Chapter 12
Generalised Estimation Equations

In this chapter, we analyse three data sets; California birds, owls, and deer. In the first data set, the response variable is the number of birds measured repeatedly over time at two-weekly intervals at the same locations. In the owl data set (Chapter 5), the response variable is the number of calls made by all offspring in the absence of the parent. We have multiple observations from the same nest, and 27 nests were sampled. In the deer data, the response variable is the presence or absence of parasites in a deer; the data are from multiple farms.

In the first instance, we apply a generalised linear model (GLM) with a Poisson distribution for the California birds and owl data and a binomial GLM for the deer data. However, such analyses violate the independence assumption; for the California bird data, there is a longitudinal aspect, we have multiple observations per nest for the owls, and multiple deer from the same farm. We therefore introduce generalised estimation equations (GEE) as a tool to include a dependence structure, discuss its underlying mathematics, and apply it on the same data sets.

GEE was introduced by Liang and Zeger (1986), and since their publication, several approaches have been developed to improve the technique. We use the original method as it is the simplest. Useful GEE references are Ziegler et al. (1996), Greene (1997), Fitzmaurice et al. (2004), and a textbook completely dedicated to GEE by Hardin and Hilbe (2002). This chapter heavily depends on the Fitzmaurice et al. (2004) book. Chapter 22 contains a binary GEE case study.

12.1 GLM: Ignoring the Dependence Structure

12.1.1 The California Bird Data

Elphick and Oring (1998, 2003) and Elphick et al. (2007) analysed time series of several water bird species recorded in California rice fields. Their main goals were to determine whether flooding fields after harvesting results in greater use by aquatic birds, whether different methods of manipulating the straw in conjunction with flooding influences how many fields are used, and whether the depth that the

A.F. Zuur et al., *Mixed Effects Models and Extensions in Ecology with R*,
Statistics for Biology and Health, DOI 10.1007/978-0-387-87458-6_12,

fields are flooded to is important. Biological details can be found in the references mentioned above.

Counts were made during winter surveys at several fields. Here, we only use data measured from one winter (1993–1994), and we use species richness to summarise the 49 bird species recorded. The sampling took place at multiple sites, and from each site, multiple fields were repeatedly sampled. Here, we only use one site (called 4mile) for illustrative purposes. There are 11 fields in this site, and each field was repeatedly sampled; see Fig. 12.1. Note that there is a general decline in bird numbers over time. One of the available covariates is water depth per field, but water depth and time are collinear (as can be inferred from making an `xyplot` of depth versus time for each field), so we avoid using them together as covariates in the models.

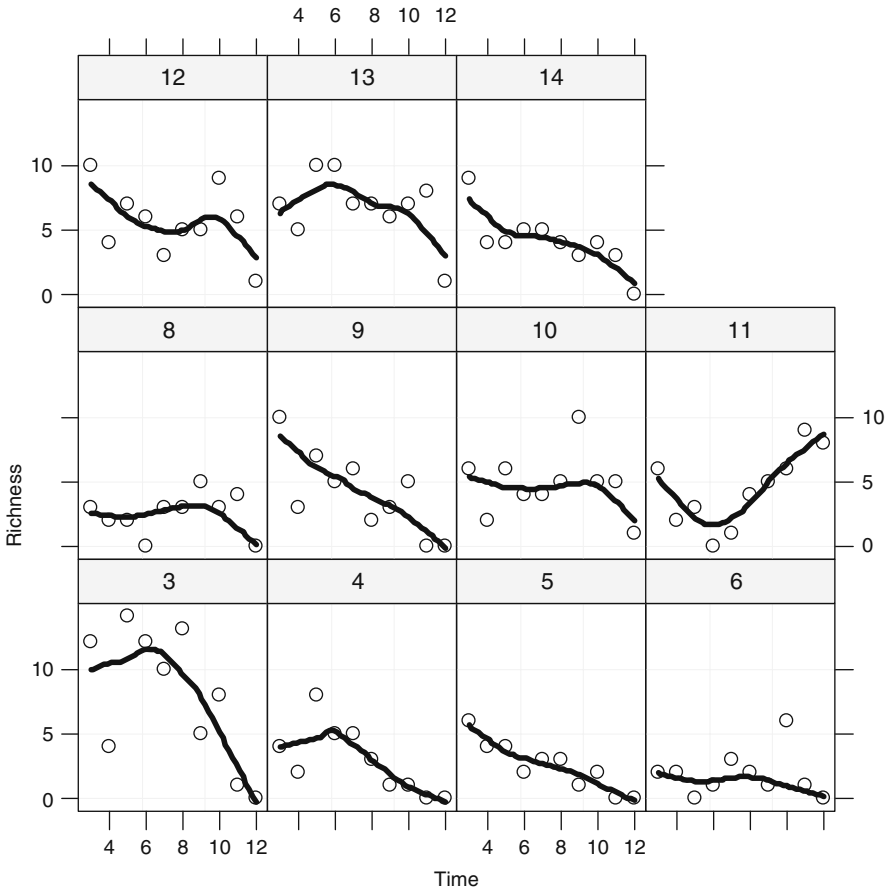

Fig. 12.1 `xyplot` of species richness plotted against time (expressed in two-weekly periods). Each panel represents a different field. A LOESS smoother was added to aid visual interpretation of the graph

The following R code reads the data, calculates the richness index, and makes the `xyplot` in Fig. 12.1.

```
> library(AED); data(RiceFieldBirds)
> RFBirds <- RiceFieldBirds           #Saves some space
> RFBirds$Richness <- rowSums(RFBirds[, 8:56] > 0)
> RBirds$fField <-factor(RFBirds$FIELD)
> library(lattice)
> xyplot(Richness ~ Time | fField, data= RFBirds,
   panel=function(x, y){
     panel.grid(h = -1, v = 2)
     panel.points(x, y, col = 1)
     panel.loess(x, y, col = 1, lwd = 2)})
```

The first few lines access the data and the object with the data is renamed into a shorter name. The `rowSums` command is used to calculate species richness (add `na.rm = TRUE` if you have missing values in your data), and the rest is a matter of some simple `xyplot` commands and options to get points and smoothers in the panels (see also Chapter 2). As always in R, things can be done in at least five different ways. Instead of the code in the panel function, you can also use:

```
> xyplot(Richness ~ Time | fField, data=  RFBirds,
       type = c ("p" , "smooth" , "grid"))
```

It gives the same graph, but the code looks a bit more cryptic. Additional parameters like the span width and line thickness for the smoother can also be specified (just add `span = 0.5, lwd = 2, col = 1` to the command above).

Counts took place approximately every two weeks. As well as species richness, we also have water depth and information on rice debris management. The aim of the analysis presented here is to explain the richness values as a function of depth and management effects. The response variable is a count, and therefore we are in the world of GLMs with a likely candidate model the GLM with a Poisson distribution and log-link function. Actually, it is a bit more complicated as the original data were densities; numbers per field and the sizes of the fields are different. This means that (the log of the) size of the field can be used as an offset variable (Chapter 9). Based on biological knowledge, and an initial analysis using generalised additive modelling (Elphick et al., 2007), the effect of the covariate depth is modelled as a quadratic term. The following three steps define the GLM.

1. Define Y_{is} as the richness measured in field i at time s. We assume that Y_{is} is Poisson distributed with mean μ_{is}. In mathematical notation, we have: $Y_{is} \sim \text{P}(\mu_{is})$. Recall that for a Poisson distribution, the mean is the variance.

