Chapter 8
Generalized Linear Models: Diagnostics

Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.
Box [1, p. 792]

8.1 Introduction and Overview

This chapter introduces some of the necessary tools for detecting violations of the assumptions in a GLM, and then discusses possible solutions. The assumptions of the GLM are first reviewed (Sect. 8.2), then the three basic types of residuals (Pearson, deviance and quantile) are defined (Sect. 8.3). The leverages are then given in the GLM context (Sect. 8.4) leading to the development of standardized residuals (Sect. 8.5). The various diagnostic tools for checking the model assumptions are introduced (Sect. 8.7) followed by techniques for identifying unusual and influential observations (Sect. 8.8). Comments about using each type of residual and the nomenclature of residuals are given in Sect. 8.6. We then discuss techniques to remedy or ameliorate any weaknesses in the models (Sect. 8.9), including the introduction of quasi-likelihood (Sect. 8.10). Finally, collinearity is discussed (Sect. 8.11).

8.2 Assumptions of GLMs

The assumptions made when fitting GLMs concern:

- Lack of outliers: All responses were generated from the same process, so that the same model is appropriate for all the observations.
- Link function: The correct link function $g()$ is used.
- Linearity: All important explanatory variables are included, and each explanatory variable is included in the linear predictor on the correct scale.
- Variance function: The correct variance function $V(\mu)$ is used.
- Dispersion parameter: The dispersion parameter ϕ is constant.
- Independence: The responses y_i are independent of each other.

P. K. Dunn, G. K. Smyth, *Generalized Linear Models with Examples in R*,
Springer Texts in Statistics, https://doi.org/10.1007/978-1-4419-0118-7_8

- Distribution: The responses y_i come from the specified EDM.

The first assumption concerns the suitability of the model overall. The other assumptions are ordered here from those that affect the first moment of the responses (the mean), to the second moment (variances) to third and higher moments (complete distribution of y_i). Generally speaking, assumptions that affect the lower moments of y_i are the most basic. Compare these to the assumptions for the (normal) linear regression model (Sect. 3.2). This chapter discusses methods for assessing the validity of these assumptions.

Importantly, the assumptions are never *exactly* true. Instead, it is important to be aware of the sensitivity of the conclusions to deviations from the model assumptions. The model assumptions should always be checked after fitting a model to identify potential problems, and this information used to improve the model where possible.

8.3 Residuals for GLMs

8.3.1 Response Residuals Are Insufficient for GLMs

The distances $y_i - \hat{\mu}_i$ are called the *response residuals*, and are the basis for residuals in linear regression. The response residuals are inadequate for assessing a fitted GLM, because GLMs are based on EDMs where (in general) the variance depends on the mean. As an example, consider the cherry tree data (Example 3.14, p. 125), and the theory-based model fitted to the data:

```
> data(trees)
> cherry.m1 <- glm( Volume ~ log(Girth) + log(Height),
                    family=Gamma(link=log), data=trees)
> coef( cherry.m1 )
(Intercept)  log(Girth) log(Height)
  -6.691109    1.980412    1.132878
```

Consider two volumes y_1 and y_2 marked on Fig. 8.1. Also shown are the modelled distributions of the observations for the corresponding fitted values $\hat{\mu}_i$ (based on the gamma distribution). Note that both observations are $y_i - \hat{\mu}_i = 7$ greater than the respective predicted means. However, observation y_1 is in the extreme tail of the fitted distribution, but observation y_2 is not in the extreme tail of the distribution, even though the response residuals $y_i - \hat{\mu}_i$ are the same for each case. A new definition of residuals is necessary.

Ideally, residuals for GLMs should behave similarly to residuals for linear regression models, because residuals in that case are familiar and easily interpreted. That is, ideally residuals for GLMs should be approximately normally distributed with mean zero and constant variance. Response residuals do not necessarily have constant variance or a normal distribution.

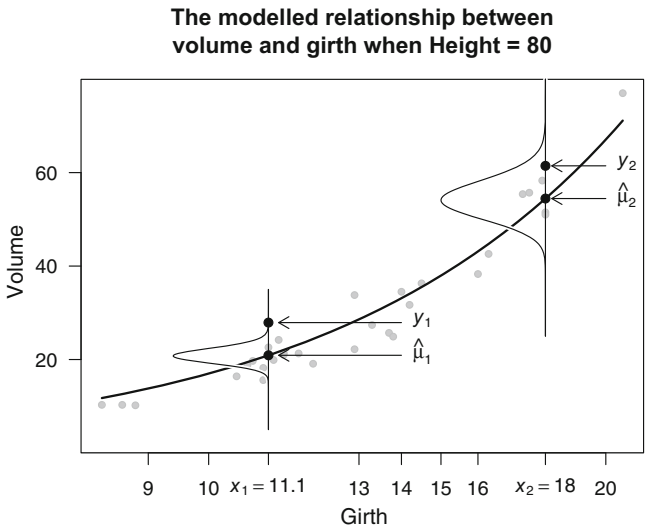

Fig. 8.1 The cherry tree data. The solid line shows the modelled relationship between `Volume` and `log(Girth)` when `Ht=80`. Two observations from the gamma GLM as fitted to the cherry tree data are also shown. Observation y_1 is extreme, but observation y_2 is not extreme, yet the difference $y_i - \hat{\mu}_i = 7$ is the same in both cases. Note that log-scale is used on the horizontal axis since the covariate is `log(Girth)` (Sect. 8.3.1)

8.3.2 Pearson Residuals

The most direct way to handle the non-constant variance in EDMs is to divide out the effect of non-constant variance. In this spirit, define *Pearson residuals* as

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

where $V()$ is the variance function. Notice that r_P is the square root of the unit Pearson statistic (Sect. 6.8.5). For a fitted GLM in R, say `fit`, the Pearson residuals are found using `resid(fit, type="pearson")`. The Pearson residuals are actually the ordinary residuals when the GLM is treated as a least-squares regression model using the working responses and weights (Sect. 6.7).

The Pearson statistic has an approximate chi-square distribution when the Central Limit Theorem applies, under the conditions given in Sect. 7.5 (p. 276). Under these same conditions, the Pearson residuals have an approximate normal distribution.

Example 8.1. For the normal distribution, $V(\mu) = 1$ (Table 5.1), and so the Pearson residuals are $r_P = (y - \hat{\mu})\sqrt{w}$. □

Example 8.2. For the Poisson distribution, $V(\mu) = \mu$ (Table 5.1), and so the Pearson residuals are $r_P = (y - \hat{\mu})/\sqrt{\hat{\mu}/w}$. □

8.3.3 Deviance Residuals

The Pearson residuals are the square root of the unit Pearson statistic. Similarly, define the deviance residuals r_D as the signed square root of the unit deviance (Sect. 5.4):

$$r_D = \text{sign}(y - \hat{\mu})\sqrt{wd(y, \hat{\mu})}. \tag{8.1}$$

(The function sign(x) equals 1 if $x > 0$; -1 if $x < 0$; and 0 if $x = 0$.) For a fitted model in R, say `fit`, the deviance residuals are found using `resid(fit)`. In other words, the deviance residuals are computed by default by `resid()`. A summary of the deviance residuals is given in the `summary()` of the output object produced by `glm()` (as seen in Fig. 6.1).

The deviance statistic has an approximate chi-square distribution when the saddlepoint approximation applies, under the conditions given in Sect. 7.5 (p. 276). Under these same conditions, the deviance residuals have an approximate normal distribution.

Example 8.3. Using the unit deviance for the normal distribution (Table 5.1), the deviance residuals are $r_D = (y - \hat{\mu})\sqrt{w}$. The deviance residuals are the same as the Pearson residuals for the normal distribution, and only for the normal distribution. □

Example 8.4. Using the unit deviance for the Poisson distribution (Table 5.1), the deviance residuals are

$$r_D = \text{sign}(y - \hat{\mu})\sqrt{2w\left\{y \log\left(\frac{y}{\hat{\mu}}\right) - (y - \hat{\mu})\right\}}.$$

□

8.3.4 Quantile Residuals

The Pearson and deviance residuals have approximate normal distributions as explained above, with the deviance residuals more likely to be more normally distributed than the Pearson residuals [12]. When the guidelines in Sect. 7.5 (p. 276) are not met, the Pearson and deviance residuals can be clearly non-normal, especially for discrete distributions.

