CHAPTER 6
Generalized Linear Models

In previous chapters, we have seen how to model a binomial or Poisson response. Multinomial response models can often be recast as Poisson responses and the standard linear model with a normal (Gaussian) response is already familiar. Although these models each have their distinctive characteristics, we observe some common features in all of them that we can abstract to form the *generalized linear model* (GLM). By developing a theory and constructing general methods for GLMs, we can are able to tackle a wider range of data with different types of response variables. GLMs were introduced by Nelder and Wedderburn (1972) while McCullagh and Nelder (1989) provides a book-length treatment.

6.1 GLM Definition

A GLM is defined by specifying two components. The response should be a member of the exponential family distribution and the link function describes how the mean of the response and a linear combination of the predictors are related.

Exponential family: In a GLM the distribution of Y is from the exponential family of distributions which take the general form:

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

The θ is called the *canonical parameter* and represents the location while ϕ is called the *dispersion parameter* and represents the scale. We may define various members of the family by specifying the functions *a, b,* and *c*. The most commonly used examples are:

1. Normal or Gaussian:

$$f(y|\theta,\phi) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$
$$= \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right]$$

So we can write $\theta = \mu$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \theta^2/2$ and $c(y,\phi) = -(y^2/\phi + \log(2\pi\phi))/2$.

2. Poisson:

$$f(y|\theta,\phi) = e^{-\mu}\mu^y/y!$$
$$= \exp(y\log\mu - \mu - \log y!)$$

So we can write $\theta = \log(\mu), \phi \equiv 1, a(\phi) = 1, b(\theta) = \exp(\theta)$ and $c(y,\phi) = -\log y!$.

3. Binomial:

$$f(y|\theta,\phi) = \binom{n}{y}\mu^y(1-\mu)^{n-y}$$

$$= \exp(y\log\mu + (n-y)\log(1-\mu) + \log\binom{n}{y}$$

$$= \exp(y\log\frac{\mu}{1-\mu} + n\log(1-\mu) + \log\binom{n}{y}$$

So we see that $\theta = \log\frac{\mu}{1-\mu}$, $b(\theta) = -n\log(1-\mu) = n\log(1+\exp\theta)$ and $c(y,\phi) = \log\binom{n}{y}$.

The gamma and inverse Gaussian are other lesser-used members of the exponential family that are covered in Chapter 7. Notice that in the normal density, the ϕ parameter is free (as it is also for the gamma density) while for the Poisson and binomial it is fixed at one. This is because the Poisson and binomial are one parameter families while the normal and gamma have two parameters. In fact, some authors reserve the term *exponential family* distribution for cases where ϕ is not used, while using the term *exponential dispersion family* for cases where it is. This has important consequences for the analysis.

Some other densities, such as the negative binomial and the Weibull distribution, are not members of the exponential family, but they are sufficiently close that the GLM can be fit with some modifications. It is also possible to fit distributions that are not in the exponential family using the GLM-style approach, but there are some additional complications.

The exponential family distributions have mean and variance:

$$EY = \mu = b'(\theta)$$
$$\text{var } Y = b''(\theta)a(\phi)$$

The mean is a function of θ only while the variance is a product of functions of the location and the scale. $b''(\theta)$ is called the *variance function* and describes how the variance relates to the mean.

In the Gaussian case, $b''(\theta)=1$ and so the variance is independent of the mean. For other distributions, this is not true, making the Gaussian case exceptional. We can introduce weights by setting:

$$a(\phi) = \phi/w$$

where w is a known weight that varies between observations.

Link function: Let us suppose we may express the effect of the predictors on the response through a *linear predictor*:

$$\eta=\beta_0+\beta_1 x_1+\ldots+\beta_p x_p=x^T\beta$$

The link function, g, describes how the mean response, $EY=\mu$, is linked to the covariates through the linear predictor:

$$\eta=g(\mu)$$

In principle, any monotone continuous and differentiable function will do, but there are some convenient and common choices for the standard GLMs.

In the Gaussian linear model, the identity link, $\eta=\mu$ is the obvious selection, but another choice would give $y=g^{-1}(x^T\beta)+\varepsilon$. This does not correspond directly to a transform on the response: $g(y)=x^T\beta+\varepsilon$ as, for example, in a Box-Cox type transformation. In a GLM, the link function is assumed known whereas in a *single index model,* g is estimated.

For the Poisson GLM, the mean μ must be positive so $\eta=\mu$ will not work conveniently since η can be negative. The standard choice is $\mu=e^{\eta}$ so that $\eta=\log\mu$ which ensures $\mu>0$. This log link means that additive effects of x lead to multiplicative effects on μ.

For the binomial GLM, let p be the probability of success and let this be our μ if we define the response as the proportion rather than the count. This requires that $0\leq p\leq 1$. There are several commonly used ways to ensure this: the logistic, probit and complementary log-log links. These are discussed in detail in Chapter 2.

The *canonical link* has g such that $\eta=g(\mu)=\theta$, the canonical parameter of the exponential family distribution. This means that $g(b'(\theta))=\theta$. The canonical links for the common GLMs are shown in Table 6.1. If a canonical link is used, X^TY is

Family	Link	Variance Function
Normal	$\eta=\mu$	1
Poisson	$\eta=\log\mu$	μ
Binomial	$\eta=\log(\mu/(1-\mu))$	$\mu(1-\mu)$
Gamma	$\eta=\mu^{-1}$	μ^2
Inverse Gaussian	$\eta=\mu^{-2}$	μ^3

Table 6.1 *Canonical links for GLMs.*

sufficient for β. The canonical link is mathematically and computationally convenient and is often the natural choice of link. However, one is not required to use the canonical link and sometimes context may compel another choice.

