Chapter 22
A Comparison of GLM, GEE, and GLMM Applied to Badger Activity Data

N.J. Walker, A.F. Zuur, A. Ward, A.A. Saveliev, E.N. Ieno, and G.M. Smith

22.1 Introduction

In this chapter, we analyse a data set consisting of signs of badger (*Meles meles*; see Fig. 22.1) activity around farms. The data are longitudinal and from multiple farms; so it is likely a temporal correlation structure is required. The response variable is binary; the presence or absence of badger activity. The dataset comes from a survey carried out on 36 farms over 8 consecutive seasons running from autumn 2003 to summer 2005. For analytical convenience, we consider these intervals to be exactly equal, which is a close enough approximation to the reality. All farms in the survey were in South-West England, which is a high-density badger country.

Fig. 22.1 Photograph of two badgers on the nightly hunt for food. The photo was taken by Dr Richard Yarnell, School of Animal, Rural and Environment Sciences, Nottingham Trent University, UK

N.J. Walker (✉)
Woodchester Park CSL, Tinkley Lane, Nympsfield, Gloucester GL10 3UJ, United Kingdom

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_22,

This work was carried out in the wider context of badgers and their possible role in transmitting bovine tuberculosis to cattle. One avenue for tackling this problem might be to reduce the rates of badger visits to farms in particular areas where they may come into contact with resident cattle. The aim of this study was to predict the occurrence of signs of badger activity on farms.

There are many different ways of measuring badger activity, but for the purposes of this chapter, we just consider one of these: 'signs of activity'. This was used as a binary variable that took the value 1 when signs of badger activity were recorded and 0 if no signs were recorded. Signs of activity included badger faeces, indications of digging, feeding evidence, etc. Several potential explanatory variables were recorded – these are detailed in Table 22.1.

Consecutive observations on badger activity at a given farm may be temporally auto-correlated. Because of this and because the data are in binary form, we

Table 22.1 List of variables with a short description. The response variable is Signs_in_yard

Variable	Description
Year	Calendar year
Season	Spring (Mar–May), Summer (Jun–Aug), autumn(Sept–Nov) and winter (Dec–Feb)
Farm_code_numeric	Blinded farm identifier
Survey	Which of the 8 survey occasions (i.e. the time indicator)
Signs_in_yard	Binary indicator of signs of badger activity
Latrines_with_farm_feed	Binary indicator – do (any) observed badger latrines contain farm feed? (This is a proxy for the fact that badgers must have been on farm).
No_latrines_with_farm_feed	The number of the above
No_scats_with_farm_feed	Number of badger faeces identified as containing farm feed
No_latrines	Number of badger latrines observed
No_setts_in_fields	Number of badger setts (i.e. homes) observed
No_active_setts_in_fields	Number of actively used setts observed
No_buildings	Number of buildings on farm
No_cattle_in_buildings_yard	Number of cattle housed in the building yard
Mode_feed_store_accessibility	Quantitative index of how easy it would be for badgers to access the farm's feed store
Accessible_feed_store_present	Binary indicator – is such a feed store present?
Mode_cattle_house_accessibility	Quantitative index of how easy it would be for badgers to access the cattle house
Accessible_cattle_house_present	Binary indicator – is such a feed store present?
Accessible_feed_present	Binary indicator – is accessible feed present
Grass_silage	Binary indicator of presence of grass silage
Cereal_silage	Binary indicator of presence cereal silage
HayStraw	Binary indicator of presence of Hay/Straw
Cereal_grains	Binary indicator of presence of cereal grains
Concentrates	Binary indicator of presence of concentrates
Proteinblocks	Binary indicator of presence of protein blocks
Sugarbeet	Binary indicator of presence of sugar beet
Vegetables	Binary indicator of presence of vegetables
Molasses	Binary indicator of presence of molasses

used generalised estimating equations (GEE) and generalised linear mixed models (GLMM). If there would be no temporal auto-correlation, then generalised linear modelling (GLM) can be applied. The underlying GLM, GEE, and GLMM theory was discussed in Chapters 9, 12, and 13.