2. The systematic part of the GLM is given by

$$\eta(\text{Depth}_{is}, \text{SPTREAT}_{is}, \text{AREA}_{is}) = \alpha + \text{offset}(\log(\text{AREA}_{is})) + \beta_1 \times \text{Depth}_{is} + \beta_2 \times \text{Depth}_{is}^2 + \beta_3 \times \text{SPTREAT}_{is}$$

   The term SPTREAT is the categorical variable defining management type. It is also possible to include an interaction between depth and the management type and also between the quadratic function of depth and management type. But to keep the models simple, we do not do this.
3. The link between the expected values and systematic component is the log-link:

$$\log(\mu_{is}) = \eta(\text{Depth}_{is}, \text{SPTREAT}_{is}, \text{AREA}_{is})$$

Full details of Poisson GLMs are given in Chapter 9. It is important to realise that the GLM assumes independence of all richness values, including those from the same field (which are separated by only two weeks). For the moment, we will ignore this problem and just carry on with the GLM. Later in this chapter, we will apply GEE to incorporate auto-correlation on the data from the same field and compare results. An initial GLM indicated overdispersion, and we therefore applied a quasi-Poisson GLM with the following R code. Results from the `summary` command are given as well and we will compare them later with the GEE results.

```
> RFBirds$LA <- log(RFBirds$AREA)
> RFBirds$fSptreat <- factor(RFBirds$SPTREAT)
> RFBirds$DEPTH2 <- RFBirds$DEPTH^2

> M0 <- glm(Richness ~ offset(LA) +  fSptreat +  DEPTH +
            DEPTH2, family = quasipoisson, data = RFBirds)
> summary(M0)

Coefficients:

                 Estimate Std. Error t value  Pr(>|t|)
(Intercept)    -0.7911754  0.2136575  -3.703   0.00034
fSptreatrlfld  -0.4931558  0.1666480  -2.959   0.00380
DEPTH           0.0690528  0.0249844   2.764   0.00674
DEPTH2         -0.0016531  0.0006732  -2.455   0.01569

Dispersion parameter for quasipoisson family taken to
be 2.392596
    Null deviance: 297.47  on 109  degrees of freedom
Residual deviance: 245.10  on 106  degrees of freedom
AIC: NA
```

Note that the overdispersion is 2.39. All terms in the model are significant at the 5% level, although the quadratic depth term is only weakly significant with a p-value of 0.015.

12.1.2 The Owl Data

In Chapters 5 and 6, we analysed data from a study on vocal begging behaviour when the owl parents bring prey to their nest. In both chapters, we used sibling negotiation as response variable. It was defined as the number of calls made by all offspring in the absence of the parents counted during 30-second time intervals before arrival of a parent divided by the number of nestlings. Just as in the previous section, we can use the (natural) logarithm of the number of nestlings as an offset variable and analyse the number of calls NCalls_{is} at time s in nest i using a Poisson GLM. Hence, we assume that $\text{NCalls}_{is} \sim \text{P}(\mu_{is})$, and therefore the mean and variance of NCalls_{is} are equal to μ_{is}. The systematic part is given by

$$\begin{aligned}\eta_{is} = \alpha &+ \log(\text{Broodsize}_i) + \beta_1 \times \text{SexParent}_{is} + \beta_2 \times \text{FoodTreatment}_{ij} \\ &+ \beta_3 \times \text{ArrivalTime}_{ij} + \beta_4 \times \text{SexParent}_{is} \times \text{FoodTreatment}_{ij} \\ &+ \beta_5 \times \text{SexParent}_{is} \times \text{ArrivalTime}_{ij}\end{aligned}$$

Recall from Chapter 5 that the sex of the parent is male or female, food treatment at a nest is deprived or satiated, and arrival time of the parent at the nest was coded with values from 21 (9.00 PM) to 30 (6.00 AM). Note that there is no regression parameter in front of the $\log(\text{Broodsize}_i)$ term; it is modelled as an offset variable. The link between the expected value of Y_{is}, μ_{is}, and the systematic component η_{is} is the log-link:

$$\log(\mu_{is}) = \eta_{is} \quad \Leftrightarrow \quad \mu_{is} = e^{\eta_{is}}$$

The model is fitted with the following R code.

```
> library(AED)  ; data(Owls)
> Owls$NCalls <- Owls$SiblingNegotiation
> Owls$LBroodSize <- log(Owls$BroodSize)
> Form <- formula(NCalls ~ offset(LBroodSize) +
                  SexParent *  FoodTreatment +
                  SexParent *  ArrivalTime)
> O1 <- glm(Form, family = poisson, data = Owls)
```

Instead of the name `SiblingNegotiation`, we used the shorter name `NCalls` as it saves some space in the code. The results of the `summary(O1)` command are not shown here, but there is overdispersion. Therefore, we refitted the model with a quasi-Poisson GLM:

```
> O2 <- glm(Form, family = quasipoisson, data = Owls)
> drop1(O2, test = "F")
```

Results of the `drop1` command are not presented here, but indicate that the two two-way interactions are not significant. Using a backwards selection, we ended up with the model containing food treatment and arrival time:

```
> Form <- formula(NCalls ~ offset(LBroodSize) +
                  FoodTreatment + ArrivalTime)
> O3 <- glm(Form, family = quasipoisson, data = Owls)
> summary(O3)

                        Estimate Std. Error t value  Pr(>|t|)
(Intercept)              3.81333    0.53946   7.069  4.39e-12
FoodTreatmentSatiated -0.53230    0.08260  -6.444  2.40e-10
ArrivalTime           -0.12924    0.02205  -5.861  7.60e-09

Dispersion parameter for quasipoisson family taken to be 6.246006
    Null deviance: 4128.3  on 598  degrees of freedom
Residual deviance: 3652.6  on 596  degrees of freedom
AIC: NA
```

All regression parameters are highly significant. We will return to these results once the GEE has been discussed.

12.1.3 The Deer Data

Vicente et al. (2006) looked at the distribution and faecal shedding patterns of the first-stage larvae (L1) of *Elaphostrongylus cervi* (*Nematoda: Protostrongylidae*) in red deer across Spain. Effects of environmental variables on *E. cervi* L1 counts were determined using generalised linear mixed modelling (GLMM) techniques. Full details on these data can be found in their paper. In this book, we use only part of their data to illustrate GEE and GLMM (Chapter 13).

In this section, we keep the analysis simple and focus on the relationship between the presence and absence of *E. cervi* L1 in deer and the explanatory variables length and sex of the host. Because the response variable is of the form 0–1, we are immediately in the world of a binomial GLM. The explanatory variables are length and sex of the deer, the first is continuous and sex is nominal. The following three steps define the GLM.

1. Define Y_{is} as 1 if the parasite *E. cervi* L1 is found in animal j at farm i, and 0 otherwise. We assume that Y_{is} is binomially distributed with probability p_{is}. In mathematical notation, we have: $Y_{is} \sim \text{B}(1, p_{is})$. Recall that for a binomial distribution, we have $\text{E}(Y_{is}) = p_{is}$ and $\text{var}(Y_{is}) = p_{is} \times (1 - p_{is})$.
2. The systematic part of the GLM is given by:

$$\eta(\text{Length}_{is}, \text{Sex}_{is}) = \alpha + \beta_1 \times \text{Length}_{is} + \beta_2 \times \text{Sex}_{is} + \beta_3 \times \text{Length}_{is} \times \text{Sex}_{is}$$

3. The link between the expected values and systematic component is the logistic link:

$$\text{logit}(p_{is}) = \eta(\text{Length}_{is}, \text{Sex}_{is}) \Leftrightarrow$$
$$p_{is} = \frac{e^{\alpha+\beta_1\times\text{Length}_{is}+\beta_2\times\text{Sex}_{is}+\beta_3\times\text{Length}_{is}\times\text{Sex}_{is}}}{1+e^{\alpha+\beta_1\times\text{Length}_{is}+\beta_2\times\text{Sex}_{is}+\beta_3\times\text{Length}_{is}\times\text{Sex}_{is}}}$$

The notation logit stands for the logistic link (Chapter 10), and p_{ij} is the probability that animal j on farm i has the parasite, Length_{ij} is the length of the deer, and Sex_{ij} tells us whether it is male or female. Instead of the subscripts i and j, we could have used one index k identifying the animal. However, with respect to the methods that are to come, it is more useful to use indices i and j.

