ı of the current mixed modeling in
ıppear over and over again, in later
ıe applied work of many quantitative

cts in WinBUGS: The WinBUGS model
ts, random-slope model (i.e., the second
nilar to the fixed-effects "version" of the
in Chapter 11. Without looking at the
ıt chapter, take the linear mixed model
current chapter and change it back to a
ation-specific intercepts and slopes, i.e.,
fit in R as `lm(mass ~ pop*length)`.
ients regression without intercept–slope
`population * year`, with year

CHAPTER

13

Introduction to the Generalized Linear Model: Poisson "t-test"

OUTLINE

13.1 INTRODUCTION

The unification of a large number of statistical methods such as regression, analysis of variance (ANOVA), and analysis of covariance (ANCOVA) under the umbrella of the *general linear model* was a big advancement for applied statistics. However, even more significant was the unification of an even wider range of statistical methods within the class of the *generalized linear model* or GLM in 1972 by Nelder and Wedderburn (also see McCullagh and Nelder, 1989). They showed that a large number of techniques previously thought of as representing quite separate types of analyses, including logistic regression, multinomial regression, Chi-square tests, log-linear models, as well as the general linear model,

Introduction to WinBUGS for Ecologists
DOI: 10.1016/B978-0-12-378605-0.00013-2

could all be represented as special cases of a generalized version of a linear model. In that way, much of what was well understood for the linear model could be carried over to that much larger class of models.

The two main ideas of the GLM are that, first, a *transformation of the expectation* of the response $E(y)$ is expressed as a linear combination of covariate effects rather than the expected (mean) response itself. And second, for the random part of the model, *distributions other than the normal* can be chosen, e.g., Poisson or binomial.

Formally, a GLM is described by the following three components:

1. a *statistical distribution* is used to describe the random variation in the response y; this is the stochastic part of the system description,
2. a so-called *link function* g, that is applied to the expectation of the response $E(y)$, and
3. a *linear predictor*, which is a linear combination of covariate effects that are thought to make up $g(E(y))$; this is the systematic or deterministic part of the system description.

Binomial, Poisson, and normal are probably the three most widely used statistical distributions in a GLM (see Chapter 6). The former two are distributions for non-negative, discrete responses and therefore suitable to describe counts. The normal is the most widely used distribution for continuous responses such as measurements. The three most widely used link functions are the identity, `logit(=log(odds)=log(x/(1-x)))`, and the `log`. For various reasons, one link function is typically advantageous, although not obligate, for each of these distributions. For instance, the normal distribution combined with an identity link yields the general linear model; the Poisson with a log link yields a log-linear model; and the binomial with a logit link yields a logistic regression. Hence, all the normal linear models seen in Chapters 4–11 are simply special cases of a GLM.

In the next nine chapters, we will go through a progression from simple to more complex models for Poisson and binomial responses. As for normal linear models, we begin again with what might be called a "Poisson t-test" in the sense that it consists of a comparison of two groups. To better see the analogy with the normal linear model, we start by writing the model for the normal two-group comparison (see Chapter 7) in GLM format:

1. Distribution: $y_i \sim \text{Normal}(\mu_i, \sigma^2)$
2. Link function: identity, i.e., $\mu_i = E(y_i) = \text{linear predictor}$
3. Linear predictor: $\alpha + \beta * x_i$

Next, we generalize this model to count data. The inferential situation considered is that of counts (C) of Brown hares (Fig. 13.1) in a sample of 10 arable and 10 grassland study areas. We wonder whether hare density depends on land-use.

FIGURE 13.1 Brown hare (*Lepus europaeus*), Germany, 2008. (*Photo N. Zbinden*)

The typical distribution assumed for such counts is a Poisson, which applies when counted things are distributed independently and randomly and samples of equal size are taken randomly. Then, the number of hares counted per study area (C) will be described by a Poisson. The Poisson has a single parameter, the expected count λ, that is often called the intensity and here represents the mean hare density. In contrast to the normal, the Poisson variance is not a free parameter but is equal to the mean λ. For a Poisson-distributed random variable C, we write $C \sim \text{Poisson}(\lambda)$.