6.2 Fitting a GLM

The parameters, β, of a GLM can be estimated using maximum likelihood. The log-likelihood for single observation, where $a_i(\phi) = \phi/w_i$, is:

$$\log L(\theta_i, \phi; y_i) = w_i[\frac{y_i\theta_i - b(\theta_i)}{\phi}] + c(y_i, \phi)$$

So for independent observations, the log-likelihood will be $\sum_i \log L(\theta_i, \phi; y_i)$. Sometimes we can maximize this analytically and find an exact solution for the MLE $\hat{\beta}$, but the Gaussian GLM is the only common case where this is possible. Typically, we must use numerical optimization. By applying the Newton-Raphson method with Fisher scoring, McCullagh and Nelder (1989) show that the optimization is equivalent to iteratively reweighted least squares (IRWLS).

The procedure can be understood intuitively by analogy to the procedure for the Gaussian linear model $Y=X\beta+\varepsilon$. Suppose $\text{var } Y \propto f(\hat{\eta})$ where $\hat{y} = \hat{\eta} = X\hat{\beta}$. We would use weights w_i where $w_i^{-1} = f(\hat{\eta})$. Since the weights are a function of $\hat{\beta}$ an iterative fitting procedure would be needed. We might set the weights all equal to one, estimate $\hat{\beta}$, use this to recompute the weights, reestimate $\hat{\beta}$, and so on until convergence.

We can use a similar idea to fit a GLM. Roughly speaking, we want to regress *g(y)* on *X* with weights inversely proportional to var *g(y)*. However, *g(y)* might not make sense in some cases—for example, in the binomial GLM. So we linearize *g(y)* as follows: Let $\eta=g(\mu)$ and $\mu=EY$. Now do a one-step expansion:

$$\begin{aligned} g(y) &\approx g(\mu) + (y-\mu)g'(\mu) \\ &= \eta + (y-\mu)\frac{d\eta}{d\mu} \\ &\equiv z \end{aligned}$$

and

$$\widehat{\text{var}}\, z = \left(\frac{d\eta}{d\mu}\right)^2 V(\hat{\mu}) = \frac{1}{w}$$

So the IRWLS procedure would be:

1. Set initial estimates $\hat{\eta}_0$ and $\hat{\mu}_0$.
2. Form the "adjusted dependent variable" $z_0 = \hat{\eta}_0 + (y-\hat{\mu}_0)\frac{d\eta}{d\mu}|_{\hat{\eta}_0}$.
3. Form the weights $w_0^{-1} = \left(\frac{d\eta}{d\mu}\right)^2 |_{\hat{\eta}_0} V(\hat{\mu}_0)$.

4. Reestimate β to get $\hat{\eta}_1$.
5. Iterate steps 2–3–4 until convergence.

Notice that the fitting procedure uses only $\eta=g(\mu)$ and $V(\mu)$, but requires no further knowledge of the distribution of y. This point will be important later in Section 7.4. Estimates of variance may be obtained from:

$$\hat{\text{var}}(\hat{\beta}) = (X^T W X)^{-1}\hat{\phi}$$

which is comparable to the form used in weighted least squares with the exception that the weights are now a function of the response for a GLM.

Let's implement the procedure explicitly to understand how the fitting algorithm works. We use the Bliss data from Section 2.7 to illustrate this. Here is the fit we are trying to match:

```
> data(bliss)
> modl <- glm(cbind(dead,alive) ~ conc,
family=binomial, bliss)
> summary(modl) $coef
            Estimate Std. Error z value   Pr(>|z|)
(Intercept)  -2.3238    0.41789 -5.5608 2.6854e-08
conc          1.1619    0.18142  6.4046 1.5077e-10
```

For a binomial response, we have:

$$\eta = \log\frac{\mu}{1-\mu} \qquad \frac{d\eta}{d\mu} = \frac{1}{\mu(1-\mu)} \qquad V(\mu) = \mu(1-\mu)/n \qquad w = n\mu(1-\mu)$$

where the variance is computed with the understanding that y is the proportion not the count. We use y for our initial guess for $\hat{\mu}$ which works here because none of the observed proportions are zero or one:

```
> y <- bliss$dead/30; mu <- y
> eta <- logit(mu)
> z <- eta + (y-mu)/(mu*(1-mu))
> w <- 30*mu*(1-mu)
> lmod <- lm(z ~ conc, weights=w, bliss)
> coef(lmod)
(Intercept)        conc
    -2.3025      1.1536
```

It is interesting how close these initial estimates are to the converged values given above. This is not uncommon. Even so, to get a more precise result, iteration is necessary. We do five iterations here:

```
> for(i in 1:5){
+ eta <- lmod$fit
+ mu <- ilogit(eta)
+ z <- eta + (y-mu)/(mu*(1-mu))
```

```
+ w <- 30*mu*(1-mu)
+ lmod <- lm(z ~ bliss$conc, weights=w)
+ cat(i,coef(lmod),"\n")
+ }
1 -2.3237 1.1618
2 -2.3238 1.1619
3 -2.3238 1.1619
4 -2.3238 1.1619
5 -2.3238 1.1619
```

We can see that convergence is fast in this case. The *Fisher scoring iterations* referred to in the output record the number of iterations. In most cases, the convergence is rapid. If there is a failure to converge, this is often a sign of some problem with the model specification or unusual feature of the data. An example of such a problem with the estimation may be seen in Section 2.8. A look at the final (weighted) linear model reveals that:

```
> summary(lmod)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.3238     0.1462   -15.9  0.00054
conc          1.1619     0.0635    18.3  0.00036
Residual standard error: 0.35 on 3 degrees of freedom
```

The standard errors are not correct and can be computed (rather inefficiently) as follows:

```
> xm <- model.matrix(lmod)
> wm <- diag(w)
> sqrt(diag(solve(t(xm) %*% wm %*% xm)))
[1] 0.41787 0.18141
```

Now $\hat{\text{var}}(\hat{\beta}) = (X^T W X)^{-1}$ because $\phi = 1$ for the binomial model but in the Gaussian linear model $\hat{\text{var}}(\hat{\beta}) = (X^T W X)^{-1}\hat{\sigma}^2$. To get the correct standard errors from the lm fit, we need to scale out the $\hat{\sigma}$ as follows:

```
> summary(lmod)$coef[,2]/summary(Imod)$sigma
(Intercept)        conc
    0.41789     0.18142
```

These calculations are shown for illustration purposes only and are done more efficiently and reliably by the glm function.

