weight	survival	pred	res
1	46	44	44.4434	-0.01001
2	55	27	42.7112	-0.42609
3	61	24	41.6295	-0.50452
4	75	24	39.3067	-0.45590
5	64	36	41.1089	-0.12984
6	75	36	39.3067	-0.08661
7	71	44	39.9435	0.09831
8	59	44	41.9839	0.04727
9	64	120	41.1089	1.30216
10	67	29	40.6013	-0.31864
11	60	36	41.8060	-0.14589
12	63	36	41.2810	-0.13383
13	66	36	40.7691	-0.12188

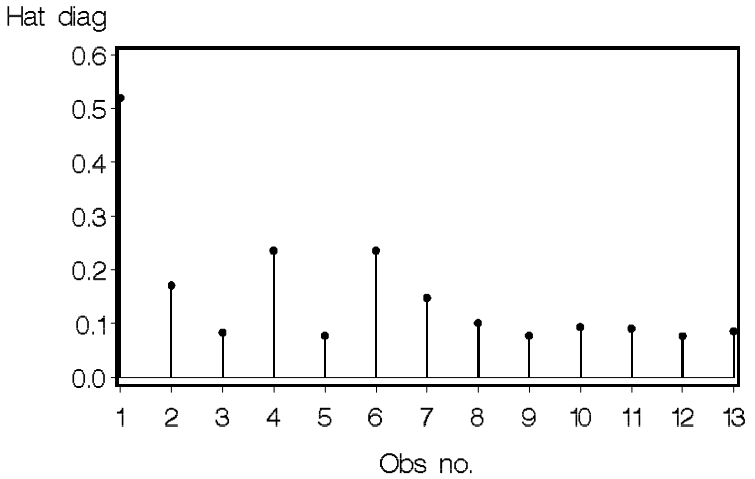
B. In the plot of residuals against fitted values, one observation stands out as a possible outlier:



C. The long-living sheep is an outlier in the Normal probability plot as well:



D. The influence of each observation can be obtained via the Insight procedure. A plot of hat diagonal values against observation number is as follows:



The "influence limit" is $2 \cdot p/n = 2 \cdot 2/13 = 0.308$. The first observation is influential according to this criterion.

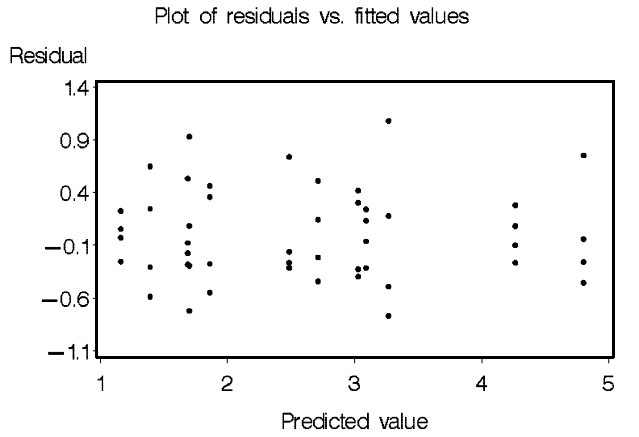
Exercise 4.1

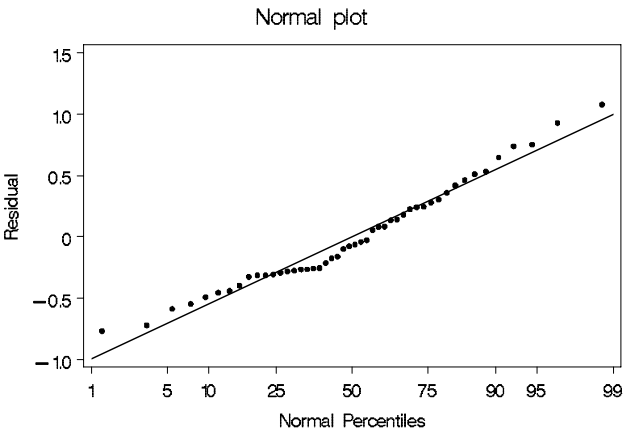
An analysis on the transformed data using a two-factor model with interaction gives the following edited output:

Source	DF	Squares	Mean Square	F Value	Pr > F
treatment	3	20.41428935	6.80476312	28.34	<.0001
poison	2	34.87711982	17.43855991	72.63	<.0001
treatment*poison	6	1.57077226	0.26179538	1.09	0.3867
Error	36	8.64308307	0.24008564		
Corrected Total	47	65.50526450			

R-Square	Coeff Var	Root MSE	z Mean
0.868055	18.68478	0.489985	2.622376

The effects of treatment and of poison are highly significant; there is no significant interaction. The residual plots ($\widehat{e}$ against $\widehat{y}$; Normal probability plot) seem to indicate that the data agree fairly well with the assumptions:





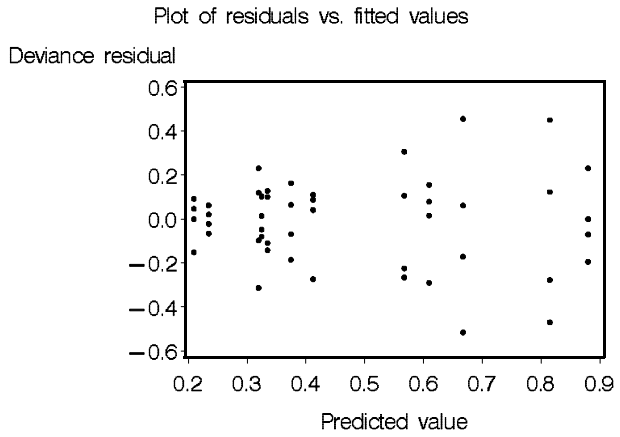
The same model analyzed as a generalized linear model with a Gamma distribution gives the following results:

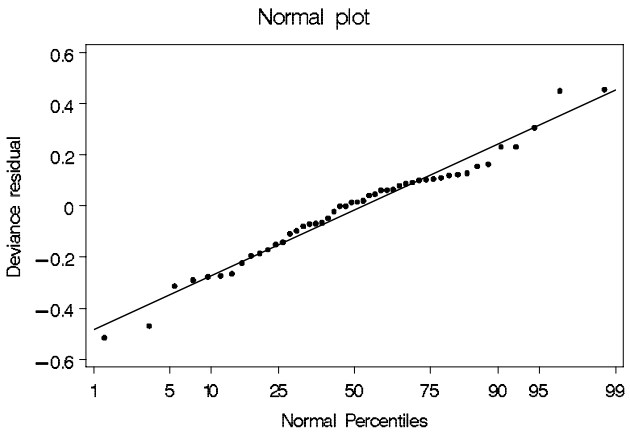
Criterion	DF	Value	Value/DF
Deviance	36	1.9205	0.0533
Scaled Deviance	36	48.3179	1.3422
Pearson Chi-Square	36	1.8755	0.0521
Scaled Pearson X2	36	47.1866	1.3107
Log Likelihood		50.0573	

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
treatment	3	43.76	<.0001
poison	2	59.31	<.0001
treatment*poison	6	10.04	0.1232

The conclusions are the same: significant effects of treatment and poison, no significant interaction. The residual plots for this model are:





The distribution of the deviance residuals is close to normal, but the Gamma model seems to produce residuals for which the variance increases slightly with increasing $\widehat{y}$.

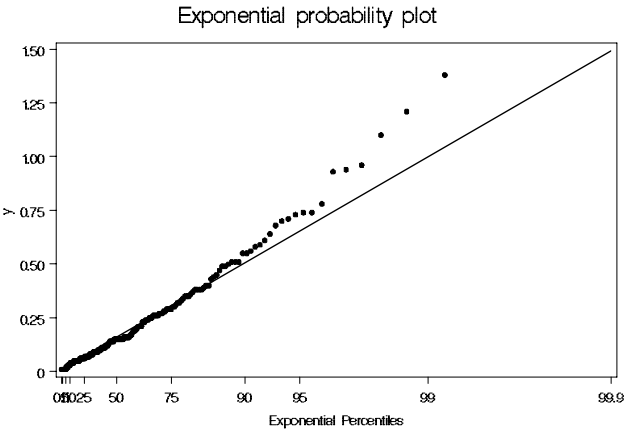
Exercise 4.2

The exponential distribution is a special case of the gamma distribution, with scale parameter equal to 1. Such a model fits these data reasonably well, according to the fit statistics:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	199	210.6205	1.0584
Scaled Deviance	199	210.6205	1.0584
Pearson Chi-Square	199	218.3507	1.0972
Scaled Pearson X2	199	218.3507	1.0972
Log Likelihood		98.8650	

An easy way of judging the fit to an exponential distribution is to ask Proc Univariate to produce an exponential probability plot:



It seems that the data deviate somewhat from an exponential distribution in the upper tail of the distribution.

Exercise 5.1

One questions with these data is whether we should include the factors Exposure, Temperature and Humidity as "class" variables or as numeric variables. One approach is to compare deviances for the different approaches, for a main effects model:

Types of terms	Deviance	df	D/df
All "class"	30.865	86	0.3582
Temperature numeric	31.2108	87	0.3587
Humidity also numeric	32.1509	89	0.3612
Also Exposure numeric	55.0698	91	0.6052

Treating temperature as numeric costs $31.2108 - 30.865 = 0.3458$ on 1 df, clearly an unsignificant loss. Similarly, adding Humidity as a numeric factor gives $32.1509 - 31.2108 = 0.9401$ on 2 df, which is clearly nonsignificant. On the other hand, when we treat Exposure as numeric and linear, we lose $55.0698 - 32.1509 = 22.919$ on 2 df, so this approximation is not worthwhile. We could use a quadratic term for exposure, but we might as well keep it as a class variable. Many models for these data that include interactions lead to a Hessian matrix that is not positive definite. However, when some factors are included as numeric variables, most interactions can indeed be estimated. p-values for two-way interactions are Exposure*Humidity ($p = 0.9834$); Species*exposure ($p = 0.9676$); Temp*exposure ($p = 0.3279$); Temp*humidity ($p = 0.6625$);

Temp*species ($p = 0.9091$); and Humidity*species ($p = 0.3006$). There does not seem to be any need to include interactions. It is interesting to note that an "old-fashioned" Anova on $\arcsin\left(\sqrt{\widehat{p}}\right)$ suggests that the interactions species*exposure, temp*exposure and humidity*exposure are indeed significant. The generalized linear model approach may suffer from the fact that 41 of the 96 observations have $\widehat{p} = 0$.