An alternative to Pearson and deviance residuals are the quantile residuals [5], which are *exactly* normally distributed apart from the sampling variability in estimating μ and ϕ, assuming that the correct EDM is used. The quantile residual r_Q for an observation has the same cumulative probability on a standard normal distribution as y does for the fitted EDM. A simple modification involving randomization is needed for discrete EDMs. For a fitted model in R, say `fit`, the quantile residuals are found using `qresid(fit)`, using the function `qresid()` from package **statmod**.

8.3.4.1 Quantile Residuals: Continuous Response

Quantile residuals are best described in the context of an example. Consider an exponential EDM (4.37) (which is a gamma EDM with $\phi = 1$) fitted to data where one observation is $y = 1.2$ with $\hat{\mu} = 3$. First, determine the cumulative probability that an observation is less than or equal to y on this fitted exponential distribution using `pexp()` (Fig. 8.2, left panel):

```
> y <- 1.2; mu <- 3
> cum.prob <- pexp(y, rate=1/mu); cum.prob
[1] 0.32968
```

Then find the value of the standard normal variate with the same cumulative probability using `qnorm()`; this is the quantile residual (Fig. 8.2, right panel):

```
> rq <- qnorm(cum.prob); rq
[1] -0.4407971
```

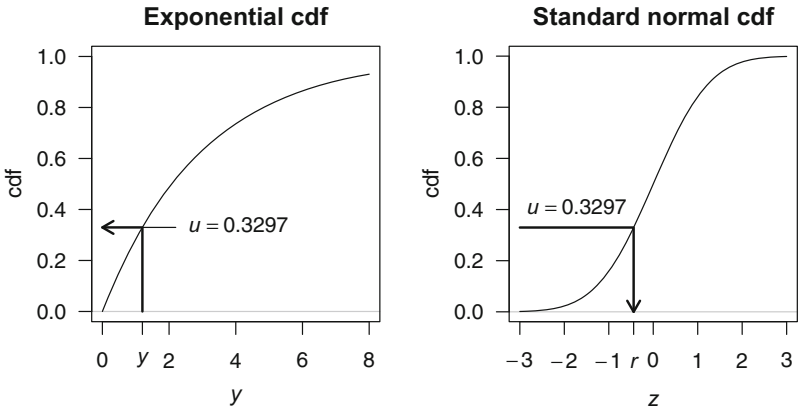

Fig. 8.2 Computing the quantile residuals for an exponential EDM for an observation $y = 1.2$, when $\hat{\mu} = 3$ (Sect. 8.3.4.2)

More formally, let $\mathcal{F}(y;\mu,\phi)$ be the cumulative distribution function (CDF) of a random variable y (it need not belong to the EDM family). The quantile residuals are

$$r_Q = \Phi^{-1}\{\mathcal{F}(y;\hat{\mu},\phi)\},$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. (For example, $\Phi^{-1}(0.975) = 1.96$ and $\Phi^{-1}(0.025) = -1.96$.) If ϕ is unknown, use the Pearson estimator of ϕ.

Example 8.5. For the exponential distribution, the probability function is given in (4.37). The CDF is

$$\mathcal{F}(y) = 1 - \frac{1}{\mu}\exp\left(-\frac{y}{\mu}\right)$$

for $y > 0$. The quantile residual is

$$r_Q = \Phi^{-1}\left\{1 - \frac{1}{\hat{\mu}}\exp\left(-\frac{y}{\hat{\mu}}\right)\right\}.$$

□

Example 8.6. For the normal distribution, $\mathcal{F}$ is the CDF of a normal distribution with mean μ and variance σ^2/w. Since $\Phi^{-1}(\cdot)$ is the *inverse* of the standard normal CDF, the quantile residuals are

$$r_Q = \frac{(y-\hat{\mu})\sqrt{w}}{s},$$

where s is the estimate of σ. For the normal distribution, $r_Q = r_P/s = r_D/s$.
□

8.3.4.2 Quantile Residuals: Discrete Response

For discrete EDMs, a simple modification is necessary to define the quantile residuals. Consider a Poisson EDM for the observation $y = 1$ when $\hat{\mu} = 2.6$.

Locate the observation $y = 1$ on the Poisson CDF (Fig. 8.3, left panel). Since the CDF is discrete at $y = 1$, the CDF makes a discrete jump between $a = 0.074$ and $b = 0.267$:

```
> y <- 1; mu <- 2.6
> a <- ppois(y-1, mu); b <- ppois(y, mu)
> c(a, b)
[1] 0.07427358 0.26738488
```

Choose a point at random from the shaded area of the plot between a and b:

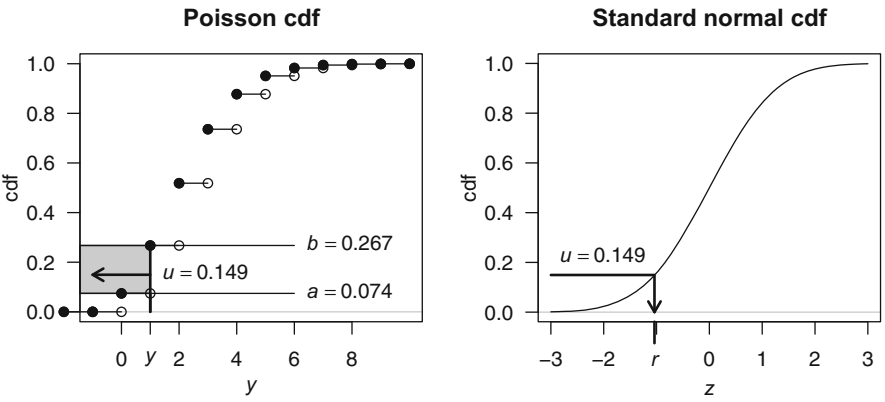

Fig. 8.3 Computing the quantile residuals for a situation where the observed value is $y = 1$ when $\hat{\mu} = 2.6$ for a Poisson distribution. The filled circles indicate the value is included, while a hollow circle indicates the value is excluded (Sect. 8.3.4.2)

```
> u <- runif(1, a, b); u
[1] 0.1494077
```

In this example, the chosen random number is $u = 0.149$. Then find the value of a standard normal variate with the same cumulative probability, as in the continuous EDM case (Fig. 8.3, right panel). This standard normal variate is the quantile residual for that observation:

```
> rq <- qnorm( u ); rq
[1] -1.038977
```

In this example, the quantile residual is $r_Q = \Phi^{-1}(0.149) = -1.039$. (Using the extremities of the interval for u_i, the quantile residual will be between approximately -0.621 and -1.445.)

This randomization is an advantage: the quantile residuals are continuous even for discrete distributions, unlike deviance and Pearson residuals (Example 8.8; Problem 8.4). As for the continuous case, the quantile residuals have an exact standard normal distribution.

Symbolically, let the lower and upper limits of the region in the CDF be $a = \lim_{\epsilon \uparrow 0} \mathcal{F}(y + \epsilon; \hat{\mu}, \phi)$ and $b = \mathcal{F}(y; \hat{\mu}, \phi)$ respectively. (The notation $\lim_{\epsilon \uparrow 0}$ means the limit as ϵ approaches 0 from *below*, so that ϵ is always negative.) Then, define randomized quantile residuals as

$$r_Q = \Phi^{-1}(u),$$

where u is a uniform random variable on the interval $(a, b]$. For the Poisson example above, $b = \mathcal{F}(y = 1; \hat{\mu} = 2.6)$, where $\mathcal{F}$ is the CDF for the Poisson distribution. The value of a is the value of the CDF as y approaches but is *less* than $y = 1$. Thus, $a = \lim_{\epsilon \uparrow 0} \mathcal{F}(y + \epsilon; \hat{\mu} = 2.6) = \mathcal{F}(y = 0.2, \hat{\mu} = 2.6)$.