The aim of this chapter is not to find the best possible model for the data, but merely to contrast GLM, GEE, and GLMM. When writing this chapter, we considered two ways to do this, namely,

1. Apply a model selection in each of the three models (GLM, GEE, and GLMM). It is likely that the optimal GLM consists of a different set of explanatory variables than the GEE and GLMM. The reason for this is the omission of the dependence structure in the data. We have seen this behaviour already in various other examples in this book with the Gaussian distribution. Also, recall the California data set that was used to illustrate GLM and GEE in Chapter 12; the p-values of the GLM were considerably smaller than those of the GEE! Therefore, in a model selection, one ends up with different models. Using this approach, the story of the chapter is then that (erroneously) ignoring a dependence structure gives you a different set of significant explanatory variables.
2. Apply the GLM, GEE, and GLMM on the same set of explanatory variables and compare the estimated parameters and p-values. If they are different (especially if the GLM p-values are much smaller), then the message of the chapter is that ignoring the dependence structure in a GLM gives inflated p-values.

Both approaches are worthwhile presenting, but due to limited space, we decided to go for option 2 and leave the first approach as an exercise to the reader. The question is then: Which GLM model should we select? We decided to adopt the role of an ignorant scientist and apply the model selection using the GLM and contrast this with the GEE and GLMM applied on the *same* selection of covariates. Note that the resulting GEE and GLMM models are not the optimal models as we are not following our protocol from Chapters 4 and 5, which stated that we should first look for the optimal random structure using a model that contained as many covariates as possible.

22.2 Data Exploration

The first problem we encountered was the spreadsheet (containing data on 282 observations), which was characterised by a lot of missing values. Most R functions used so far have options to remove missing values automatically. In this section, we will use the `geepack` package, and its `geeglm` function requires the removal of all missing values.

Rows with missing values in the response variable were first removed. Some of the explanatory variables had no missing values at all and other explanatory variables had 71 missing values! Removing every row (observation) that contains a

Table 22.2 Number of missing values per variable. The data set contains 288 rows (observations). The notation '# NAs' stands for the number of missing values. The response variable is Signs_in_yard and contains 6 missing values

Variable	# NAs	Variable	# NAs
Year	0	Accessible_feed_store_present	6
Season	0	Mode_cattle_house_accessibility	71
Farm_code_numeric	0	Accessible_cattle_house_present	6
Survey	0	Accessible_feed_present	6
Signs_in_yard	6	Grass_silage	6
Latrines_with_farm_feed	33	Cereal_silage	6
No_latrines_with_farm_feed	34	HayStraw	6
No_scats_with_farm_feed	59	Cereal_grains	6
No_latrines	30	Concentrates	6
No_setts_in_fields	10	Proteinblocks	6
No_active_setts_in_fields	15	Sugarbeet	6
No_buildings	6	Vegetables	6
No_cattle_in_buidlings_yard	6	Molasses	6
Mode_feed_store_accessibility	38		

missing value reduces the sample size. Therefore, it is perhaps better to remove entirely explanatory variables with several missing values. This is an arbitrary process; where do you draw the line when you stop removing explanatory variables? The answer should be based on biological knowledge and common sense (drop the variables with lots of missing values and that you also think are the least important). Table 22.2 shows the number of missing values per variable. The explanatory variable `Mode_cattle_house_accessibility` has 71 missing values. If we insist on using it, we end up removing 71 observations or 24% of the data! To avoid such a situation, we decided to omit all explanatory variables with more than 15 missing values from the analysis. From the remaining data, we removed all rows where there was at least one observation missing, ending up with 273 observations for analysis.

Table 22.2 was obtained with the following R code.

```
> library(AED); data(BadgersFarmSurveys.WithNA)
> Badgers.NA <- BadgersFarmSurveys.WithNA  #Saves space
> colSums(sapply(Badgers.NA, FUN = is.na)
```

The `sapply` function creates a matrix of length 288 by 27 with the elements FALSE (corresponding element in `Badger.NA` is not a missing value) and TRUE (corresponding element is a missing value). The function `colSums` converts each FALSE into a 0 and TRUE into a 1 and takes the sum per column: the number of missing values per variable.

The number of explanatory variables is very large, and using a data exploration, we tried to find collinear explanatory variables. Pairplots (for the continuous

variables), Pearson correlation coefficients and variance inflation factors indicated that No_setts_in_fields and No_active_setts_in_fields are collinear; they have a correlation of 0.86.

We decided to drop the variable No_active_setts_in_fields. The variables No_buildings and No_cattle_in_buildings_yard have a correlation of 0.53. We decided to drop the second one. The explanatory variables Proteinblocks and Vegetables had only a few values of 1; the majority of observations had a 0 value. Including them caused numerical problems and we decided to drop them.