The model with only main effects, and with humidity and temperature used as numeric variables, gives the following results:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	89	32.1509	0.3612
Scaled Deviance	89	32.1509	0.3612
Pearson Chi-Square	89	27.9761	0.3143
Scaled Pearson X2	89	27.9761	0.3143
Log Likelihood		-534.9172	

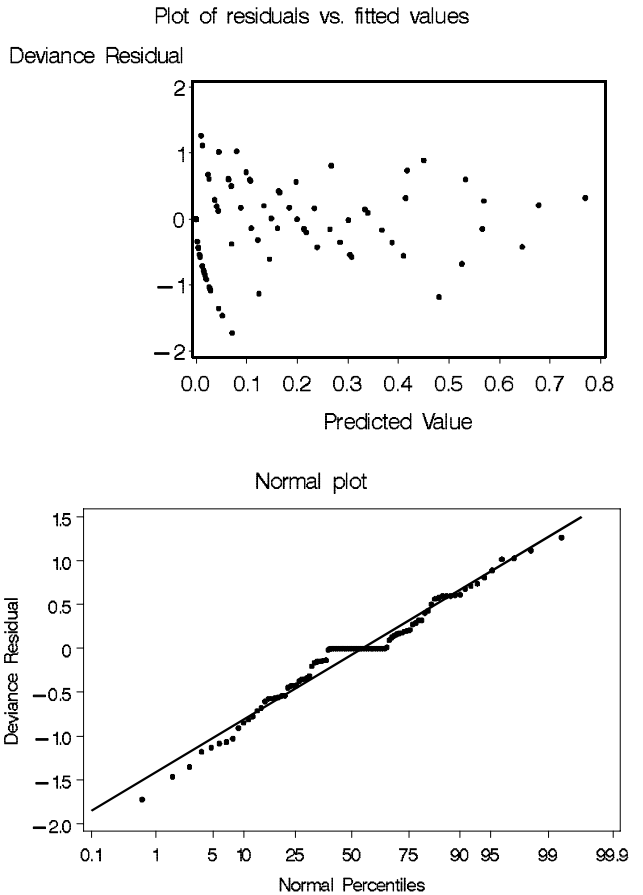
Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	5.6733	0.9636	3.7847	7.5618	34.67	<.0001
Species	A	1	-1.2871	0.1616	-1.6038	-0.9703	63.44	<.0001
Species	B	0	0.0000	0.0000	0.0000	0.0000	.	.
Exposure	1	1	-26.6441	32311.10	-63355.2	63301.95	0.00	0.9993
Exposure	2	1	-3.1793	0.2988	-3.7649	-2.5937	113.24	<.0001
Exposure	3	1	-0.9434	0.1633	-1.2635	-0.6232	33.35	<.0001
Exposure	4	0	0.0000	0.0000	0.0000	0.0000	.	.
Humidity		1	-0.1054	0.0138	-0.1324	-0.0784	58.56	<.0001
Temp		1	0.0930	0.0192	0.0555	0.1305	23.59	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Species	1	69.34	<.0001
Exposure	3	385.00	<.0001
Humidity	1	63.52	<.0001
Temp	1	24.38	<.0001

The survival is highly related to all four factors. Residual plots for this model are as follows:



Leverage diagnostics, in terms of diagonal elements of the Hat matrix, can be obtained e.g. from the Insight procedure but are not listed here in order to save space.

Exercise 5.2

The inferential aspects of this exercise are interesting: to which population could we generalize the results? However, we set this question aside. One question in this data set is how to model Age. The relation between age and survival can be explored by plotting proportion survival against sex and age (in 10-year intervals). The resulting plot is as follows:

It seems that survival probability for women is high, and increases with age, whereas only the young boys were rescued ("women and children first"). One possibility to modeling is to use a dummy variable for children under 10, and to use a linear age relation for ages above 10 years. If the dummy variable for childhood is d, a model for these data can be written as

$$\begin{aligned}\text{logit}(\widehat{p}) &= \beta_0 + \beta_1 \cdot sex + \beta_2 \cdot pclass \\ &\quad + d\left(\beta_2 + \beta_3 \cdot age + \beta_4 \cdot sex + \beta_5 \cdot age \cdot sex + \beta_6 \cdot pclass \cdot sex\right).\end{aligned}$$

This model assumes a separate survival probability for boys and girls below 10, and a linear change in survival probability (different for males and females) for persons above 10 years. This model fits fairly well to the data, as judged by Deviance/df:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	744	619.9224	0.8332
Scaled Deviance	744	619.9224	0.8332
Pearson Chi-Square	744	732.8209	0.9850
Scaled Pearson X2	744	732.8209	0.9850
Log Likelihood		-309.9612	

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Pclass	2	20.00	<.0001
Sex	1	0.50	0.4777
d	1	3.27	0.0705
d*Age	1	4.94	0.0262
d*Sex	1	7.72	0.0055
d*Age*Sex	1	2.47	0.1158
d*Pclass*Sex	4	33.19	<.0001

As an interpretation of the parameter estimates: there is a highly significant effect of passenger class, as well as an interaction between class and sex for persons above 10 years. Sex (which actually should be interpreted as sex of a child) is not significant: young boys and girls have similar survival probabilities. The fact that d*Sex is significant means that there are differences in survival for males and females above 10 years. In this analysis, passengers with missing age data have been excluded. However, there seems to be a relation between missing age and passenger class: age data are missing for 30% of first class passengers, 24% of second class passengers but 55% of third class passengers, so the analysis should be interpreted with care.

Exercise 5.3

A binomial model with treatment, ln(dose) and their interaction as factors produces the following results:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	11	22.7228	2.0657
Scaled Deviance	11	22.7228	2.0657
Pearson Chi-Square	11	20.5940	1.8722
Scaled Pearson X2	11	20.5940	1.8722
Log Likelihood		-368.7886	

LR Statistics For Type 3 Analysis

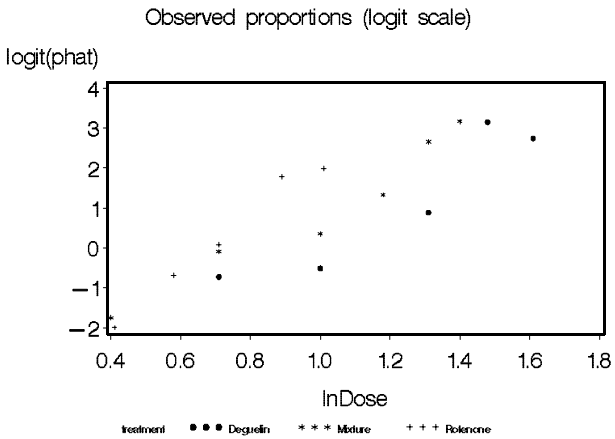
Source	DF	Chi-Square	Pr > ChiSq
treatment	2	3.66	0.1601
lnDose	1	287.20	<.0001
lnDose*treatment	2	9.25	0.0098