Four replications of the quantile residuals are recommended [5] when used with discrete distributions because quantile residuals for a discrete response

have a random component. Any features not preserved across all four sets of residuals are considered artifacts of the randomization. In the discrete case, quantile residuals are sometimes called *randomized* quantile residuals, for obvious reasons.

Quantile residuals are best used in residual plots where trends and patterns are of interest, because $y - \hat{\mu} < 0$ does not necessarily imply $r_Q < 0$ (Problem 8.7). Quantile residuals are strongly encouraged for discrete EDMs (Example 8.8).

8.4 The Leverages in GLMs

8.4.1 Working Leverages

As previously explained in Sect. 6.7, a GLM can be treated locally as a linear regression model with working responses z_i and working weights W_i. The working responses and weights are functions of the fitted values $\hat{\mu}_i$, but, if we treat them as fixed, we can compute leverages (or hat values) for each observation exactly as for linear regression (Sect. 3.4.2).

The ith leverage h_i is the weight that observation z_i receives when computing the corresponding value of the linear predictor $\hat{\eta}_i$. If the leverage is small, this is evidence that many observations, not just one, are contributing to the estimation of the fitted value. In the extreme case that $h_i = 1$, the ith fitted value will be entirely determined by the ith observation, so that $\hat{\eta}_i = z_i$ and $\hat{\mu}_i = y$.

The variance of the working residuals $e_i = z_i - \hat{\eta}_i$ can be approximated by (see Sect.6.7)

$$\text{var}[e_i] \approx \phi V(\hat{\mu}_i)(1 - h_i).$$

If ϕ is unknown, a suitable estimate is used to give $\widehat{\text{var}}[e_i]$. As in linear regression, the leverages are computed using `hatvalues()` in R.

* *8.4.2 The Hat Matrix*

In the context of GLMs, the *hat matrix* is

$$\text{H} = \text{W}^{1/2}\text{X}(\text{X}^T\text{WX})^{-1}\text{X}^T\text{W}^{1/2}, \tag{8.2}$$

where W is the diagonal matrix of weights from the final iteration of the fitting algorithm (Sect. 6.3). The form is exactly the same as used in linear regression (Sect. 3.4.2), except in the GLM case W depends on the fitted values

$\hat{\boldsymbol{\mu}}$. The leverages (or hat diagonals) h_i are the diagonal elements of H, and are found in R using `hatvalues()`.

8.5 Leverage Standardized Residuals for GLMs

The Pearson, deviance and quantile residuals discussed in Sect. 8.3 are the basic types of residuals (called *raw residuals*). As with linear regression, standardized residuals have approximately constant variance, and are defined analogously:

$$\begin{aligned} r'_P &= \frac{r_P}{\sqrt{\phi(1-h)}} = \frac{y-\hat{\mu}}{\sqrt{\phi V(\hat{\mu})(1-h)/w}} \\ r'_D &= \frac{r_D}{\sqrt{\phi(1-h)}} = \frac{\text{sign}(y-\hat{\mu})\sqrt{wd(y,\hat{\mu})}}{\sqrt{\phi(1-h)}} \\ r'_Q &= \frac{r_Q}{\sqrt{1-h}}, \end{aligned} \tag{8.3}$$

where h are the leverages. If ϕ is unknown, use an estimate of ϕ (R uses the Pearson estimate $\bar{\phi}$). The standardized deviance residuals are found directly using `rstandard()`; the standardized Pearson and quantile residuals must be computed in R using the formulae above.

The standardized deviance residuals have a useful interpretation. The square of the standardized deviance residuals is approximately the reduction in the residual deviance when Observation i is omitted from the data, scaled by ϕ (Problem 8.6).

Observe that division by ϕ (or its estimate) is not needed for the quantile residuals as the quantile residuals are transformed to the standard normal distribution with variance one.

Example 8.7. For the model `cherry.m1` fitted to the cherry tree data (Sect. 8.3; data set: `trees`), compute the three types of raw residuals in R as follows:

```
> library(statmod)          # Provides qresid()
> rP <-  resid( cherry.m1, type="pearson" )
> rD <-  resid( cherry.m1 ) # Deviance resids are the default
> rQ <- qresid( cherry.m1 )
```

Then compute the standardized residuals also:

```
> phi.est <- summary( cherry.m1 )$dispersion  # Pearson estimate
> rP.std <- rP / sqrt( phi.est*(1 - hatvalues(cherry.m1)) )
> rD.std <- rstandard(cherry.m1)
> rQ.std <- rQ / sqrt( 1 - hatvalues(cherry.m1) )
> all.res <- cbind( rP, rP.std, rD, rD.std, rQ, rQ.std )
> head( all.res )  # Show the first six values only
```

```
           rP     rP.std          rD     rD.std         rQ     rQ.std
1  0.01935248  0.2620392  0.01922903  0.2603676  0.2665369  0.2893348
2  0.03334904  0.4558288  0.03298537  0.4508579  0.4380951  0.4800656
3  0.01300934  0.1811459  0.01295335  0.1803663  0.1882715  0.2101705
4 -0.01315583 -0.1691519 -0.01321397 -0.1698994 -0.1380666 -0.1423184
5 -0.04635977 -0.6169148 -0.04709620 -0.6267146 -0.5606192 -0.5980889
6 -0.04568564 -0.6188416 -0.04640051 -0.6285250 -0.5519432 -0.5993880
> apply( all.res, 2, var ) # Find the variance of each column
         rP      rP.std          rD      rD.std          rQ      rQ.std
0.005998800 1.013173741 0.006113175 1.032103295 0.950789672 1.031780512
```

The variance of the quantile residuals is near one since they are mapped to a standard normal distribution. The standardized residuals are all similar for this example. □

8.6 When to Use Which Type of Residual

Quantile, deviance and Pearson residuals all have exact normal distributions when the responses come from a normal distribution, apart from variability in $\hat{\mu}$ and $\hat{\phi}$. The deviance residuals are also exactly normal for inverse Gaussian GLMs. However, in many cases neither the Pearson nor deviance residuals can be guaranteed to have distributions close to normal, especially for discrete EDMs. The simple rules in Sect. 7.5 (p. 276) can be used to determine when the normality can be expected to be sufficiently accurate.

Quantile residuals are especially encouraged for discrete EDMs, since plots using deviance and Pearson residuals may contain distracting patterns (Example 8.8). Furthermore, standardizing or Studentizing the residuals is encouraged, as these residuals have more constant variance. For some specific diagnostic plots, special types of residuals are used, such as partial residuals and working residuals (Sect. 8.7.3).

8.7 Checking the Model Assumptions

8.7.1 Introduction

As with linear regression models, plots involving the residuals are used for assessing the validity of the model assumptions for GLMs. These plots are discussed in this section. Remedies for any identified problems follow in Sect. 8.9.

A strategy similar to that used for linear regression is adopted for assessing assumptions with GLMs. First, check independence when possible (Sect. 8.7.2). Then, use plots of the residuals against $\hat{\mu}$ and residuals against each explanatory variable to identify structural problems in the model. In

all these situations, the ideal plots contain no patterns or trends. Finally, plotting residuals in a Q–Q plot (Sect. 8.8) is convenient for detecting large residuals.

8.7.2 Independence: Plot Residuals Against Lagged Residuals

Independence of the responses is the most important assumption. Independence of the responses is usually a result of how the data are collected, so is often impossible to detect using residuals. As for linear regression, independence is, in most cases, best assessed from understanding the process by which the data were collected. However, if the data are collected over time, independence can be checked by plotting residuals against the previous residual in time. Ideally, the plots show no pattern under independence. If the data are spatial, independence can be checked by plotting the residuals against spatial explanatory variables (such as latitude and longitude). Again, the ideal plots show no pattern under independence.

