Models for discrete longitudinal data
Analysing discrete correlated data using R

Christophe Lalanne

Novembre 2007

Preface

This textbook is intended to the R user already accustomed to the Generalized Linear Model and wants to extend his knowledge with the analysis of correlated discrete data. Without special help, a new R user should easily get inside the code, although we assume a general background in statistics, and a specific understanding of GLM.

The two books of Molenberghs & Verbeke sound as though they really were wrotten for becoming familiar with GLMM without any slow and painful mathematics. They emphasize a general modelling approach, from theoretical concept to algorithmic considerations, still retaining practical issues at hand.

I choose to present the second of their books, *Models for Discrete Longitudinal Data*, as I am working with this kind of data most of the time. However, they use the SAS System to carry their numerical applications. Every statistician would acknowledge the notoriety of this piece of software, but now R has come to a degree of sophistication and notoriety that make unlikely not to find a dedicated function to kindly do the job you ask for. So, why not switch to R ? Did you ever seen something as harder to do as plotting a simple graph with SAS ? Why borrying with the `GPLOT` syntax since R provides much than we need to plot longitudinal data ? Furthermore, apart from the fact that the syntax changes from one procedure to the other (e.g. specifying a model largely differs between `PROC MIXED` and `PROC GENMOD` despite the fact that we only treat the data as discrete in the latter case...), I found that SAS scripts are not fully understandable at first sight. They merely seem to be extract from the obfuscated C coding challenge (`www.ioccc.org`). However, one must agree that SAS offers some standardized and well recognized statistical procedures. Well, R too...

Finally, I also choose to organize this textbook with the practitioner approach in mind. Thus, each chapter corresponds to a given dataset and is divised into several section, each tackling a different question about the data.

A brief description of the motivating study used throughout the textbook of Molenberghs & Verbeke is provided at the beginning of each chapter.

All materials related to the two textbooks published by these authors can be found on the website : `www.censtat.uhasselt.be/software/`.

I have translated the datasets from SAS format (`sas7bdat`) into csv files as this the best choice for dealing with data into R. For those who want to analyse the data using the SAS System, the original datasets are also provided. All analyses were done with R version 2.6.0 (2007-10-03). Some R packages are mandatory for replicating the results (`gee`, `lme4`), while others might be very useful for going beyond the simple exploration of the data structure and provide exciting graphical displays (e.g. `lattice`).

Contents

Chapter 1

A brief overview

This chapter aims at describing the basic properties of Generalized Linear Mixed Models. We will try to provide an as comprehensive as possible overview of the statistical properties of the various models helded under this model family. For this purpose, we will follow the organization of Molenberghs & Verbeke's book. First, after recalling the reader to the basics of Linear Mixed Models, we will describe the two general modelling approaches: the subject-specific centred or *conditional model*, and the population-averaged centred or *marginal model*. Highlighting the strength of each one, we will also present their derivatives, ranging from...

1.1 A review of the basic properties of GLM

This section provides a refresher about the classical Generalized Linear Model, where we assume that collected observations are all independant and identically distributed according to a distribution that is a member of the exponential family. This includes, but is not limited to:

- the binomial, derived from the Bernoulli schema and typically used to model binary data (presence/absence, correct/false, etc.);
- the Poisson, involved in the modeling of rare events;
- the negative binomial, most often used as an alternative to the Poisson distribution when we want to take into account over-dispersion (free and eventually larger scale parameter than in the Poisson case, where $\mathrm{E}(Y) = \mathrm{V}(Y) = \mu$);
- the hypergeometric, used in the light of sampling without replacement.

Basically, a random variable Y is said to follow a distribution belonging to the exponential family if its density can be expressed as a function of both a natural or canonical parameter, denoted θ, and a scale parameter,

ϕ. Usually, this function require two additional functions, $\psi(\cdot)$ and $c(\cdot,\cdot)$, which link these parameters together with the observations. Thus, we have the general formulation of the density function:

$$f(y) \equiv f(y \mid \theta, \phi) = \exp\{\phi^{-1}[y\theta - \psi(\theta)] + c(y, \phi)\} \tag{1.1}$$

It can be shown very easily that

$$\begin{aligned} \mathrm{E}(Y) &= \psi'(\theta) \\ \mathrm{V}(Y) &= \phi\psi''(\theta) \end{aligned}$$

but the most important relation stands for the mean $\mu = \mathrm{E}(Y)$ and the variance which are related through

$$\sigma^2 = \phi\psi''[\psi'^{-1}(\mu)] = \phi\nu(\mu),$$

with $\nu(\mu)$ called the variance function. An important issue is whether we choose to specify $\nu(\mu)$ as belonging to the exponential family, in which case standard likelihood estimation techniques are still avalaible, or if we use instead a set of estimating equations for $\nu(\mu)$ in order to get so-called *quasi-likelihood* estimates.

Many statistical textbook relate the previously discrete distributions to the exponential family, and one can easily express each of these distribution under the form 1.1. For instance, the density function of a binomial process can be rewritten as

$$f(y) = \exp\left[y \ln\left(\frac{\pi}{1-\pi}\right) + \ln(1-\pi)\right],$$

with $\Pr(Y = 1) = \pi$. Here, the natural parameter θ is the logit of π ($\ln[\pi/(1-\pi)]$), the scale parameter equals unity, $\phi = 1$, and the mean and variance are $\mu = \pi$ and $\nu(\pi) = \pi(1-\pi)$. Of course, we can use a different link function (probit, log-log, complementary log), and we know that there is little difference between logit and probit, except for the tails of the distribution (but see 2.2.

In order to get a linear model, we just have to write the link function as a linear combination of explanatory variables, for instance

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = x_i'\beta, \tag{1.2}$$

or equivalently

$$\pi_i = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}.$$

Now that we have set up the general working scenario, we could rely on maximum likelihood estimation to draw inference about the parameters

of interest. The likelihood is simply the product of the density function and conversely, the log-likelihood equals the sum of log densities. With binomially i.i.d. observations, we have

$$\ell(x_i) = \dots$$

In the general case, and using expression 1.1, we usually have

$$\ell(\beta, \phi) = \frac{1}{\phi} \sum_{i=1}^{N} y_i \theta_i - \psi(\theta_i) + \sum_{i=1}^{N} c(y_i, \phi). \tag{1.3}$$

Several techniques can be used to test for the significance of a formal hypothesis (e.g. Wald test). Before that, we have to calculate the estimates of our parameters. Score equations are obtained from equating the first-order derivatives of $\ell(\beta, \phi)$ to zero and are

$$S(\beta) = \sum_i \frac{\partial \theta_i}{\partial \beta}(y_i - \psi'(\theta)) = 0.$$

Since $\mu_i = \psi'(\theta)$ and $\nu_i = \nu(\mu_i) = \psi''(\theta_i)$, it follows that

$$\frac{\partial \mu_i}{\partial \beta} = \psi''(\theta_i)\frac{\partial \theta_i}{\partial \beta} = \nu_i \frac{\partial \theta_i}{\partial \beta}$$

which implies

$$S(\beta) = \sum_i \frac{\partial \mu_i}{\partial \beta} \nu_i^{-1}(y_i - \mu_i) = 0.$$

Such equations have to be solved using iterative algorithms, such as iteratively reweighted least squares, Newton-Raphson or Fisher scoring. Most of the time (that is, except in the logistic case), ϕ is unknown and has to be estimated too. This involves estimation of the standard errors of the elements in β. As $\mathrm{V}(Y_i) = \phi\nu_i$, we can use

$$\hat{\phi} = \frac{1}{N - p} \sum_i (y_i - \hat{\mu}_i)^2 / \nu_i(\hat{\mu}_i)$$

as a consistent estimator of ϕ. The interested reader can refer to Cullagh and Nelder [1989] for further considerations on estimation in the GLM framework.

1.2 The Linear Mixed Models

Note to the reader: As this section provides a very elegant and useful summary of inference based on Linear Mixed Model, I here reproduce most of the text of Molenberghs & Verbeke.

1.2.1 The gaussian case

As Molenberghs & Verbeke stated at the beginning of Chapter 4, the LMM for gaussian correlated data yet provides some clues to get a better understanding of the problems yieded by correlated data. So, before investigating the case of categorical repeated measurements, let's start with the somewhat "easier" gaussian case.

For continuous data, we already know that the linear model allows us to fit a linear combination of predictors to the measured response variable, under normality assumptions. To take into account the presence of correlated mesurements, two simple approaches can be undertaken. First, we can formulated a multivariate analog to the univariate linear model, used for instance in linear regression. In this approach, each component is modelled as a univariate linear regression model, together with the association structure which is specified through a marginal covariance matrix. Second, we can use a random-effect approach, in which we fit a separate intercept for each subject (considered here as a cluster or level 2 structure). In this latter scheme, also called a conditional approach, we focus on individual subject rather than directly assessing the mean response as in the multivariate or marginal model.

Let's take an example which shall be more illustrative. Suppose we have a sample of N measurements, assumed to be the realization of a random variable Y. Our dataset consists in a set of y_i responses, $i = 1 \dots N$, some of which belong to the same subject.

A first approach would consist in a multivariate regression model, where we assume that each component of the response vector Y, denoted as $Y_i = (Y_i 1, \dots, Y_{in_i})'$, enters in a linear relationship with the predictor(s). Writting down the i components with j repetaed measurements as $Y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij}$, and assuming that $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, we then have a set of i regression equations (one for each subject). In matrix notation, we can always find a design matrix such that $Y_i = X_i\beta + \varepsilon_i$. In this case, $\beta' = (\beta_0, \beta_1)$ is our object of interest. We achieve the specification of this model by specifying a covariance matrix V_i for the residuals ε_i and we have our marginal multivariate model:

$$Y_i \sim \mathcal{N}(X_i\beta, V_i) \tag{1.4}$$

where $V_i = \sigma^2 I_{n_i}$, with I_{n_i} denoting the identity matrix for subject i who has been measured n_i times. Of course, we have completely discarded the fact that repeated measures within the same subunit (i.e. a given subject in our case) might be positively correlated. We could also constraint V_i such as to include a first-order auto-regressive process, whereby correlation between consecutive measurements decrease as the time between two measurements increases. In all cases, we always assume a symmetric, definite positive covariance matrix, and we use a single regression model for each subject.

In another approach, we would prefer to account for the possible difference between both the intercept and slope of each subject under investigation. In this latter scheme, the outcome is still assumed to be of the form $Y_{ij} = \tilde{\beta}_{i0} + \tilde{\beta}_{i1} t_{ij} + \varepsilon$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma_i)$, but now we shall view our set of subjects as a random sample taken from a larger population. Therefore, the subject-specific regression coefficients $\tilde{\beta}_i = (\tilde{\beta}_{i0}, \tilde{\beta}_{i1})'$ are random samples themselves, and we can postulate that they are drawn from a gaussian distribution of regression coefficients. Our last model can be rewritten as

$$Y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) t_{pi} + \varepsilon_{ij}, \tag{1.5}$$

with $\tilde{\beta}_{i0} = \beta_0 + b_{i0}$ and $\tilde{\beta}_{i1} = \beta_1 + b_{i1}$. The population-averaged profile remains linear, with intercept and slope β_0 and β_1, although we also have subject-specific profile.

In its general formulation, the GLMM takes the form

$$\begin{aligned} Y_i \mid b_i &\sim \mathcal{N}(X_i\beta + Z_i b_i, \sigma_i) && (1.6) \\ b_i &\sim \mathcal{N}(0, D), && (1.7) \end{aligned}$$

where D stands for the $n \times n$ covariance matrix for the random effect. The fixed part of the model, $X_i\beta$, let us assess the mean profile, i.e. the population-averaged evolution, while the random effects contained in b_i allow for estimable variation of subject-specific profiles.

For inference, tests are based on on the marginal distribution for the response Y_i. Giving the marginal density function

$$f_i(y_i) = \int f_i(y_i \mid b_i) f(b_i)\, db_i,$$

it can be shown that it is the density function of an n_i-dimensional normal distribution with mean $X_i\beta$ and covariance matrix $V_i = Z_i D Z_i' + \sigma_i$.

Since the mixed model is defined through the distribution of both the fixed effects, $f_i(y_i \mid b_i)$ and the random ones, $f(b_i)$, it is called the *hierarchical* formulation of the LMM, while the preceding marginal normal distribution corresponds to the *marginal* formulation of the LMM. As underlined by Molenberghs & Verbeke these models are not strictly equivalent as different random-effects models can induce the same marginal distribution. As an example, consider the case where every subject is measured twice ($n_i = 2$). Assume that the random-effects structure is confined to a random intercept (b_i is thus a scalar), and that the residual error structure is defined by $\Sigma_i = \Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$ (Model I). The resulting marginal covariance matrix is then:

$$V = \begin{pmatrix} 1 \\ 1 \end{pmatrix} (d)(1\ 1) + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} d + \sigma_1^2 & d \\ d & d + \sigma_2^2 \end{pmatrix}. \tag{1.8}$$

If on the contrary we consider both a random intercept and a random slope for the random-effects, we have $b_i = (b_{0i}, b_{1i})'$, mutually uncorrelated (but not necessarily independent!). The resiudal error structure is $\Sigma_i = \Sigma = \sigma^2 I_2$. This is what Molenberghs & Verbeke called Model II. The resulting covariance matrix is now:

$$\begin{aligned} V &= \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \\ &= \begin{pmatrix} d_1 + \sigma^2 & d_1 \\ d_1 & d_1 + d_2 + \sigma^2 \end{pmatrix}. \end{aligned} \tag{1.9}$$

Parametrization 1.8 and 1.9, which belong to thwo different hierarchical model, leads to the same marginal model, with $d_1 = d$, $d_2 = \sigma_2^2 - \sigma_1^2$ and $\sigma^2 = \sigma_1^2$. Furthermore, it should be noted that some marginal models are not necessarily implied by a mixed model. Considering a model with compound symmetry: if the within-subject correlation is positive ($\gamma \geq 0$), we have a mixed model with random intercept $b_i \sim \mathcal{N}(0, \gamma)$ and uncorrelated errors with common variance σ^2 (because of compound symmetry hypothesis). However, this does not hold anymore if $\gamma < 0$.