This analysis suggests that there is a significant interaction between treatment and ln(dose), i.e. that the slopes may be different. However, the fit of the model is not perfect: Deviance/df=2.07. If we fit the same model, but this time allowing the program to estimate the scale parameter, we get:

LR Statistics For Type 3 Analysis

Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
treatment	2	11	0.89	0.4395	1.77	0.4120
lnDose	1	11	139.03	<.0001	139.03	<.0001
lnDose*treatment	2	11	2.24	0.1529	4.48	0.1066

The p-values are rather sensitive to overdispersion. This analysis suggests that the interaction is not significant, i.e. that the slopes may be equal. The observed proportions (on a logit scale) plotted against ln(dose) are:



Exercise 5.4

A. $H_0 : \beta_{p*a} = 0$ against $H_1 : \beta_{p*a} \neq 0$ can be tested using the deviances. The test statistic is $(D_1 - D_2) / (df_1 - df_2)$ which, under H_0, is asymptotically distributed as χ^2 on $(df_1 - df_2)$ degrees of freedom. The condition that model 1 is nested within model 2 is fulfilled. Assumptions: Independent observations; large sample. Result: $(226.5177 - 226.4393) / (8 - 7) = 0.0784$ which should be compared with χ^2 on 1 d.f. The 5% limit is 3.841; the 1% limit is 6.635 and the 0.1% limit is 10.828. Our result is clearly non-significant; H_0 cannot be rejected.

B. $H_0 : \beta_{p*r} = 0 \quad H_1 : \beta_{p*r} \neq 0$ can similarly be tested using the deviances. The test statistic is $(D_1 - D_3) / (df_1 - df_3)$ which, under H_0, is asymptotically distributed as χ^2 on $(df_1 - df_3)$ degrees of freedom. The condition that model 1 is nested within model 3 is fulfilled. Assumptions: Independent observations; large sample. Result:

$(226.5177 - 216.4759) / (8 - 7) = 10.042$ which should be compared with χ^2 on 1 d.f. The 5% limit is 3.841; the 1% limit is 6.635 and the 0.1%

limit is 10.828. Our result is significant at the 1% level but not at the 0.1% level. H_0 is rejected.

C. The logit link is $\log \frac{p}{(1-p)}$. When p is zero (or one) this is not defined. The four cells with observed count =0 do not contribute to the likelihood. When we replace 0 with 0.5 in these cells they are included, so we get an extra four d.f. compared with model 3.

D. The odds ratios of not being infected are calculated as e^{β}. The corresponding odds ratios of being infected are the inverses of these. This gives:

Planned: OR=$e^{-0.8311} = 0.436$; OR of infection=$1/0.436 = 2.294$ Antibio: OR=$e^{3.4991} = 33.086$;OR of infection=$1/33.086 = 0.030$ Risk: OR=$e^{-3.7172} = 0.024$; OR of infection=$1/0.024 = 41.667$ Planned*Risk: $e^{2.4394} = 11.466$; OR of infection=$1/11.466 = 0.087$.

In the presence of interactions, raw Odds ratios are not very informative. One might consider to calculate odds ratios separately for each cell of the $2 \cdot 2 \cdot 2$ cross-table. All odds ratios take one cell as the baseline, with OR=1. We might use the cell Planned=0, Risk=0, Antibio=0 as a baseline. The remaining cell odds ratios (of no infection) compared to this baseline are:

	Planned			
	1		0	
	Risk		Risk	
Antibio	1	0	1	0
1	4.02	14.41	0.80	33.09
0	0.12	0.43	0.02	1.00

E. Remember that the observations, in this example, are binary, i.e. $y = 1$ and $y = 0$ are the only possible values of y. The first data line has Planned=1, Antibio=1, Risk=1 and Infection=1. Using the parameter estimates, we get for this observation $\text{logit}(\mu) = 2.1440 - 0.8311 + 3.4991 - 3.7172 = 3.5342$. Using the inverse logit transformation $\frac{e^x}{1+e^x}$ this corresponds to $\widehat{y} = \frac{e^{3.5342}}{1+e^{3.5342}} = 0.9717$ which is the predicted value. The raw residual is $y - \widehat{y} = 1 - 0.9716 = 0.0284$.

For the second observation the predictors have the same value but $y = 0$ so the raw residual is $0 - 0.9716 = -0.9716$.

The third and fourth observations have the same predicted values, obtained through $\text{logit}(\mu) = 2.1440 - 0.8311 + 3.4991 = 4.812$ which gives predicted value $\widehat{y} = \frac{e^{4.812}}{1+e^{4.812}} = 0.9919$ and raw residuals $1 - 0.9919 = 0.0081$ and $0 - 0.9919 = -0.9919$, respectively.

Note that the counts (Wt) are not the values to predict!