The following code accesses the data from our AED package, defines Sex as a nominal variable, and converts the *E. cervi* count data into presence and absence.[1]

```
> library(AED); data(DeerEcervi)
> DeerEcervi$Ecervi.01 <- DeerEcervi$Ecervi
> DeerEcervi$Ecervi.01[DeerEcervi$Ecervi >0 ] <- 1
> DeerEcervi$fSex <- factor(DeerEcervi$Sex)
> DeerEcervi$CLength <- DeerEcervi$Length -
                            mean(DeerEcervi$Length)
```

Note that we centred length. If you do not centre the length, the intercept represents the probability that a deer of length 0 has the parasite. This of course is nonsense as there cannot be any deer of length 0. By centring length, the intercept has the more meaningful interpretation of the probability that an animal of average length has the parasite. The code below applies a GLM on the selected data, drops each allowable term in turn, from the model, and applies a likelihood ratio test that is Chi-square distributed. Note that because the interaction between length and sex is included, we cannot drop the main terms `Length` and `Sex`.

```
< DE.glm<-glm(Ecervi.01 ~ Length * fSex,
              data = DeerEcervi, family = binomial)
> drop1 (DE.glm, test = "Chi")

Single term deletions. Model: Ecervi.01 ~ CLength*fSex
             Df Deviance    AIC    LRT Pr(Chi)
<none>            1003.7 1011.7
CLength:fSex  1   1008.1 1014.1    4.4   0.036

> summary(DE.glm)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.652409   0.109602   5.953 2.64e-09
CLength      0.025112   0.005576   4.504 6.68e-06
```

[1]The motivation for this is purely pedagogical; we want to present three GEE examples, one of which is a binomial GEE.

```
fSex2          0.163873   0.174235   0.941   0.3469
CLength:fSex2 0.020109   0.009722   2.068     0.0386

Dispersion parameter for binomial family taken to be 1
    Null deviance: 1073.1  on 825  degrees of freedom
Residual deviance: 1003.7  on 822  degrees of freedom
AIC: 1011.7
```

The output from a `drop1` function was discussed in Chapter 10. Recall that it compares the deviance of the specified model with that of nested models. The difference between these two deviances is Chi-square distributed. The Length–Sex interaction term is significant at the 5% level. We will return to the numerical output once the GEE has been discussed.

The problem with this analysis is that the data were obtained from 24 farms. This means that we are sampling deer that may have been in contact with each other, and we can therefore not assume that the presence or absence of parasites on deer from the same farm are independent.

12.2 Specifying the GEE

12.2.1 Introduction

The GLMs presented in the previous section are potentially flawed because the data are longitudinal (California birds) or we have repeated measurements from the same nest (owls) or farm (deer). Hence, the assumption of independence is invalid. We could just ignore the potential existence of dependence and present the analyses of the data obtained by GLM, but this will tend to increase the risk of a Type I error, particularly where within-subject (auto-) correlation is strong.

In Chapters 8, 9, and 10, we have seen how GLM gives us a framework for analysing response data whose inherent stochasticity can be modelled with any one of a number of probability distributions belonging to the exponential family, i.e. can be expressed in the form

$$\exp\left\{\frac{Y \times \theta - b(\theta)}{a(\phi)} - c(Y, \phi)\right\}$$

where Y is the response variable. Additionally, in Chapters 5, 6, and 7, we looked at ways to model within-subject correlation, and incorporate it into the analysis through, for example, mixed modelling. Liang and Zeger (1986) set out to establish an algorithm that combined these two methodologies.

The modelling of correlation structures is relatively easily managed with normally distributed response data. Although the mathematics appear involved to the non-technical reader, the mechanics of optimisation are computationally trivial, particularly with powerful modern computers. However, the complications

multiply when we are modelling auto-correlation where the data are clearly non-normal and cannot be transformed.

Although this normally means binary (presence/absence), proportional, and count data, the same general arguments can be applied to any response that can be modelled using GLMs, e.g. overdispersed count data using a negative binomial type variance structure. We specifically write 'negative binomial type' as we are not going to make any distributional assumptions in the GEE.

Although not yet discussed, we can use generalised linear mixed modelling (Chapter 13) to account for within-subject 'compound-symmetry' type correlation. This is the simplest mixed model structure where all we saying is that all observations from a given source (subject) are correlated. No allowance can be made within this procedure for correlation patterns between the observations from the same source, e.g. temporal auto-correlation. One important philosophical and methodological difference from (generalised linear) mixed modelling is that GEEs do not estimate the distributional properties of the subjects themselves. In a mixed model setting, if there are a sufficient number of subjects (or fields, nest, trawls, etc.), we can estimate the variance of the distribution that their effects are drawn from. This is usually taken to be a Normal distribution.

We now specify a GEE, and broadly follow Chapter 11 in Fitzmaurice et al. (2004). GEE models are also called 'marginal' models, but this is slightly confusing. In previous chapters, we used the word 'marginal' in the context of conditional models (i.e. conditional on a random effect). Here, it means that the model for the mean response only depends on covariates and not on random effects.

12.2.2 Step 1 of the GEE: Systematic Component and Link Function

Suppose we have a response variable Y_{is} and one explanatory variable X_{is}. Y_{is} can be the number of birds in field i at time s, the number of sibling calls in nest i at time s, or the presence or absence of the parasite *E. cervi* in deer j sampled at farm i. The systematic part in all these models is given by

$$\eta = \alpha + \beta_1 \times X_{is}$$

It is also possible to have more explanatory variables. The relationship between the conditional mean and the systematic component has the same structure as in GLM models. Hence, for count data we use

$$E(Y_{is}) = e^{\eta} = e^{\alpha+\beta_1 \times X_{is}}$$

and for the $0-1$ data

$$E(Y_{is}) = \frac{e^{\alpha+\beta_1 \times X_{is}}}{1 + e^{\alpha+\beta_1 \times X_{is}}}$$

The word conditional is with respect to the explanatory variables (we pretend we know what they are). We should, therefore, write the first part of the equation as $E(Y_{is}|X_{is})$. This reads as follows: The expected value of Y_{is} for given X_{is}. A more general notation for the relationship between the conditional mean and the explanatory variables for the count data is

$$E(Y_{is}|X_{is}) = \mu_{is} \text{ and } g(\mu_{is}) = \alpha + \beta_1 \times X_{is}$$

12.2.3 Step 2 of the GEE: The Variance

For count data, the easiest conditional variance structure of Y_{is} is given by

$$\text{var}(Y_{is}|X_{is}) = \mu_{is}$$

Obviously, we can also opt for a negative binomial type variance structure (Chapter 11), but for the moment we will keep it simple. The notation for a more general model is: $\text{var}(Y_{is} \mid X_{is}) = \phi \times v(\mu_{is})$, where $v()$ is the variance function, and ϕ is the scale parameter (overdispersion), which we need to estimate or simply set to 1. Choosing $\phi = 1$ and $v(\mu_{is}) = \mu_{is}$ gives a identical variance structure to the one used in Poisson GLM. For the absence–presence deer data, we can choose a binomial variance structure (Chapter 10).

You may wonder why we do not just assume that count data Y_{is} is Poisson distributed with mean μ_{is}, or for 0–1 data a binomial distribution? After all, these give the same mean and variance relationships as specified above. The reason is because we can do the GEE without having to specify a distribution. Furthermore, in the next step, we have to specify a correlation structure between the observations. Assuming that Y_{is} is Poisson or binomial distributed makes this step awkward, we will explain below why.

Basically, all we have done so far is follow the quasi-GLM route by specifying the relationship between the mean and the explanatory variables and the variance structure. The next step specifies the association between the observations. Note we carefully wrote 'association' and not correlation.

12.2.4 Step 3 of the GEE: The Association Structure

Now we have to specify an association structure between Y_{is} and Y_{it}, where s and t are two different sampling days on the same field i, two observations from the same nest, or two deer from the same farm. There are many ways of doing this, and the

type of data (e.g. continuous, binary or counts) also affects how the association is defined.