The discussion for linear regression is still relevant (Sect. 3.5.5, p. 106), including the typical plots in Fig. 3.8.

8.7.3 Plots to Check the Systematic Component

Plots of the residuals against the fitted values $\hat{\mu}$ and the residuals against x_j are the main tools for diagnostic analysis. Using either the standardized deviance or quantile residuals is preferred in these plots because they have approximately constant variance. Quantile residuals are especially encouraged for discrete EDMs to avoid distracting patterns in the residuals (Example 8.8).

Two features of the plots are important:

- Trends: Any trends appearing in these plots indicate that the systematic component can be improved. This could mean changing the link function, adding extra explanatory variables, or transforming the explanatory variables.
- Constant variation: If the random component is correct (that is, the correct EDM is used), the variance of the points is approximately constant.

The plots can be constructed in R using `plot()`, or using `scatter.smooth()` which also adds a smoothing curve to the plots which may help detect trends. Detecting trends in the plots is often easier if the fitted values $\hat{\mu}$ are spread out more evenly horizontally. This is achieved by using the appropriate variance-stabilizing transformation of $\hat{\mu}$ (Table 5.2), often called the constant-information scale in this context (Table 8.1).

Table 8.1 The constant-information scale transformations of $\hat{\mu}$ for common EDMs for use in residual plots (Sect. 8.7.3)

EDM Scale	EDM Scale
Binomial: $\sin^{-1}\sqrt{\hat{\mu}}$	Inverse Gaussian: $1/\sqrt{\hat{\mu}}$
Poisson: $\sqrt{\hat{\mu}}$	Tweedie ($V(\mu) = \mu^{\xi}$): $\hat{\mu}^{(2-\xi)/2}$
Gamma: $\log \hat{\mu}$	

If the evidence shows problems with the systematic component, then the cause may be an incorrect link function, or an incorrect linear predictor (for example, important explanatory variables are missing, or covariates should be transformed), or both. To further examine the link function, an informal check is to plot the *working responses* (6.9)

$$z_i = \hat{\eta}_i + \frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i}(y_i - \hat{\mu}_i)$$

against $\hat{\eta}_i$. If the link function is appropriate, then the plot should be roughly linear [10, §12.6.3]. If a noticeable curvature is apparent in the plot, then another choice of link function should be considered. The working responses z_i are found in R using that $z_i = e_i + \hat{\eta}_i$, where e_i are the working residuals (Sect. 6.7), found in R using `resid(fit, type="working")`. Other methods also exist for evaluating the choice of link function [2, 13].

To determine if covariate x_j is included on the incorrect scale, use *partial residuals*

$$u_j = e_i + \hat{\beta}_j x_j, \tag{8.4}$$

found in R using `resid(fit, type="partial")`. This command produces an $n \times p$ array holding the partial residuals for each explanatory variable x_j in the p columns. A plot of u_j against x_j (called a *component-plus-residual plot* or *partial residual plot*) is linear if x_j is included on the correct scale. The R function `termplot()` can also be used to produce partial residual plots, as in linear regression. If many explanatory variables are included on the incorrect scale, the process of examining the partial residual plots for each explanatory variables is iterative: one covariate at a time is fixed, and the partial residual plots re-examined.

Example 8.8. A binomial GLM with a logit link function was used to model 60 observations each with a sample size of 3 (that is, $m = 3$). The systematic component of the fitted model assumed $\eta = \log\{\mu/(1-\mu)\} = \beta_0 + \beta_1 x$ for the covariate x. After fitting the model, the plot of quantile residuals against x shows a curved trend (Fig. 8.4, top left panel), indicating that the model is inadequate. Interpreting the deviance residuals is difficult (Fig. 8.4, top right panel), as the data lie on parallel curves, corresponding to the four possible values of y.

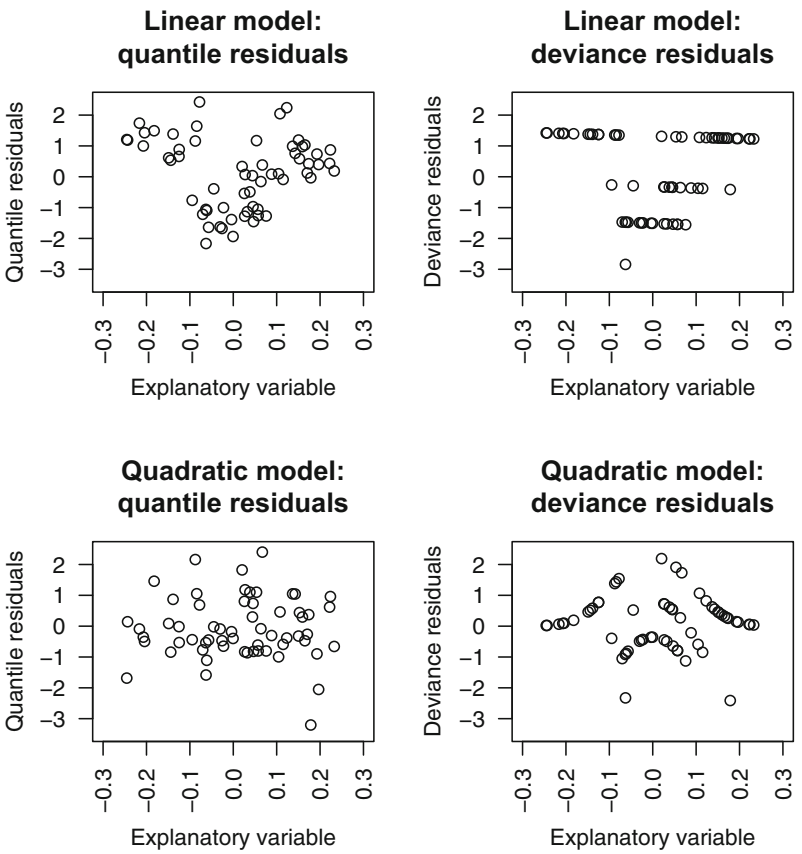

Fig. 8.4 The residuals from a fitted binomial GLM. Top panels: the binomial GLM with a linear systematic component plotted against the explanatory variable; bottom panels: the binomial GLM with a quadratic systematic component plotted against the explanatory variable; left panels: the quantile residuals; right panel: the deviance residuals (Example 8.8)

After fitting the systematic component $\eta = \log\{\mu/(1-\mu)\} = \beta_0 + \beta_1 x + \beta_2 x^2$, the plot of quantile residuals against x (Fig. 8.4, bottom left panel) shows no trend and indicates the model now fits well. The deviance residuals still contain distracting parallel curves (Fig. 8.4, bottom right panel) that make any interpretation difficult. The data actually are randomly generated from a binomial distribution so that η truly depends quadratically on x. (This example is based on [5].) □

Example 8.9. Consider the model `cherry.m1` fitted to the cherry tree data (Example 3.14; data set: `trees`). We now examine the plots of r'_D against $\hat{\mu}$, against `log(Girth)` and against `log(Height)` (Fig. 8.5, top panels):