6.3 Hypothesis Tests

When considering the choice of model for some data, we should define the range of possibilities. The *null* model is the smallest model we will entertain while the *full* or *saturated* model is the most complex.

The null model represents the situation where there is no relation between the predictors and the response. Usually this means we fit a common mean μ for all y, that is, one parameter only. For the Gaussian GLM, this is the model $y=\mu+\varepsilon$. For some contingency table models, there will be additional parameters that represent row or column totals or other such constraints. In these cases, the null model will have more than one parameter.

In the saturated model, the data is explained exactly. Typically, we need to use n parameters for n data points. This can often be achieved by fitting a sufficiently high-order polynomial or by treating the numerical values of quantitative predictors as codes, thereby changing them into qualitative predictors. If enough interactions are included, the model will be saturated. This model tells us no more than the data itself and is usually uninformative.

A statistical model describes how we partition the data into systematic structure and random variation. The null model represents one extreme where the data is represented entirely as random variation, while the saturated or full model represents the data as being entirely systematic.

The full model does give us a measure of how well any model could possibly fit and so we might consider the difference between the log-likelihood for the full model, $l(y,\phi|y)$, and that for the model under consideration, $l(\hat{\mu},\phi|y)$, expressed as a likelihood ratio statistic:

$$2(l(y,\phi|y)-l(\hat{\mu},\phi|y))$$

Provided that the observations are independent and for an exponential family distribution, when $a_i(\phi)=\phi/w_i$, this simplifies to:

$$\sum_i 2w_i(y_i(\tilde{\theta}_i-\hat{\theta}_i)-b(\tilde{\theta}_i)+b(\hat{\theta}_i))/\phi$$

where $\tilde{\theta}$ are the estimates under the full (saturated) model and $\hat{\theta}$ are the estimates under the model of interest. The above can be written simply as $D(y,\hat{\mu})/\phi$ where $D(y,\hat{\mu})$ is called the deviance and $D(y,\hat{\mu})/\phi$ is called the scaled deviance. Deviances for the common GLMs are shown in Table 6.2.

GLM	Deviance
Gaussian	$\sum_i(y_i-\hat{\mu}_i)^2$
Poisson	$2\sum_i[y_i\log(y_i/\hat{\mu}_i)-(y_i-\hat{\mu}_i)]$
Binomial	$2\sum_i[y_i\log(y_i/\hat{\mu}_i)+(m-y_i)\log((m-y_i)/(m-\hat{\mu}_i))]$
Gamma	$2\sum_i[-\log(y/\hat{\mu})+(y-\hat{\mu})/\hat{\mu}]$

Inverse Gaussian	$\sum_i (y-\hat{\mu})^2/(\hat{\mu}^2 y)$

Table 6.2 *For the binomial yi ~ B(m, p_i) and μi = mpi that is μ is the count and not proportion in this formula. For the Poisson, the deviance is known as the G-statistic. The second term* $\sum_i (y_i - \hat{\mu}_i)$ *is usually zero if an intercept term is used in the model.*

Pearson's X^2 statistic:

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

where $V(\hat{\mu}) = \text{var}\,(\hat{\mu})$ is an alternative measure of discrepancy that is sometimes used in place of the deviance.

There are two main types of hypothesis test we shall employ. The goodness of fit test simply asks whether the current model fits the data. The other type of test compares two nested models where the smaller model represents a linear restriction on the parameters of the larger model. The goodness of fit test can be viewed as model comparison test if we identify the smaller model with the model of interest and the larger model with the full or saturated model.

For the goodness of fit test, we use the fact that, under certain conditions, provided the model is correct, the scaled Deviance and the Pearson's X^2 statistic are both asymptotically χ^2 with degrees of freedom equal to the number of identifiable parameters. For GLMs such as the Gaussian, we usually do not know the value of the dispersion parameter, ϕ, and so this test cannot be used. For the binomial and the Poisson, $\phi = 1$, and so the test is practical. However, the accuracy of the asymptotic approximation is dubious for smaller datasets. For a binary, that is a 0-1 response, the approximation is worthless.

For comparing a larger model, Ω, to a smaller nested model, ω the difference in the scaled deviances, $D_\omega - D_\Omega$ is asymptotically χ^2 with degrees of freedom equal to the difference in the number of identifiable parameters in the two models. For the Gaussian model and other models where the dispersion ϕ is usually not known, this test cannot be directly used. However, if we insert an estimate of ϕ we may compute an *F*-statistic of the form:

$$\frac{(D_\omega - D_\Omega)/(df_\omega - df_\Omega)}{\hat{\phi}}$$

where $\hat{\phi} = X^2/(n-p)$ is a good estimate of the dispersion. For the Gaussian model, $\hat{\phi} = RSS_\Omega/df_\Omega$, and the resulting *F*-statistic has an exact *F* distribution for the null. For other GLMs with free dispersion parameters, the statistic is only approximately *F* distributed.

For every GLM except the Gaussian, an approximate null distribution must be used whose accuracy may be in doubt particularly for smaller samples. However, the approximation is better when comparing models than for the goodness of fit statistic.

Let's consider the possible tests on the Bliss insect data:

```
> summary(modl)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.324      0.418   -5.56  2.7e-08
conc          1.162      0.181    6.40  1.5e-10
(Dispersion parameter for binomial family taken to be
1)
    Null deviance: 64.76327 on 4 degrees of freedom
Residual deviance:  0.37875 on 3 degrees of freedom
```

We are able to make a goodness of fit test by examining the size of the residual deviance compared to its degrees of freedom:

```
> 1-pchisq(deviance(modi),df.residual(modi))
[1] 0.9446
```

where we see the *p*-value is large indicating no evidence of a lack of fit. As with lack of fit tests for Gaussian linear models, this outcome does not mean that this model is correct or that no better models exist. We can also quickly see that the null model would be inadequate for the data since the null deviance of 64.7 is very large for four degrees of freedom.