Exercise 5.5

A. The model is $g = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma z_{ijk} + (\beta\gamma)_j z_{ijk} + e_{ijk}$, $i = 1, 2$; $j = 1, 2, 3$; $k = 1, 2, 3$. This gives the model in matrix terms as $\mathbf{y} = \mathbf{XB} + \mathbf{e}$, where

$$\mathbf{X} = \begin{bmatrix}
1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & z_1 & z_1 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & z_2 & z_2 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & z_3 & z_3 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & z_4 & 0 & z_4 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & z_5 & 0 & z_5 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & z_6 & 0 & z_6 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & z_7 & 0 & 0 & z_7 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & z_8 & 0 & 0 & z_8 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & z_9 & 0 & 0 & z_9 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & z_{10} & z_{10} & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & z_{11} & z_{11} & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & z_{12} & z_{12} & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & z_{13} & 0 & z_{13} & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & z_{14} & 0 & z_{14} & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & z_{15} & 0 & z_{15} & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & z_{16} & 0 & 0 & z_{16} \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & z_{17} & 0 & 0 & z_{17} \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & z_{18} & 0 & 0 & z_{18}
\end{bmatrix};$$

$$\mathbf{B} = \begin{bmatrix}
\mu \\
\alpha_1 \\
\alpha_2 \\
\beta_1 \\
\beta_2 \\
\beta_3 \\
(\alpha\beta)_{11} \\
(\alpha\beta)_{12} \\
(\alpha\beta)_{13} \\
(\alpha\beta)_{21} \\
(\alpha\beta)_{22} \\
(\alpha\beta)_{23} \\
\gamma \\
(\beta\gamma)_1 \\
(\beta\gamma)_2 \\
(\beta\gamma)_3
\end{bmatrix}$$

B. The inverse of the logit link $g(p) = \log \frac{p}{1-p}$ is $g^{-1} = \frac{e^p}{e^p+1}$.

Exercise 6.1

A model with a Poisson distribution and a log link gives the following model information:

Criteria For Assessing Goodness Of Fit

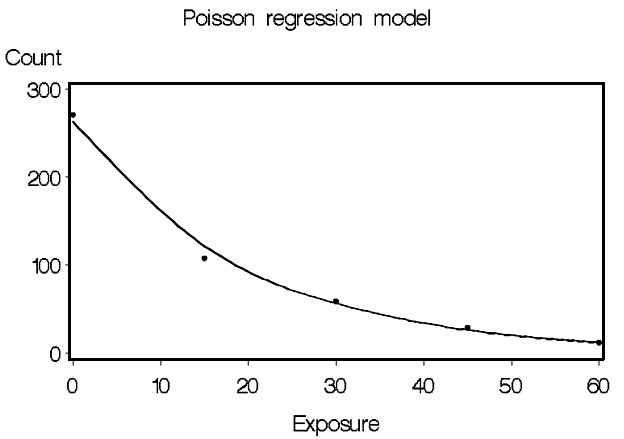
Criterion	DF	Value	Value/DF
Deviance	3	2.2906	0.7635
Scaled Deviance	3	2.2906	0.7635
Pearson Chi-Square	3	2.2453	0.7484
Scaled Pearson X2	3	2.2453	0.7484
Log Likelihood		1911.7443	

The fit of the model to the data is good, as judged by deviance/df. The parameter estimates are as follows:

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	5.5713	0.0567	5.4602	5.6825	9650.28	<.0001
exposure	1	-0.0513	0.0030	-0.0572	-0.0455	298.25	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

A plot of observed counts along with the fitted function indicates a good fit:



Exercise 6.2

Two Poisson models were fitted to the data: one with a log link, another with an identity link. The model with a log link fitted marginally better, as judged by the deviance/df criterion. First the log link results:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	6	4.0033	0.6672
Scaled Deviance	6	4.0033	0.6672
Pearson Chi-Square	6	3.9505	0.6584
Scaled Pearson X2	6	3.9505	0.6584
Log Likelihood		362.7354	

Analysis Of Parameter Estimates

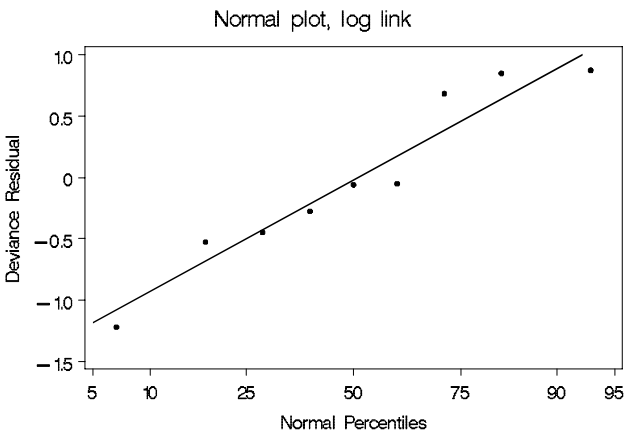
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	2.1752	0.2555	1.6745	2.6759	72.50	<.0001
mode1	1	0.0070	0.0024	0.0023	0.0118	8.34	0.0039
mode2	1	0.0025	0.0028	-0.0030	0.0081	0.81	0.3685
Scale	0	1.0000	0.0000	1.0000	1.0000		

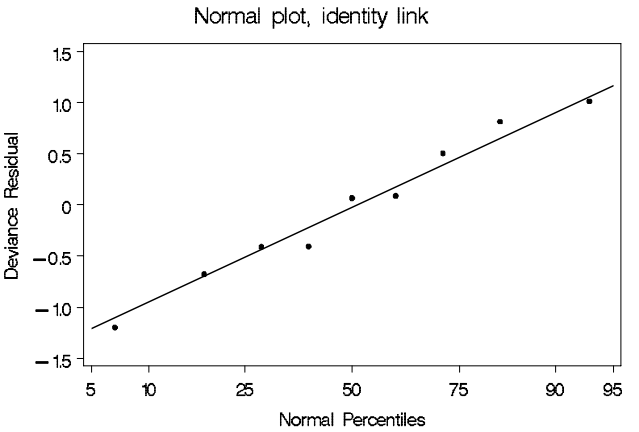
The model fit for the Identity link model:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	6	4.1971	0.6995
Scaled Deviance	6	4.1971	0.6995
Pearson Chi-Square	6	4.1567	0.6928
Scaled Pearson X2	6	4.1567	0.6928
Log Likelihood		362.6385	

Both models show a good fit; slightly better for the log link. Plots of residuals vs. fitted values are similar for the two models. The normal probability plot is slightly better for the identity link model:





In all, it is difficult to judge which of the two models is "best", based on statistical criteria.