1.2.2 Estimation and inference for the Marginal Model

Estimation of the parameters of the LMM is done via maximum likelihhod techniques, based on the marginal model which is multivariate normal with mean $X_i\beta$ and covariance $V_i(\alpha) = Z_i D Z_i' + \Sigma_i$, for subject i. Assuming independence between subjects, the likelihood is:

$$\begin{aligned} \ell(\theta) &= \prod_{i=1}^{N} (2\pi)^{-n_i/2} |V_i(\alpha)|^{-\frac{1}{2}} \\ &\quad \times \exp\left[-\frac{1}{2} (Y_i - X_i\beta)' V_i^{-1}(\alpha)(Y_i - X_i\beta) \right. \end{aligned} \tag{1.10}$$

Estimation of $\theta' = (\beta', \alpha')$ which requires the joint maximization of 1.10 with respect to all elements of θ involves numerical algorithm.

Conditionally on α, the MLE of β is found to be [Laird and Ware, 1982]:

$$\hat{\beta}(\alpha) = \left(\sum_{i=1}^{N} X_i' W_i X_i \right)^{-1} \sum_{i=1}^{N} X_i' W_i Y_i, \tag{1.11}$$

where $W_i = V_i^{-1} = (Z_i D Z_i' + \Sigma_i)^{-1}$. Usually, α is unknown and need to be estimated too. An M-Estimator can be used although common practice is oriented toward the use of a restricted MLE, the so-called REML [Harville, 1974]: this allows to estimate α without having to estimate first the fixed part of the model held in β.

For inference purpose, of course, we are much interested in β as fixed effects describe the average evolution. Conditionally on α, the MLE of β is given by 1.11, which is normally distributed with mean given by

$$\mathrm{E}\left[\hat{\beta}(\alpha)\right] = \left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1} \sum_{i=1}^{N} X_i' W_i E[Y_i] = \beta, \tag{1.12}$$

and covariance

$$\begin{aligned} \mathrm{Var}\left[\hat{\beta}(\alpha)\right] &= \left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1} \\ &\quad \times \left(\sum_{i=1}^{N} X_i' W_i \mathrm{Var}[Y_i] W_i X_i\right) \\ &\quad \times \left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1} \\ &= \left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1}. \end{aligned} \tag{1.13}$$

This relation holds provided that the mean and covariance has been correctly specified, i.e. $\mathrm{E}(Y_i) = X_i\beta$ and $\mathrm{Var}(Y_i) = V_i = Z_i D Z_i' + \Sigma_i$. Hypothesis testing regarding β components can be done using Wald test.

Il should be emphasized that Wald tests are based on standard errors computed by replacing α in 1.13 by its ML or REML estimate. Therefore, they underestimate the true variability for $\hat{\beta}$. It is the reason why we don't rely on normal or χ^2 distribution anymore but use a t or F distribution instead. In these latter cases, the denominator degree of freedom has to be estimated from the data. One commonly uses the so-called Satterthwaite-type approximations [Satterthwaite, 1941]. Approximate p-value can be computed and are close to the ones obtained with classical MLE because degrees of freedom are large enough to yield reliable estimation. Recall that longitudinal data bring several independent information mediated by different subjects which prove to be sufficient for correct inference in this case. This is very different from a crossed random-effects design, where subjects are crossed with items for instance, which in this case yield very different p-value according the method we choose.

Another point of interest lies in the fact that standard errors are valid only if the mean and covariance are correctly specified, as stated above. Since it is often difficult to assess whether the covariance matrix has been correctly specified, we often prefer to rely on the so-called "sandwich" estimator given in 1.13 (the first right-member expression, indeed). These robust, or empirical standard errors are computed using an estimate of $\mathrm{Var}(Y_i)$ given by

$(y_i - X_i\hat{\beta})(y_i - X_i\hat{\beta})'$. These estimators are consistent provided the mean and variance arehas been correctly specified. The interested reader is referred to classical GEE theory for further development on these computation [Liang and Zeger, 1986].

1.2.3 Inference for the random effects

When we have estimated the systematic part of the model, we often are interested in estimating random-effects b_i as well. Indeed, random-effects reflect variation between subject-specific profiles around the overall mean profile. They can be considered like residuals, and as such can be used to detect outlyiers.

For this purpose, it is no longer sufficient to rely on the marginal model $\mathcal{N}(X_i\beta, V_i)$ and one has to resort to one of tow hierarchical formulation provided in 1.8 and 1.9. This assumption proved to be sufficient in the case where between-subjects variance is larger than within-subject variance.

Because the subject-specific parameters b_i are random variates, Bayesian techniques should be used [Gelman et al., 1995]. Conditional on b_i, Y_i follows a multivariate normal distribution with mean vector $X_i\beta + Z_i b_i$ and with covariance matrix Σ_i. One can show that, conditionally on $Y_i = y_i$, b_i follows a multivariate normal posterior distribution with mean $\hat{b_i}(\theta) = DZ_i'V_i^{-1}(\alpha)(y_i - X_i\beta)$. This expression is used to estimate b_i. Its covariance estimator is

$$\mathrm{Var}[\hat{b_i}(\theta)] = DZ_i' \left\{ V_i^{-1} - V_i^{-1}X_i \left(\sum_{i=1}^{N} X_i'V_i^{-1}X_i \right)^{-1} X_i'V_i^{-1} \right\} Z_i D. \tag{1.14}$$

The estimator given by 1.14 underestimates the variability in $\hat{b_i}(\theta) - b_i$ because it does not take into account the variation of b_i. Therefore, we ususally base inference for b_i on [Laird and Ware, 1982]:

$$\mathrm{Var}[\hat{b_i}(\theta) - b_i] = D - \mathrm{Var}[\hat{b_i}(\theta)] \tag{1.15}$$

In practice, the unknown parameters β and α in 1.14 and 1.15 are replaced by their MLE or REML estimates. The resulting estimates for the b_i are called "Empirical Bayes" (EB) estimates and will be denoted by $\hat{b_i}$. Again, 1.14 and 1.15 underestimate the true variability in the estimated $\hat{b_i}$ because they do not account for the variability introduced by replacing the unknown parameters θ by its estimate. As for the fixed effects, inference is therefore based on approximate t or F-tests, rather than on Wald tests.

From 1.15, it follows that for any linear combination λb_i of the random effects, $\mathrm{Var}(\lambda'\hat{b_i}) \leq \mathrm{Var}(\lambda' b_i)$ which means that EB estimates show less variance than actually present in the random-effects population. This is known as a *shrinkage* effect. Shrinkage is also seen in the prediction $\hat{y_i} \equiv$

$X_i\hat{\beta} + Z_i\hat{b_i}$ of the ith profile, which can be rewritten as $\hat{y_i} = \Sigma_i V_i^{-1}\hat{\beta} + [I_{n_i} - \Sigma_i V_i^{-1}]y_i$. Therefore, $\hat{y_i}$ can be interpreted as a weighted average of the population-averaged profile $X_i\hat{\beta}$ and the observed data y_i, with weights $\Sigma_i V_i^{-1}$ and $I_{n_i} - \Sigma_i V_i^{-1}$. Severe shrinkage is to be expected when the residual variability is large in comparison to the between-subject variability (i.e. the random effects).

1.3 An introduction to the Marginal Model

1.3.1 Analyzing 2-way contingency tables

Following the notation introduced by Molenberghs & Verbeke, we will consider a contingency table whereby ordinary multinomial cell counts and their cumulative counterparts are defined as

$$Z^*_{ijr} = \begin{cases} 1 & \text{if } Y_{1r} = i \text{ and } Y_{2r} = j, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$Z_{ijr} = \begin{cases} 1 & \text{if } Y_{1r} \leq i \text{ and } Y_{2r} \leq j, \\ 0 & \text{otherwise.} \end{cases} \tag{1.16}$$

Here, the subsript r indexes the rth individual. The corresponding probabilities are thus defined as $\mu^*_{ij} = \Pr(Z^*_{ijr} = 1)$ and $\mu_{ij} = \Pr(Z_{ijr} = 1)$.

Based on the Multivariate Logistic Model [Cullagh and Nelder, 1989], whereby the vector link function is expressed in terms of the joint probabilities and allows to model marginal means in terms of base-line category, adjacent category logits, continuation-ratio logits or cumulative logits, we will consider two kind of models widely used in the case of association study. First, in the case of $I \times J$ tables, *Goodman's Local Association Model*, also called the RC model, considers log local cross-ratios of the form

$$\ln \theta^*_{ij} = \ln \left(\frac{\mu^*_{ij}\mu^*_{i+1,j+1}}{\mu^*_{i,j+1}\mu^*_{i+1,j}} \right).$$

It includes main effects together with parameters describing the association structure. For instance, from $\mu^*_{ij} = \alpha_i\beta_j e^{\phi\lambda_i\nu_j}$, which is the closed form for the joint cell probabilities, we can derive the local cross-ratios $\ln \theta^*_{ij} = \phi(\lambda_i - \lambda_{i+1})(\nu_j - \nu_{j+1})$. This model can be extended to include additive effects on the association and is known as the R+C+RC model [Goodman, 1981]. Second, *Dale's Marginal Model* is a marginal model for ordinal variable that are modeled through marginal cumulative logits and global cross-ratios [Dale, 1986]. Cumulative logits are expressed as

$$\eta_{1i} = \text{logit}\left[\Pr(Y_1 \leq 1)\right] = \ln(\mu_{iJ}) - \ln(1 - \mu_{iJ})$$
$$\eta_{2i} = \text{logit}\left[\Pr(Y_2 \leq 1)\right] = \ln(\mu_{Ij}) - \ln(1 - \mu_{Ij})$$

and global cross-ratios used to define the joint probabilities are computed as

$$\ln \psi_{ij} = \ln \left(\frac{\mu_{ij}(1 - \mu_{Ij} - \mu_{iJ} + \mu_{ij})}{(\mu_{iJ} - \mu_{ij})(\mu_{Ij}\mu_{ij})} \right).$$

Note, however, that local cross-ratios might be used instead of global ones if required. For the particular case of binary variables, both lead to the same results. This model which can be expressed in the following way

$$\ln \psi_{ij} = \phi + \rho_{1i} + \rho_{2j} + \sigma_{1i}\sigma_{2j}$$

includes a constant association parameter (ϕ), both row and colum effects as well as interactions between the rows and columns. As such, this model resembles Goodman's R+C+RC model. Model fitting is done via Newton-Raphson or Fisher scoring [a].

The following examples shall be used to illustrate some of the previously discussed ideas. Further developments will be provided in the next chapter.

British Occupational Status Study

This study which was already analysed in Goodman [1979] comprises a sample of subjects cross-classified according to their father's occupational status and their own status membership. Data are available in the R package `gnm` as `occupationalStatus` and are reproduced in Tab. 1.1. We shall notice, however, that Molenberghs & Verbeke's data considers 7 categories only while our R dataset includes 8 categories. We could aggregate row and column 5 and 6 in our dataset to match that of Molenberghs & Verbeke, but the file `occupationalStatus.dat` should fix the problem.

```
> occupationalStatus <- matrix(scan("occupationalStatus.dat", sep = ","),
+     nr = 7)
> dimnames(occupationalStatus) <- c(list(1:7), list(1:7))
```

	1	2	3	4	5	6	7
1	50	19	26	8	18	6	2
2	16	40	34	18	31	8	3
3	12	35	65	66	123	23	21
4	11	20	58	110	223	64	32
5	14	36	114	185	714	258	189
6	0	6	19	40	179	143	71
7	0	3	14	32	141	91	106

Table 1.1: British Occupational Study. Cross-classification of male sample according to each subject's occupational status category (column) and his father's occupational category (row).

Let's apply the preceding models.

To fit an RC model, we need the `VGAM` package which provides the `grc` function. First, we can specify a constant association in the following way:

```
> library(VGAM)
> options(contrasts = c("contr.treatment", "contr.poly"))
> BOSS.rc <- grc(occupationalStatus, Rank = 1)
> BOSS.rc

Call:
rrvglm(formula = as.formula(str2), family = poissonff, data = .grc.df,
    control = myrrcontrol, constraints = cms)

Coefficients:

Residual Deviance: 75.58977
Log-likelihood: -164.9052
```

We find a residual deviance ($\chi^2(25)$) of 75.59, indicating that this model does not provide a very good fit. If we allow for interaction between rows and columns, by specifying `Rank=2`, residual deviance now is 36.05 with 16 degrees of freedom.

With Dale's model, we need to resort to

The Caithness Data

In another study, Goodman [1981] uses an association model for two-way contingency tables with ordered categories (eye and hair color of 5387 children, see Tab. 1.2). Data also are available in the package `MASS` as `caith`.

	fair	red	medium	dark	black
blue	326	38	241	110	3
light	688	116	584	188	4
medium	343	84	909	412	26
dark	98	48	403	681	85

Table 1.2: Caithness Data. Eye color (rows) and hair color (columns) of 5387 children in Caithness.

Other useful graphical summary for such a table are provided by mosaic plots, which divide the plotting surface recursively according to the proportions of each factor in turn. The following snippet is taken from Venables and Ripley [2002].

```
> caith1 <- as.matrix(caith)
> names(dimnames(caith1)) <- c("eyes", "hair")
> mosaicplot(caith1, color = T, main = "")
```

Before applying a marginal model such as the one decribed in the preceding section, let's take a look at the simplest model of nominal association [see e.g. Venables and Ripley, 2002, p. 326].

```
> corresp(caith)

First canonical correlation(s): 0.4463684

 Row scores:
      blue      light     medium       dark
-0.89679252 -0.98731818  0.07530627  1.57434710

 Column scores:
      fair        red     medium       dark      black
-1.21871379 -0.52257500 -0.09414671  1.31888486  2.45176017
```

Raw and column scores are scaled by ρ, the first canonical correlation, and are (by construction) maximally correlated. Indeed, if we consider an $r \times c$ table of counts (N) and choose R and C as matrices of indicators for the rows and columns, such that $R^T C = N$, we can apply a singular value decomposition to their correlation matrix

$$X_{ij} = \frac{n_{ij}/n - (n_{i\cdot}/n)(n_{\cdot j}/n)}{\sqrt{(n_{i\cdot}/n)(n_{\cdot j}/n)}} = \frac{n_{ij} - n r_i c_j}{n\sqrt{r_i c_j}}$$

where $r_i = n_{i\cdot}/n$ and $c_j = n_{\cdot j}/n$) are the proportions in each row and column. According to Venables and Ripley [2002], let D_r and D_c be the diagonal matrices of r and c. Then, correspondence analysis corresponds to selecting the first singular value and left and right singular vectors of X_{ij} and rescaling by $D_r^{-1/2}$ and $D_c^{-1/2}$.