If hare density depends on land-use, i.e., is different in arable and grassland areas, the assumption of a constant mean density across all 20 study areas is not realistic. And in a "Poisson t-test" we are specifically interested in whether hare density differs between grassland and arable areas. Therefore, here is a model for hare count C_i in area i:

1. Distribution: $C_i \sim \text{Poisson}(\lambda_i)$
2. Link function: log, i.e., $\log(\lambda_i) = \log(E(C_i)) = \text{linear predictor}$
3. Linear predictor: $\alpha + \beta * x_i$

In words, hare count C_i in area i is distributed as a Poisson random variable with mean $E(C_i) = \lambda_i$. The log-transformation of λ_i is assumed to be a linear function $\alpha + \beta * x_i$, where α and β are unknown constants and x_i is the value of an area-specific covariate. If x_i is an indicator for arable areas, then α becomes the mean hare density on a log-scale in grassland areas and β, again on a log-scale, is the difference in mean density between the two land-use types.

13.2 AN IMPORTANT BUT OFTEN FORGOTTEN ISSUE WITH COUNT DATA

Whenever we interpret λ as the mean *hare density*, we make the implicit assumption that *every individual hare is indeed seen*, i.e., that detection probability (p) is equal to 1. This is not very likely for hares nor indeed for any wild animal because typically some individuals are overlooked (Yoccoz et al., 2001; Kéry, 2002; Williams et al., 2002; Kéry and Juillerat, 2004; Schmidt, 2005, 2008). Alternatively, we may assume that the *proportion of hares overlooked per area is the same*, on average, in both land-use types. In that case, counts are considered just an index to absolute density, i.e., a measure for *relative density*, and what we model as the Poisson parameter λ_i is in reality the product between absolute hare density and the proportion p of hares seen. Only by making the assumption that p is identical, on average, in both land-use types may we validly interpret a mean *difference* between counts in arable and grassland areas as an indication of a difference in true hare density. See Chapters 20 and 21 for more on this important topic, the distinction between the imperfectly observed true state and the observed data, or, between the ecological and the observation processes underlying ecological field data.

13.3 DATA GENERATION

For now, we simulate and analyze hare counts under the assumption that detectability is perfect. First we need an indicator for land-use:

```
n.sites <- 10
x <- gl(n = 2, k = n.sites, labels = c("grassland", "arable"))
n <- 2*n.sites
```

Let the mean hare density in grassland and arable areas be 2 and 5 hares, respectively. Then, $\alpha = \log(2) = 0.69$ and $\log(5) = \alpha + \beta$, thus, $\beta = \log(5) - \log(2) = 0.92$. Therefore, the expected density λ_i is given by:

```
lambda <- exp(0.69 + 0.92*(as.numeric(x)-1)) # x has levels 1 and 2, not 0 and 1
```

We add the noise that comes from a Poisson distribution and inspect the hare counts we've thus generated (Fig. 13.2):

```
C <- rpois(n = n, lambda = lambda)          # Add Poisson noise
aggregate(C, by = list(x), FUN = mean)      # The observed means
boxplot(C ~ x, col = "grey", xlab = "Land-use", ylab = "Hare count", las = 1)
```

Again, we can get a feel for the strong effects of chance (sampling variation) by repeatedly generating hare counts and observing by how much they vary from one sample of 20 counts to another sample of 20 counts.

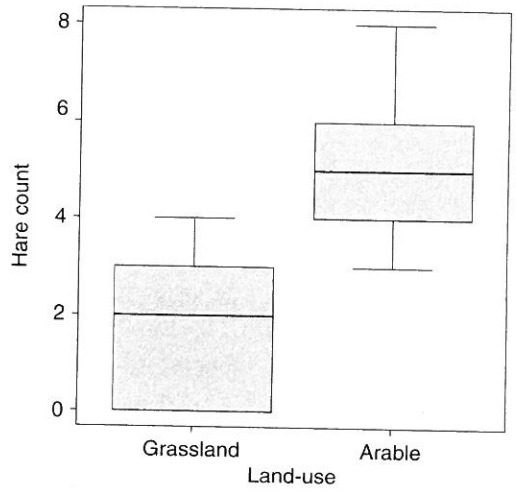

FIGURE 13.2 Relationship between hare count and land-use.