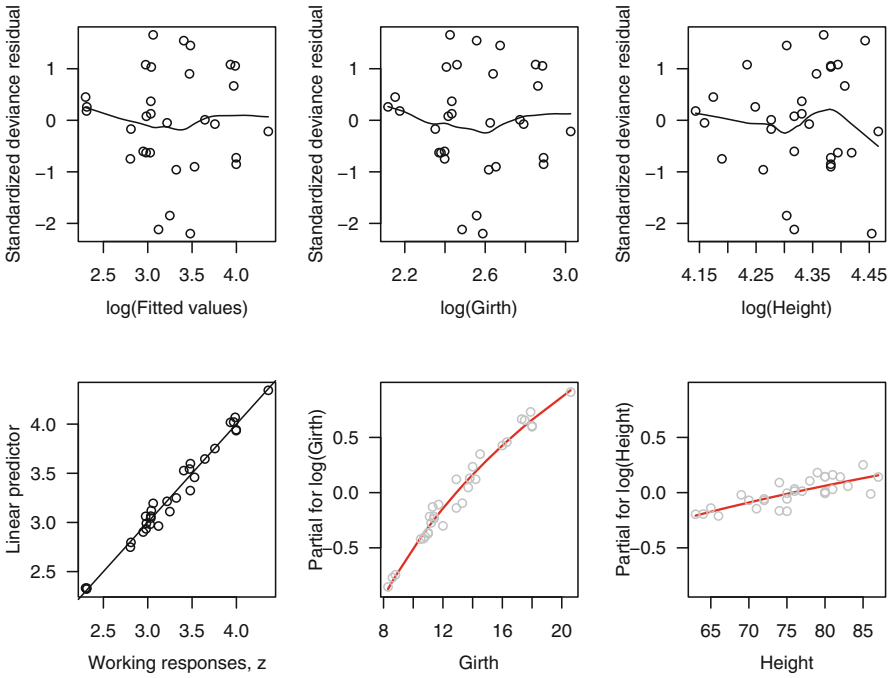

Fig. 8.5 Diagnostic plots for Model `cherry.m1` fitted to the cherry tree data. Top left panel: r'_D against $\log \hat{\mu}_i$; top centre panel: r'_D against `log(Girth)`; top right panel: r'_D against `log(Height)`; bottom left panel: $\hat{\eta}$ against z; bottom centre panel: the partial residual plot for girth; bottom right panel: the partial residual plot for height (Example 8.9)

```
> scatter.smooth( rstandard(cherry.m1) ~ log(fitted(cherry.m1)), las=1,
    ylab="Standardized deviance residual", xlab="log(Fitted values)" )
> scatter.smooth( rstandard(cherry.m1) ~ log(trees$Girth), las=1,
    ylab="Standardized deviance residual", xlab="log(Girth)" )
> scatter.smooth( rstandard(cherry.m1) ~ log(trees$Height), las=1,
    ylab="Standardized deviance residual", xlab="log(Height)" )
```

(The constant-information scale (Table 8.1) is the logarithmic scale for the gamma distribution, as used in the top left panel.) The plots appear approximately linear, but the variance of the residuals for smaller values of $\hat{\mu}$ may be less than for larger values of $\hat{\mu}$. The plot of z_i against $\hat{\eta}_i$ is also approximately linear (Fig. 8.5, bottom left panel) suggesting a suitable link function:

```
> z <- resid(cherry.m1, type="working") + cherry.m1$linear.predictor
> plot( z ~ cherry.m1$linear.predictor, las=1,
    xlab="Working responses, z", ylab="Linear predictor")
> abline(0, 1)                    # Adds line of equality
```

The plot of the partial residual (Fig. 8.5, bottom centre and right panels) suggest `Girth` and `Height` are included on the appropriate scale:

```
> termplot(cherry.m1, partial.resid=TRUE, las=1)
```

The line shown on each `termplot()` represents is the ideal relationship, so in both cases the plots suggest the model is adequate. □

8.7.4 Plots to Check the Random Component

The choice of random component for a GLM is usually based on an understanding of the data type: proportions of cases are modelled using binomial GLMs, and counts by a Poisson GLM, for example. However, Q–Q plots may be used to determine if the choice of distribution is appropriate [5]. Quantile residuals are used for these plots, since quantile residuals have an exact normal distribution (apart from sampling variability in estimating μ and ϕ) if the correct EDM has been chosen.

Example 8.10. Consider the model `cherry.m1` (Sect. 8.3) fitted to the cherry tree data (Example 3.14; data set: `trees`). A Q–Q plot of the quantile residuals (Fig. 8.6) shows that using a gamma GLM seems reasonable.

```
> qr.cherry <- qresid( cherry.m1 )
> qqnorm( qr.cherry, las=1 ); qqline( qr.cherry)
```

□

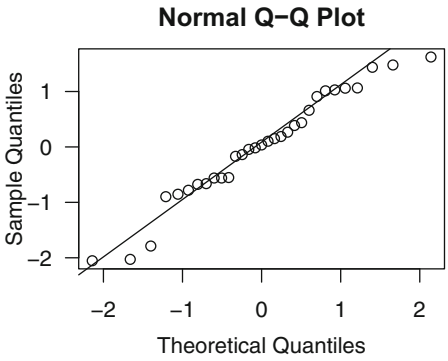

Fig. 8.6 The Q–Q plot of quantile residuals for Model `cherry.m1` fitted to the cherry tree data (Example 8.10)

8.8 Outliers and Influential Observations

8.8.1 Introduction

As for linear regression models, outliers are observations inconsistent with the rest of the data, and influential observations are outliers that substantially change the fitted model when removed from the data set. The tools used to identify outliers (Sect. 3.6.2) and influential observations (Sect. 3.6.3) in linear regression models are also used for GLMs, using results from the final step of the IRLS algorithm (Sect. 6.3), as discussed next.

8.8.2 Outliers and Studentized Residuals

For GLMs, as with linear regression models, outliers are identified as observations with unusually large residuals (positive or negative); the Q–Q plot is often convenient for doing this. Standardized deviance residuals are commonly used, though the use of quantile residuals are strongly encouraged for discrete data.

As for linear regression, *Studentizing* the residuals may also be useful (Sect. 3.6.2). For GLMs, computing Studentized deviance residuals requires refitting the original model n further times, when each observation is omitted one at a time. For each model without Observation i, the reduction in the deviance is computed. Fitting $n + 1$ models is necessary to do this, which is computationally expensive, and is avoided by approximating the Studentized residuals [18] by using

$$r_i'' = \text{sign}(y_i - \hat{\mu}_i)\sqrt{\frac{1}{\phi}\left(r_{D,i}^2 + \frac{h_i}{1-h_i}r_{P,i}^2\right)}.$$

If ϕ is unknown, estimate ϕ using

$$\bar{\phi}_{(i)} = \frac{D(y,\hat{\mu}) - r_{D,i}^2/(1-h_i)}{n - p' - 1},$$

which approximates the mean deviance estimate of ϕ in the model without Observation i (written $\bar{\phi}_{(i)}$). The approximate Studentized deviance residuals can be found in R using `rstudent()`, as used for linear regression models.

Example 8.11. Consider the cherry tree data and the model `cherry.m1` fitted in Sect. 8.3 (data set: `trees`). Compute the raw quantile residuals, raw deviance residuals, standardized deviance residuals, and Studentized residuals:

```
> library( statmod )   # To compute quantile residuals
> rs <- cbind( rD=resid(cherry.m1), "r'D"=rstandard(cherry.m1),
               "r''"=rstudent(cherry.m1), rQ=qresid(cherry.m1))
> head(rs)
           rD        r'D        r''         rQ
1  0.01922903  0.2603676  0.2537382  0.2665369
2  0.03298537  0.4508579  0.4408129  0.4380951
3  0.01295335  0.1803663  0.1756442  0.1882715
4 -0.01321397 -0.1698994 -0.1652566 -0.1380666
5 -0.04709620 -0.6267146 -0.6125166 -0.5606192
6 -0.04640051 -0.6285250 -0.6140386 -0.5519432
> apply( abs(rs), 2, max)   # The maximum absolute for each residual
      rD      r'D      r''       rQ
0.166763 2.197761 2.329122 2.053011
```

Since ϕ is small in this case, the saddlepoint approximation is suitable (Sect. 5.4.4), and the quantile, standardized and Studentized residuals are very similar. No large residuals exist. □

8.8.3 Influential Observations

Influential observations are outliers with high leverage. The measures of influence used for linear regression models, such as Cook's distance D, DFFITS, DFBETAS and the covariance ratio, are approximated for GLMs by using results from the final iteration of the IRLS algorithm (Sect. 6.7).