Option 1: The Unstructured Correlation

For continuous data, the obvious tool to define association between the two observations Y_{is} and Y_{ik} is the Pearson correlation coefficient. Just as in Chapter 6, we have various options to parameterise the correlation. The most flexible choice is the so-called unstructured correlation, which is given by

$$\text{cor}(Y_{is}, Y_{it}) = \alpha_{st}$$

This correlation structure can be easily understood if we imagine temporally sequential observations coming from the same source, for example, a blood pressure reading taken from the same patient/animal at regular (e.g. hourly) intervals. The correlation matrix can be expressed thus:

$$\begin{pmatrix}
1 & \alpha_{12} & \alpha_{13} & \alpha_{14} & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
\alpha_{12} & 1 & \alpha_{23} & \alpha_{24} & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
\alpha_{13} & \alpha_{23} & 1 & \alpha_{34} & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
\alpha_{14} & \alpha_{24} & \alpha_{34} & 1 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & 0 & 0 & 0 & 1 & \alpha_{12} & \alpha_{13} & \alpha_{14} & & & & & 0 \\
0 & 0 & 0 & 0 & \alpha_{12} & 1 & \alpha_{23} & \alpha_{24} & & & & & 0 \\
0 & 0 & 0 & 0 & \alpha_{13} & \alpha_{23} & 1 & \alpha_{34} & & & & & 0 \\
0 & 0 & 0 & 0 & \alpha_{14} & \alpha_{24} & \alpha_{34} & 1 & & & & & 0 \\
\vdots & \vdots & \vdots & \vdots & & & & & \ddots & & & & \vdots \\
\vdots & \vdots & \vdots & \vdots & & & & & & 1 & \alpha_{12} & \alpha_{13} & \alpha_{14} \\
\vdots & \vdots & \vdots & \vdots & & & & & & \alpha_{12} & 1 & \alpha_{23} & \alpha_{24} \\
\vdots & \vdots & \vdots & \vdots & & & & & & \alpha_{13} & \alpha_{23} & 1 & \alpha_{34} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \alpha_{14} & \alpha_{24} & \alpha_{34} & 1
\end{pmatrix}$$

The upper left 4×4 block is the correlation matrix for the first patient, the second block for the second patient, etc. In each block, α_{st} is the correlation between observations s and t. We use the different blocks to estimate these parameters. No correlation between the parameters α_{12}, α_{13}, α_{14}, α_{23}, α_{24}, and α_{342} is assumed, and they are estimated completely independently. Note that this is based on 4 observations per subject. The number of independent parameters to be estimated rapidly increases as the number of within-subject observations increases with all the attendant problems related to matrix inversion.

This is the most general correlation model and perhaps the least intuitively appealing. Essentially, all correlations between within-subject observations are

estimated independently; thus a lot more parameters need to be estimated. Because of this complexity, the GEE algorithm can break down because the correlation matrix cannot be inverted. However, it can be a useful approach if no obvious correlation structure suggests itself and can be a useful exploratory step to help arrive at a final choice of correlation structure.

Let us discuss the applicability of the unstructured correlation for the response variables in the California bird data set, owl data set, and deer data set. For the moment, we ignore that these response variables are not continuous. For the California bird data, each block of correlation in the matrix above is for a field; for the owl data each block is a nest; and for the deer data, each block is a farm. For the deer data, it does not make sense to use the unstructured correlation because there is no relationship between animals 1 and 2 at farm 1, and animals 1 and 2 at farm 2. For the California bird data, it may be an option to use this correlation structure as observations 1 and 2 in field 1, and observations 1 and 2 in field 2 both tells us something about the temporal relationship at the start of the experiment. On the down side, 10 temporal observations per field mean that we have to estimate $10 \times 9/2 = 45$ correlation parameters, which is a lot! The unstructured correlation may be an option for these data if you have hundreds of fields, but not with only 12 fields. For the owl data, it is a bit more complicated. If we just analyse the number of calls sampled at the nests without a time order, then the set up of the data is similar to that of the deer data. Hence, in this case, we cannot use the unstructured correlation. But we also know the arrival time of the parents at the nest, which unfortunately, is irregularly spaced. However, in Chapter 6, we argued that based on biology, we could assume that owl parents chose the arrival time, and therefore, from their point of view, the data are regularly spaced. Hence, if we use the unstructured correlation, then α_{12} represents the correlation between arrivals 1 and 2, α_{13} the correlation between arrivals 1 and 3, etc. This would make sense, but unfortunately, this approach requires an enormous amount of correlation parameters as some nests contain more than 50 observations. Hence, it is not practical.

Option 2: AR-1 Correlation

Another option for continuous data is to say that the correlation between two observations from the same patient, field, nest, or farm i is

$$\text{cor}(Y_{is}, Y_{it}) = \alpha^{|s-t|}$$

This type of auto-regressive correlation structure was also used in Chapter 6 (using the `corAR1` function). Autoregressive correlation is observed when correlation between within-subject observations can be modelled directly as a function of the 'distance' between the observations in question. Using the same example as above, the following correlation matrix is used.

$$\begin{pmatrix}
1 & \alpha & \alpha^2 & \alpha^3 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
\alpha & 1 & \alpha & \alpha^2 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
\alpha^2 & \alpha & 1 & \alpha & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
\alpha^3 & \alpha^2 & \alpha & 1 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & 0 & 0 & 0 & 1 & \alpha & \alpha^2 & \alpha^3 & & & & & 0 \\
0 & 0 & 0 & 0 & \alpha & 1 & \alpha & \alpha^2 & & & & & 0 \\
0 & 0 & 0 & 0 & \alpha^2 & \alpha & 1 & \alpha & & & & & 0 \\
0 & 0 & 0 & 0 & \alpha^3 & \alpha^2 & \alpha & 1 & \cdots & \cdots & \cdots & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & & & & & \ddots & & & & \vdots \\
\vdots & \vdots & \vdots & \vdots & & & & & & 1 & \alpha & \alpha^2 & \alpha^3 \\
\vdots & \vdots & \vdots & \vdots & & & & & & \alpha & 1 & \alpha & \alpha^2 \\
\vdots & \vdots & \vdots & \vdots & & & & & & \alpha^2 & \alpha & 1 & \alpha \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \alpha^3 & \alpha^2 & \alpha & 1
\end{pmatrix}$$

Again, each block refers to the same patient, field, nest or farm. The above correlation matrix assumes regular distances (or time interval) between observations. The parameterisation is rather more involved where the distances are uneven. The advantage of this correlation structure is that only one correlation parameter needs estimated, i.e. α.

For the California birds, it is a good option; the correlation between observations separated by one time unit (2 weeks) is likely to be more similar than those separated by larger time units. For the deer that, it would not make any sense as there is no time order in the sampled animals per farm. For the owl data, it only makes sense if we consider the time order in the data.

The AR-1 correlation can be used for any data set in which there is a time order, although instead of time, depth or age gradients can also be used. This means that it can be used for the California bird data and for the owl data (using the arrival time). There are several books and papers that discuss how to use GEE for spatial data; see for example Diggle and Ribeiro (2007) and especially Pebesma et al. (2000).