The package `ca` provides more detailed summary and useful graphical maps of row/columns scores, which I find more pretty than the `biplot` function of `MASS` (see Fig. 1.1, right).

```
> library(ca)
> ca(caith)

 Principal inertias (eigenvalues):
           1        2        3
Value      0.199245 0.030087 0.000859
Percentage 86.56%   13.07%   0.37%


 Rows:
             blue     light    medium     dark
Mass     0.133284  0.293299  0.329311 0.244106
ChiDist  0.437855  0.450620  0.247359 0.715398
Inertia  0.025553  0.059557  0.020149 0.124932
Dim. 1  -0.896793 -0.987318  0.075306 1.574347
Dim. 2   0.953623  0.510004 -1.412478 0.772036
```

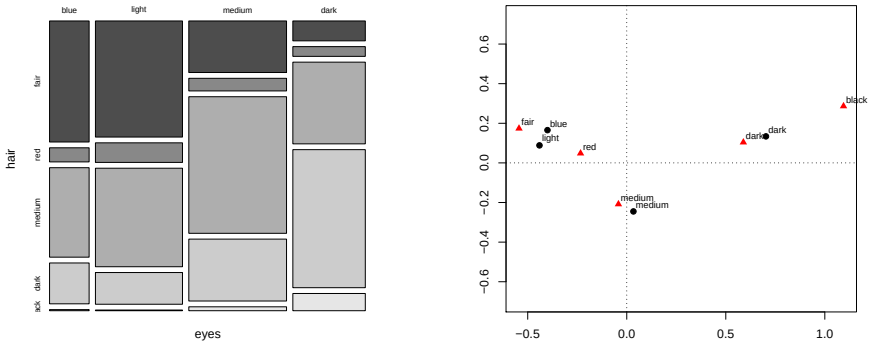

Figure 1.1: Correspondence analysis of Caithness Data.

```
 Columns:
              fair       red    medium     dark    black
Mass      0.270095  0.053091  0.396696 0.258214 0.021905
ChiDist   0.571235  0.265854  0.212526 0.597901 1.132193
Inertia   0.088134  0.003752  0.017918 0.092308 0.028079
Dim. 1   -1.218714 -0.522575 -0.094147 1.318885 2.451760
Dim. 2    1.002243  0.278336 -1.200909 0.599292 1.651357
```

```
> plot(ca(caith))
```

In that case, we found again row and column scores reported on the `Dim. 1` row in the summary output. There are obviously the same as those computed with the `corresp` function.

Going back to our marginal model, we may first fit a simple model which assumes independance between both responses but Goodman has shown that this does not provide a good fit. We try ...

1.3.2 Analyzing 3-way contingency tables

Dale's Model has been extended to tables with arbitrary dimensions. Details can be found in Molenberghs and Lesaffre [1994]. Here is a brief outline of the basics of multi-way analysis: let Y_1, Y_2 and Y_3 be our three variables with I, J and K levels respectively. Now, we can define the cumulative three-way probabilities as μ_{ijk} $(i = 1, \ldots, I,\ j = 1, \ldots, J\ ,\ k = 1, \ldots, K)$. Marginal paramaters are defined as

$$\begin{aligned} \eta_{1i} &= \text{logit}[\Pr(Y_1 \leq i)] = \ln(\mu_{iJK}) - \ln(1 - \mu_{iJK}), \\ \eta_{2j} &= \text{logit}[\Pr(Y_2 \leq j)] = \ln(\mu_{IjK}) - \ln(1 - \mu_{IjK}), \\ \eta_{3k} &= \text{logit}[\Pr(Y_3 \leq k)] = \ln(\mu_{IJk}) - \ln(1 - \mu_{IJk}). \end{aligned} \tag{1.17}$$

while pairwise association parameters are defined as

$$\begin{aligned}
\ln \psi_{12,ij} &= \ln \frac{\mu_{ijK}(1-\mu_{IjK}-\mu_{iJK}+\mu_{ijK})}{(\mu_{iJK}-\mu_{ijK})(\mu_{IjK}-\mu_{ijK})}, \\
\ln \psi_{13,ik} &= \ln \frac{\mu_{iJk}(1-\mu_{IJk}-\mu_{iJK}+\mu_{iJk})}{(\mu_{iJK}-\mu_{iJk})(\mu_{IJk}-\mu_{iJk})}, \\
\ln \psi_{23,jk} &= \ln \frac{\mu_{Ijk}(1-\mu_{IJk}-\mu_{IjK}+\mu_{Ijk})}{(\mu_{IJk}-\mu_{Ijk})(\mu_{IjK}-\mu_{Ijk})}.
\end{aligned} \tag{1.18}$$

1.4 Likelihood-based Marginal Models

We will here focus our attention on the Bahadur Model and extend the Dale's Model based on global odds ratio. These models not only allow us to study association structures, as before, but also permit to analyse the joint probability of a number of events occurring at two consecutive measurements. Furthermore, they form the basis of other methods not encompassed by the likelihood approach and which shall be described later on.

1.4.1 The Bahadur Model

In the case of binary data, the model proposed by Bahadur [1961], in the context of clustered data, aims at describing the probability of observing a positive response taking into account the within cluster correlation.

In its elementary form, the Bahadur's model considers that the marginal distribution Y_{ij} observed for measurement j on subject i can be considered as a Bernoulli experiment, such that $E(Y_{ij}) = \Pr(Y_{ij} = 1) \equiv \pi_{ij}$. First, we start by conditionning expectation upon any covariates X_i, and the association is simply described by the pairwise probability

$$\Pr(Y_{ij_1} = 1, Y_{ij_2} = 1) = E(Y_{ij_1} Y_{ij_2}) \equiv \pi_{ij_1 j_2}.$$

In other words, the "sucess probability" of two measurements from the same subject can be modeled in terms of marginal probabilities, the π_{ij}, together with the marginal correlation coefficient which can be expressed as:

$$\text{corr}(Y_{ij_1}, Y_{ij_2}) \equiv \rho_{ij_1 j_2} = \frac{\pi_{ij_1 j_2} - \pi_{ij_1}\pi_{ij_2}}{[\pi_{ij_1}(1-\pi_{ij_1})\pi_{ij_2}(1-\pi_{ij_2})]^{1/2}}. \tag{1.19}$$

The pairwise probability $\pi_{ij_1 j_2}$ (i.e. the second moment) can be shown to be

$$\pi_{ij_1 j_2} = \pi_{ij_1}\pi_{ij_2} + \rho_{ij_1 j_2}[\pi_{ij_1}(1-\pi_{ij_1})\pi_{ij_2}(1-\pi_{ij_2})]^{1/2}. \tag{1.20}$$

In addition to the first two moments, Bahadur considers third and higher order correlation coefficients, $\rho_{ij_1 j_2}, \rho_{ij_1 j_2 j_3}, \ldots, \rho_{i12\ldots n_i}$, thus completely specifying the joint distribution. With this approach, the general Bahadur model

takes the form $f(y) = f_1(y_i)c(y_i)$, where

$$\begin{aligned} f_1(y_i) &= \prod_{j=1}^{n_i} \pi_{ij}^{y_{ij}} (1-\pi_{ij}^{y_{ij}})^{1-y_{ij}} \\ c(y_i) &= 1 + \sum_{j_1<j_2} \rho_{ij_1j_2} e_{ij_1} e_{ij_2} + \sum_{j_1<j_2<j_3} \rho_{ij_1j_2j_3} e_{ij_1} e_{ij_2} e_{ij_3} \\ &\quad + \cdots + \rho_{i12\ldots n_i} e_{i1} e_{i2} \ldots e_{in_i}. \end{aligned} \tag{1.21}$$

We see that the probability mass function, $f(y)$, is nothing else than the product of an independence model ($f_1(y_i)$) and a correction factor ($c(y_i)$), which can be used to account for over-dispersion.

With (exchangeably) clustered data, the Bahadur model can be extended quite naturally. Details can be found in Molenberghs & Verbeke, but it can be shown that knowing $Z_i = \sum_{j=1}^{n_i} Y_{ij}$, the number of successes within a unit, with realized value z_i, is sufficient. This leads to the following model:

$$\begin{aligned} f_1(y_i) &= \pi_i^{z_i}(1-\pi_i)^{n_i-z_i} \\ c(y_i) &= 1 + \sum_{r=2}^{n_i} \rho_i(r) \sum_{s=0}^{r} \binom{z_i}{s} \binom{n_i-z_i}{r-s} (-1)^{s+r} \lambda_i^{r-2s}, \end{aligned} \tag{1.22}$$

with $\lambda_i = \sqrt{\pi_i/(1-\pi_i)}$. The probability mass function of Z_i is then

$$f(z_i) = \binom{n_i}{z_i} f(y_i).$$

If three and higher order correlations are assumed to be zero, this simplifies further to

$$\begin{aligned} f(z_i) &\equiv f(z_i \mid \pi_i, \rho_{i(2)}, n_i) = \binom{n_i}{z_i} \pi_i^{z_i}(1-\pi_i)^{n_i-z_i} \\ &\quad \times \left[1 + \rho_{i(2)} \left\{ \binom{n_i-z_i}{2} \frac{\pi_i}{1-\pi_i} - z_i(n_i-z_i) + \binom{z_i}{2} \frac{1-\pi_i}{\pi_i} \right\}\right]. \end{aligned} \tag{1.23}$$

This formulation is more pretty than the Dale's model, which has no closed form (i.e. free of integrals), but note that correlation between two responses are constrained by the fact that higher order correlations have been removed. Such a restriction on correlation parameters is discussed by Molenberghs & Verbeke (page 89).

The marginal parameters π_i and $\rho_{i(2)}$ have to be modelled using a composite link. For π_i, the logistic function arises naturally as Y_{ij} are binary, while for $\rho_{i(2)}$ a Fisher's z-transform has to be applied. This leads to the following *generalized linear regression*:

$$\begin{pmatrix} \ln\left(\frac{\pi_i}{1-\pi_i}\right) \\ \ln\left(\frac{1+\rho_{i(2)}}{1-\rho_{i(2)}}\right) \end{pmatrix} \equiv \eta_i = X_i\beta. \tag{1.24}$$

If we denote the log-likelihood contribution of the ith unit by

$$\ell_i = \ln f(z_i \mid \pi_i, \rho_{i(2)}, n_i),$$

the MLE of β, $\hat{\beta}$, is computed by nulling the score equations which are

$$U(\beta) = \sum_{i=1}^{N} X_i'(T_i')^{-1} L_i \tag{1.25}$$

where

$$\begin{aligned} T_i &= \frac{\partial \eta_i}{\partial \Theta_i} = \begin{pmatrix} \frac{\partial \eta_{i1}}{\partial \pi_1} & \frac{\partial \eta_{i2}}{\partial \pi_1} \\ \frac{\partial \eta_{i1}}{\partial \rho_{i(2)}} & \frac{\partial \eta_{i2}}{\partial \rho_{i(2)}} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\pi_i(1-\pi_i)} & 0 \\ 0 & \frac{2}{(1-\rho_{i(2)})(1+\rho_{i(2)})} \end{pmatrix}, \end{aligned}$$

and

$$L_i = \frac{\partial \ell_i}{\partial \Theta_i} = \begin{pmatrix} \frac{\partial \ell_i}{\partial \pi_i} \\ \frac{\partial \ell_i}{\partial \rho_{i(2)}} \end{pmatrix}.$$

Here, $\Theta_i = \begin{pmatrix} \pi_i \\ \rho_{i(2)} \end{pmatrix}$ is the set of natural parameters. Resolution of $U(\beta) = 0$ requires numerical algorithm, such as Newton-Raphson.

1.4.2 Fully-specified Marginal Models

1.4.3 The Multivariate Probit Model

1.5 The GEE approach

1.5.1 Advantages of the GEE over likelihood-based approaches

Though full likelihood approaches allow one to draw inference about the parameters of interest, together with the joint probabilities (as for the Bahadur Model), computational complexity increase with the number of repeated measurements while choosing the right full distributional specification is never very easy. When one is only interested in the marginal parameter and pairwise associations, quasi-likelihood methods provide an alternative way to model such correlated data. Mean response is modelled as a parametric function of covariates and variance is viewed as a function of the mean.

There is another alternative for studying both the marginal profile and the association structure: the Generalized Estimating Equations, first proposed by Liang and Zeger [1986].

1.5.2 Theoretical framework

First, we note that the score equations for a multivariate marginal normal model $Y_i \sim \mathcal{N}(X_i\beta, V_i)$ can be expressed as

$$\sum_{i=1}^{N} X_i'(A_i^{1/2} R_i A_i^{1/2})^{-1}(y_i - X_i, beta) = 0, \tag{1.26}$$

with V_i being the marginal covariance matrix. This covariance matrix has been decomposed such that

$$V_i = A_i^{1/2} R_i A_i^{1/2} \tag{1.27}$$

where A_i is a diagonal matrix of variances and R_i is the marginal correlation matrix.

1.5.3 Other GEE methods

1.6 Conditional Model

1.6.1 Transition Models

Transition models are a specific class of conditional models. A measurement Y_{ij} is modelled as a function of the previous outcomes or past events, denoted $h_{ij} = (Y_{i1}, \ldots, Y_{i,j-1})$ (see Diggle et al. [2002] for further description of this model). Such a model can be rewritten as a regression model with outcome Y_{ij} and predictors h_{ij}, or expressing the error term ε_{ij} as a function of the past error terms. This sounds like the time series approach. Note that Markov models also are subsumed in this general approach.

A stationary first-order autoregressive model for continuous data can be formulated in the following way (here, the order relates to the number of previous measurements that are considered to influence the current one):

$$Y_{i1} = x_{i1}'\beta + \varepsilon_{i1}, \tag{1.28}$$

$$Y_{ij} = x_{i1}'\beta + \alpha Y_{i,j-1} + \varepsilon_{i1}. \tag{1.29}$$

Such a model produces a marginal multivariate normal model with AR(1) covariance matrix, and is convenient for equally spaced outcomes. This can be shown by noting that $\mathrm{Var}(Y_{ij}) = \sigma^2$ and $\mathrm{cov}(Y_{ij}, Y_{ij'}) = \alpha^{|j'-j|}\sigma^2$, if we assume that $\varepsilon_{i1} \sim \mathcal{N}(0, \sigma^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2(1-\alpha^2))$.