An approximation to Cook's distance for GLMs is

$$D \approx \left(\frac{r_P}{1-h}\right)^2 \frac{h}{\phi p'} = \frac{(r'_P)^2}{p'} \frac{h}{1-h} \tag{8.5}$$

as computed by the function `cooks.distance()` in R, where the Pearson estimator $\bar{\phi}$ of ϕ is used if it is unknown. Thus, Cook's distance is a combination of the size of the residual (measured by r'_P) and the leverage (measured by a monotonic function of h). Applying (8.5) for a linear regression model produces the same formula for Cook's distance given in (3.6) (p. 110).

DFBETAS, DFFITS, and the covariance ratio CR are computed using the same formulae as those used in linear regression (Sect. 3.6.3, p. 110), using the deviance residuals and using $\bar{\phi}_{(i)}$ in place of $s^2_{(i)}$. As for linear regression models, these statistics can be computed in R using `dffits()` (for DFFITS), `dfbetas()` (for DFBETAS), and `covratio()` (for CR). The function `influence.measures()` returns DFBETAS, DFFITS, CR, D, and the leverages h, flagging which are deemed influential (or high leverage in the case of h) according to the criteria in Sect. 3.6.3.

Example 8.12. For the model `cherry.m1` fitted to the cherry tree data (Sect. 8.3; data set: `trees`), influential observations are identified using `influence.measures()`:

```
> im <- influence.measures(cherry.m1); names(im)
[1] "infmat" "is.inf" "call"
> im$infmat <- round(im$infmat, 3 ); head( im$infmat )
  dfb.1_ dfb.l(G) dfb.l(H)  dffit cov.r cook.d   hat
1  0.015   -0.083    0.005  0.107 1.305  0.004 0.151
2  0.120   -0.082   -0.090  0.197 1.311  0.014 0.167
3  0.065   -0.021   -0.054  0.087 1.385  0.003 0.198
4 -0.011    0.021    0.004 -0.041 1.181  0.001 0.059
5  0.145    0.171   -0.170 -0.228 1.218  0.018 0.121
6  0.186    0.191   -0.212 -0.261 1.261  0.023 0.152
> colSums( im$is.inf )
  dfb.1_ dfb.l(G) dfb.l(H)    dffit    cov.r   cook.d      hat
       0        0        0        0        3        0        0
```

Three observations are identified as influential, but only by CR. Since none of the other measures identify these observations as influential, we should not be too concerned. Sometimes, plots of the influence statistics are useful (Fig. 8.7):

```
> cherry.cd <- cooks.distance( cherry.m1)
> plot( cherry.cd, type="h", ylab="Cook's distance", las=1)
> plot( dffits(cherry.m1), type="h", las=1, ylab="DFFITS")
> infl <- which.max(cherry.cd)  # The Observation number of largest D
> infl                          # Which observation?
18
18
> cherry.cd[infl]               # The value of D for that observation
       18
0.2067211
```

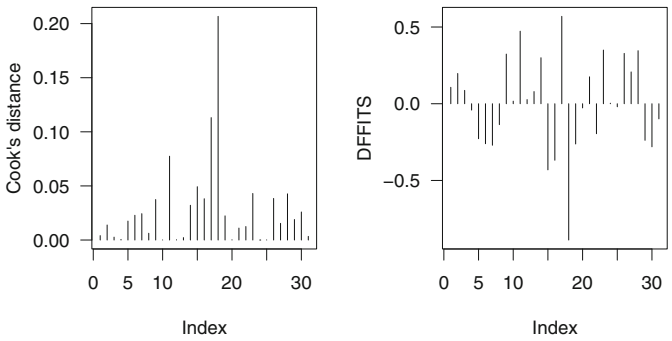

Fig. 8.7 Identifying influential observations for model `cherry.m1` fitted to the cherry tree data. Left panel: Cook's distance; right panel: DFFITS (Example 8.12)

The value of Cook's distance for Observation 18 is much larger than any others, but the observation is not identified as significantly influential. To demonstrate, we fit the model without Observation 18, then compare the estimated coefficients:

```
> cherry.infl <- update(cherry.m1, subset=(-infl) )
> coef(cherry.m1)
(Intercept)  log(Girth) log(Height)
  -6.691109    1.980412    1.132878
> coef(cherry.infl)
(Intercept)  log(Girth) log(Height)
  -7.209148    1.957366    1.267528
```

(The negative sign in `subset=(-infl)` *omits* Observation `infl` from the data set for this fit only.) The changes are not substantial, apart perhaps from the intercept. Contrast to the changes in the coefficients when another observation with a smaller value of D is omitted:

```
> cherry.omit1 <- update(cherry.m1, subset=(-1) ) # Omit Obs. 1
> coef(cherry.m1)
(Intercept)  log(Girth) log(Height)
  -6.691109    1.980412    1.132878
> coef(cherry.omit1)
(Intercept)  log(Girth) log(Height)
  -6.703461    1.986711    1.131840
```

The coefficients are very similar to those from model `cherry.m1` when Observation 1 is omitted: Observation 1 is clearly not influential. □

8.9 Remedies: Fixing Identified Problems

The techniques of Sects. 8.7 and 8.8 identify weaknesses in the fitted model. This section discusses possible remedies for these weaknesses. The following strategy can be adopted:

- If the responses are not independent (Sect. 8.7.2), use other methods, such as generalized estimating equations [7], generalized linear mixed models [2, 11] or spatial GLMs [4, 6]. These are beyond the scope of this book.
- Ensure the correct EDM is used (Sect. 8.7.3); that is, ensure the random component is adequate. For GLMs, the response data usually suggest the EDM:
  - Proportions of totals may be modelled using a binomial EDM (Chap. 9).
  - Count data may be modelled using a Poisson or negative binomial EDM (Chap. 10).

  - Positive continuous data may be modelled using a gamma or inverse Gaussian EDM (Chap. 11). In some cases, a Tweedie EDM may be necessary (Sect. 12.2.3).
  - Positive continuous data with exact zeros may be modelled using a Tweedie EDM (Sect. 12.2.4).

  Occasionally, a mean–variance relationship may be suggested that does not correspond to an EDM. In these cases, quasi-likelihood may be used (Sect. 8.10), or a different model may be necessary.
- Ensure the systematic component is correct (Sect. 8.7.3):
  - The link function may need to change. Changing the link function may be undesirable, because this changes the relationship between y and every explanatory variable, and because only a small number of link functions are useful for interpretability.
  - Important explanatory variables may be missing.
  - The covariates may need to be transformed. Partial residual plots may be used to determine if the covariates are included on the correct scale (and can be produced using `termplot()`). Simple transformations, polynomials in covariates (Sect. 3.10) or data-driven systematic components based on regression splines (Sect. 3.12) may be necessary in the model. R functions such as `poly()`, `bs()` and `ns()` are used for GLMs in the same way as for linear regression models.

Outliers and influential observations also may be remedied by making structural changes to the model. Sometimes, other strategies are needed to accommodate outliers and influential observations, including (under appropriate circumstances) omitting these observations; see Sect. 3.13.