Option 3: Exchangeable Correlation

This is the most easily understood and most easily estimated form of within-subject correlation. The correlation between two observations from the same field i is assumed to be

$$\text{cor}(Y_{is}, Y_{it}) = \alpha$$

If, for example, we take body weights of a batch of roe deer (say 4) from 5 different sites across the country, it is probably sufficient to just say that bodyweights from a given site are correlated. We do not need to consider temporal or sequential correlation (we ignore here the potential issue of within-site spatial correlation which is

deliberately vague in this example). But it is reasonable to expect that bodyweights from a given site will be more similar, on average, than those from other sites (availability of resources, genetic similarity, etc). It can be imagined that the correlation between bodyweights Y_{ij}, where i denotes the area and j the animal within the area, may take the form

$$\begin{pmatrix}
1 & \alpha & \alpha & \alpha & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
\alpha & 1 & \alpha & \alpha & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
\alpha & \alpha & 1 & \alpha & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
\alpha & \alpha & \alpha & 1 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & 0 & 0 & 0 & 1 & \alpha & \alpha & \alpha & & & & & 0 \\
0 & 0 & 0 & 0 & \alpha & 1 & \alpha & \alpha & & & & & 0 \\
0 & 0 & 0 & 0 & \alpha & \alpha & 1 & \alpha & & & & & 0 \\
0 & 0 & 0 & 0 & \alpha & \alpha & \alpha & 1 & & & & & 0 \\
\vdots & \vdots & \vdots & \vdots & & & & & \ddots & & & & \vdots \\
\vdots & \vdots & \vdots & \vdots & & & & & & 1 & \alpha & \alpha & \alpha \\
\vdots & \vdots & \vdots & \vdots & & & & & & \alpha & 1 & \alpha & \alpha \\
\vdots & \vdots & \vdots & \vdots & & & & & & \alpha & \alpha & 1 & \alpha \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \alpha & \alpha & \alpha & 1
\end{pmatrix}$$

So we now have a new term α which expresses the correlation between bodyweights from animals in the same area. In the context of GEE, this is referred to as exchangeable correlation, but in other settings, it is often referred to as 'compound-symmetry'. We have also seen this correlation structure in Chapters 5 and 6 with the linear mixed effects model. There, we had the compound symmetry correlation due to a random intercept, but saw a similar correlation structure in the time series chapter. In the first case, the correlation is always positive; see also Pinheiro and Bates (2000, pp. 227–228).

Option 4: Another Correlation Structure – Stationary Correlation

One interesting case is where within-subject correlation exists up to a given distance and then stops completely. Although this is not an obvious choice of correlation structure, we may happen to know this is a good model in advance or it may be indicated from an exploratory analysis of the correlation structure. If we imagine again the situation of four consecutive blood readings taken, once per hour, as in the hypothesised scenario for autoregressive correlation.Under a model where correlation is autoregressive up to a time lag of 2 hours, but ceases thereafter, the correlation matrix will take this general form

$$\begin{pmatrix}
1 & \alpha & \alpha^2 & 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
\alpha & 1 & \alpha & \alpha^2 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
\alpha^2 & \alpha & 1 & \alpha & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & \alpha^2 & \alpha & 1 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & 0 & 0 & 0 & 1 & \alpha & \alpha^2 & 0 & & & & & 0 \\
0 & 0 & 0 & 0 & \alpha & 1 & \alpha & \alpha^2 & & & & & 0 \\
0 & 0 & 0 & 0 & \alpha^2 & \alpha & 1 & \alpha & & & & & 0 \\
0 & 0 & 0 & 0 & 0 & \alpha^2 & \alpha & 1 & & & & & 0 \\
\vdots & \vdots & \vdots & \vdots & & & & & \ddots & & & & \vdots \\
\vdots & \vdots & \vdots & \vdots & & & & & & 1 & \alpha & \alpha^2 & 0 \\
\vdots & \vdots & \vdots & \vdots & & & & & & \alpha & 1 & \alpha & \alpha^2 \\
\vdots & \vdots & \vdots & \vdots & & & & & & \alpha^2 & \alpha & 1 & \alpha \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & \alpha^2 & \alpha & 1
\end{pmatrix}$$

Note that unlike the other correlation structures described, we see cases of zero correlation within-subject.

This gives a flavour of the various correlation structures available. There are several others (e.g. non-stationary auto-correlation and ante-dependence), and you can impose correlation estimates a priori if this information is known in advance.

The number of unknown parameters in the auto-regressive and compound symmetric correlation structures was only 1, but with the unstructured correlation we have $t \times (t-1)/2$ parameters. This is potentially difficult to estimate, especially if we have a relatively large number of observations over time; 10 longitudinal observations means we already need to estimate 45 association parameters!

12.3 Why All the Fuss?

We saw in Chapters 5, 6, and 7 how data from the same source (e.g. all readings taken from the same beach) can be correlated and the implications this has for the variance-covariance structure, which in turn informs the error associated with the parameter estimates. In the simplest scenario, we can imagine a situation where there is no within-subject correlation and the correlation matrix for the data Y_{ij} is simply diagonal.

$$\begin{pmatrix}
1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 0 & & 0 \\
0 & 0 & 1 & & 0 \\
\vdots & & & \ddots & 0 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}$$

Indeed, we no longer need to consider the problem in terms of Y_{ij} and i subjects, rather just Y_j with no subject index. This is the correlation structure adopted implicitly in GLM.

In Chapters 9 and 10, we discussed the mathematical background of GLMs. Recall that we started with a distribution, specified a likelihood function L, and then found the parameters that maximised the likelihood function. The matrix of standard errors is essential as the basis of statistical inference, and typically, this is estimated as the inverse of the matrix of second derivatives of the GLM log-likelihood L such that:

$$V_H(\hat{\beta}) = \{(-\frac{\partial^2 L}{\partial \beta_u \partial \beta_v})\}_{p \times p}^{-1}$$

A different approach, which is asymptotically equivalent, (i.e. tends towards an equivalent solution with increasing sample size) is based on the *expectation* of the second derivative which comes from the result

$$E\left(\frac{\partial^2 L}{\partial \beta_u \partial \beta_v}\right) = -\frac{\partial L}{\partial \beta_u} \times \frac{\partial L}{\partial \beta_v}$$

This second approach based on the expectation of the second derivatives is usually referred to as Fisher scoring. Although these two approaches will tend towards the same solution, discrepancies can occur, particularly where the sample size is small. Statistical tests are then based on the recognised t-test formulation, i.e.

$$\frac{\hat{\beta}}{s.e.(\hat{\beta})}$$

The standard errors are taken from the diagonal elements of the V_H or the Fisher information matrix. The problem with this approach is that the underlying statistical theory assumes that the observations are independent. And this is where GEE provides a solution to cases where we might be violating the independence assumption.

Basically, GEE uses the same equations as generalised least squares (GLS) and GLM, but instead of using a diagonal matrix for the covariance matrix (implying independence), we replace it by an association matrix, as defined in the previous section. If you are not familiar with the regression, GLS and GLM maths, you can skip a couple of paragraphs, as it is not essential for *using* GEE. We stress that we only present the principle; the reader interested in full mathematical details is advised to read Liang and Zeger (1986).

12.3.1 A Bit of Maths

In linear regression, the following criteria (which is the residual sum of squares) is minimised to find the optimal regression parameters.

$$\sum_{i=1}^{N} (\mathbf{Y}_i - \mathbf{X}_i \times \boldsymbol{\beta})' \times (\mathbf{Y}_i - \mathbf{X}_i \times \boldsymbol{\beta})$$

In the context of the California bird data, i is the index for fields, $N = 11$, $\mathbf{Y}_i$ contains all the longitudinal data from field i, and $\mathbf{X}_i$ denotes the associated explanatory variables. The regression parameters in $\boldsymbol{\beta}$ are obtained by minimising this expression by taking derivates with respect to $\boldsymbol{\beta}$, setting them to 0, and solving the resulting equations. In GLS, we use a similar optimisation criterion, namely,

$$\sum_{i=1}^{N} (\mathbf{Y}_i - \mathbf{X}_i \times \boldsymbol{\beta})' \times \boldsymbol{\Sigma}_i^{-1} \times (\mathbf{Y}_i - \mathbf{X}_i \times \boldsymbol{\beta})$$

The matrix $\boldsymbol{\Sigma}_i$ is a covariance matrix which can either have different diagonal elements (to model heterogeneity) or non-zero off-diagonal elements to allow for temporal or spatial correlation. In Chapters 5, 6, and 7, we used $\boldsymbol{\Sigma}_i$ to describe the within-field correlation structure. Taking derivatives of this optimisation criterion with respect to $\boldsymbol{\beta}$ and setting them to 0 give

$$\sum_{i=1}^{N} \mathbf{X}_i \times \boldsymbol{\Sigma}_i^{-1} \times (\mathbf{Y}_i - \mathbf{X}_i \times \boldsymbol{\beta}) = \mathbf{0}$$