Chapter 2

The Toenail data

2.1 The data

The data were obtained from a randomized, double-blind, parallel group, multicenter study for the comparison of two oral treatments (in what follows coded as A and B) for toenail dermatophyte onychomycosis (TDO), described in full detail by Backer et al. [1996]. TDO is a common toenail infection, difficult to treat, affectinf more than 2 out of 100 persons. Antifungal compounds, classically used for treatment of TDO, need to be taken until the whole nail has grown out healthy. The development of new compounds, however, has reduced the treatment duration to 3 months. The aim at the present study was to compare the efficacy and safety of 12 weeks of continuous therapy with treatment A or with treatment B.

In total, 2 x 189 patients were randomized, distributed over 36 centers. Subjects were followed during 12 weeks (3 months) of treatment and followed further, up to a total of 48 weeks (12 months). Measurements were taken at baseline, every month during treatment, and every 3 months afterwards, resulting in a maximum of 7 measurements per subject. At the firts occasion, the treating physician indicates one of the affected toenails as the target nail, the nail which will be followed over time. The analysis will be restricted to only those patients for which the target nail was one of the two big toenails. This reduces the sample under consideration to 146 and 148 subjects, in group A and group B, respectively.

One of the responses of interest was the unaffected nail length, measured from the nail bed to the infected part of the nail, which is always at the free end of the nail, expressed in mm. This outcome has been extensively studied in the other textbook dealing with continuous measurements data (Verbeke and Molenberghs, 2000). Another important outcome in this study was the severity of the infection, coded as 0 (not severe) or 1 (severe). The question of interest was whether the percentage of severe infection decreased over time, and whether that evolution was different for the two treatment groups.

Due to a variety of reasons, the outcome has been measured at all 7 scheduled time points, for only 224 (76%) out of the 298 participants. It should be noted thta the occurence of missingness is similar in both treatment groups.

2.2 The basics

First, we load the data (stored as a csv file) and recode some of the variables, for clarity purpose:

```
> toenail <- toenail.raw <- read.csv("dataset/toenail02.csv", header = TRUE,
+     sep = ",")
> toenail$idnum <- as.factor(toenail$idnum)
> toenail$idnew <- as.factor(toenail$idnew)
> toenail$treatn <- as.factor(toenail$treatn)
> levels(toenail$treatn) <- LETTERS[1:2]
> toenail$time <- factor(toenail$time, ordered = TRUE)
```

The variable `y` is our response variable (presence or absence of toenail dermatophyte onychomycosis, TDO), while `treatn` and `time` stand for the treatment group and follow-up timeline, respectively. Let's check the structure and provide a very short summary of the data:

```
> head(toenail)
```

```
  idnum time treatn y idnew
1     1    0      B 1     1
2     1    1      B 1     1
3     1    2      B 1     1
4     1    3      B 0     1
5     1    6      B 0     1
6     1    9      B 0     1
```

```
> summary(toenail)
```

```
     idnum       time     treatn        y               idnew      
 1      :   7   0 :294   A:937   Min.   :0.0000   1      :   7  
 3      :   7   1 :288   B:970   1st Qu.:0.0000   3      :   7  
 4      :   7   2 :283           Median :0.0000   4      :   7  
 6      :   7   3 :272           Mean   :0.2139   5      :   7  
 7      :   7   6 :263           3rd Qu.:0.0000   6      :   7  
 9      :   7   9 :243           Max.   :1.0000   7      :   7  
 (Other):1865   12:264                            (Other):1865  
```

By simply looking at the raw proportion, we can see that there seems to be little difference between treatment A and treatment B.

```
> with(toenail, table(treatn, y))
```

```
      y
treatn   0   1
     A 723 214
     B 776 194
```

We shall confirm this intuition by examining row and column profiles, which can be computed as follows:

```
> trt.tab <- with(toenail, table(treatn, y))
> row.pct <- sweep(trt.tab, 1, apply(trt.tab, 1, sum), "/")
> col.pct <- sweep(trt.tab, 2, apply(trt.tab, 2, sum), "/")
> round(row.pct, 2)

      y
treatn    0    1
     A 0.77 0.23
     B 0.80 0.20

> round(col.pct, 2)

      y
treatn    0    1
     A 0.48 0.52
     B 0.52 0.48
```

However, a more useful summary would consist in summarizing the mean result observed per condition, i.e. including time variable, which can be done as follow (using the fact that for a binary variable, the mean equals the empirical frequency):

```
> with(toenail, round(tapply(y, list(treatn, time), mean), 2))

     0    1    2    3    6    9   12
A 0.37 0.35 0.32 0.22 0.11 0.09 0.11
B 0.37 0.33 0.28 0.21 0.06 0.06 0.05
```

There, we see that the frequency of severe infection evolves over time, compared to the baseline condition (`time=0`). The decline of TDO frequency might be more apparent with the help of a simple figure (Fig. 2.1).

```
> with(toenail, interaction.plot(time, treatn, y, ylim = c(0, 0.5),
+     ylab = "Frequency", legend = FALSE, main = "Raw data"))
> legend("topright", paste("Treatment", levels(toenail$treatn)),
+     lty = 1:2)
```

Now, we can start with a simple logistic regression model, discarding the correlated structure. We will thus express the logit of the proportion of positive response as a linear combination of the predictors, including the 2nd order interaction term. This model can be expressed as

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 T_i + \beta_2 t_i + \beta_3 T_i t_i, \qquad Y_i \sim \mathcal{B}(\pi_i)$$

where we assume that individual responses Y_i follow a Bernoulli distribution and T and t represent the treatment and time variables. Here, we also assume specific linear time trends for both treatment groups. The intercept β_0 would represent the mean baseline response.

This can be done as follow with R:

```
> toenail.glm <- glm(y ~ treatn * time, data = toenail.raw, family = binomial)
> summary(toenail.glm)
```

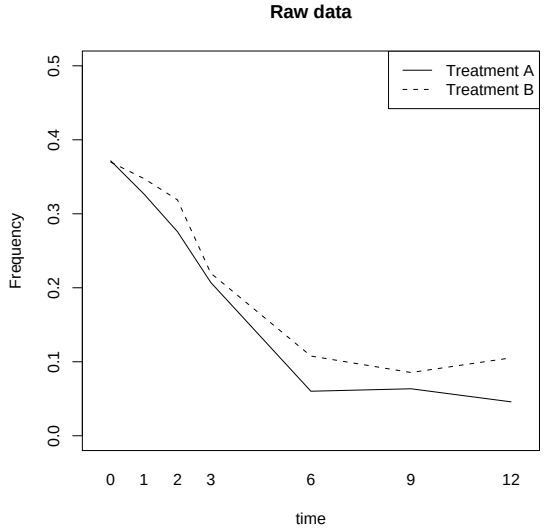

Figure 2.1: The Toenail data. Mean responses in each condition.

```
Call:
glm(formula = y ~ treatn * time, family = binomial, data = toenail.raw)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9608  -0.7769  -0.4891  -0.2331   2.6905

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.55706    0.10904  -5.109 3.24e-07 ***
treatn       0.02358    0.15648   0.151    0.880
time        -0.17693    0.02456  -7.205 5.82e-13 ***
treatn:time -0.07798    0.03944  -1.977    0.048 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1980.0  on 1906  degrees of freedom
Residual deviance: 1811.7  on 1903  degrees of freedom
AIC: 1819.7

Number of Fisher Scoring iterations: 5
```

There is a significant interaction ($p = 0.048$) which means that trends might be different in the two groups. We get some clues about the data, but we obviously are on the wrong way since we have not taken into account the multiple responses per subject. As can be seen from the output, we have considered 1907 independant observations (residual degrees of freedom + 1).

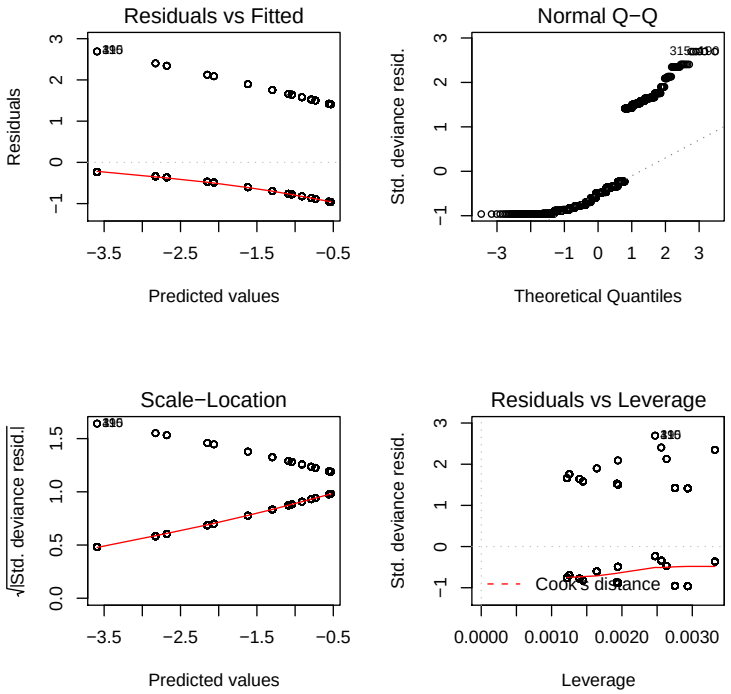

Figure 2.2: The Toenail data. Diagnostic plots for the logistic regression.

```
> par(mfrow = c(2, 2))
> plot(toenail.glm, cex = 0.8)
```

Some useful plots for model diagnostics are shown in Figure 2.2. From left tor right, and from top to bottom, it shows: (a) residuals vs. fitted values, (b) a normal Q-Q plot for standardized residuals, (c) standardized residuals vs. fitted values, and (d) leverage effect for each observation.

As we already pointed out, using a logit link rather than a probit doe not make a great difference, except for extreme values. Indeed, the following piece of code, adapted from Faraway [2006], shows how the choice of the link function affects the probability distribution. For clarity purpose, we use a reduced model, including only the `time` dependent variable. We thus fit the data assuming binomially distributed errors, with logit, probit and complementary log-log. The `logit` function is provided by the `VGAM` package.

```
> toenail.glml <- glm(y ~ time, data = toenail.raw, family = binomial(logit))
> toenail.glmp <- glm(y ~ time, data = toenail.raw, family = binomial(probit))
> toenail.glmc <- glm(y ~ time, data = toenail.raw, family = binomial(cloglog))
> x <- seq(0, 12, 0.2)
> pl <- logit(toenail.glml$coef[1] + toenail.glml$coef[2] * x,
+     inverse = TRUE)
> pp <- pnorm(toenail.glmp$coef[1] + toenail.glmp$coef[2] * x)
```

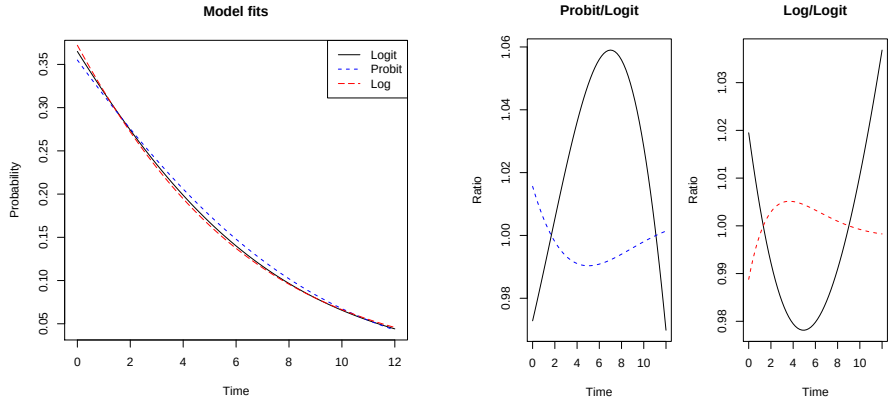

Figure 2.3: The Toenail data. (Left) Logistic regression with three different link functions. (Right) Relative ratio between the predicted probabilities according to selected models.

```
> pc <- 1 - exp(-exp((toenail.glmc$coef[1] + toenail.glmc$coef[2] *
+     x)))
```

We then plot the predicted probability of a positive response depending on time value.

```
> plot(x, pl, type = "l", ylab = "Probability", xlab = "Time",
+     main = "Model fits")
> lines(x, pp, lty = 2, col = "blue")
> lines(x, pc, lty = 5, col = "red")
> legend("topright", c("Logit", "Probit", "Log"), lty = c(1, 2,
+     5), col = c("black", "blue", "red"))
```

As can be seen, there is very littele difference between the three fits. To show how the choice of link function affects the tails of the distribution, we can plot the relative ratio between each link function for the varying time (Fig. 2.3).

```
> op <- par(mfrow = c(1, 2), mar = c(5, 4, 4, 1))
> matplot(x, cbind(pp/pl, (1 - pp)/(1 - pl)), type = "l", xlab = "Time",
+     ylab = "Ratio", main = "Probit/Logit", col = c("black", "blue"))
> matplot(x, cbind(pc/pl, (1 - pc)/(1 - pl)), type = "l", xlab = "Time",
+     ylab = "Ratio", main = "Log/Logit", col = c("black", "red"))
> par(op)
```

Difference between model fits can be found along all the domain of the dependent variable. However, they are more important for the lowest and highest values of `time`. Indeed, as the underlying distributions differ in their respective tail, for instance the standard normal and the logistic, it should not be too surprising.

Finally, we can compare the deviance of these three models:

```
> toenail.glml$deviance
```

```
[1] 1818.263
> toenail.glmp$deviance
[1] 1821.744
> toenail.glmc$deviance
[1] 1816.785
```

Very little variation are observed between these models suggesting that the logit link provides sufficient information.

2.3 Marginal model

The following model will be used throughout this section:

$$\log(\frac{\pi_{ij}}{1-\pi_{ij}}) = \beta_0 + \beta_1 T_i + \beta_2 t_{ij} + \beta_3 T_i t_{ij}$$

and we assume that $Y_{ij} \sim \mathcal{B}(\pi_{ij})$, that is the outcome follow a Bernoulli distribution with parameter $p = \pi_{ij}$. The variable t_{ij} represent the jth measurement taken from subject i.