13.4 ANALYSIS USING R

We fit the "Poisson t-test" using the R function `glm(..., family=poisson)`. To test whether mean density in grassland differs from that in arable areas, we use the t-test provided by the function `summary()` or a likelihood ratio test from `anova()`. There is no big difference here in terms of the inferences.

```
poisson.t.test <- glm(C ~ x, family = poisson)   # Fit the model
summary(poisson.t.test)                          # t-Test
anova(poisson.t.test, test = "Chisq")            # Likelihood ratio test (LRT)
```

13.5 ANALYSIS USING WinBUGS

Let's now fit the "Poisson t-test" in WinBUGS. To do this, we will take the code from the normal t-test (Chapter 7) and adapt it to the Poisson GLM case. In addition, we will do two more things in the WinBUGS program below:

1. compute Pearson residuals to assess model fit and
2. conduct a posterior predictive check including a Bayesian p-value as we did for normal linear regression in Chapter 8 (and will do for a "generalized" Poisson regression in Chapter 21).

```
# Define model
sink("Poisson.t.test.txt")
cat("
model {

# Priors
 alpha ~ dnorm(0,0.001)
 beta ~ dnorm(0,0.001)

# Likelihood
 for (i in 1:n) {
    C[i] ~ dpois(lambda[i])
    log(lambda[i]) <- alpha + beta *x[i]

# Fit assessments
    Presi[i] <- (C[i] - lambda[i]) / sqrt(lambda[i])          # Pearson residuals
    C.new[i] ~ dpois(lambda[i])                               # Replicate data set
    Presi.new[i] <- (C.new[i] - lambda[i]) / sqrt(lambda[i])      # Pearson resi
    D[i] <- pow(Presi[i], 2)
    D.new[i] <- pow(Presi.new[i], 2)
}

# Add up discrepancy measures
fit <- sum(D[])
fit.new <- sum(D.new[])
}
",fill=TRUE)
sink()

# Bundle data
win.data <- list(C = C, x = as.numeric(x)-1, n = length(x))

# Inits function
inits <- function(){ list(alpha=rlnorm(1), beta=rlnorm(1))}

# Parameters to estimate
params <- c("lambda","alpha", "beta", "Presi", "fit", "fit.new")

# MCMC settings
nc <- 3
ni <- 3000
nb <- 1000
nt <- 2

# Start Gibbs sampler
out <- bugs(data=win.data, inits=inits, parameters.to.save=params,
model.file="Poisson.t.test.txt", n.thin=nt, n.chains=nc, n.burnin=nb, n.iter=ni,
debug = TRUE)
```


13.5.1 Check of Markov Chain Monte Carlo Convergence and Model Adequacy

The first two things to do before even looking at the estimates of a model fitted using MCMC should really be to check (1) that the Markov chains have converged and (2) that the fitted model is adequate for the data set. We do both here in an exemplary manner.

Convergence—Again, we can assess convergence by graphical means (typically directly within WinBUGS) or using a numerical summary, the Brooks–Gelman–Rubin statistic, which R2WinBUGS calls Rhat. Rhat is about 1 at convergence, with 1.1 often taken an acceptable threshold. We will look at the Rhat values first.

```
print(out, dig = 3)
```

If we briefly look at the second to last column in the table, we see that the chains for all parameters seem to have converged admirably. For larger models with many more parameters, a summary of this summary table may be useful. For instance, we may ask which (if any) parameters have a value of Rhat greater than 1.1. Or we can draw a histogram of the Rhat values.

```
which(out$summary[,8] > 1.1)     # which value in the 8th column is > 1.1 ?
> which(out$summary[,8] > 1.1)
named integer(0)                 # So here we have none

hist(out$summary[,8], col = "grey", main = "Rhat values")
```

So, as expected in this simple model fitted to a "good" data set, there is no problem with convergence.

Residuals and posterior predictive check—We do the analogous to what we did in the normal linear regression example in Chapter 8. That is, we plot the residuals first and then plot the two fit statistics (for the actual data set and for the perfect, new, data sets) against each other and compute the Bayesian p-value as a numerical summary of overall lack of fit. The fit statistic for the new data sets represents, in a way, the reference distribution for the chosen test statistic, here, the sum of squared Pearson residuals.