It is also common notation to replace $\mathbf{X}_i\boldsymbol{\beta}$ by $\boldsymbol{\mu}_i$. For a GEE, we follow the same procedure, and the starting point is again

$$\sum_{i=1}^{N} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \times' \boldsymbol{\Sigma}_i^{-1} \times (\mathbf{Y}_i - \boldsymbol{\mu}_i)$$

Again, the optimal regression parameter are obtained by taking derivatives and solving the generalised estimation equations

$$\sum_{i=1}^{N} \mathbf{D}_i \times \boldsymbol{\Sigma}_i^{-1} \times (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \tag{12.1}$$

The matrix $\mathbf{D}_i$ contains first-order derivatives of the $\boldsymbol{\mu}_i$ with respect to the regression parameters. $\boldsymbol{\Sigma}_i$ is the covariance matrix, and it can be written as

$$\boldsymbol{\Sigma}_i = \mathbf{A}_i^{\frac{1}{2}} \times \text{cor}(\mathbf{Y}_i) \times \mathbf{A}_i^{\frac{1}{2}}$$

This looks complicated, but the matrices $\mathbf{A}_i$ are diagonal matrices containing the variances. So, basically this is just matrix notation for the definition of the covariance: Correlation multiplied with the square root of the variances. Again, in ordinary GLMs, both $\boldsymbol{\Sigma}_i$ and $\text{cor}(\mathbf{Y}_i)$ are diagonal matrices, because we assume independence.

The problem is that in reality we have to estimate the covariance matrix $\boldsymbol{\Sigma}_i$, and this can be quite expensive in terms of numbers of parameters. GEE applies a clever trick by replacing the inner part, $\text{cor}(\mathbf{Y}_i)$, by an estimate correlation matrix $\mathbf{R}(\alpha)$ so that we get

$$\mathbf{V}_i = \phi \times \mathbf{A}_i^{\frac{1}{2}} \times \mathbf{R}_i(\alpha) \times \mathbf{A}_i^{\frac{1}{2}}$$

The ϕ allows for extra variation such as in a quasi-Poison model. The $\mathbf{V}_i$ is an estimate of $\mathbf{\Sigma}_i$. The better you choose the correlation structure, the closer the estimated covariance matrix $\mathbf{V}_i$ (also called the working covariance matrix) is to the real covariance matrix. This means that we have to determine what form $\mathbf{R}(\alpha)$ takes, or more precisely, to choose a correlation structure that closely describes what is observed in the response data. This can be any of the correlation structures discussed in the previous section.

But we have still not answered the question in the title of this section. Well, here it comes. We want to estimate the values of the $\boldsymbol{\beta}$s and their confidence intervals and then apply statistical tests. To estimate the $\boldsymbol{\beta}$s, an iterative algorithm is applied that consists of the following steps:

1. For given ϕ and α (and therefore $\mathbf{V}_i$), obtain an estimate for the regression parameters.
2. Given the regression parameters, update ϕ and α (and therefore $\mathbf{V}_i$). Pearson residuals are used for this.
3. Iterate between steps 1 and 2 until convergence.

At convergence, the estimated regression parameters are consistent[2] and asymptotically normally distributed with mean $\boldsymbol{\beta}$ and covariance matrix: $\mathbf{B}^{-1} \times \mathbf{M} \times \mathbf{B}^{-1}$, where

$$\mathbf{B} = \sum_{i=1}^{N} \mathbf{D}_i \times \mathbf{\Sigma}_i^{-1} \times \mathbf{D}_i$$
$$\mathbf{M} = \sum_{i=1}^{N} \mathbf{D}_i \times \mathbf{\Sigma}_i^{-1} \times \text{cov}(\mathbf{Y}_i) \times \mathbf{\Sigma}_i^{-1} \times \mathbf{D}_i$$

This statement also holds true, even if your specification of the correlation structure is not correct. We used the same notation as Fitzmaurice et al. (2004). And once we have calculated the covariance matrix, we can use its diagonal elements to obtain standard errors and confidence intervals. Hence, the last thing we have to do is explain how to get the $\mathbf{B}$ and $\mathbf{M}$. This is a matter of replacing $\mathbf{\Sigma}_i$ by its estimate $\mathbf{V}_i$ and $\text{cov}(\mathbf{Y}_i)$ by the covariance matrix $(\mathbf{Y}_i - \boldsymbol{\mu}_i) \times (\mathbf{Y}_i - \boldsymbol{\mu}_i)'$. Your chosen correlation $\mathbf{R}(\alpha)$ structure is then used in the covariance term in the inner part of the matrix $\mathbf{M}$, resulting in the so-called sandwich estimator. GEE is robust against misspecification of the correlation structure (it still provides valid standard errors). This does not mean you do not have to bother about choosing a good correlation structure; the better your choice, the better the standard errors. And, it is only a characteristic of large sample sizes.

[2]Consistent means that estimated parameters are nearly equal to the population parameters.

All the complications involved in choosing the within-subject correlation structure are essentially a means to an end. Usually we are interested in the significance of covariates or so-called fixed effects. In a sense, the correlation structure can be seen as an inconvenience that needs to be accounted for before making meaningful inferences about the parameters we are primarily interested in.

In summary, to answer the title of this section, GEE incorporates a correlation structure on the data from the same field, and as a result, we obtain consistent estimators.

In the second step of the two-step algorithm described above, for given regression parameters update ϕ, α, and $\mathbf{V}_i$, things are more complicated. Depending on the type of correlation structure you use, e.g. AR1, unstructured or exchangeable, the software will use different expressions for these two parameters.

12.4 Association for Binary Data

We can easily extend the idea above to deal with similar count data problems such as, for example, the number of ticks found on the same deer measured for bodyweight. Alternatively, in Chapter 22, we use a case study where the response variable is the presence or absence of badger activity at farms: A binary variable. The same holds for the deer data. Various statistical textbooks contain phrases like: 'the correlation is modelled at the level of the linear predictor rather than at the scale of the raw data'. The underlying idea is that for binary data, the correlation coefficient is not the most natural tool to define association. Using some basic definitions like P(A and B) = P(B) × P($A \mid B$), the definition of the expectation of discrete random variable, the mean and variance of a binary variable, and the definition of the correlation and covariance, we can easily show that the correlation between two binary variables with means μ_1 and μ_2 ($\mu_1 \geq \mu_2$) is smaller than $\sqrt{(\mu_2 - \mu_1\mu_2)/(\mu_1 - \mu_1\mu_2)}$. If, for example, E($Y_1$) = μ_1 = 0.7 and E(Y_2) = μ_2 = 0.3, then the correlation between Y_1 and Y_2 is smaller than 0.49. To overcome this, Fitzmaurice et al. (2004) used odds ratios to define the (unstructured) association as $\log(\text{OR}(Y_{is},Y_{ik})) = \alpha_{sk}$, where

$$OR\,(Y_{is}, Y_{ik}) = \frac{\Pr(Y_{is} = 1 \text{ and } Y_{ik} = 1) \times \Pr(Y_{is} = 0 \text{ and } Y_{ik} = 0)}{P(Y_{is} = 1 \text{ and } Y_{ik} = 0) \times P(Y_{is} = 0 \text{ and } Y_{ik} = 1)}$$

However, this association parameterisation is not available in the main GEE packages in R (it is in SAS), but you could program it yourself using the option for the user specified correlation structure in the GEE functions. In the GEE functions, we use in R in the next section, and in Chapter 22, we specify a correlation structure at the level of the raw data.