In order to fit a GEE model, we need to load the `gee` library. We can formulate the model as is within R except for the repeated statement which has to be reported in the `id` parameter.

```
> toenail.gee <- gee(y ~ treatn * time, id = idnum, data = toenail.raw,
+     family = binomial, scale.fix = TRUE)
(Intercept)      treatn        time treatn:time
-0.55705808  0.02357666 -0.17692959 -0.07797568
```

Here, we use the option `scale.fix=TRUE` to allow comparison with the SAS output, but there is no really need to constrain the scale to be 1. We also use an *independance* working correlation matrix (the default in R). We can check that correlated data have been taken into account, unlike in the previous logistic approach:

```
> toenail.gee

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Logit
 Variance to Mean Relation: Binomial
 Correlation Structure:     Independent

Call:
gee(formula = y ~ treatn * time, id = idnum, data = toenail.raw,
    family = binomial, scale.fix = TRUE)

Number of observations :  1907
```

```
Maximum cluster size   :  7


Coefficients:
(Intercept)      treatn        time treatn:time
-0.55705808  0.02357666 -0.17692959 -0.07797568

Estimated Scale Parameter:  1
Number of Iterations:  1

Working Correlation[1:4,1:4]
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1


Returned Error Value:
[1] 0
```

Of course, displaying the working correlation is not very useful since this the identity matrix because we specify an independance hypothesis. We are now working with correlated measurements and the maximum cluster size is 7. More précisely, we can compute the number of independant observations:

```
> length(unique(toenail$idnum))

[1] 294
```

It could be useful to check the minimum cluster size, which is

```
> min(with(toenail, table(y, idnum))) + 1

[1] 1

> sum(table(toenail$idnum) == 1)

[1] 5
```

Thus, the number of measurements per subject varies from 1 to 7, and there are in fact 5 individuals that have only one measurement. This explains why we don't get perfect a correspondance between the number of measurements (1907) and the number of clusters (294). We can identify the individual with only one measurement, because we might want to remove them later on:

```
> names(table(toenail$idnum)[table(toenail$idnum) == 1])

[1] "45"  "48"  "63"  "99"  "377"

> toenail[toenail$idnum == 45, ]

    idnum time treatn y idnew
207    45    0      A 1    34
```

For instance, subject whose `id` is 45 has been measured only at time 0. This subject has perhabs leave after the beginning the study.

We obtain quite comparable results as those displayed by Molenberghs & Verbeke (pp. 206–212):

```
> summary(toenail.gee)

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Logit
 Variance to Mean Relation: Binomial
 Correlation Structure:     Independent

Call:
gee(formula = y ~ treatn * time, id = idnum, data = toenail.raw,
    family = binomial, scale.fix = TRUE)

Summary of Residuals:
        Min          1Q      Median          3Q         Max
-0.36970527 -0.26051531 -0.11275408 -0.02679591  0.97320409


Coefficients:
               Estimate Naive S.E.    Naive z Robust S.E.    Robust z
(Intercept) -0.55705808 0.10903934 -5.1087806  0.17133792 -3.25122476
treatn       0.02357666 0.15648046  0.1506684  0.25055984  0.09409591
time        -0.17692959 0.02455777 -7.2046266  0.03016880 -5.86465399
treatn:time -0.07797568 0.03943713 -1.9772150  0.05459973 -1.42813317

Estimated Scale Parameter:  1
Number of Iterations:  1

Working Correlation
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]    1    0    0    0    0    0    0
[2,]    0    1    0    0    0    0    0
[3,]    0    0    1    0    0    0    0
[4,]    0    0    0    1    0    0    0
[5,]    0    0    0    0    1    0    0
[6,]    0    0    0    0    0    1    0
[7,]    0    0    0    0    0    0    1
```

First of all, we shall note that the initial estimates are comparable to those found with the logistic regression. This is because initial estimates are computed by fitting an ordinary logistic regression, ignoring the correlation structure. R also computes two kinds of estimates for standard errors: the first column displays the standard errors based on the assumption of uncorrelated measurements. Obviously, these are the same as in the logistic case since the first step in GEE fitting assumes independent observations. The sandwich estimators are found under the column headed `Robust S.E.`. We can see that these empirical corrected SE are quite a bit larger than the model-based ones: this implies that ignoring the correlation in these data could lead to invalid conclusions. With the model-based SE, we would report the treatment by time interaction as marginally significant ($z = -1.977$, `2*pnorm(-1.9772150)` gives a p-value of 0.048), while using robust SE yields

a different picture ($z < 1.96$). We can also see differences between both model when inspecting the full asymptotic covariance matrices[1]:

```
> round(toenail.gee$naive.variance, 4)

            (Intercept)  treatn    time treatn:time
(Intercept)      0.0119 -0.0119 -0.0018      0.0018
treatn          -0.0119  0.0245  0.0018     -0.0041
time            -0.0018  0.0018  0.0006     -0.0006
treatn:time      0.0018 -0.0041 -0.0006      0.0016

> round(toenail.gee$robust.variance, 4)

            (Intercept)  treatn    time treatn:time
(Intercept)      0.0294 -0.0294 -0.0023      0.0023
treatn          -0.0294  0.0628  0.0023     -0.0067
time            -0.0023  0.0023  0.0009     -0.0009
treatn:time      0.0023 -0.0067 -0.0009      0.0030
```

Recall that the postulated variance matrix is used to get final estimates of the regression coefficients. In the next phase of the analysis, we will consider *exchangeable* and *unstructured* working assumptions. This is done by using the option `corstr="exchangeable"` and `corstr="unstructured"` when calling the `gee` function. In the case of the exchangeable structure, we get

```
> toenail.gee.exch <- gee(y ~ treatn * time, id = idnum, data = toenail.raw,
+     family = binomial, scale.fix = TRUE, corstr = "exchangeable")

(Intercept)      treatn        time treatn:time
-0.55705808  0.02357666 -0.17692959 -0.07797568

> toenail.gee.exch$coefficients

 (Intercept)       treatn         time  treatn:time
-0.584063732  0.009965551 -0.177032576 -0.086677389
```

With the unstructured variance matrix, the results are quite similar:

```
> toenail.gee.unstr <- gee(y ~ treatn * time, id = idnum, data = toenail.raw,
+     family = binomial, scale.fix = TRUE, corstr = "unstructured")

(Intercept)      treatn        time treatn:time
-0.55705808  0.02357666 -0.17692959 -0.07797568

> toenail.gee.unstr$coefficients

(Intercept)      treatn        time treatn:time
-0.68976394  0.08281554 -0.14825606 -0.10430442
```

[1] R clearly produces shorter output than SAS, but sometimes we might only be interested in a specific result, such as a p-value or a regression coefficient, and don't need the rest of the information. So, don't forget to look at the structure of the working R object with the `str` function, in order to be able to extract the specific piece of information you want. It saves time... and paper when we produce a report like this one.

Step	0	1	2	3
β_0	-0.5840	-0.5841	-0.5841	-0.5841
β_1	0.0144	0.0104	0.0100	0.0100
β_2	-0.1771	-0.1770	-0.1770	-0.1770
β_3	-0.0861	-0.0867	-0.0867	-0.0867

Table 2.1: Estimates obtained in 3 iterations from an GEE model with exchangeable correlation structure.

Before going more in details within these two models, we can see that the estimates are quite different depending on the covariance structure we specify as initial guesses. Furthermore, the estimates differ from what has been observed in the preceding case (independance assumption) and are now different from the initial conditions (computed as standard regression coefficients form a GLM). This is because several iterations are performed before converging to stable estimates. We can check this by displaying the number of iterations needed to achieve convergence. For instance, with the first model:

```
> toenail.gee.exch$iter

[1] 4
```

To get an idea of the values of the estimates at each step, we have to add the option `silent=FALSE`:

```
> gee(y ~ treatn * time, id = idnum, data = toenail.raw, family = binomial,
+     scale.fix = TRUE, corstr = "exchangeable", silent = F)
```

The step by step estimates are reported in Tab. 2.1 (step 0 means starting values).

Now, let's look at each model separately. With the exchangeable hypothesis, we assume that observations (ranging from 1 to 7) within a given cluster correlate in the same manner with each other. The treatment by time interaction still is non-significant, and now the asymptotic variance-covariance matrices are have changed to

```
> round(toenail.gee.exch$naive.variance, 4)

            (Intercept)  treatn   time treatn:time
(Intercept)      0.0181 -0.0181 -2e-04      0.0002
treatn          -0.0181  0.0349  2e-04     -0.0001
time            -0.0002  0.0002  4e-04     -0.0004
treatn:time      0.0002 -0.0001 -4e-04      0.0013

> round(toenail.gee.exch$robust.variance, 4)

            (Intercept)  treatn    time treatn:time
(Intercept)      0.0301 -0.0301 -0.0025      0.0025
treatn          -0.0301  0.0680  0.0025     -0.0078
time            -0.0025  0.0025  0.0010     -0.0010
treatn:time      0.0025 -0.0078 -0.0010      0.0032
```

Note, however, that these matrices are still far apart one from the other, suggesting to switch to an unstructured covariance matrix. Indeed, with such a large sample size but small cluster size, efficiency is questionable. If we take a look at the exchangeable working correlation, we see that it is estimated as

```
> toenail.gee.exch$working.correlation[1, 2]
```

```
[1] 0.4211848
```

We have already computed the coefficients (see Tab. 2.1, last column), and they differ from the previous ones, as a result of the constraints imposed on the correlation structure.

With the unstructured working assumptions, the parameters estimates also differ from what was obtained earlier. Here is the whole output:

```
> summary(toenail.gee.unstr)
```

```
 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Logit
 Variance to Mean Relation: Binomial
 Correlation Structure:     Unstructured

Call:
gee(formula = y ~ treatn * time, id = idnum, data = toenail.raw,
    family = binomial, corstr = "unstructured", scale.fix = TRUE)

Summary of Residuals:
        Min          1Q      Median          3Q         Max
-0.35275562 -0.24748530 -0.11669895 -0.02563883  0.97436117


Coefficients:
               Estimate Naive S.E.    Naive z Robust S.E.   Robust z
(Intercept) -0.68976394 0.16694967 -4.1315681  0.16787094 -4.1088942
treatn       0.08281554 0.23608105  0.3507929  0.24304125  0.3407469
time        -0.14825606 0.02748639 -5.3937997  0.02826793 -5.2446729
treatn:time -0.10430442 0.04541538 -2.2966758  0.05141970 -2.0284914

Estimated Scale Parameter:  1
Number of Iterations:  4

Working Correlation
          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
[1,] 1.0000000 0.8626806 0.6718167 0.4742889 0.2330428 0.1464755 0.1016829
[2,] 0.8626806 1.0000000 0.7783627 0.5629794 0.2523393 0.2185336 0.1226081
[3,] 0.6718167 0.7783627 1.0000000 0.7293826 0.2615086 0.1987628 0.1515076
[4,] 0.4742889 0.5629794 0.7293826 1.0000000 0.3388481 0.2596779 0.1973358
[5,] 0.2330428 0.2523393 0.2615086 0.3388481 1.0000000 0.4590733 0.3776915
[6,] 0.1464755 0.2185336 0.1987628 0.2596779 0.4590733 1.0000000 0.6075596
[7,] 0.1016829 0.1226081 0.1515076 0.1973358 0.3776915 0.6075596 1.0000000
```

One should notice that the treatment by time interaction is now significant, within both models, and the associated p-value can be computed, in the case of the model-based result, as:

```
> 2 * pnorm(-2.2966758)

[1] 0.02163727
```

We might prefer to make use of the robust SE and z statistics, but if we carefully look at the two sets of estimates, we would see that there is little differences between the two. This is at variance with what was observed in our last two models (independence and exchangeable assumptions). We could check that our model specification isn't too bad by again looking at the working correlation matrices:

```
> round(toenail.gee.unstr$naive.variance, 4)

            (Intercept)  treatn    time treatn:time
(Intercept)      0.0279 -0.0279 -0.0024      0.0024
treatn          -0.0279  0.0557  0.0024     -0.0048
time            -0.0024  0.0024  0.0008     -0.0008
treatn:time      0.0024 -0.0048 -0.0008      0.0021

> round(toenail.gee.unstr$robust.variance, 4)

            (Intercept)  treatn    time treatn:time
(Intercept)      0.0282 -0.0282 -0.0022      0.0022
treatn          -0.0282  0.0591  0.0022     -0.0058
time            -0.0022  0.0022  0.0008     -0.0008
treatn:time      0.0022 -0.0058 -0.0008      0.0026
```

The robust matrix is very close to the naive one, which indicate that the specifications we use to constrain the model are quite in agreement with the observed data. However, one should not forget that no inference can be made about this correlation structure. We may say that, within the unstructured assumption, there seems to be a decrease of the correlation between two consecutive measurements as the distance between them grows up. That's all.

Figure 2.4 (left) shows the fit provided by this latest model. As can be seen, curves are much "smoother" than those of the raw data shown in Fig. 2.1. Indeed, the negative slopes are smallest between 0 and 6, for both treatments, than there were in the raw data. In other words, the observed decrease in the frequency of toenail infections is higher than what would be expected according to this model.

```
> interaction.plot(toenail$time, toenail$treatn, toenail.gee.unstr$fitted.values,
+     ylim = c(0, 0.5), legend = FALSE, main = "Fitted data", xlab = "time",
+     ylab = "Expected frequency")
> legend("topright", paste("Treatment", levels(toenail$treatn)),
+     lty = 1:2)
```

Residuals can be plotted as follows :

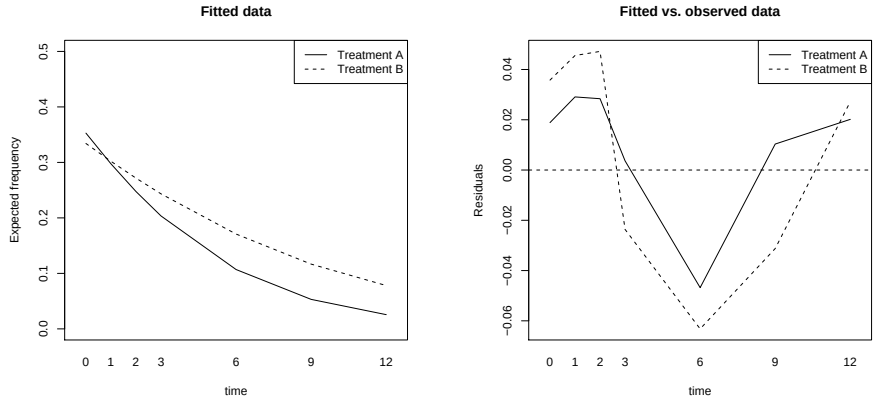

Figure 2.4: The Toenail data. (Left) Expected mean responses in each condition following an GEE fit with unstructured working covariance matrix. (Right) Residuals of the fit.

```
> interaction.plot(toenail$time, toenail$treatn, toenail.gee.unstr$residuals,
+     legend = FALSE, main = "Fitted vs. observed data", xlab = "time",
+     ylab = "Residuals")
> legend("topright", paste("Treatment", levels(toenail$treatn)),
+     lty = 1:2)
> abline(h = 0, lty = 2)
```

So, what can be concluded from this fit? Are there any problems with the predicted pattern observed between succesive measurements as compared to what was observed in the present study? Not at all. We are only faced with a slight difference between observed and expected results, which usually is the case when modeling any set of observed responses. Residuals can be seen in Figure 2.4 (right) and there is a higher (negative) deviation for the central time points. A more useful diagnostic plot would consist in plotting Pearson residuals against fitted values. Recall that Pearson residuals are defined as

$$r_j = \frac{y_j - m_j\hat{\pi}}{\sqrt{m_j\hat{\pi}(1-\hat{\pi})}}$$

where m_j is the number of trials with the jth covariate pattern, $\hat{\pi}$ is the expected proportional response (here, it is the fitted values from the model) and y_j is the number of successes with the jth covariate. We shall use such a plot when fitting the data using an autoregressive dependence between subunits (Sec. 2.5). Note that we can also use standardized Pearson residuals defined in the usual way as

$$r_{sj} = \frac{r_j}{\sqrt{1-h_j}}$$

where h_j stands for the leverage for the jth covariate.