For GLMs other than the normal linear model, the variability of the response depends on the mean response. To get residuals with approximately constant variance, Pearson residuals are often computed. They are obtained by dividing the raw residuals $(y_i - \bar{y})$ by the standard deviation of y; see also WinBUGS code above.

```
plot(out$mean$Presi, las = 1)
abline(h = 0)
```

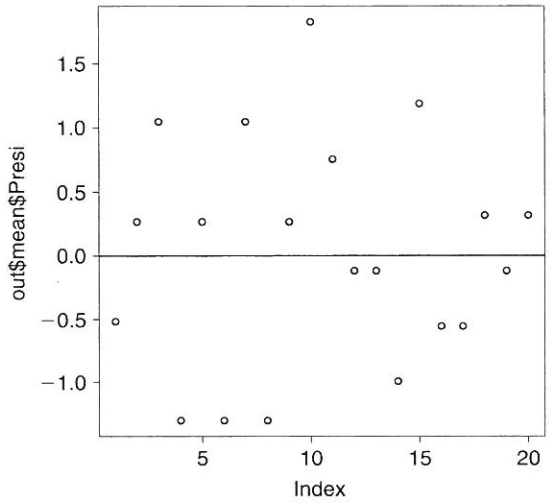

FIGURE 13.3 Pearson residuals for the hare counts.

There is no obvious sign of lack of fit for any particular data point (Fig. 13.3). Next, we conduct a posterior predictive check (Fig. 13.4):

```
plot(out$sims.list$fit, out$sims.list$fit.new, main = "Posterior predictive check
\nfor sum of squared Pearson residuals", xlab = "Discrepancy measure for actual data set",
ylab = "Discrepancy measure for perfect data sets")
abline(0,1, lwd = 2, col = "black")
```

Of course, this looks perfect and computation of the Bayesian p-value (below) confirms this impression. Here, we compute the Bayesian p-value outside WinBUGS in R. This is easier, but of course, we need to have saved the Markov chains for both `fit` and `fit.new`.

```
mean(out$sims.list$fit.new > out$sims.list$fit)
> mean(out$sims.list$fit.new > out$sims.list$fit)
[1] 0.624
```

13.5.2 Inference Under the Model

Now that we are convinced that the model is adequate for these data, we inspect the estimates and compare them with what we put into the data set, as well as what the frequentist analysis in R tells us.

```
print(out, dig = 3)
```

A comparison of the Bayesian solution with the input values that were used for generating the data set ($\alpha = 0.69, \beta = 0.92$) and the solution given by `glm()` shows a reasonably decent consistency (in view of the small sample size).

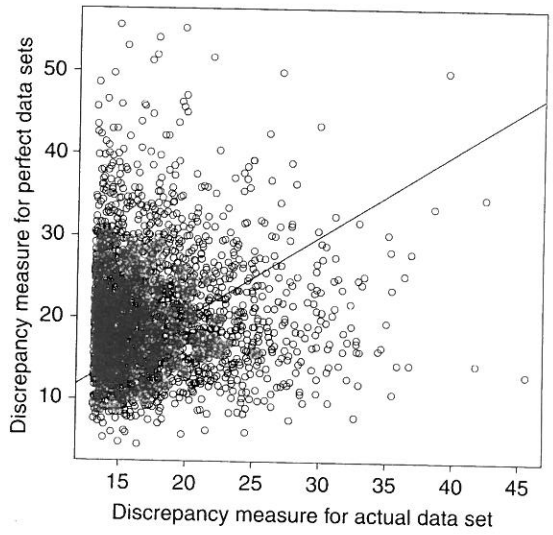

FIGURE 13.4 Graphical posterior predictive check based on the sum of squared Pearson residuals. The Bayesian p-value (0.62) is the proportion of points above the line. The hard boundary on the left is because of the fact that the discrepancy measure cannot be smaller than that corresponding to the maximum likelihood estimate.

```
summary(poisson.t.test)
```

So is there a difference in hare density according to land-use? Let's look at the posterior distribution of the coefficient for arable (Fig. 13.5).

```
hist(out$sims.list$beta, col = "grey", las = 1, xlab = "Coefficient for arable",
main = "")
```

The posterior distribution does not overlap zero, so arable sites really do appear to have a different hare density than grassland sites. The same conclusion is arrived at when looking at the 95% credible interval of β in the summary of the analysis mentioned earlier: (0.64–1.72).

Finally, we will form predictions for presentation. Predictions are the expected values of the response under certain conditions, such as for particular covariate values. We have seen earlier that predictions are a valuable means for synthesizing the information that a model extracts from a data set. In a Bayesian analysis, forming predictions is easy. Predictions are just another form of unobservables, such as parameters or missing values. Therefore, we can base inference about predictions on their posterior distributions.

To summarize what we have learned about the differences in hare densities in grassland and arable study areas, we plot the posterior distributions of the expected hare counts (λ) for both habitat types (Fig. 13.6). We obtain the expected hare counts by exponentiating parameter α and $\alpha + \beta$, respectively.