12.5 Examples of GEE

12.5.1 A GEE for the California Birds

In this section we revisit the California bird data, and apply GEE. The first two steps of GEE were presented in Section 12.1, but are repeated here. In the first step, we specify the relationship between the mean μ_{is} and the covariates:

$$E(Y_{is}) = \mu_{is} = e^{\alpha+\beta_1\times\text{Depth}_{is}+\beta_2\times\text{Depth}_{is}^2+\beta_3\times\text{Sptreat}_{is}}$$

In the second step, we specify the variance of the observed data:

$$\text{Var}(Y_{is}) = \phi \times \mu_{is}$$

Hence, we use $v(\mu_{is}) = \phi\mu_{is}$, which is in line with the characteristics of a (quasi-) Poisson GLM, but keep in mind we do not specify any distribution here. In the third step, we need to specify a correlation structure. One option is to use biological knowledge and argue that the number of birds in a field i at time s depends on those measured at time $s - 1$, and also, although less strong, on $s - 2$, etc., in the same field. Accepting this approach suggests using an auto-regressive correlation structure. We could also make an auto-correlation function for the data of each field, and investigate whether there is a significant auto-correlation. And if this shows no correlation, then we can apply a GLM.

The alternative option of a compound correlation is unlikely to be appropriate here. Why would bird numbers separated by 2 weeks (1 sampling unit) have the same correlation as those separated by 20 weeks (10 sampling units)?

There are various packages for GEE in R, but we only use the `geeglm` function from the `geepack` package in this book. The `gee` function from the `gee` package is also useful and so is the package `yags`. These packages are not part of the base installation of R; so you will need to download and install them. We use the `geepack` package as it is slightly more advanced than the others, e.g. it allows for a Wald test via the `anova` command.

The following code loads the `geepack` package (assuming it has been downloaded and installed) and applies the GEE (you also need to run the code from Section 12.1 for the data preparation).

```
> library(geepack)
> M.gee1 <- geeglm(Richness ~ offset(LA) + DEPTH +
         DEPTH2 + fSptreat, data = RFBirds,
         family = poisson, id = fField, corstr = "ar1")
> summary(M.gee1)
```

Note that this function wants us to specify a distribution with the `family` option, even though we are not assuming any distribution directly.

The grouping structure is given by the `id` option; this specifies which bird observations form a block of data. The `corstr` option specifies the type of correlation. This correlation is applied on each block of data. We argued above that the AR-1 auto-correlation structure should be used; hence `corstr = "ar1"`. Alternatives are `unstructured` (multiple αs), `exchangeable` (one α), `independence` (this gives the same results as the ordinary GLM), and `userdefined` (for the braves; you can program your own correlation structure). Our data does not contain missing values and were sorted along time within a field. If this is not the case, you need to use the `waves` option; see also the `geeglm` help file. This option ensures that R does not mess up the order of the observations. The `summary` command gives the following output.

```
  Coefficients:
                   Estimate      Std.err     Wald      p(>W)
(Intercept)    -0.678203399 0.3337043786 4.130438 0.04211845
fSptreatrlfld  -0.522313667 0.2450125672 4.544499 0.03302468
DEPTH           0.049823774 0.0287951864 2.993874 0.08358002
DEPTH2         -0.001141096 0.0008060641 2.004033 0.15688129

Estimated Scale Parameters:
             Estimate   Std.err
(Intercept)  2.333533 0.3069735

Correlation: Structure = ar1  Link = identity
Estimated Correlation Parameters:
       Estimate   Std.err
alpha 0.4215071 0.1133636
Number of clusters:   11   Maximum cluster size: 10
```

The correlation between two sequential observations in the same field is 0.42; if the time lag is two units (4 weeks), the correlation is $0.421^2 = 0.177$, between observations separated by three units (6 weeks), it is $0.421^3 = 0.075$, etc. The scale parameter is 2.333, which is similar to the over-dispersion parameter of the quasi-Poisson model applied on the same data in Section 12.1. There is a weak but significant treatment effect of the straw. Hence, the following model was fitted on the bird data.

$$
\begin{aligned}
E[Y_{is}] &= \mu_{is} = e^{-0.678+0.049\times \text{Depth}_{is}-0.001\times \text{Depth}_{is}^2-0.522\times \text{Sptreat}_{is}} \\
\text{var}(Y_{is}) &= 2.333 \times \mu_{is} \\
\text{cor}(Y_{is}, Y_{it}) &= 0.421^{|s-t|}
\end{aligned}
$$

This relationship is not conditional on random effects, only on the explanatory variables. For this reason, it is called a marginal model. Hardin and Hilbe (2002) called it the population average GEE, abbreviated as PA-GEE.

Note that in the GLM in Section 12.1 both the straw management variable and the depth variables are significant. In the GEE, which takes into account temporal correlation, only the straw management variable is significant!

The nice thing of the `geepack` package is that it allows for a Wald test, which can be used to test the significance of nominal variables with more than two levels. This is not the case here, but for illustrative purposes, we show how it can be used to decide whether we need any of the depth terms. The code below fits a GEE without any of the depth terms and applies a Wald test using the `anova` command. The output suggests that we only need fSptreat.

```
> M.gee2 <- geeglm(Richness ~ offset(LA) +  fSptreat,
          data = RFBirds, family = poisson, id = FIELD,
          corstr =  "ar1")
> anova(M.gee1, M.gee2)
Analysis of 'Wald statistic'  Table
Model 1 Richness ~ offset(LA) +  DEPTH + DEPTH2 + fSptreat
Model 2 Richness ~ offset(LA) +  fSptreat
  Df     X2 P(>|Chi|)
1  2 3.9350    0.1398
```

12.5.2 A GEE for the Owls

So, what is an appropriate correlation structure for the owl data? We could use the compound correlation structure, which is called 'exchangeable' within the context of the GEE. This assumes that all observations from the nest are correlation with the value of α. Code to do this is given by

```
> library(geepack)
> Form <- formula(NCalls ~ offset(LBroodSize) +
                    SexParent * FoodTreatment +
                    SexParent * ArrivalTime)
> O4 <- geeglm(Form, data = Owls, family = poisson,
               id = Nest, corstr = "exchangeable")
```

The results of the `summary(O4)` command are not given here, but show that the estimated value of α is 0.058, which is rather small.

In Chapters 5 and 6, we analysed the *average* sibling negotiation. Recall that we have multiple observations from the same nest, but that these were obtained during two nights. The food treatment was swapped during the second night. In Chapter 5, the compound correlation was imposed by using nest as a random intercept. In Chapter 6, we continued the analysis by arguing that there may be autoregressive correlation between the observations made in the same night from the same nest. The only thing is arrival times of the birds are not regularly spaced in

time, but we argued that from the owls' point of view, time may be regularly spaced (this was a biological assumption). We can do the same here, except that we use the number of calls.

The problem is that the data file does not contain a column that identifies the group of observations from the same night and nest; hence we have to make it.

```
> N <- length(Owls$Nest)
> NLev <- c(paste(unique(Owls$Nest), ".Dep", sep = ""),
            paste(unique(Owls$Nest), ".Sat", sep = ""))
> Owls$NestNight <- factor(levels = NLev)
> for (i in 1:N){
    if (Owls$FoodTreatment[i] == "Deprived") {
        Owls$NestNight[i] <-
               paste(Owls$Nest[i], ".Dep", sep = "")}
    if (Owls$FoodTreatment[i] == "Satiated") {
        Owls$NestNight[i] <-
               paste(Owls$Nest[i], ".Sat", sep = "")}}
```

This is a bit of tedious programming, and instead of explaining it in detail, let us show the results of the code:

```
> Owls[1 : 10, c(1, 2, 4, 10)]
         Nest FoodTreatment ArrivalTime      NestNight
1  AutavauxTV      Deprived       22.25 AutavauxTV.Dep
2  AutavauxTV      Satiated       22.38 AutavauxTV.Sat
3  AutavauxTV      Deprived       22.53 AutavauxTV.Dep
4  AutavauxTV      Deprived       22.56 AutavauxTV.Dep
5  AutavauxTV      Deprived       22.61 AutavauxTV.Dep
6  AutavauxTV      Deprived       22.65 AutavauxTV.Dep
7  AutavauxTV      Deprived       22.76 AutavauxTV.Dep
8  AutavauxTV      Satiated       22.90 AutavauxTV.Sat
9  AutavauxTV      Deprived       22.98 AutavauxTV.Dep
10 AutavauxTV      Satiated       23.07 AutavauxTV.Sat
```

The variable `NestNight` tells us which observations are from the same night and same nest. The column `ArrivalTime` shows at what time an observation was made, but as we already discussed, we will consider the arrivals as regularly spaced in time. So, the `for` loop with the `if` statement was only used to make the variable `NestNight`. You could also have done this in Excel. As always in R, things can be done in multiple ways. Here is an alternative piece of R code to obtain exactly the same `NestNight`.

```
> Owls$NestNight <- factor(
      ifelse(Owls$FoodTreatment ==  "Deprived",
         paste(Owls$Nest, ".Dep", sep=""),
         paste(Owls$Nest, ".Sat", sep="")))
```

The `ifelse` executes the first paste command if an observation is food deprived, and as the name already suggests, the second paste command otherwise. No need for a loop. Elegant, but it takes a bit more time to see what it does.