Exercise 6.3

Models with a Poisson distribution, a log link and using log(mon_serv) as an offset were fitted to the data. Some of the models were:

Model	deviance	df
1. Main effects only	38.695	25
1. +type*yr_c	14.587	13
1. +type*per_op	33.756	21
1. +yr_c*per_op	36.908	23

It seems that a model with main effects, plus the type*yr_c interaction, would describe the data well. However, this model produces a near-singular Hessian matrix.

Exercise 6.4

The model analyzed in the text is saturated, which means that the data should agree perfectly with the model. The model for males is $\log(\widehat{\mu}) = \log(t) + \widehat{\beta}_0$ which gives $\log(320) = \log(21.4) + \widehat{\beta}_0$ i.e. $\beta_0 = \log(320) - \log(21.4) = 2.7049$. The model for females is $\log(\widehat{\mu}) = \log(t) + \widehat{\beta}_0 + \widehat{\beta}_1$. We get $\widehat{\beta}_1 = \log(175) - \log(17.3) - 2.7049 = -0.39082$. These results agree with the computer outputs in the text.

Exercise 6.5

A. The LR test of the hypothesis H_0: $\mu_A = \mu_B$ is obtained by comparing the deviances of the two models:

$(D_1 - D_2) / (df_1 - df_2) = (27.857 - 16.2676) / (19 - 18) = 11.589$ which is used as an asymptotic χ^2 on 1 d.f. The 5% limit is 3.841; the 1% limit is 6.635 and the 0.1% limit is 10.828. Our observed value is even larger than 10.828; the result is clearly significant and H_0 can be rejected at the 0.1% level.

The Wald test of the same hypothesis uses the test statistic $\frac{\widehat{\beta}-0}{s.e.(\widehat{\beta})} = \frac{0.5878}{0.1764} = 3.3322$. This is compared to appropriate limits of a standard normal variate z. Limits: 5%: 1.96; 1%: 2.576; 0.1%: 3.291. Our observed test statistic is (numerically) larger than the 0.1% limit: we reject the null hypothesis.

B. The model is $g(\mu_B) = \beta_0 + \beta_1 x$, where x is a dummy variable ($x = 1$ for treatment A). For treatment A the model is $g(\mu_A) = \beta_0 + \beta_1$, and for treatment B the model is $g(\mu_B) = \beta_0$. The link function g is a log function. Thus, it holds that $g(\mu_B) - g(\mu_A) = (\beta_0) - (\beta_0 + \beta_1) = -\beta_1$. Therefore $\log(\mu_B) - \log(\mu_A) = -\beta_1$, which means that $\log\left(\frac{\mu_B}{\mu_A}\right) = -\beta_1$. A 95% Wald confidence interval for β_1 is $0.5878 \pm 1.96 \cdot 0.1764$; 0.5878 ± 0.3457; the limits are $[0.2421 \ldots 0.9335]$. Taking antilogs of minus these limits, approximate 95% limits for $\frac{\mu_B}{\mu_A}$ are obtained as $e^{-0.2421} = 0.785$ and $e^{-0.9335} = 0.393$.

Exercise 6.6

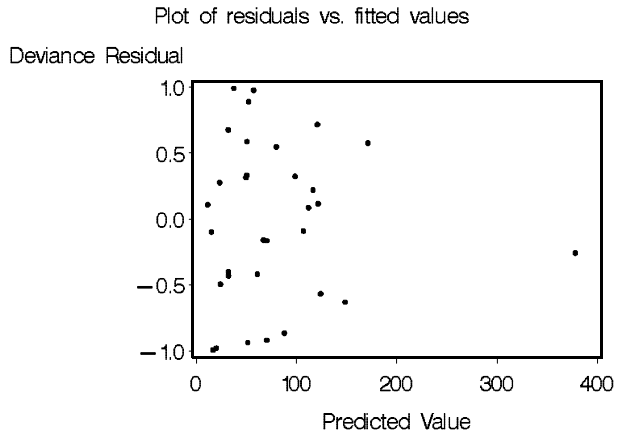
A model with a Poisson distribution, a log link, and no offset variable gives a deviance of 15.9371 on 13 df, deviance/df=1.2259. Inclusion of log(miles) as an offset gives the deviance 16.0602 on 13 df, deviance/df=1.2354. Inclusion of the offset does not affect the fit very much, possibly because the values for miles are rather similar for the different years.

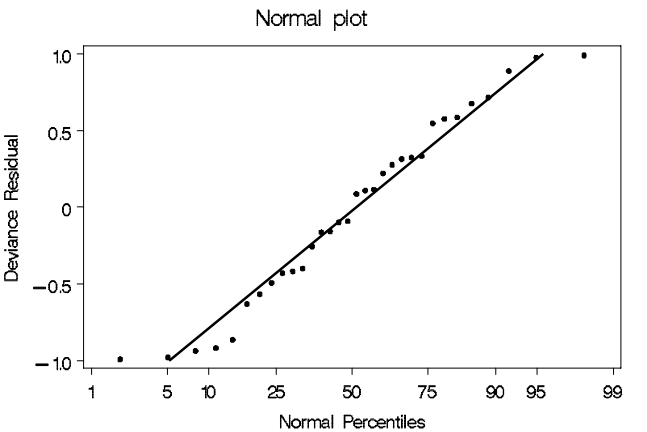
Exercise 6.7

A. In the full model, the three-way interaction is not significant ($p = 0.2639$). The model with all main effects and two-way interactions gives deviance 11.1746 on 9 df, deviance/df=1.2416. All main effects and interactions are highly significant ($p < 0.0001$).