We can also test for the significance of the linear concentration term by comparing the current model to the null model:

```
> anova(modl,test="Chi")
Analysis of Deviance Table
Model: binomial, link: logit
Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     4        64.8
conc  1     64.4         3         0.4    1e-15
```

We see that the concentration term is clearly significant. We can also fit and test a more complex model:

```
> mod12 <- glm(cbind(dead, alive) ~ conc+I(conc^2),
family=binomial,bliss)
> anova(mod1,mod12,test="Chi")
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         3      0.379
2         2      0.195  1    0.183     0.669
```

We can see that there is no need for a quadratic term in the model. The same information could be extracted with:

```
> anova(mod12,test="Chi")
```

We may also take a Wald test approach. We may use the standard error of the parameter estimates to construct a *z*-statistic of the form $\hat{\beta}/se(\hat{\beta})$. This has an asymptotically normal null distribution. For the Bliss data, for the concentration term, we have z=1.162/0.181=6.40. Thus the (approximate) *p*-value for the Wald test of the concentration parameter being equal to zero is $1.5e^{-10}$ and thus we clearly reject the null here. Remember that this is again only an approximate test except in the special case of the Gaussian GLM where the *z*-statistic is the *t*-statistic and has an exact *t*-distribution. The difference of deviances test is preferred to the Wald test due, in part, to the problem noted by Hauck and Donner (1977).

6.4 GLM Diagnostics

As with standard linear models, it is important to check the adequacy of the assumptions that support the GLM. The diagnostic methods for GLMs mirror those used for Gaussian linear models. However, some adaptations are necessary and, depending on the type of GLM, not all diagnostic methods will be applicable.

Residuals: Residuals represent the difference between the data and the model and are essential to explore the adequacy of the model. In the Gaussian case, the residuals are $\hat{\varepsilon} = y - \hat{\mu}$. These are called response residuals for GLMs, but since the variance of the response is not constant for most GLMs, some modification is necessary. We would like residuals for GLMs to be defined such that they can be used in a similar way as in the Gaussian linear model.

The *Pearson residual* is comparable to the standardized residuals used for linear models and is defined as:

$$r_P = \frac{y - \hat{\mu}}{\sqrt{V(\hat{\mu})}}$$

where V(μ)=b''(θ). These are just a rescaling of $y - \hat{\mu}$. Notice that $\sum r_P^2 = X^2$ and hence the name. Pearson residuals can be skewed for nonnormal responses.

The *deviance residuals* are defined by analogy to Pearson residuals. The Pearson residual was r_P such that $\sum r_P^2 = X^2$, so we set the deviance residual as r_D such that $\sum r_D^2 = \text{Deviance} = \sum d_i$. Thus:

$$r_D = sign(y - \hat{\mu})\sqrt{d_i}$$

For example, in the Poisson:

$$r_D = sign(y - \hat{\mu})[2(y \log y/\hat{\mu} - y + \hat{\mu})]^{1/2}$$

Let's examine the types of residuals available to us using the Bliss data. We can obtain the deviance residuals as:

```
> residuals(modl)
[1] -0.451015 0.359696 0.000000 0.064302 -0.204493
```

These are the default choice of residuals. The Pearson residuals are:

```
> residuals(modl,"pearson")
        1         2         3         4         5
-0.432523 0.364373 0.000000 0.064147 -0.208107
```

which are just slightly different from the deviance residuals. The response residuals are:

```
> residuals(modl,"response")
          1         2         3         4          5
-0.0225051 0.0283435 0.0000000 0.0049898 -0.0108282
```

which is just the response minus the fitted value:

```
> bliss$dead/30 - fitted(modl)
          1         2         3         4          5
-0.0225051 0.0283435 0.0000000 0.0049898 -0.0108282
```

Finally, the so-called working residuals are:

```
> residuals(modl,"working")
        1         2         3         4         5
-0.277088 0.156141 0.000000 0.027488 -0.133320
> modl$residuals
        1         2         3         4         5
-0.277088 0.156141 0.000000 0.027488 -0.133320
```

Note that it is important to use the residuals () function to get the deviance residuals which are most likely what is needed for diagnostic purposes. Using $residuals gives the working residuals which is not usually needed for diagnostics. We can now identify the working residuals as a by-product of the IRWLS fitting procedure:

```
> residuals(lmod)
          1           2           3           4
 5
-2.7709e-01 1.5614e-01 -3.8463e-16 2.7488e-02 -1.3332e-
01
```

Leverage and influence: For a linear model, $\hat{y} = Hy$, where H is the *hat matrix* that projects the data onto the fitted values. The leverages *hi* are given by the diagonal of H and represent the potential of the point to influence the fit. They are solely a function of X

and whether they are in fact influential will also depend on y. Leverages are somewhat different for GLMs. The IRWLS algorithm used to fit the GLM uses weights, *w*. These weights are just part of the IRWLS algorithm and are not user assigned. However, these do affect the leverage. We form a matrix W=diag(w) and the hat matrix is:

$H=W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$

We extract the diagonal elements of H to get the leverages *hi*. A large value of *hi* indicates that the fit may be sensitive to the response at case *i*. Large leverages typically mean that the predictor values are unusual in some way. One important difference from the linear model case is that the leverages are no longer just a function of X and now depend on the response through the weights W. The leverages may be calculated as:

```
> influence(mod1)$hat
      1       2       3       4       5
0.42550 0.41331 0.32238 0.41331 0.42550
```

As in the linear model case, we might choose to studentize the residuals as follows:

$$r_{SD} = \frac{r_D}{\sqrt{\hat{\phi}(1-h_i)}}$$

or compute jacknife residuals representing the difference between the observed response for case *i* and that predicted from the data with case *i* excluded, scaled appropriately. These are expensive to compute exactly and so an approximation due to

Williams (1987) can be used:

$$sign(y-\hat{\mu})sqrt(1-h_i)r_{SD}^2+h_ir_{SP}^2$$

where $r_{SP}=r_P/\sqrt{1-h_i}$. These may be computed as:

```
> rstudent(modl)
        1         2         3         4         5
-0.584786 0.472135 0.000000 0.083866 -0.271835
```

Outliers may be detected by observing particularly large jacknife residuals.