22.3 GLM Results Assuming Independence

The following code accesses the data (we removed the missing values in Excel and created a new data file), renames some of the longer variable names, and applies a GLM assuming independence. We could have renamed the variables in the data file, but the code below shows you the coding misery due to having long variable names (let it be a warning!). Always try to choose the names as short as possible when you create the data file. Most of the nominal variables are binary with values 0 (representing no) and 1 (representing yes), and for these, the factor command can be avoided because this is exactly what it does: making columns with zeros and ones. However, we decided to use it as it is too easy to make a mistake. The drop1 function applies an analysis of deviance (Chapter 9).

```
> library(AED); data(BadgersFarmSurveysNoNA)
> Badgers <- BadgersFarmSurveysNoNA
> Badgers$fSeason <- factor(Badgers$Season)
> Badgers$fFeed.store <-
          factor(Badgers$Accessible_feed_store_present)
> Badgers$fCattle.house <-
        factor(Badgers$Accessible_cattle_house_present)
> Badgers$fFeed.present <-
        factor(Badgers$Accessible_feed_present)
> Badgers$fGrass.silage <- factor(Badgers$Grass_silage)
> Badgers$fCereal.silage <- factor(Badgers$Cereal_silage)
> Badgers$fHayStraw <- factor(Badgers$HayStraw)
> Badgers$fCereal.grains <- factor(Badgers$Cereal_grains)
> Badgers$fConcentrates <- factor(Badgers$Concentrates)
> Badgers$fSugarbeet <- factor(Badgers$Sugarbeet)
> Badgers$fMolasses <- factor(Badgers$Molasses)
> B.glm <- glm(Signs_in_yard ~ fSeason+
      No_setts_in_fields + No_buildings + fFeed.store +
      fCattle.house + fFeed.present + fGrass.silage +
      fCereal.silage + fHayStraw + fCereal.grains +
```

```
      fConcentrates + fSugarbeet + fMolasses,
      family = binomial, data = Badgers)
> drop1(B.glm, test = "Chi")
```

The results are not presented here, but most explanatory variables are not significant at the 5% level. We decided to drop the least significant explanatory variable, refit the model, reapply the `drop1` command, and continue to drop explanatory variables until all remaining variables in the model are significant. The final model contains `No_setts_in_fields` and `fFeed.store`. Applying this model in R and an analysis of deviance with the `drop1` function gave

```
> B2.glm <- glm(Signs_in_yard ~ No_setts_in_fields +
        fFeed.store, family = binomial, data = Badgers)
> drop1(B2.glm, test = "Chi")

Single term deletions. Model:
Signs_in_yard ~ No_setts_in_fields + fFeed.store

                   Df Deviance     AIC     LRT    Pr(Chi)
<none>                  182.509 188.509
No_setts_in_fields   1  234.107 238.107  51.597 6.813e-13
fFeed.store          1  187.307 191.307   4.798   0.02849
```

The number of setts in the field is highly significant, and the presence of accessible feed store is weakly significant. The parameter estimates are obtained with the `summary(B2.glm)` command and are given below.

```
Coefficients:
                   Estimate Std. Error z value  Pr(>|z|)
(Intercept)         -2.5891     0.4602  -5.626 1.85e-08
No_setts_in_fields   0.2862     0.0457   6.262 3.79e-10
fFeed.store1        -1.0341     0.4587  -2.254   0.0242

Dispersion parameter for binomial family taken to be 1
    Null deviance: 237.79  on 272  degrees of freedom
Residual deviance: 182.51  on 270  degrees of freedom
AIC: 188.51
```

Note the number of setts in the field has a positive effect on the probability of finding badger activity. The nominal variable `fFeed.store` has values 0 and 1; hence, the `summary` output shows that on the linear predictor scale, for observations that have accessible feed storage, the intercept is lower by -1.03. Using the definition of the logit link function, this can be translated into probabilities (Chapter 10). The final GLM model is given by the following three steps.

1. Let Y_{is} be the value of the variable Signs_in_yard for farm i at time s. We assume that Y_{is} follows a binomial distribution with probability p_{is}; hence, $E(Y_{is}) = p_{is}$ and $var(Y_{is}) = p_{is} \times (1 - p_{is})$.
2. The systematic component is given by: $\eta_{is} = -2.58 + 0.28 \times \text{No\_setts\_in\_field}_{is} - 1.03 \times \text{fFeed.store}_{is}$.
3. The link between the expected value of Y_{is} and η_{is} is the logistic link: $\text{logit}(p_{is}) = \eta_{is}$.
4. All observations are independent.