Applying the GEE is now simple:

```
> O3 <- geeglm(Form, data = Owls, family = poisson,
                id = NestNight, corstr = "ar1")
```

To figure out whether we need the two two-way interactions, we can drop each of them in turn, apply the Wald test, and remove the least significant variable:

```
> O3.A <- geeglm(NCalls ~ off-set(LBroodSize) +
      SexParent + FoodTreatment +
      SexParent * ArrivalTime, data = Owls,
      family = poisson, id = NestNight, corstr = "ar1")
> O3.B <- geeglm(NCalls ~ off-set(LBroodSize) +
      SexParent * FoodTreatment +
      SexParent + ArrivalTime, data = Owls,
      family = poisson, id = NestNight, corstr = "ar1")
> anova(O3, O3.A)

Analysis of 'Wald statistic' Table
Model 1 NCalls ~ offset(LBroodSize) + SexParent * Food-Treatment +
                 SexParent * ArrivalTime
Model 2 NCalls ~ offset(LBroodSize) + SexParent + FoodTreatment +
                 SexParent * ArrivalTime
  Df      X2  P(>|Chi|)
1  1 0.23867   0.62517

> anova(O3, O3.B)

Analysis of 'Wald statistic' Table
Model 1 NCalls ~ offset(LBroodSize) + SexParent * Food-Treatment +
                 SexParent * ArrivalTime
Model 2 NCalls ~ offset(LBroodSize) + SexParent * Food-Treatment +
                 SexParent + ArrivalTime
  Df      X2 P(>|Chi|)
1  1 0.40269  0.52570
```

The sex of the parent and food treatment interaction is the least significant term and was dropped. This process can then be repeated a couple of times until all terms in the model are significant. The final model and its output are given by:

```
> O6 <- geeglm(NCalls ~ off-set(LBroodSize) +
      FoodTreatment + ArrivalTime, data = Owls,
      family = poisson, id = NestNight, corstr = "ar1")
> summary(O6)

Call:
geeglm(formula = NCalls ~ offset(LBroodSize) + FoodTreatment +
    ArrivalTime, family = poisson, data = Owls, id = NestNight,
    corstr = "ar1")
```

```
 Coefficients:
                          Estimate    Std.err     Wald        p(>W)
(Intercept)             3.5927875 0.67421928 28.39623 9.885749e-08
FoodTreatmentSatiated -0.5780999 0.11507976 25.23527 5.074576e-07
ArrivalTime           -0.1217358 0.02725415 19.95133 7.943886e-06

Estimated Scale Parameters:
            Estimate   Std.err
(Intercept) 6.639577 0.5234689
Correlation: Structure = ar1  Link = identity

Estimated Correlation Parameters:
       Estimate    Std.err
alpha 0.5167197 0.06830255
Number of clusters:   277   Maximum cluster size: 18
```

The correlation of the calls between two sequential arrivals is 0.51, which is relatively high. The overdispersion is 6.6, which is similar to that of the quasi-Poisson GLM. The estimated regression parameters are similar to those of the quasi-Poisson GLM, but the p-values are considerably larger (at least for the slopes). However, the biological conclusions are the same; there is a food treatment effect (lower number of calls from food satiated observations) and later the night, the less calls. The final GEE is given by

$$E(\text{NCalls}_{is}) = \mu_{is} \quad \text{and} \quad \text{var}(\text{NCalls}_{is}) = 6.6 \times \mu_{is}$$
$$\text{cor}(\text{NCalls}_{is}, \text{NCalls}_{it}) = 0.51^{|t-s|}$$

12.5.3 A GEE for the Deer Data

The required correlation structure for the deer data is obvious; it has to be the compound correlation, alias the exchangeable correlation because there is no specific (e.g. time) order between the observations from the same farm. The code and numerical output to fit this model is as follows. The exchangeable correlation is selected using the `corstr = "exchangeable"` bit, and `id = Farm` tells the `geeglm` function which observations are from the same farm.

```
> library(geepack)
> DE.gee <- geeglm(Ecervi.01 ~ CLength * fSex,
               data = DeerEcervi,family = binomial,
               id = Farm, corstr = "exchangeable")
> summary(DE.gee)

Call:
geeglm(formula = Ecervi.01 ~ CLength * fSex, family = binomial,
    data = DeerEcervi, id = Farm, corstr = "exchangeable")

 Coefficients:
               Estimate     Std.err      Wald       p(> W)
(Intercept)   0.73338099 0.280987616  6.812162 9.053910e-03
```

```
CLength       0.03016867 0.006983758 18.660962 1.561469e-05
fSex2         0.47624445 0.217822972  4.780271 2.878759e-02
CLength:fSex2 0.02728028 0.014510259  3.534658 6.009874e-02

Estimated Scale Parameters:
             Estimate   Std.err
(Intercept)  1.145337 0.4108975

Correlation: Structure = exchangeable  Link = identity
Estimated Correlation Parameters:
       Estimate    Std.err
alpha 0.3304893 0.04672826
Number of clusters:   24   Maximum cluster size: 209
```

Note that a scale parameter is used. For a fair comparison with the binomial GLM (which does not contain a dispersion parameter), you can use the option `scale.fix = TRUE` in the `geeglm` command. Because the estimated dispersion parameter is only 1.14, we did not do this here. The correlation parameter is 0.33, which is moderate. The two-way interaction term is not significant ($p = 0.06$) at the 5% level, where in the binomial GLM it was! Hence, by including the compound correlation, the biological conclusions have changed! Perhaps we should re-phrase the last sentence a little bit as it suggests that both models are valid. The GLM without the correlation structure is potentially flawed as it ignores the correlation structure in the data. Therefore, only the GEE should be used for biological interpretation!

12.6 Concluding Remarks

GLS is a special case of GEE if we specify a Normal distribution and the identity link function. But we do not recommend running the GLS with GEE software as most existing GEE functions in R are less flexible in the sense of allowing for multiple variances ϕ for modelling heterogeneity.

For longitudinal data, GEE is useful if you have many fields or nest and relatively few longitudinal observations per field or nest *i*. If it is the other way around, standard errors produced by the sandwich estimator are less good.

Hardin and Hilbe (2002) used an AIC-type criterion to compare models with different correlation structures. It is called *quasilikelihood under the independence model information criterion* (QIC) after a paper from Pan (2001). A similar criterion is also used for selection explanatory variables. The `geeglm` function does not produce the QIC; hence, you have to program this yourself. The appendix in Hardin and Hilbe (2002) gives Stata code for this. The R package `yags` does produce the QIC. It is open code, which means that you can easily see how the programmer of `yags` implemented it. The problem that you may encounter with the QIC is that not every referee may have heard of it or agree with it.

We have not mentioned the word model validation yet. Hardin and Hilbe (2002) dedicate a full chapter to this; they present a couple of tests to detect patterns in residuals, and also graphical model validation tools. The graphical validation uses

Pearson residuals and follows the model validation steps of GLM; see also Chapters 9 and 10. We strongly suggest that after reading this chapter, you consult Hardin and Hilbe (2002). However, you have to either use Stata to follow their examples or read over the Stata code and use any of the R packages to do the same.