Leverage only measures the potential to affect the fit whereas measures of influence more directly assess the effect of each case on the fit. We can examine the change in the fit from omitting a case by looking at the changes in the coefficients:

```
> influence(mod1)$coef
  (Intercept)        conc
1  -0.2140015  0.0806635
2   0.1556719 -0.0470873
3   0.0000000  0.0000000
4  -0.0058417  0.0084177
5   0.0492639 -0.0365734
```

Alternatively, we can examine the Cook statistics:

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T (X^T W X)(\hat{\beta}_{(i)} - \hat{\beta})}{p\hat{\phi}}$$

which may be calculated as:

```
> cooks.distance(modi)
        1         2         3         4         5
0.1205927 0.0797100 0.0000000 0.0024704 0.0279174
```

We can see that the biggest change would occur by omitting the first observation. However, since this is a very small dataset with just five observations, we would not contemplate dropping cases. In any event, we see that the change in the coefficients would not qualitatively change the conclusion.

Model diagnostics: We may divide diagnostic methods into two types. Some methods are designed to detect single cases or small groups of cases that do not fit the pattern of the rest of the data. Outlier detection is an example of this. Other methods are designed to check the assumptions of the model. These methods can be subdivided into those that check the structural form of the model, such as the choice and transformation of the predictors, and those that check the stochastic part of the model, such as the nature of the variance about the mean response. Here, we focus on methods for checking the assumptions of the model.

For linear models, the plot of residuals against fitted values is probably the single most valuable graphic. For GLMs, we must decide on the appropriate scale for the fitted values. Usually, it is better to plot the linear predictors $\hat{\eta}$ rather than the predicted responses $\hat{\mu}$. We revisit the model for Galápagos data first presented in Section 3.1. Consider first a plot using $\hat{\mu}$ presented in the first panel of Figure 6.1:

```
> data(gala)
> gala <- gala[,-2]
> modp <- glm(Species ~ .,family=poisson,gala)
> plot(residuals(modp) ~ predict
(modp,type="response"),
xlab=expression(hat(mu)),ylab="Deviance residuals")
```

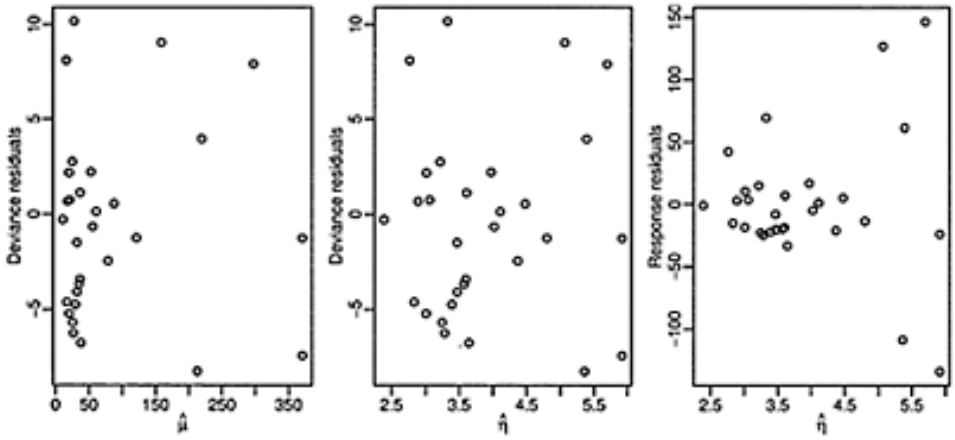

Figure 6.1 *Residual vs. fitted plots for the Galápagos model. The first uses fitted values in the scale of the response while the second uses fitted values in the scale of the linear predictor. The third plot uses response residuals while the first two use deviance residuals.*

There are just a few islands with a large predicted number of species while most predicted response values are small. This makes it difficult to see the relationship between the residuals and the fitted values because most of the points are compressed on the left of the display. Now we try plotting $\hat{\eta}$:

```
> plot(residuals(modp) ~ predict(modp,type="link"),
  xlab=expression(hat(eta)),ylab="Deviance residuals")
```

Now the points, shown in the second panel of Figure 6.1, are more evenly spaced in the horizontal direction. We are looking for two main features in such a plot. Is there any nonlinear relationship between the predicted values and the residuals? If so, this would be an indication of a lack of fit that might be rectified by a change in the model. For a linear model, we might consider a transformation of the response, but this is usually impractical for a GLM since it would change the assumed distribution of the response. We might also consider a change to the link function, but often this is undesirable since there a few choices of link function that lead to easily interpretable models. It is best if a change in the choice of predictors or transformations on these predictors can be made since this involves the least disruption to the GLM. For this particular plot, there is no evidence of nonlinearity.

The variance of the residuals with respect to the fitted values should also be inspected. The assumptions of the GLM would require constant variance in the plot and, in this case, this appears to be the case. A violation of this assumption would prompt a change in the model. We might consider a change in the variance function $V(\mu)$, but this would

involve abandoning the Poisson GLM since this specifies a particular form for the variance function. We would need to use a quasi-likelihood GLM described in Section 7.4. Alternatively, we could employ a different GLM for a count response such as the negative binomial. Finally, we might use weights if we could identify some feature of the data that would suggest a suitable choice.