Example 8.13. A suitable model for the cherry tree data was found in Sect. 8.3 (data set: `trees`). However, as an example we now consider residual plots from fitting a naive gamma GLM using the default (reciprocal) link function (Fig. 8.8):

```
> m.naive <- glm( Volume ~ Girth + Height, data=trees, family=Gamma)
> scatter.smooth( rstandard(m.naive) ~ log(fitted(m.naive)), las=1,
    xlab="Fitted values", ylab="Standardized residuals")
> scatter.smooth( rstandard(m.naive) ~ trees$Girth, las=1,
    xlab="Girth", ylab="Standardized residuals")
> scatter.smooth( rstandard(m.naive) ~ trees$Height, las=1,
    xlab="Height", ylab="Standardized residuals")
> eta <- m.naive$linear.predictor
> z <- resid(m.naive, type="working") + eta
> plot( z ~ eta, las=1,
    xlab="Linear predictor, eta", ylab="Working responses, z")
> abline(0, 1, col="grey")
> termplot(m.naive, partial.resid=TRUE, las=1)
```

(The constant-information scale (Table 8.1) is the logarithmic scale for the gamma distribution, as used in the top left panel.) The plots of r'_D against

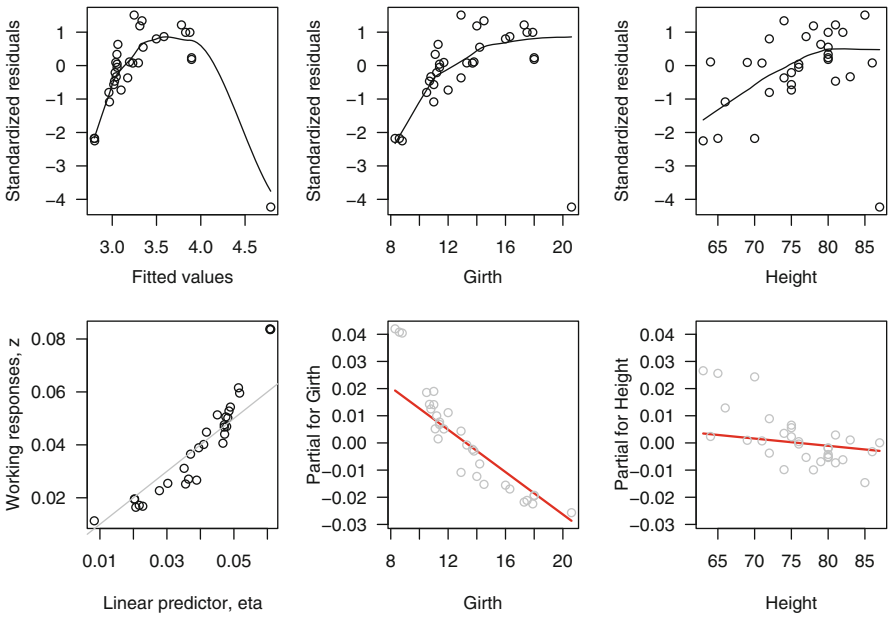

Fig. 8.8 Diagnostic plots for Model `m.naive` fitted to the cherry tree data. Top left panel: r'_D against $\log \hat{\mu}_i$; top centre panel: r'_D against `Girth`; top right panel: r'_D against `Height`; bottom left panel: z against $\hat{\eta}$; bottom centre panel: the partial residual plot for girth; bottom right panel: the partial residual plot for height (Example 8.13)

$\log \hat{\mu}$ (Fig. 8.8, top left panel) and r'_D against the covariates (top centre and top right panels) show an inadequate systematic component as shown by the trends and patterns. The plot of z_i against $\hat{\eta}_i$ (bottom left panel) suggests an incorrect link function. The partial residual plots (bottom centre and bottom right panels) suggest the covariates are included in the model incorrectly. In response to these diagnostic plots, consider the same model but with the more usual logarithmic link function (Fig. 8.9):

```
> m.better <- update(m.naive, family=Gamma(link="log"))
> scatter.smooth( rstandard(m.better) ~ log(fitted(m.better)), las=1,
    xlab="log(Fitted values)", ylab="Standardized residuals")
> scatter.smooth( rstandard(m.better) ~ trees$Girth, las=1,
    xlab="Girth", ylab="Standardized residuals")
> scatter.smooth( rstandard(m.better) ~ trees$Height, las=1,
    xlab="Height", ylab="Standardized residuals")
> eta <- m.better$linear.predictor
> z <- resid(m.better, type="working") + eta
> plot( z ~ eta, las=1, las=1,
    xlab="Linear predictor, eta", ylab="Working residuals, z")
> abline(0, 1, col="grey")
> termplot(m.better, partial.resid=TRUE, las=1)
```

The partial residual plots are much improved (Fig. 8.9, bottom centre and bottom right panels), and the plot of z_i against $\hat{\eta}$ (bottom left panel) suggests

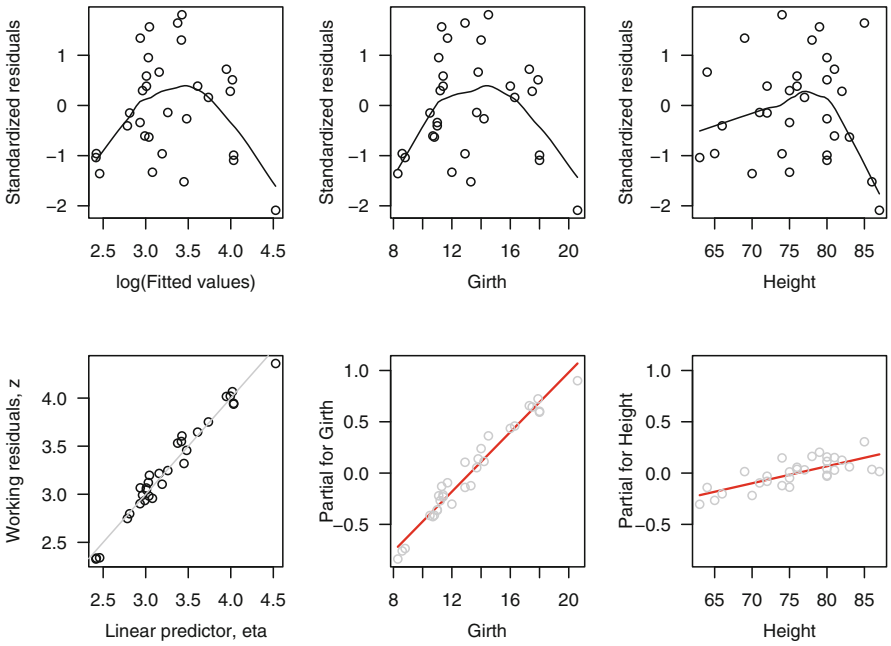

Fig. 8.9 Diagnostic plots for Model `m.better` fitted to the cherry tree data. Top left panel: r'_D against $\log \hat{\mu}_i$; top centre panel: r'_D against `Girth`; top right panel: r'_D against `Height`; bottom left panel: z against $\hat{\eta}$; bottom centre panel: the partial residual plot for girth; bottom right panel: the partial residual plot for height (Example 8.13)

the correct link function is used. However, the plots of r'_D against $\log \hat{\mu}$ (top left panel) and r'_D against the covariates (top centre and top right panels) still suggest a structural problem with the model.

In response to these diagnostic plots, model `cherry.m1` could be adopted. The residual plots from model `cherry.m1` then show an adequate model (Fig. 8.5, p. 310). In any case, `cherry.m1` has sound theoretical grounds, and should be preferred anyway. □

8.10 Quasi-Likelihood and Extended Quasi-Likelihood

In rare cases, sometimes the mean–variance relationship for a data set suggests a distribution that is not an EDM. However, the theory developed for GLMs is all based on distributions in the EDM family. However, note that for EDMs, the log-probability function has the neat derivative (Sect. 6.2)

$$\frac{\partial \log \mathcal{P}(\mu, \phi; y)}{\partial \mu} = \frac{y - \mu}{\phi V(\mu)}. \tag{8.6}$$

This relationship is used in fitting GLMs to find the estimates $\hat{\beta}_j$ (Sect. 6.2); the estimates of β_j and the standard errors $\text{se}(\hat{\beta}_j)$ are consistent given only the mean and variance information.

Motivated by these results, consider a situation where only the form of the mean and the variance are known, but no distribution is specified. Since no distribution is specified, no log-likelihood exists. However, analogous to (8.6), some *quasi-probability function* $\bar{\mathcal{P}}$ exists which satisfies

$$\frac{\partial \log \bar{\mathcal{P}}(y; \mu, \phi)}{\partial \mu} = \frac{y - \mu}{\phi V(\mu)}, \tag{8.7}$$

when only the variance function $V(\cdot)$ is known. On integrating,

$$\log \bar{\mathcal{P}}(y; \mu, \phi) = \int^{\mu} \frac{y - u}{\phi V(u)}\, du.$$

Suppose we have a series of observations y_i, for which we assume $\text{E}[y_i] = \mu_i$, and $\text{var}[y_i] = \phi V(\mu_i)/w_i$. Suppose a link-linear predictor for μ_i in terms of regression coefficients β_j, as for a GLM. Then the *quasi-likelihood function* (more correctly, the quasi-log-likelihood) is defined by

$$\mathcal{Q}(y; \mu) = \sum_{i=1}^{n} \log \bar{\mathcal{P}}(y_i; \mu_i, \phi/w_i).$$

The quasi-likelihood $\mathcal{Q}$ behaves like a log-likelihood function, but does not correspond to any probability function. As a result, the AIC and related statistics (Sect. 7.8) are not defined for quasi-models. In addition, quantile residuals (Sect. 8.3.4) are not defined for quasi-likelihood models since the quantile residuals require the CDF to be defined.