Finally, this is not the end of the story, and, obviously, other models could be used, as we shall see in the next sections, but recall from Section 1.4 that we are modeling the mean responses (marginal approach).

2.4 Alternating logistic regression

Alternating logistic regression can be used to fit a marginal model when considering odds-ratio rather than raw frequency.

We will need the package `alr`, which can be downloaded from V. Carey's homepage. However, the package has been built for Windows. For those who like me cannot imagine installing this kind of operating system, you can find the same package targeted at Un*x platform at the following webpage: `www.aliquote.org/articles/tech/MDLD/`. In comparison to the SAS procedure `GENMOD`, the `alr` function only allows to fit binary correlated responses within three kind of dependence model: general, exchangeable, and 1-nested (see the help page). We will only compare the first two with the results obtained from SAS.

Here is what we obtain with this kind of model, after having set initial values for α (`ainit`):

```
> toenail.alr <- alr(y ~ treatn * time, id = idnum, ainit = 0.01,
+     data = toenail.raw)
> summary(toenail.alr)
```

2.5 Conditional model

The data can be analysed using a transition model by considering the model:

$$\text{logit}(\frac{\mu_{ij}}{1-\mu_{ij}}) = \beta_0 + \beta_1 T_i + \beta_2 t_{ij} + \beta_3 T_i t_{ij} + \alpha_1 y_{i,j-1}$$

where we still assume $Y_{ij} \sim \mathcal{B}(\mu_{ij})$, α being used to introduce the stationary autoregressive dependence. Here, we consider a first-order AR model. We can similarly introduce an AR(2) formulation within the same expression by just including an additional term $\alpha_2 y_{i,j-2}$.

Comparison with an GEE analysis including an AR(1) structure. At the time of this writtings, I cannot run successfully such a model with the `gee` package. The basic formula interface would look like:

```
> gee(y ~ treatn * time, id = idnum, data = toenail.raw, family = binomial,
+     corstr = "AR-M", Mv = 1)
```

but I get an error message due to cluster size. In fact, as checked page 28, we get a minimum cluster size of 1. I thus use `geepack`[2] which provides

[2]Replicating the previous analyses with this package lead to the same results compared to Carey's original package `gee`, and we can be quite confident in the computational com-

quite the same functionnality. However, there is some limitations in this case, and we cannot fit a second-order autoregressive model with `geepack`. Nonetheless, we could specify the correlation structure in the `zcor` optional parameter.

```
> toenail.gee.ar1 <- geeglm(y ~ treatn * time, data = toenail.raw,
+     id = idnum, family = binomial, corstr = "ar1")
> summary(toenail.gee.ar1)

Call:
geeglm(formula = y ~ treatn * time, family = binomial, data = toenail.raw,
    id = idnum, corstr = "ar1")

 Coefficients:
             Estimate    Std.err       Wald        p(>W)
(Intercept) -0.6451858 0.16997404 14.4080216 1.471740e-04
treatn       0.1168047 0.24964689  0.2189113 6.398707e-01
time        -0.1426984 0.02854287 24.9944092 5.749679e-07
treatn:time -0.1172862 0.05519906  4.5147140 3.360450e-02

Estimated Scale Parameters:
            Estimate   Std.err
(Intercept) 1.022523 0.4532843

Correlation: Structure = ar1  Link = identity

Estimated Correlation Parameters:
      Estimate   Std.err
alpha 0.6847294 0.1608756
Number of clusters:   294   Maximum cluster size: 7
```

As can be seen from the summary output, the estimated correlation is found to be 0.68. Again, the fitted values can be plotted against the explanatory variable as

```
> interaction.plot(toenail$time, toenail$treatn, toenail.gee.ar1$fitted.values,
+     ylim = c(0, 0.5), legend = FALSE, main = "Fitted data", xlab = "time",
+     ylab = "Expected frequency")
> legend("topright", paste("Treatment", levels(toenail$treatn)),
+     lty = 1:2)
```

but the `geepack` package offers an additional way to inspect graphically the fit (but see `?plot.geeglm`). However, due to some conflict between all fitted methods called by the various packages we are using, I adapt a little bit the `plot.geeglm` function and load it into the workspace before calling it[3]:

parability between the two packages. As quoted from a discussion found on R-help archive (2003), the two packages are using different estimators for the correlation parameter, and therefore different weights for the observations. This is a widespread issue with GEE.

[3] a nicer way to solve this problem would be to detach the conflicting package(s), but as I'm compiling this document using Sweave, this would force me to play with `attach` and `detach` function all the time...

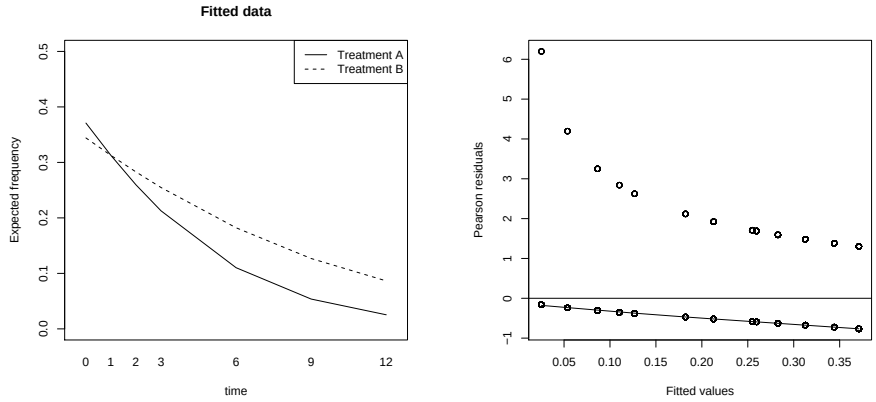

Figure 2.5: The Toenail data. (Left) Expected mean responses in each condition following an GEE fit with an embedded AR1 process. (Right) Fitted values and residuals plotted against each other.

```
> my.plot <- function(x, ...) {
+     xx <- x$fitted.values
+     rp.raw <- x$residuals
+     rp <- rp.raw/sd(xx)
+     plot(xx, rp, ylab = "Pearson residuals", xlab = "Fitted values")
+     abline(h = 0)
+     m <- lowess(rp ~ xx)
+     lines(m)
+ }
```

Note that there is also an `anova` method for R `geeglm` object. Using this method, we get

```
> anova(toenail.gee.ar1)

Analysis of 'Wald statistic' Table
Model: binomial, link: logit
Response: y
Terms added sequentially (first to last)

            Df     X2 P(>|Chi|)
treatn       1  0.771     0.380
time         1 56.459  5.74e-14
treatn:time  1  4.515     0.034
```

from which we found that time effect is highly significant, which is not very surprising given the observed slopes (see Fig. 2.1). Treatment by time interaction also appears to be significant at the .05 level.

One should, however, not forget to keep an eye open when interpreting the p-value and the Wald or t statistic resulting from GLMs[b]. We here recall the reader to the so-called Hauck-Donner phenomenon [Hauck and Dooner, 1977]. In the case of logistic regression, a small t-value indicates

either an insignificant or a very significant effect, but `step.glm` assumes the first. This is not very likely to happen in our context, but keep in mind that statistics give in essence numbers that have to be interpreted, and one should not accept significant results without investigating further the data. Quoting Brian Ripley (*s-news mailing list*, 1998):

> There is a little-known phenomenon for binomial GLMs that was pointed out by Hauck & Donner (1977: JASA 72:851-3). The standard errors and t values derive from the Wald approximation to the log-likelihood, obtained by expanding the log-likelihood in a second-order Taylor expansion at the maximum likelihood estimates. If there are some $\hat{\beta}_i$ which are large, the curvature of the log-likelihood at $\hat{\vec{\beta}}$ can be much less than near $\beta_i = 0$, and so the Wald approximation underestimates the change in log-likelihood on setting $\beta_i = 0$. This happens in such a way that as $|\hat{\beta}_i| \to \infty$, the t statistic tends to zero. Thus highly significant coefficients according to the likelihood ratio test may have non-significant t ratios.
>
> To expand a little, if |t| is small it can EITHER mean than the Taylor expansion works and hence the likelihood ratio statistic is small OR that $|\hat{\beta}_i|$ is very large, the approximation is poor and the likelihood ratio statistic is large. (I was using 'significant' as meaning practically important.) But we can only tell if $|\hat{\beta}_i|$ is large by looking at the curvature at $\beta_i = 0$, not at $|\hat{\beta}_i|$.
>
> There is one fairly common circumstance in which both convergence problems and the Hauck-Donner phenomenon (and trouble with `step`) can occur. This is when the fitted probabilities are extremely close to zero or one. Consider a medical diagnosis problem with thousands of cases and around fifty binary explanatory variables (which may arise from coding fewer categorical factors); one of these indicators is rarely true but always indicates that the disease is present. Then the fitted probabilities of cases with that indicator should be one, which can only be achieved by taking $\hat{\beta}_i = \infty$. The result from `glm` will be warnings and an estimated coefficient of around ± 10 (and an insignificant t value).

Chapter 3

The Epilepsy data

3.1 The data

The data are obtained from a randomized double-blind, parallel group multicenter study for the comparison of placebo with a new anti-epileptic drug (AED), in combination with one or two other AED's. The study is described in full detail in Faught et al. [1996]. The randomization of epilepsy patients took place after a 12-week baseline period that served as a stabilization period for the use of AED's, and during wich the number of seizures were counted. After that period, 45 patients were assigned to the placebo group, 44 to the active (new) treatment group. Patients were then measured seekly. Patients were followed (double-blind) during 16 weeks, after which they were entered into a long-term open-extension study. Some patients were followed for up to 27 weeks. The outcome of interest is the number of epileptic seizures experienced during the last week, i.e., since the last time the outcome was measured. The key research question is whether or not the additional new treatment reduces the number of epileptic seizures.

There is an unstable behavior that could be applained by the presence of extreme values, but it is also accounted by the fact that very little obervations are available. There is also a serious drop in the the number of measurements past the end of the actual double-blinde period, i.e., past week 16.

3.2 The basics

First, we load the data and get the things right.

```
> epilepsy <- read.csv("dataset/epilepsy.csv", header = TRUE, sep = ",")
> epilepsy$id <- as.factor(epilepsy$id)
> epilepsy$trt.fac <- as.factor(epilepsy$trt)
> epilepsy$sex <- as.factor(epilepsy$sex)
> epilepsy$race <- as.factor(epilepsy$race)
```

First, we can take a look at the first observations and display a simple summary of the data:

```
> head(epilepsy)
```

```
  trt   id sex age race height weight bserate    date0 studyweek nseizw trt.fac
1   0 1204   1  30    2     71    168     4.3 26/09/88         1      1       0
2   0 1204   1  30    2     71    168     4.3 26/09/88         2      0       0
3   0 1204   1  30    2     71    168     4.3 26/09/88         3      3       0
4   0 1204   1  30    2     71    168     4.3 26/09/88         4      1       0
5   0 1204   1  30    2     71    168     4.3 26/09/88         5      0       0
6   0 1204   1  30    2     71    168     4.3 26/09/88         6      1       0
```

```
> summary(epilepsy)
```

```
      trt               id       sex           age        race
 Min.   :0.0000   601501 :  27   1:1161   Min.   :19.00   1:1268
 1st Qu.:0.0000   602423 :  27   2: 258   1st Qu.:29.00   2: 151
 Median :1.0000   611602 :  27            Median :33.00
 Mean   :0.5095   601206 :  23            Mean   :34.91
 3rd Qu.:1.0000   601801 :  22            3rd Qu.:40.00
 Max.   :1.0000   601618 :  21            Max.   :68.00
                  (Other):1272
     height          weight         bserate            date0
 Min.   :59.00   Min.   : 88.0   Min.   :  4.00   08/11/88:  71
 1st Qu.:67.00   1st Qu.:155.0   1st Qu.:  6.00   16/07/90:  53
 Median :69.00   Median :176.0   Median :  9.80   10/06/88:  35
 Mean   :68.88   Mean   :175.3   Mean   : 19.14   10/08/89:  33
 3rd Qu.:72.00   3rd Qu.:196.0   3rd Qu.: 20.70   23/09/88:  33
 Max.   :76.00   Max.   :270.0   Max.   :198.30   24/05/90:  33
                                                  (Other) :1161
   studyweek          nseizw       trt.fac
 Min.   : 1.000   Min.   : 0.000   0:696
 1st Qu.: 5.000   1st Qu.: 0.000   1:723
 Median : 9.000   Median : 1.000
 Mean   : 9.118   Mean   : 3.177
 3rd Qu.:13.000   3rd Qu.: 4.000
 Max.   :27.000   Max.   :73.000
```

A quick numerical summary indicate that patients are 35 ± 10 years old, with half the participants being less than 33 years old. There is about 80% of men among the participants, but a careful inspection of the data indicates that women are all 33 years old.

```
> summary(epilepsy$age[unique(epilepsy$id)])
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.00   33.00   35.00   37.82   38.00   51.00
```

```
> sd(epilepsy$age[unique(epilepsy$id)])
```

```
[1] 7.646392
```

```
> table(epilepsy$sex[unique(epilepsy$id)])
```

```
 1  2
73 16
```

```
> table(epilepsy$sex[unique(epilepsy$id)], epilepsy$age[unique(epilepsy$id)],
+     dnn = list("Sex", "Age"))

   Age
Sex 30 33 35 38 40 51
  1 19  0 16 16  2 20
  2  0 16  0  0  0  0

> table(epilepsy$race[unique(epilepsy$id)])

 1  2
34 55

> table(epilepsy$sex[unique(epilepsy$id)], epilepsy$race[unique(epilepsy$id)],
+     dnn = list("Sex", "Race"))

   Race
Sex  1  2
  1 34 39
  2  0 16
```

Again, we can take a look at the other univariate distributions.

```
> summary(epilepsy$height[unique(epilepsy$id)])

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  65.00   66.00   68.00   67.83   68.00   72.00

> sd(epilepsy$height[unique(epilepsy$id)])

[1] 2.154467

> summary(epilepsy$weight[unique(epilepsy$id)])

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  135.0   150.0   155.0   157.8   168.0   179.0

> sd(epilepsy$weight[unique(epilepsy$id)])

[1] 14.38967
```

To get an idea of the follow-up period, we can look at a few subjects:

```
> tab.trt <- table(epilepsy$trt, epilepsy$id)
> tab.trt[, 1:5]

    1204 1208 1213 1305 1307
  0   19   20   16   16   16
  1    0    0    0    0    0
```

and we can check that some of the patients were followed far longer than the 16 weeks as initially planned (e.g. patient whose `id` is 1208 was followed for 20 weeks, with placebo). Indeed, the range of the follow-up duration is

```
> range(tab.trt[tab.trt != 0])

[1]  2 27

> sum(tab.trt >= 16)

[1] 77
```

Figure 3.1: The Epilepsy data. Age and sex of the participants.