For all GLMs but the Gaussian, we have a nonconstant variance function. However, by using deviance residuals, we have already scaled out the variance function and so, provided the variance function is correct, we do expect to see constant variance in the plot. If we use response residuals, that is $y - \hat{\mu}$, as seen in the third panel of Figure 6.1:

```
> plot (residuals (modp,type="response") ~ predict
(modp, type="link"),
  xlab=expression(hat(eta)),ylab="Response residuals")
```

We see a pattern of increasing variation consistent with the Poisson.

In some cases, plots of the residuals are not particularly helpful. For a binary response, the residual can only take two possible values for given predicted response. This is the most extreme situation, but similar discreteness can occur for binomial responses with small group sizes and Poisson responses that are small. Plots of residuals in these cases tend to show curved lines of points corresponding to the limited number of observed responses. Such artifacts can obscure the main purpose of the plot. Difficulties arise for binomial data where the covariate classes have very different sizes. Points on plots may represent just a few or a large number of individuals.

Investigating the nature of the relationship between the predictors and the response is another primary objective of diagnostic plots. Even before a model is fit to the data, we might simply plot the response against the predictors. For the Galápagos data, consider a plot of the number of species against the area of the island shown in the first panel of Figure 6.2:

```
➢  plot (Species ~ Area, gala)
```

We see that both variables have skewed distributions. We start with a log transformation on the predictor as seen in the second panel of Figure 6.2:

```
> plot(Species ~ log(Area), gala)
```

We see a curvilinear relationship between the predictor and the response. However, the default Poisson GLM uses a log link which we need to take into account. To allow for the choice of link function, we can plot the linearized response:

$$z = \eta + (y - \mu)\frac{d\eta}{d\mu}$$

as we see in the third panel of Figure 6.2:

```
> mu <- predict (modp, type="response")
> z <- predict (modp)+(gala$Species-mu)/mu
```

```
> plot(z ~ log(Area), gala,ylab="Linearized Response")
```

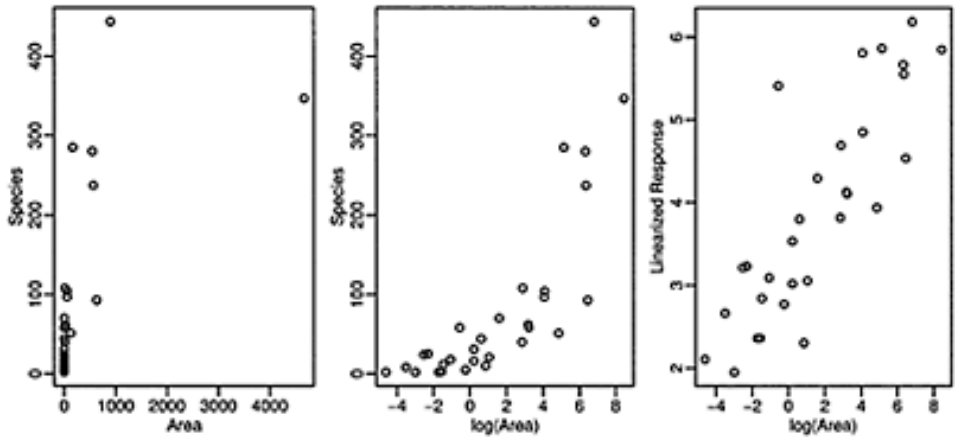

Figure 6.2 *Plots of the number of species against area for the Galápagos data. The first plot clearly shows a need for transformation, the second shows the advantage of using logged area, while the third shows the value of using the linearized response.*

We now see a linear relationship suggesting that no further transformation of area is necessary. Notice that we used the current model in the computation of z. Some might prefer to use an initial guess here to avoid presuming the choice of model. For this dataset, we find that a log transformation of all the predictors is helpful:

```
> modpl <- glm(Species ~ log(Area) + log(Elevation) +
log(Nearest) +
  log(Scruz+0.1) + log(Adjacent), family=poisson, gala)
> c(deviance(modp),deviance(modpl))
[1] 716.85 359.12
```

We see that this results in a substantial reduction in the deviance.

The disadvantage of simply examining the raw relationship between the response and the other predictors is that it fails to take into account the effect of the other predictors. Partial residual plots are used for linear models to make allowance for the effect of the other predictors while focusing on the relationship of interest. These can be adapted for use in GLMs by plotting $z - \hat{\eta} + \beta_j x_j$ versus x_j. The interpretation is the same as in the linear model case. We compute the partial residual plot for the (now logged) area, as shown in the first panel of Figure 6.3:

```
> mu <- predict (modpl,type="response")
> u <- (gala$Species-mu)/mu + coef(modpl)
[2]*log(gala$Area)
> plot(u ~ log(Area), gala,ylab="Partial Residual")
> abline(0,coef(modpl)[2])
```

In this plot, we see no reason for concern. There is no nonlinearity indicating a need to transform nor are there any obvious outliers or influential points. Partial residuals can also be obtained from residuals (., type="partial") although an offset will be necessary if you want the regression line displayed correctly on the plot.

One can search for good transformations of the predictors in nongraphical ways. Polynomials terms or spline functions of the predictors can be experimented with, but generalized additive models, described in Chapter 12, offer a more direct way to discover some good transformations.