B. The model with scores 0, 1, 2, 3 for coffee and 0, 1, 2, 3 for cigarettes is not a good model: deviance=301.08 on 25 df.

C. Of the two models, the model in A. is to be preferred because of a much better fit. Residuals plots for this model are as follows:





The Residual vs. Fits plots shows some tendency towards an "inverse trumpet" shape, with a decreasing variance for increasing $\widehat{y}$. The Normal plot is rather straight, with a couple of deviating observations at each end.

Exercise 6.8

A. The test of the hypothesis of no relation between temperature and probability of failure is obtained by calculating the difference in deviance between the null model and the estimated model. These differences can be interpreted as χ^2 variates on 1 d.f., for which the 5% limit is 3.841 and the 1% limit is 6.635.

i) Poisson model: $\chi^2 = 22.434 - 16.8337 = 5.600$; this is significant at the 5% level ($p = 0.018$).

ii) Binomial model: $\chi^2 = 24.2304 - 18.0863 = 6.1441$; again significant at the 5% level ($p = 0.0132$).

Both models indicate a significant relationships between failure risk and temperature. Note, however, that the number of observations and, in particular, the number of failures, is so small that the asymptotics may not work.

B. Predicted values at 31°F are

i) Poisson model: $g(\widehat{\mu}) = 5.9691 - 0.1034 \cdot 31 = 2.7637$. Since $g(\cdot)$ is a log link, this gives $\widehat{\mu} = \exp(2.7637) = 15.858$.

ii) Binomial model: $g(\widehat{p}) = 5.0850 - 0.1156 \cdot 31 = 1.5014$. $g(\cdot)$ is a logit link which has inverse $\frac{e^x}{1+e^x}$ so $\widehat{p} = \frac{e^{1.5014}}{1+e^{1.5014}} = 0.81778$. With $n = 6$ O-rings on board, we would expect $n\widehat{p} = 6 \cdot 0.81778 = 4.9$ of them to fail!

C. The Poisson model has the disadvantage that the expected number of failing O-rings is actually larger than the total number on board: we predict 16 failures among 6 O-rings. The Binomial model is more reasonable.

D. Using the Binomial model with $n = 6$ and $p = 0.8179$, $P(x \geq 3) = 1 - P(x \leq 2) = 1 - 0.0121 = 0.9879$.

Exercise 6.9

The odds ratios can be calculated as $\exp\left(\widehat{\beta}_i\right)$. In the presence of interactions the main effect odds ratios are not very illuminating, so we only consider the interactions. In the table we abbreviate Gender=G; Location=L; Injury=I and Belt use=B. We interpret the parameters by the ordered values in the SAS printout. Since N is (alphabetically) before Y, the odds ratio for, for example, B*I means that persons with "B=No" have "I=No" less often. This, of course, could be stated as "Users of seat belts are injured less often". For the different interaction effects in the model we get:

Term	OR	Comment
G*L	$\exp(-0.2099) = 0.811$	Females traveled in rural areas less often than males.
G*B	$\exp(-0.4599) = 0.631$	Females avoided belt use less often than males.
G*I	$\exp(-0.5405) = 0.582$	Females were uninjured less often than males.
L*B	$\exp(-0.0849) = 0.919$	Belts are avoided less often in rural areas.
L*I	$\exp(-0.7550) = 0.470$	Passengers are uninjured less often in rural areas
B*I	$\exp(-0.8140) = 0.443$	Non-users of belts are uninjured less often

Exercise 7.1

A generalized linear model with a multinomial distribution and a cumulative logit link gave the following result:

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept1		1	-1.1607	0.1814	-1.5163	-0.8051	40.93	<.0001
Intercept2		1	-0.6222	0.1705	-0.9564	-0.2881	13.32	0.0003
Intercept3		1	1.1782	0.1817	0.8220	1.5344	42.03	<.0001
treatment	BP	1	0.3219	0.2216	-0.1124	0.7563	2.11	0.1462
treatment	CP	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

According to this analysis, there is no significant association between treatment and response ($p = 0.1462$). A standard Chi-square test of independence gives $\chi^2 = 4.6$ on 3 df, $p = 0.20$.

Exercise 7.2

The standard χ^2 test of independence gives $\chi^2 = 24.1481$ on 4 df, $p < 0.0001$. An ordinal model for prediction of the attitude towards detection of cancer gives a Type 3 $\chi^2 = 25.86$ on 2 df, $p < 0.0001$. The parameter estimates for this model are as follows:

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept1		1	-2.7703	0.2500	-3.2602	-2.2803	122.80	<.0001
Intercept2		1	-0.4759	0.1337	-0.7380	-0.2137	12.66	0.0004
mammo	< 1 year	1	-1.4753	0.3247	-2.1117	-0.8388	20.64	<.0001
mammo	> 1 year	1	-0.4926	0.2928	-1.0664	0.0812	2.83	0.0924
mammo	Never	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

An alternative model is the linear by linear association model. For these data, this model gives:

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	3.6821	0.0707	3.5435	3.8206	2711.93	<.0001
cancer	0	1	-1.0564	0.1036	-1.2595	-0.8533	103.95	<.0001
cancer	1	1	0.0171	0.0411	-0.0633	0.0976	0.17	0.6763
cancer	2	0	0.0000	0.0000	0.0000	0.0000	.	.
mammo	< 1 year	1	-3.0357	0.1375	-3.3052	-2.7662	487.41	<.0001
mammo	> 1 year	1	-2.2606	0.0688	-2.3954	-2.1258	1079.83	<.0001
mammo	Never	0	0.0000	0.0000	0.0000	0.0000	.	.
c*m		1	0.6437	0.0348	0.5756	0.7119	343.06	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

The $c * m$ association is highly significant. All three analyses suggest a strong relationship between mammography experience and attitude towards cancer detection.