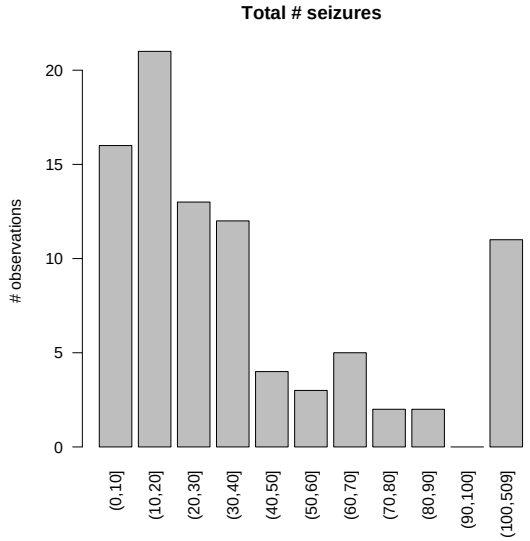

Figure 3.2: The Epilepsy data. Mean responses in each condition.

We can see that there are several dropouts ($n = 89 - 77 = 12$ patients), as can be expected for such longitudinal study.

Then, we shall summarize the total number of seizures per subject with an histogram, after having aggregated the data in a more convenient way than that used by default by R.

```
> epilepsy.nseiz <- with(epilepsy, tapply(nseizw, id, sum))
> epilepsy.nseiz.aggr <- cut(epilepsy.nseiz, breaks = c(seq(0,
+     100, by = 10), max(epilepsy.nseiz)))
> (res <- table(epilepsy.nseiz.aggr))

epilepsy.nseiz.aggr
   (0,10]   (10,20]   (20,30]   (30,40]   (40,50]   (50,60]   (60,70]   (70,80]
       16        21        13        12         4         3         5         2
  (80,90]  (90,100] (100,509]
        2         0        11

> barplot(res, las = 2, ylab = "# observations", main = "Total # seizures")
```

We will first fit a Poisson regression model to the response variable, the number of seizures `nseizw`. Following the convention of Molenberghs & Verbeke, our model takes the form:

$$\ln(\lambda_i/n_i) = \beta_0 + \beta_1 \text{Baseline}_i + \beta_2 T_i$$

where Y_i is assumed to follow a Poisson distribution $\mathcal{P}\lambda_i$, and n_i is the number of weeks subject i has been followed for. $\mathsf{Baseline}_i$ stands for the

baseline seizure rate for subject i and is coded as `bserate` in our `data.frame`. We can check that this model is strictly equivalent to one where $\ln(\lambda_i) = \ln(n_i) + \beta_0 + \beta_1 \text{Baseline}_i + \beta_2 T_i$, where $\ln(n_i)$ is most often called an 'offset' term.

We can fit the model as follow with R:

```
> epilepsy.glm <- glm(nseizw ~ bserate + trt, data = epilepsy,
+     offset = log(studyweek), family = poisson)
> summary(epilepsy.glm)

Call:
glm(formula = nseizw ~ bserate + trt, family = poisson, data = epilepsy,
    offset = log(studyweek))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.4043  -1.8558  -0.7565   0.9356  15.8580

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.3036182  0.0217910  -59.82   <2e-16 ***
bserate      0.0175103  0.0002459   71.20   <2e-16 ***
trt         -0.5755265  0.0340664  -16.89   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 11943.3  on 1418  degrees of freedom
Residual deviance:  8513.8  on 1416  degrees of freedom
AIC: 11330

Number of Fisher Scoring iterations: 6
```

Be careful with the offset: we need to specify the log component, and not only the variable name. We can see from the results of the fit that the treatment significantly reduces the average weekly number of epileptic seizures.

3.3 Marginal model

In this section, we will start our analyses by considering a Poisson model expressed as

$$\log(\lambda_{ij}) = \begin{cases} \beta_0 + \beta_1 t_{ij} & \text{if placebo,} \\ \beta_0 + \beta_2 t_{ij} & \text{if treated,} \end{cases}$$

with $Y_{ij} \sim \mathcal{P}(\lambda_{ij})$. We can see that this model assumes a common intercept for both treatment groups. This approach is similar to the one we used in Section 2.3, except we now use a log link.

```
> epilepsy.gee <- gee(nseizw ~ trt + studyweek, id = id, data = epilepsy,
+     family = poisson)
```

Parameter	Estimate	SE	z value
β_0	1.3581	0.1729	7.8549
β_1	0.0232	0.3150	0.0736
β_2	-0.0244	0.0128	-1.8994

Table 3.1: Estimates obtained from the GEE1 analysis. (z values are based on robust SE)

```
(Intercept)         trt   studyweek
 1.35811275  0.02317964 -0.02439452
```

A summary of the call to the `gee` function yields the estimates reported in Tab. 3.1.

Chapter 4

The fluvoxamine trial

4.1 The data

Accumulated experience with fluvoxamine, a serotonin reuptake inhibitor, in controlled clinical trials has show it to be as effective as conventional antidepressant drugs and more effective than placebo in the treatment of depression [Burton, 1991]. However, many patients who suffer from depression have concomitant morbidity with conditions such as obsessive-compulsive disorder, anxiety disorders and, to some extent, panic disorders. In most trials, patients with comorbidity are excluded, and therefore, it is of interest to gather evidence as to the importance of such factors, with a view on improved diagnosis and treatment. The general aim of this study was to determine the profile of fluvoxamine in ambulatory clinical psychiatric practice.

A total of 315 patients were enrolled with one or more of the following diagnoses: depression, obsessive, compulsive disorder, and panic disorder. Several covariates were recorded, such as gender and initial severity on a 5-point ordinal scale, where severity increases with category. After recruitment of the patient in the study, he or she was investigated at four visits (weeks 2, 4, 8, and 12). On the basis of about twenty psychiatric symptoms, the therapeutic effect and the side-effects were scored at each visit in an ordinal manner. Side-effect is coded as (1) = no; (2) = not interfering with functionality of patient; (3) = interfering significantly with functionality of patient; (4) = the side-effect surpasses the therapeutic effect. Similarly, the effect of therapy is recorded on a four-point ordinal scale: (1) = no improvement over baseline or worsening; (2) = minimal improvement (not changing functionality); (3) = moderate improvement (partial disappearance of symptoms); and (4) = important improvement (almost disappearance of symptoms). Thus a side effect occurs if new symptoms occur while there is therapeutic effect if old symptoms disappear.

There is also a lot a non negligeable amount of missing data but a much

larger fraction is fully observed than in the analgesic trial. Among the incomplete sequences, dropout is much more common than intermittent missingness, the latter type confined to two sequences only. It should also be noted that there are subjects, 14 in total without any follow-up measurements. This group of subjects is still an integral part of the trial, as they contain baseline information, includinf covariate information and baseline assessement of severity of the mental illness.

4.2 Summary of the data

Before running into detailed and purposeful association models, it is always a good idea to start with the basics, as already done with the preceding datasets.

We happen to setup the data as follow:

```
> fluvox <- read.csv("dataset/placpape.csv", header = TRUE, sep = ",")

> table(fluvox$SEX)

  1   2
112 203

> summary(fluvox$AGE)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  16.00   32.00   41.00   42.41   52.00   80.00    3.00

> sd(fluvox$AGE, na.rm = TRUE)

[1] 13.32038
```

As can be seen, we have twice as much women (coded as '2') as men and participants are 41 ± 13.3 years old on average. Next, we can summarize in a convenient table the diagnosed patients. We want to know how many of them have been classified as relevant to depressive, obsessive or panic category, as well as the number of patients who are showing two or more of these symptomatic behavior (see for instance the result of `table(fluvox`$DEPRES, fluvox$`OBSESSI)`).

```
> par(mfrow = c(2, 2))
> barplot(table(fluvox$SEX), names.arg = c("Men", "Women"), ylab = "Frequency",
+     las = 1)
> hist(fluvox$AGE, prob = TRUE, main = "", xlab = "Age")
> lines(density(fluvox$AGE, na.rm = T), col = "red")
```

4.3 Looking at usual Association models

Let's look first at the results of fitting the data with Goodman's conditional model and Dale's marginal model which were discussed in Section. 1.4.

First, we set up Table 6.6 of Molenberghs & Verbeke as follow:

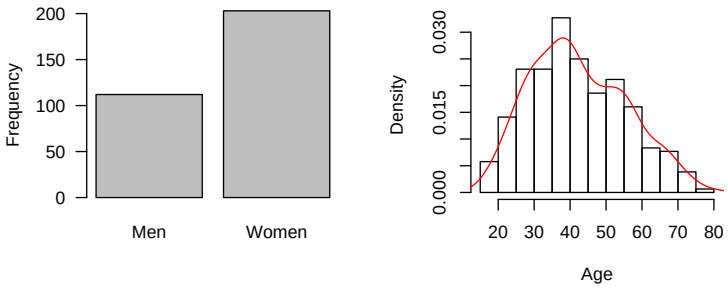

Figure 4.1: The fluvoxamine data. Participants' overview.

```
> fluvox.tab6.6 <- matrix(c(105, 34, 2, 3, 14, 80, 7, 1, 0, 7,
+     10, 0, 0, 1, 2, 2), nc = 4)
```

Recall that this table results from cross-classifying side effects at the second and third occasion.

The following function allows us to compute local and global cross-ratio coefficients.

```
> global.cr <- function(x, i, j) {
+     if (any(x > 1))
+         x <- x/sum(x)
+     pr <- apply(x, 1, sum)
+     pc <- apply(x, 2, sum)
+     res <- (x[i, j] * (1 - x[nrow(x), j] - x[i, ncol(x)] + x[i,
+         j]))/((x[i, ncol(x)] - x[i, j]) * (x[nrow(x), j] - x[i,
+         j]))
+     return(res)
+ }
> local.cr <- function(x, i, j) {
+     if (any(x > 1))
+         x <- x/sum(x)
+     stopifnot(i < nrow(x) - 1, j < ncol(x) - 1)
+     res <- (x[i, j] * x[i + 1, j + 1])/(x[i + 1, j] * x[i, j +
+         1])
+     return(res)
+ }
> global.cr.tab <- function(x) {
+ }
```

Chapter 5

Other datasets

5.1 Datasets

5.1.1 The analgesic trial

The data come from a single-arm clinical trial in 395 patients who are given analgesic treatment for pain caused by chronic nonmalignant disease. Treatment was to be administered for 12 months and assessed by means of a 'Global Satisfaction Assessment' (GSA) scale, rated on a five-point scale: (1) = very good; (2) = good; (3) = indifferent; (4) = bad; (5) = very bad.

Apart from the outcome of interest, number of covariates are available, such as age, sex, weight, duration of pain in years prior to the start of the study, type of pain, physical functioning, psychiatric condition, respiratory probelms, etc. GSA was rated by each person four times during the trial, at months 3, 6, 9 and 12. It should be noted that there are numerous missing values. Not only monotone or dropout occurs, there are also subjects with intermittent values.

5.1.2 The epilepsy data

5.1.3 The POPS study

The Project on Preterm and Small-for-gestational age infants (POPS) collected information on 1338 infants born in The Netherlands in 1983 and having gestational age less than 32 weeks and/or birthweight less than 1500g (Verloove et al., 1988). In total, 133 clinics were involved. The study population represents 94% of the births in that year with similar gestational age and birthweight characteristics. Prenatal, perinatal, and postnatal information as well as two year follow-up data were collected. Furthermore, the data base contains information on the delivery and specific details of the infant. After two years the child was reexamined.

Three ability scores were measured at the age of two, and risks factors were measured at delivery. All ability scores were recorded in a dichotomous

manner. They were available for 799 children. The first score (ABIL1) checks whether the child can pile three bricks, ABILS1 = 1 corresponds to 'no', whereas ABIL1 = 2 to 'yes'. The second score (ABIL2) measures whether the physical movements of the child are natural, ABIL2 = 1 (no) and ABIL2 = 2 (yes). Although ABIL2 is a purely physical ability score, ABIL1 is a combination of physical and mental qualities. The third ability score, ABIL3, expressess whether or not the child is able to put a ball in a box if he or she is asked to do so. The problem is to determine the risk factors for low performance at the three tests. Further it is of interest to compare the predicted probabilities taking into account the relationship between the responses to those calculated under the assumption of independent responses.

On the 1338 subjects, 818 (61.1%) have all three ability scores observed, and 471 (35.2%) have none of them. Only 49 (3.7%) have partial information. The latter is not unexpected, since two years lapsed between enrollment and the assessment of the ability scores.

References:

Verloove, S. P. and Verwey, R. Y. (1988). *Project on preterm and small-for-gestational age infants in the Netherlands*, 1983 (Thesis, University of Leiden). University Microfilms International, Ann Arbor, Michigan, USA, no. 8807276.