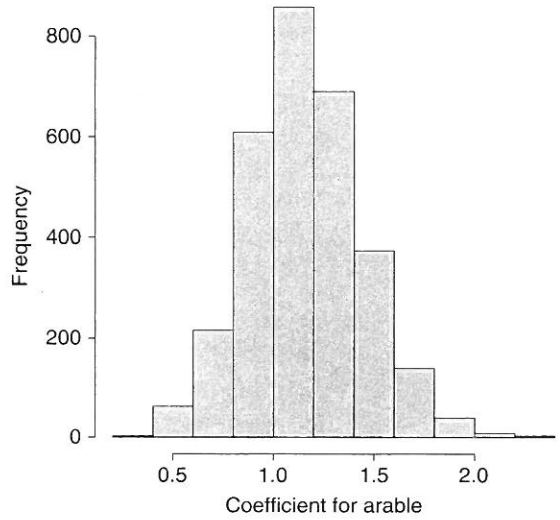

FIGURE 13.5 Posterior distribution of the coefficient of arable.

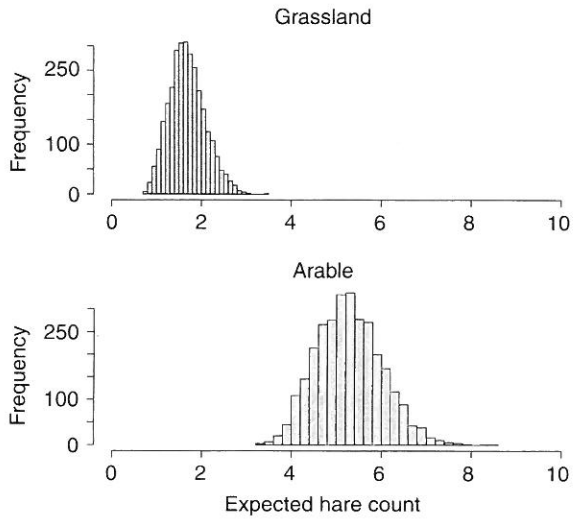

FIGURE 13.6 Posterior distribution of mean hare density in grassland (top) and arable areas (bottom).

```
par(mfrow = c(2,1))
hist(exp(out$sims.list$alpha), main = "Grassland study areas", col = "grey", xlab =
"", xlim = c(0,10), breaks = 20)
hist(exp(out$sims.list$alpha + out$sims.list$beta), main = "Arable study areas",
col = "grey", xlab = "Expected hare count", xlim = c(0,10), breaks = 20)
```


13.6 SUMMARY

We have introduced the generalized linear model, or GLM, where effects of covariates are linear in the transformed expectation of a response, which may come from a distribution other than the normal. The GLM is another key concept that appears over and over again in modern applied statistics in empirical sciences such as ecology. Therefore, we will deepen our understanding of this essential model class in subsequent chapters. Furthermore, we will combine the GLM and the mixed model to arrive at the most complex model considered in this book, the generalized linear mixed model, in Chapters 16 and 19–21.

EXERCISES

1. *Predictions*: Within the WinBUGS code, add a line that directly computes the mean hare density in arable areas.
2. *Derived quantities*: Summarize the posterior distribution for the difference in mean hare density in grassland and arable areas.
3. *Zeroes (migrating raptors)*: This fine example is borrowed from Bernardo (2003). It beautifully illustrates the power of Bayesian inference based on the posterior distribution of the unobservables (parameters, etc.). Ornithologists frequently count migrating birds of prey at places where they concentrate in spring or autumn, e.g., along coasts, mountain ridges, or on isthmuses. Assume that at a certain place, one rare raptor species had not been seen during 10 consecutive days. What is the probability that we see at least one on day 11? What is the probability that we see two or more? In your solution, make explicit your reasoning for using the particular statistical model you choose and discuss a few of its assumptions that may not hold in reality (e.g., serial independence, constancy of rates).
4. *Zeroes (contrast estimate)*: Assume that no hare was ever observed in grassland areas, i.e., that the counts in all 10 grassland areas were zero. Try to fit the Poisson t-test using R and using WinBUGS.
5. *Swiss hare data*: Compare mean counts (not density) in arable and in grassland areas in one selected year (e.g., 2000; take the smallest counts when there is more than one per year). When taking these counts as observations from an identical Poisson distribution, is there anything that strikes you as inadequate?