22.4 GEE Results

This time around, we call GEE and fit, in addition to the previous model, an auto-regressive structure to the within-farm observations. As discussed in the introduction, we deliberately choose the same set of explanatory variables for optimal comparison between the statistical methods. An alternative option is to start from scratch with all explanatory variables and apply a new model selection. The following R code was used.

```
> B.gee <- geeglm(Signs_in_yard ~ No_setts_in_fields +
              fFeed.store, family = binomial,
              id = farm_code_numeric, corstr = "ar1",
              waves = Survey, data = Badgers)
> summary(B.gee)

Mean Model:
 Mean Link:                 logit
 Variance to Mean Relation: binomial

 Coefficients:
                      estimate     san.se         wald            p
(Intercept)        -2.97581231 0.53278887 31.196134543 2.332300e-08
No_setts_in_fields  0.21951360 0.06936777 10.013994983 1.553552e-03
fFeed.store1        0.01389024 0.40863960  0.001155416 9.728840e-01

Scale is fixed.
Correlation Model:
 Correlation Structure:     ar1
 Correlation Link:          identity

 Estimated Correlation Parameters:
       estimate    san.se     wald            p
alpha 0.4901059 0.1137123 18.57656 1.632153e-05

Returned Error Value:    0
Number of clusters:   36   Maximum cluster size: 8
```

Note that the package `geepack` is not part of the base installation of R, and you need to download and install it. The `summary` command shows that the number of setts in the field is significant.

For optimal comparison with the GLM, we set the scale parameter ϕ equal to 1 (for binary data it does not make sense to correct for overdispersion). Note that the presence of the accessible feed store is no longer significant. Also the number of setts in the field is less significant ($p = 0.0015$ for the GEE and $p = 6.81 \times 10^{-13}$ for the GLM). The auto-correlation is moderate with a value of 0.49. Its standard error is small, indicating that the correlation is significant. However, the literature is not clear on the use of this standard error, and some packages will not print it. References were given in Chapter 12 that can be used to compare GEE models with and without a correlation structure.

Summarising, we can see that in comparison with the GLM approach, the GEE gives a more conservative result. Both models find the variable 'no. setts in fields' to be significant (although the p-value is lower, i.e. stronger association in the GLM). However, the GEE finds this to be the only significant variable, the GLM also gives 'accessible feed store present' as a significant predictor. This highlights the general effect to be expected by adjusting for inherent auto-correlation, i.e. more conservative results. This is particularly important in this example because the stepwise regression has an inherently high risk of including spurious explanatory variables in the final model. This result is not surprising given the multiple testing involved. (We could get round this by making some kind of adjustment in terms of significance thresholds, e.g. Bonferroni correction.)

The estimated correlation parameter indicates the presence of auto-correlation between within-farm observations, justifying our decision to use GEE. This is further evidence that the association between 'accessible feed store present' and 'signs of badger activity' as indicated by the original GLM model was statistically spurious. The final GEE model is given by the following three steps.

1. Let Y_{is} be the value of the variable Signs_in_yard for farm i at time s. We assume that $\mathrm{E}(Y_{is}) = p_{is}$ and $\mathrm{var}(Y_{is}) = p_{is} \times (1 - p_{is})$.
2. The systematic component is given by $\eta_{is} = -2.97 + 0.21 \times \mathrm{No\_setts\_in\_field}_{is} + 0.01 \times \mathrm{fFeed.store\_present}_{is}$. The link between the expected value of Y_{is} and η_{is} is the logistic link: $\mathrm{logit}(p_{is}) = \eta_{is}$.
3. The correlation between Y_{is} and Y_{ik} is given by $\mathrm{cor}(Y_{is}, Y_{ik}) = 0.49^{|s-k|}$.

22.5 GLMM Results

To compare results obtained with GEE and GLMM, which we also applied, the following R code was used. Again, for optimal comparison of the statistical methods, we used the same set of explanatory variables.

```
> library(lme4)
> B.glmm <- lmer(Signs_in_yard ~ No_setts_in_fields +
                 fFeed.store + (1 | farm_code_numeric),
                 family = binomial, data = Badgers)
```

The results obtained by the `summary(B.glmm)` command are given below.

```
Random effects:
 Groups             Name        Variance  Std.Dev.
 farm_code_numeric (Intercept)      5.32      2.30

Estimated scale (compare to 1)  0.77

Fixed effects:
                          Estimate     SE z-val   p-val
(Intercept)                 -5.34   0.98 -5.40 <0.001
No_setts_in_fields           0.37   0.10  3.38  0.0007
fFeed.store1                 0.28   0.70  0.39  0.69
```

If we are happy to accept the random effect structure used in this model, then we again arrive at the same conclusion that number of setts in the fields is an important predictor of signs of badger activity. The *p*-value is slightly lower here, suggesting that in this instance at least, the GEE was the most conservative approach.