The link function is a fundamental assumption of the GLM. Quite often the choice of link function is set by the characteristics of the response, such as positivity, or by ease of interpretation, as with logit link for binomial GLMs. It is often difficult to contemplate alternatives. Nevertheless, it is worth checking to see whether the link assumption is not grossly wrong. Before doing this, it is important to eliminate other simpler violations of the assumptions that are more easily rectified such as outliers

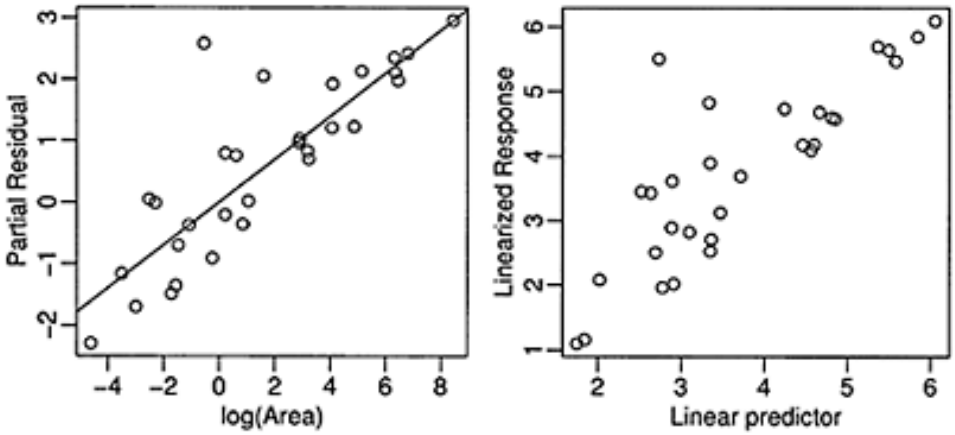

Figure 6.3 *A partial residual plot for log(Area) is shown on the left while a diagnostic for the link function is shown on the right.*

or transformations of the predictors. After these concerns have been eliminated, one can check the link assumption by making a plot of the linearized response z against linear predictor $\hat{\eta}$. An example of this is shown in the second panel of Figure 6.3:

```
> z <- predict(modpl)+(gala$Species-mu)/mu
> plot(z ~ predict(modpl), xlab="Linear predictor",
  ylab="Linearized Response")
```

In this case, we see no indication of a problem.

An alternative approach to checking the link function is to propose a family of link functions of which the current choice is a member. A range of links can then be fit and compared to the current choice. The approach is analogous to the Box-Cox method used for linear models. Alternative choices are easier to explore within the quasi-likelihood framework described in Section 7.4.

Unusual Points

We have already described the raw material of residuals, leverage and influence measures that can be used to check for points that do not fit the model or influence the fit unduly. Let's now see how to use graphical methods to examine these quantities.

The Q-Q plot of the residuals is the standard way to check the normality assumption on the errors typically made for a linear model. For a GLM, we do not expect the residuals to be normally distributed, but we are still interested in detecting outliers. For this purpose, it is better to use a half-normal plot that compares the sorted absolute residuals and the quantiles of the half-normal distribution:

$$\Phi^{-1}\left(\frac{n+i}{2n+1}\right) \quad i=1,\ldots,n$$

The residuals are not expected to be normally distributed, so we are not looking for an approximate straight line. We only seek outliers which may be identified as points off the trend. A half-normal plot is better for this purpose because in a sense the resolution of the plot is doubled by having all the points in one tail.

Since we are more specifically interested in outliers, we should plot the jacknife residuals. An example for the Galápagos model is shown in the first panel of Figure 6.4:

```
> halfnorm(rstudent(modpl))
```

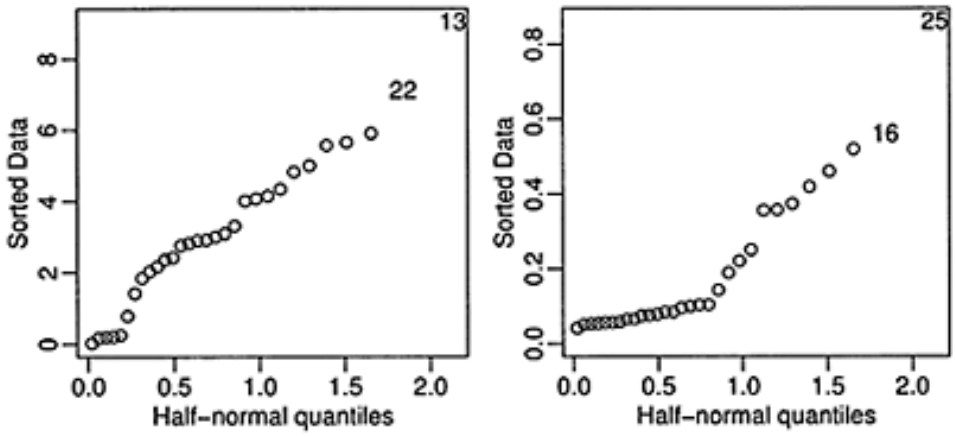

Figure 6.4 *Half-normal plots of the jacknife residuals on the left and the leverages on the right.*

We see no sign of outliers in the plot. The half-normal plot is also useful for positive-valued diagnostics such as the leverages and the Cook statistics. A look at the leverages is shown in the second panel of Figure 6.4:

```
> gali <- influence(modpl)
> halfnorm(gali$hat)
```

There is some indication that case 25, Santa Cruz island, may have some leverage. The predictor Scruz is the distance from Santa Cruz island which is zero for this case. This posed a problem for making the log transformation and explains why we added 0.1 to this variable. However, there is some indication that this inelegant fix may be causing some difficulty.

Moving on to influence, a half-normal plot of the Cook statistics is shown in the first panel of Figure 6.5:

```
> halfnorm(cooks.distance(modpl))
```

Again we have some indication that Santa Cruz island is influential. We can examine the change in the fitted coefficients. For example, consider the change in the Scruz coefficient as shown in the second panel of Figure 6.5:

```
> plot(gali$coef[, 5],ylab="Change in Scruz
coef",xlab="Case no.")
```

We see a substantial change for case 25. If we compare the full fit to a model without this case, we find:

```
> modplr <- glm(Species ~ log(Area) + log(Elevation) +
log(Nearest)
  + log(Scruz+0.1) + log(Adjacent), family=poisson,
gala, subset=-25)
➢  cbind(coef(modpl),coef(modplr))
(Intercept)       3.287941  3.050699
log(Area)         0.348445  0.334530
log(Elevation)    0.036421  0.059603
log(Nearest)     -0.040644 -0.052548
log(Scruz + 0.1) -0.030045  0.015919
log(Adjacent)    -0.089014 -0.088516
```

We see a sign change for the Scruz coefficient. This is interesting since in the full model, the coefficient is more than twice the standard error way from zero indicating some significance. A simple solution is to add a larger amount, say 0.5, to Scruz.