5.1.4 National toxicology program data

The developmental toxicity studies introduced in this section are conducted at the Research Triangle Institute, which is under contract to the National Toxicology Program of the United States (NTP data). These studies investigate the effects in mice of five chemicals: ethylene glycol (Price et al., 1985), diethylene glycol dimethyl ether (Price et al., 1987), and di(2-ethylhexyl)phthalate (Tyl et al., 1988).

a) Ethylene Glycol. Ethylene glycol (EG) is also called 1,2-ethanediol and can be represented by the chemical formula HOCH2CH2OH. It is a high-volume industrial chemical with many applications. EG is used as an antifreeze in cooling and heating systems, as one of the components of hydraulic brake fluids, as an ingredient of electrolytic condensers, and as a solvent in the paint and plastics industries. Furthermore, EG is employed in the formulation of several types of inks, as a softening agent for cellophane, and as a stabilizer for soybean foam used to extinguish oil and gasoline fires. Also, one uses EG in the synthesis of various chemical products, such as plasticizers, synthetic fibers, and waxes.

EG may represent little hazard to human health in normal industrial handling, except possibly when used as an aerosol or at elevated tempera-

tures. EG at ambient temperatures has a low vapor pressure and is not very irritating to the eyes or skin. However, accidental or intentional ingestion of antifreeze products, of which approximately 95% is EG, is toxic and may result in death.

Price et al. (1985) describe a study in which timed-pregnant CD-1 mice were dosed by gavage with EG in distilled water. Dosing occurred during the period of organogenesis and structural development of the foetuses (gestational days 8 through 15). The doses selected for the study were 0, 750, 1500, or 3000 mg/kg/day. Available data are: the number of dams containing at least one implant, the number of dams having at least one viable fetus, the number of live foetuses, the mean litter size, and the percentage of malformation for three different classes: external malformations, visceral malformations, and skeletal malformations. While for EG, skeletal malformations are substantial in the highest dose group, external and visceral malformations show only slight dose effects.

b) Di(2-ethylhexyl)Phthalate. Di(2-ethylhexyl)phthalate (DEHP) is also called octoil, dicotyl phthalate, or 1,2-benzenedicarboxylic acid bis(2-ethylhexyl) ester. It can be represented by C24H38O4. DEHP is used in vacuum pumps. Furthermore, this ester as well as other phthalic acid esters are used extensively as plasticizers for numerous plastic devices made of polyvinyl chloride. DEHP provides the finished plastic products with desirable flexibility and clarity.

It has been well documented that small quantities of phthalic acid esters may leak out of polyvinyl chloride plastic containers in the presence of food, milk, blood, or various solvents. Due to their ubiquitous distribution and presence in human and animal tissues, considerable concern has developed as to the possible toxic effects of the phthalic acid esters.

In particular, the developmental toxicity study described by Tyl et al. (1988) has attracted much interest in the toxicity of DEHP. The doses selected for the study were 0, 0.025, 0.05, 0.1, and 0.15%, corresponding to a DEHP consumption of 0, 44, 91, 191, and 292 mg/kg/day, respectively. Females were observed daily during treatment, but no maternal deaths or distinctive clinical signs were observed. The dams were sacrificed, slightly prior to normal delivery, and the status of uterine implantation sites recorded. A total of 1082 live foetuses were dissected from the uterus, anesthetized, and examined for external, visceral, and skeletal malformations.

There is a clear dose-related trends in the malformation rates. The average litter size (number of viable animals) decreases with increased levels of exposure to DEHP, a finding that is attributable to the dose-related increase in fetal deaths.

c) Diethylene Glycol Dimethyl Ether. Other names for diethylene glycol dimethyl ether (DYME) are diglyme and bis(2-methoxyethyl) ether. DYME has as its chemical formula: CH3O(CH2)2O(CH2)2OCH3. It is a component of industrial solvents. These are widely used in the manufacture of protective coatings such as lacquers, metal cotings, baking enamels, etc. Although to date, several attempts have proven inadequate to evaluate the potential of glycol ethers to produce human reproductive tocicity, structurally related compounds have been identified as reproductive toxicants in several mammalian species, producing (1) testicular toxicity and (2) embryotoxicity.

Price et al. (1987) describe a study in which timed-pregnant mice were dosed with DYME throughout major organogenesis (gestational days 8 through 15). The doses selected for the study were 0, 62.5, 125, 250 and 500 mg/kg/day.

References:

Price, C. J., Kimmel, C. A., Tyl, R. W. and Marr, M. C. (1985). The developmental toxicity of ethylene glycol in mice. *Toxicology and Applied Pharmacology*, *81*, 113–127.

Price, C. J., Kimmel, C. A., George, J. D. and Marr, M. C. (1987). The developmental toxicity of diethylene glycol dimethyl ether in mice. *Fundamental and Applied Toxicology*, *8*, 115–126.

Tyl, R. W., Price, C. J., Marr, M. C. and Kimmel, C. A. (1988). Developmental toxicity evaluation of dietary di(2-ethylhexyl)phthalate in Fischer 344 rats and CD-1 mice. *Fundamental and Applied Toxicology*, *10*, 395–412.

5.1.5 The sports injuries trial

These data come from a randomized, parallel group, double-blind trial in men comparing the effect of an active treatment to placebo on post-operative shivering and per-operative hemodynamics. The primary responses of interest were severity of post-operative shivering measured from the end of anesthesia every 5 minutes during 30 minutes as none (0), mild (1), moderate (2), or severe (3), and effect of treatment on overall consciousness assessed from the end of anesthesia at 10, 20, 30, 45, 60, 90 and 120 minutes as impossible to awake (0), difficult to awake (1), easy to awake (2), and awake, eyes open (3). One hundred forty patients were assigned to each treatment group.

Since this trial occurred in a very short time period, there is very little missing data. There was one patient who had no response information for either variable, so this patient is excluded from all analyses. There were also 3 patients with some missing information on shivering or overall consciousness, leaving 138 patients with complete information.

One interesting feature of these data is that there are structural zeros

in the awake variables. A patient could never become less awake over time, thus the cross-tabulation of the score over time contains zeros in the lower left corner.

5.1.6 Age related macular degeneration trial

These data arise from a randomized multi-centric clinical trial comparing an experimental treatment (interferon-alpha) to a corresponding placebo in the treatment of patients with age-related macular degeneration. Throughout the analyses done, we focus on the comparison between placebo and the highest dose (6 million units daily) of interferon-alpha (Z), but the full results of this trial have been reported elsewhere (Pharmacological Therapy for Macular Degeneration Study Group 1997). Patients with macular degeneration progessively lose vision. In the trial, the patients' visual acuity was assessed at different time points (4 weeks, 12 weeks, 24 weeks, and 52 weeks) through their ability to read lines of letters on standardized vision charts. These charts display lines of 5 letters of decreasing size, which the patient must read from top (largest letters) to bottom (smallest letters). Each line with at least 4 letters correctly read is called one 'line of vision'. The patient's visual acuity is the total number of letters correctly read. The primary endpoint of the trial was the loss of at least 3 lines of vision at 1 year, compared to their baseline performance (a binary endpoint). The secondary endpoint of the trial was the visual acuity at 1 year (treated as a continuous endpoint). Buyse and Molenberghs [1998] examined whether the patient's performance at 6 months could be used as a surrogate for their performance at 1 year with respect to the effect of interferon-alpha. They looked at whether the loss of 2 lines of vision at 6 months could be used as a surrogate for the loss of at least 3 lines of vision at 1 year. They also looked at whether visual acuity at 6 months could be used as a surrogate for visual acuity at 1 year.

Visual acuity can be measured in several ways. First, one can record the number of letters read. Alternatively, dichotomized versions (at most 3 lines of vision lost, or at least 3 lines of vision lost) can be used as well. Therefore, these data will be useful to illustrate methods for the joint modeling of continuous and binary outcomes, with or without taking the longitudinal nature into account. In addition, though there are 190 subjects with both month 6 and month 12 measurements available, the total number of longitudinal profiles is 240, but only for 188 of these have the four follow-up measurements been made.

Notes

[a]Recall from Linear Models that a response variable Y could be described as a linear combination of predictors, and such a model can be written as $Y = X\beta + \varepsilon$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ (i.i.d.). We usually want to estimate β and it can be shown that $\beta^* \sim \mathcal{N}(\beta, \sigma^2\mathcal{I}^{-1}(\beta))$, where β^* is the M.L.E. of β and $\mathcal{I}$ is the Fisher information.

First of all, let's consider the score function $V(\theta)$. It is the partial derivative, with respect to the parameter of interest, say θ, of the log likelihood, and can be found using the chain rule:

$$V(\theta) = \frac{\partial}{\partial\theta} \log \mathcal{L}(\theta; X) = \frac{1}{\mathcal{L}(\theta; X)} \frac{\partial \mathcal{L}(\theta; X)}{\partial\theta}.$$

One can show that

$$E(V \mid \theta) = \int_{[0,1]} \frac{f'_\theta(x;\theta)}{f(x;\theta)}\, dF(x;\theta) = \int_X \frac{f'_\theta(x;\theta)}{f(x;\theta)} f(x;\theta)\, dx = \int_X \frac{\partial f(x;\theta)}{\partial\theta}\, dx,$$

and that $E(V \mid \theta) = 0$ if some differentiability conditions are met. Thus, the expected value of the score is zero. In other words, if one were to repeatedly sample from some distribution, and repeatedly calculate the score with the true θ, then the mean value of the scores would tend to zero as the number of repeat samples approached infinity. The variance of the score simply is the Fisher information, $\mathcal{I}(\theta)$, also written as

$$\mathcal{I}(\theta) = E\left\{ \left[\frac{\partial}{\partial\theta} \log \mathcal{L}(\theta; X) \right]^2 \mid \theta \right\}.$$

Note that the Fisher information, as defined above, is not a function of any particular observation, as the random variable X has been averaged out.

If we define our starting point, θ_0, for computing the score function, we can develop $V(\theta)$, about θ_0, through a Taylor expansion such that:

$$V(\theta) \approx V(\theta_0) - \mathcal{J}(\theta_0)(\theta - \theta_0)$$

where

$$\mathcal{J}(\theta) = -\sum_{i=1}^{n} \nabla\nabla^T|_{\theta=\theta_0} \log f(Y_i; \theta)$$

is the observed information matrix at θ_0. Now, setting $\theta = \theta^*$, using that $V(\theta^*) = 0$ and rearranging gives us:

$$\theta^* = \theta_0 + \mathcal{J}^{-1}(\theta_0)V(\theta_0).$$

By recurence, we can therefore use

$$\theta_{m+1} = \theta_m + \mathcal{J}^{-1}(\theta_m)V(\theta_m).$$

[b]Here is a short summary of the construction and properties of Wald statistics. Let $Y_1, \ldots, Y_N$ be i.i.d. Bernoulli variables whose probability functions are defined such that

$$\ln\left(\frac{\Pr_i}{1 - \Pr_i}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} = f_\beta(X_i),$$

where $\Pr_i = \Pr(Y_i = 1) = [1 + \exp(-f_\beta(X_i))]^{-1}$ and $X_{i1}, \ldots, X_{ik}$ are observations taken from k independent variables.

For inference purpose, we can select β_k as the parameter of interest, without any loss of generality. Our hypothesis could be, for instance,

$$H_0 : \beta_k = \beta_{k0} \text{ vs. } H_a : \beta_k \neq \beta_{k0}.$$

If we consider both the MLE of β_k, $\hat{\beta}_k$, and H the inverse of the empirical information matrix, Wald's statistic is defined by

$$\xi^W = \frac{(\hat{\beta}_k - \beta_{k0})^2}{H_{kk}},$$

where H_{kk} is the estimated variance of $\hat{\beta}_k$. Under H_0, ξ^W asymptotically follows a $\chi^2(1)$ distribution (like the LR statistic, $-2\ln\lambda$).

Bibliography

M De Backer, P De Keyser, C De Vroey, and E Lesaffre. A 12-week treatment for dermatophyte toe onychomycosis: terbinafine 250mg/day vs. itraconazole 200mg/day—a double-blind comparative trial. *British Journal of Dermatology*, 134:16–17, 1996.

R R Bahadur. *Studies in Item Analysis and Prediction*, chapter A representation of the joint distribution of responses to n dichotomous items. Stanford, CA: Stanford University Press, 1961.

S W Burton. A review of fluvoxamine and its uses in depression. *International Clinical Psychopharmacology*, 6(Suppl. 3):1–17, 1991.

M Buyse and G Molenberghs. The validation of surrogate end-points in randomized experiments. *Biometrics*, 54:1014–1029, 1998.

P Mc Cullagh and J A Nelder. *Generalized Linear Models*. Chapman & Hall, 1989.

J R Dale. Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, 42:721–727, 1986.

P J Diggle, P J Heagerty, K-Y Liang, and S L Zeger. *Analysis of Longitudinal Data*. Oxford Science Publications. Oxford: Clarendon Press, 2nd edition, 2002.

J F Faraway. *Extending the Linear Model with* R. Chapman & Hall, 2006.

E Faught, B J Wilder, R E Ramsey, R A Reife, L D Kramer, G W Pledger, and R M Karim. Topiramate placebo-controlled dose-ranging trial in refractory partial epilepsy using 200-, 400-, and 600-mg daily dosages. *Neurology*, 46:1684–1690, 1996.

A Gelman, J B Carlin, H S Stern, and D B Rubin. *Bayesian Data Analysis*. Texts in Statistical Science. London: Chapman & Hall, 1995.

L A Goodman. Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74:537–552, 1979.

L A Goodman. Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *Journal of the American Statistical Association*, 76:320–334, 1981.

D A Harville. Bayesian inference for variance components using only error contrasts. *Biometrika*, 61:383–385, 1974.

W W Hauck and A Dooner. Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, 72(360):851–853, 1977.

N M Laird and J H Ware. Random effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

G Molenberghs and E Lesaffre. Marginal modelling of correlated ordinal data using a multivariate plackett distribution. *Journal of the American Statistical Association*, 89:633–644, 1994.

F E Satterthwaite. Synthesis of variance. *Psychometrika*, 6:309–316, 1941.

W N Venables and B D Ripley. *Modern Applied Statistics with* S. Springer, 2002.

Appendices

The SAS programming language

The R formula interface for `glm` and the like

List of Figures

List of Tables

Index