In both these models (GEE and GLMM), the coefficient for the relationship between number of setts and probability of observing signs of badger activity is positive, but note that the GLMM result was stronger (+0.21 for the GEE and +0.37 for the GLMM).

Note that the GLMM does estimate an overdispersion parameter ϕ, and the software does not allow you to set it to 1 (as would be normal for binary data). The final GLMM model is given by the following three steps.

1. Let Y_{is} be the value of the variable Signs_in_yard for farm i at time s. We assume that Y_{is} is binomial distributed with $\mathrm{E}(Y_{is}) = p_{is}$ and $\mathrm{var}(Y_{is}) = p_{is} \times (1 - p_{is})$.
2. The systematic component is given by: $\eta_{is} = -5.34 + 0.37 \times \mathrm{No\_setts\_in\_field}_{is} + 0.28 \times \mathrm{fFeed.store}_{is} + \varsigma_i$, where ς_i is a random intercept with mean 0 and variance ${\sigma_\varsigma}^2$, which is estimated as 5.32.
3. The link between the expected value of Y_{is} and η_{is} is the logistic link: $\mathrm{logit}(p_{is}) = \eta_{is}$.

22.6 Discussion

This simple example highlights how three different approaches can give three similar, but different results – and different in important respects. As stated earlier, by ignoring the inherent within-farm auto-correlation, we increase the risk of type I error. This is probably why 'accessible feed store present' was significant only in the first under-specified model.

In terms of inference, if we are happy to choose the GEE from the approaches tried here, we can say first of all that there is auto-correlation between within-farm observations with respect to observing signs of badger activity. This is not

surprising. On average, we are talking about a 3-month separation in time. So, if signs of badger activity are observed at one visit – it is easy to imagine that there will be a good chance of making the same observation 3 months later (and vice versa for non-observations). But the probability of making the same finding diminishes with time; so if we go back to the same farm, maybe 18 months later, then the chance of observing the same result is less compelling. Hence, the choice of the 1st-order autoregressive structure.

Having chosen what we hope is a suitable auto-correlation structure, we find that 'number of setts in fields' is a significant predictor and in a positive direction ($\beta = 0.21$, s.e. $= 0.06$). This is of course an intuitive result, i.e. the more badger setts observed close to the farm, the more likely that badger signs will be observed on the farm. This may seem at first glance an obvious conclusion. However, it offers support to our choice of model, and of equal importance, it gives insight into the variables not important in predicting badger activity on farms.

We should not forget that the correlation structure may be due to a missing covariate or interaction. The problem is that it is rather difficult to decide which interaction term to include as there are so many options. Good biological knowledge is required when considering which interactions to fit.

As stated in the introduction, the GEE and GLMM were applied on a selection of covariates that was determined by the GLM model selection. This is against our protocol presented in Chapters 4 and 5. Our motivation for this approach was explained in Section 22.1: to show that GLM gives a model with inflated p-values. If you want to find the optimal GEE (or GLMM) model, you should apply the model selection using these models! Because we were curious ourselves, we applied a model selection using GEE and GLMM. With the GEE, we ended up with a model that only contains the covariates 'number of sets in a field' and 'presence/absence of sugar beets'. The GLMM picked only 'number of sets in a field'. Hence, adding a dependence structure on the data gives a different set of covariates in the model selection, and the type of dependency (auto-regressive correlation from the GEE versus the symmetrical compound correlation from GLMM) also plays a role. Thus, it is important to give careful consideration to choice of correlation structure in advance of any analysis.

22.7 What to Write in a Paper

This depends of course on the journal and the audience. In general, most readers of ecological journals will not be interested in the more technical details of procedures such as GEE. A line or two on the reason for using a GEE (auto-correlation, non-standard data, e.g. binary or count) needs to be included, even when submitting to the most non-technical of journals.

All relevant parameters are given, by convention, along with the standard errors. This includes the auto-correlation parameter.