Other than this user-introduced anomaly, we find no difficulty. Using our earlier discovery of the log transformation, some variable selection and allowing for remaining overdispersion, our final model is:

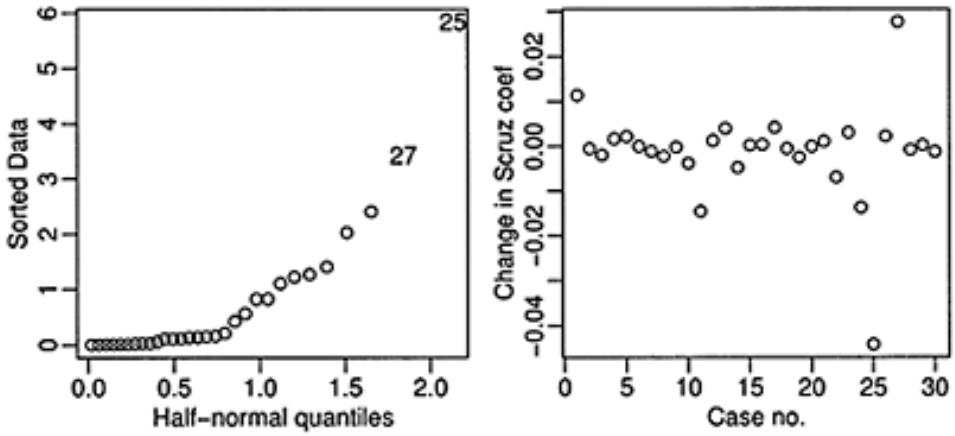

Figure 6.5 *Half-normal plot of the Cook statistics is shown on the left and an index plot of the change in the Scruz coefficient is shown on the right.*

```
                [,1]      [,2]

> modpla <- glm(Species ~ log(Area)+log(Adjacent),
family=poisson, gala)
> dp <-
sum(residuals(modpla,type="pearson")^2)/modpla$df.res
> summary(modpla,dispersion=dp)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.2767     0.1794   18.26  < 2e-16
log(Area)       0.3750     0.0326   11.50  < 2e-16
log(Adjacent)  -0.0957     0.0249   -3.85  0.00012
(Dispersion parameter for poisson family taken to be
16.527)
    Null deviance: 3510.73 on 29 degrees of freedom
Residual deviance:  395.54 on 27 degrees of freedom
```

Notice that the deviance is much lower and the elevation variable is not used when compared with our model choice in Section 3.1.

This example concerned a Poisson GLM. Diagnostics for binomial GLMs are similar, but see Pregibon (1981) and Collett (2003) for more details.

Further Reading: The canonical book on GLMs is McCullagh and Nelder (1989). Other books include Dobson (1990), Lindsey (1997), Myers, Montgomery, and Vining (2002), Gill (2001) and Fahrmeir and Tutz (2001). For a Bayesian perspective, see Dey, Ghosh, and Mallick (2000).

Exercises

1. Consider the or ings data from Chapter 2. Suppose that, in spite of all the drawbacks, we insist on fitting a model with an identity link, but with the binomial variance. Show how this may be done using a quasi family model using the glm function. (You will need need to consult the help pages for quasi and glm and in particular you will need to set good starting values for beta—if it doesn't work at the first attempt, try different values.) Describe how the fitted model differs from the standard logistic regression and give the predicted response at a temperature of 31°F.
2. Fit the orings data with a binomial response and a logit link as in Chapter 2.

   (a) Construct the appropriate test statistic for testing the effect of the temperature. State the appropriate null distribution and give the *p*-value.
   (b) Generate data under the null distribution for the previous test. Use the rbinom function with the average proportion of damaged O-rings. Recompute the test statistic and compute the *p*-value.
   (c) Repeat the process of the previous question 1000 times, saving the test statistic each time. Compare the empirical distribution of these simulated test statistics with the nominal null distribution stated in the first part of this question. Compare the critical values for a 5% level test computed using these two methods.

3. Fit the orings data with a binomial response and a logit link as in Chapter 2.

   (a) Construct the appropriate test statistic for testing the effect of the temperature. State the appropriate null distribution and give the *p*-value.
   (b) Generate a random permutation of the responses using sample and recompute the test statistic and compute the *p*-value.
   (c) Repeat the process of the previous question 1000 times, saving the test statistic each time. Compare the empirical distribution of these permuted data test statistics with the nominal null distribution stated in the first part of this question. Compare the critical values for a 5% level test computed using these two methods.

4. Data is generated from the exponential distribution with density $f(y)=\lambda\exp(-\lambda y)$ where $\lambda, y>0$.

   (a) Identify the specific form of $\theta, \phi, a(), b()$ and c() for the exponential distribution.
   (b) What is the canonical link and variance function for a GLM with a response following the exponential distribution?
   (c) Identify a practical difficulty that may arise when using the canonical link in this instance.
   (d) When comparing nested models in this case, should an F or χ^2 test be used? Explain.
   (e) Express the deviance in this case in terms of the responses yi and the fitted values $\hat{\mu}_i$.

5. The Conway-Maxwell-Poisson distribution has probability function:

$$P(Y=y)=\frac{\lambda^y}{(y!)^\nu}\frac{1}{Z(\lambda,\nu)} \qquad y=0,1,2,\ldots$$

where

$$Z(\lambda, \nu) = \sum_{i=0}^{\infty} \frac{\lambda^i}{(i!)^\nu}$$

Place this in exponential family form, identifying all the relevant components necessary for use in a GLM.