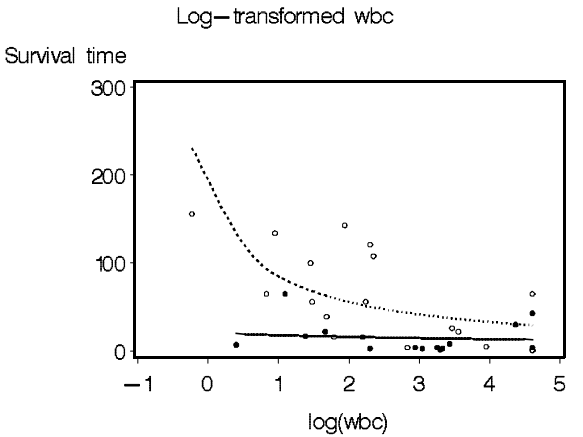
Exercise 8.1

A gamma model with log-transformed wbc values was tried. The wbc*ag interaction was far from significant so it was excluded. The suggested model produces a deviance of 38.2342 on 30 df; deviance/df=1.2745. The output is:

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	0.0057	0.0036	-0.0014	0.0128	2.44	0.1180
ag	0	1	0.0431	0.0174	0.0089	0.0773	6.09	0.0136
ag	1	0	0.0000	0.0000	0.0000	0.0000	.	.
lwbc		1	0.0061	0.0024	0.0014	0.0109	6.37	0.0116
Scale		1	0.9968	0.2160	0.6518	1.5242		

A plot of observed survival times for the two groups, along with the survival times predicted by the model, is as follows:



Exercise 8.2

The data were run using the macro for variance heterogeneity listed in the Genmod manual. The results for the mean value model was as follows:

Mean model

Obs	Parameter	Level1	DF	Estimate	StdErr	LowerCL	UpperCL	ChiSq	Prob ChiSq
1	Intercept		1	19.7200	7.6955	4.6370	34.8030	6.57	0.0104
2	group	a	1	-16.7533	7.7032	-31.8514	-1.6553	4.73	0.0296
3	group	b	1	-11.5400	7.7790	-26.7866	3.7066	2.20	0.1379
4	group	c	0	0.0000	0.0000	0.0000	0.0000	.	.
5	Scale		0	1.0000	0.0000	1.0000	1.0000	_	_

There is a significant difference ($p = 0.0296$) between groups a and c. The results for the variance model are:

Variance model

Obs	Parameter	Level1	DF	Estimate	StdErr	LowerCL	UpperCL	ChiSq	Prob ChiSq
1	Intercept		1	5.6907	0.6325	4.4511	6.9303	80.96	<.0001
2	group	a	1	-6.0301	0.8563	-7.7085	-4.3517	49.58	<.0001
3	group	b	1	-3.1318	0.7746	-4.6500	-1.6136	16.35	<.0001
4	group	c	0	0.0000	0.0000	0.0000	0.0000	.	.
5	Scale		0	0.5000	0.0000	0.5000	0.5000	_	_

The results indicate significant differences in variance between the groups.

Exercise 8.3

A SAS program for analysis of these data using the Glimmix macro is as follows. Note that the macro itself must be run before this program is

submitted.

```
%glimmix(
data=labexp,
stmts=%str(
class pot var1 var2;
model x2/n2 = var1 var2 var1*var2;
random pot*var1*var2;
),
error=binomial, link=logit );
run;
```

Some of the output is as follows:

Solution for Fixed Effects

Effect	Var1	Var2	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept			1.6849	0.2299	64	7.33	<.0001
Var1	A		-0.1987	0.3172	64	-0.63	0.5332
Var1	F		0.4159	0.3447	64	1.21	0.2320
Var1	H		-0.1333	0.3194	64	-0.42	0.6779
Var1	K		0	.	.	.	.
Var2		A	-0.6850	0.3047	64	-2.25	0.0280
Var2		F	0.1817	0.3324	64	0.55	0.5865
Var2		H	-0.4186	0.3107	64	-1.35	0.1826
Var2		K	0	.	.	.	.
Var1*Var2	A	A	0.9072	0.4402	64	2.06	0.0434
Var1*Var2	A	F	-0.9360	0.4423	64	-2.12	0.0382
Var1*Var2	A	H	-0.3074	0.4266	64	-0.72	0.4737
Var1*Var2	A	K	0	.	.	.	.
Var1*Var2	F	A	0.2112	0.4581	64	0.46	0.6464
Var1*Var2	F	F	-0.5441	0.4800	64	-1.13	0.2612
Var1*Var2	F	H	-0.6812	0.4501	64	-1.51	0.1351
Var1*Var2	F	K	0	.	.	.	.
Var1*Var2	H	A	0.4641	0.4323	64	1.07	0.2870
Var1*Var2	H	F	-0.07013	0.4600	64	-0.15	0.8793
Var1*Var2	H	H	0.4596	0.4426	64	1.04	0.3030
Var1*Var2	H	K	0	.	.	.	.
Var1*Var2	K	A	0	.	.	.	.
Var1*Var2	K	F	0	.	.	.	.
Var1*Var2	K	H	0	.	.	.	.
Var1*Var2	K	K	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Var1	3	64	3.22	0.0285
Var2	3	64	4.37	0.0074
Var1*Var2	9	64	3.05	0.0042

There is a significant interaction between varieties, i.e. some combinations of varieties are more palatable than others to the lice. This conclusion may be followed up by comparing least squares mean values for the different combinations.

Index