The unit deviance can be defined for quasi-likelihoods. First, notice that the unit deviance in (5.12) can be written as

$$\begin{aligned} d(y, \mu) &= 2\left\{t(y, y) - t(y, \mu)\right\} \\ &= 2\frac{\phi}{w}\left\{\log \mathcal{P}(y; y, \phi/w) - \log \mathcal{P}(y; \mu, \phi/w)\right\}. \end{aligned}$$

Using the quasi-likelihood in place of the log-likelihood,

$$\begin{aligned} d(y, \mu) &= 2\frac{\phi}{w}\left\{\log \bar{\mathcal{P}}(y; y, \phi/w) - \log \bar{\mathcal{P}}(y; \mu, \phi/w)\right\} \\ &= 2 \times \frac{\phi}{w} \int_{\mu}^{y} \frac{y - u}{\phi V(u)/w}\, du \\ &= 2 \int_{\mu}^{y} \frac{y - u}{V(u)}\, du. \end{aligned} \tag{8.8}$$

In this definition, the unit deviance depends only on the mean and variance. The total deviance is the (weighted) sum of the unit deviances as usual:

$$D(y, \mu) = \sum_{i=1}^{n} w_i d(y_i, \mu_i).$$

If there exists a genuine EDM for which $V(\mu)$ is the variance function, then the unit deviance and all other quasi-likelihood calculations derived from $V(\mu)$ reduce to the usual likelihood calculations for that EDM. This has the interesting implication that estimation and inference for GLMs depends only on the mean μ and the variance function $V(\mu)$. Since quasi-likelihood estimation is consistent, it follows that estimation for GLMs is robust against mis-specification of the probability distribution, because consistency of the estimates and tests is guaranteed as long as the first and second moment assumptions (means and variances) are correct.

Quasi-likelihood gives us a way to conduct inference when there is no EDM for a given mean–variance relationship. To specify a quasi-type model structure, write quasi-GLM($V(\mu)$; Link function), where $V(\mu)$ is the identifying variance function.

The most commonly-used quasi-models are for overdispersed Poisson-like or overdispersed binomial-like counts. These models vary the usual variance functions in some way, often by assuming a value for the dispersion ϕ greater than one, something which is not possible with the family of EDMs.

We discuss models for overdispersed Poisson-like counts, called quasi-Poisson models, at some length in Sect. 10.5.3. Quasi-Poisson models are specified in R using `glm()` with `family=quasipoisson()`. We discuss models for overdispersed binomial-like counts, called quasi-binomial models, at some length in Sect. 9.8. Quasi-binomial models are specified in R using `glm()` with `family=quasibinomial()`. Other quasi-models are specified in R using `family=quasi()`. For more details, see Sect. 8.13.

Inference for these quasi-models uses the same functions as for GLMs: `summary()` shows the results of the Wald tests, and `glm.scoretest()` in package **statmod** performs a score test. `anova()` performs the equivalent of likelihood ratio tests for comparing nested models by comparing the quasi-likelihood, which essentially compares changes in deviance. Analysis of deviance tests are based on the F-tests since ϕ is estimated for the quasi-models.

Example 8.14. For a Poisson distribution, $\text{var}[y] = \mu$ so that $V(\mu) = \mu$. However, in practice, often the variation in the data exceeds μ. This is called *overdispersion* (Sect. 10.5). One solution is to propose the variance structure $\text{var}[y] = \phi\mu$, but this variance structure does not correspond to any discrete EDM. Using quasi-likelihood,

$$\log \bar{\mathcal{P}}(y; \mu, \phi) = \int^{\mu} \frac{y - u}{\phi u}\, du = \frac{y \log \mu - \mu}{\phi}.$$

The same algorithms for fitting GLMs can be used to fit the model based on this quasi-likelihood. The unit deviance is

$$d(y,\mu) = 2\int_{\mu}^{y} \frac{y-u}{u}\,du = 2\left\{y\log\frac{y}{\mu} - (y-\mu)\right\},$$

identical to the unit deviance for the Poisson distribution (Table 5.1, p. 221). □

In defining the quasi-likelihood, we considered the derivative of $\log\bar{\mathcal{P}}$ with respect to μ but not ϕ. Hence the quasi-probability function is defined only up to terms not including μ. To deduce a complete quasi-probability function, the saddlepoint approximation can be used. This gives

$$\log\tilde{\mathcal{P}}(y;\mu,\phi) = -\frac{1}{2}\log\{2\pi\phi V(y)\} - \frac{d(y,\mu)}{2\phi},$$

which we call the *extended quasi-log-probability function*. Then

$$\mathcal{Q}^{+}(y;\mu,\phi/w) = \sum_{i=1}^{n}\log\tilde{\mathcal{P}}(y_i;\mu_i,\phi/w_i)$$

defines the *extended quasi-likelihood*. Solving $\mathrm{d}\mathcal{Q}^{+}(y;\mu,\phi/w)/\mathrm{d}\mu = 0$ shows that the solutions regarding μ are the same as for the quasi-likelihood and hence the log-likelihood. However, the extended quasi-likelihood has the advantage that solving $\mathrm{d}\mathcal{Q}^{+}(y;\mu,\phi/w)/\mathrm{d}\phi = 0$ produces the mean deviance estimate of ϕ.

The key use of extended quasi-likelihood is to facilitate the estimation of extended models which contains unknown parameters in the variance function $V()$, or which model some structure for the dispersion ϕ in terms of covariates.

8.11 Collinearity

As in linear regression (Sect. 3.14), *collinearity* occurs when at least some of the covariates are highly correlated with each other, implying they measure almost the same information.

As discussed in Sect. 3.14, collinearity causes no problems in prediction, but the parameter estimates $\hat{\beta}_j$ are hard to estimate with precision. Several equations may be found from which to compute the predictions, all of which may be effective but which produce different interpretations.

Collinearity is most easily identified by examining the correlations between the covariates. Any correlations greater than some (arbitrary) value, perhaps 0.7, are of concern. Other methods also exist for identifying collinearity. The same remedies apply as for linear regression (Sect. 3.14):

- Omitting some explanatory variables from the analysis.
- Combine explanatory variables in the model provided the combination makes sense.
- Collect more data, if there are observations that can be made that better distinguish the correlated covariates.
- Use special methods, such as ridge regression [17, §11.2], which are beyond the scope of this book.

Example 8.15. For the cherry tree data (Example 3.14; data set: `trees`), the two explanatory variables are correlated:

```
> cor( trees$Girth, trees$Height)
[1] 0.5192801
> cor( log(trees$Girth), log(trees$Height) )
[1] 0.5301949
```

Although correlated (that is, taller trees tend to have larger girths), collinearity is not severe enough to be a concern. □

8.12 Case Study

The noisy miner data [9] have been used frequently in this book (Example 1.5; `nminer`). The GLM fitted to model the number of noisy miners `Minerab` from the number of eucalypt trees `Eucs` is:

```
> library(GLMsData); data(nminer)
> nm.m1 <- glm( Minerab ~ Eucs, data=nminer, family=poisson)
> printCoefmat(coef(summary(nm.m1)))
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -0.876211   0.282793 -3.0984  0.001946 **
Eucs         0.113981   0.012431  9.1691 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The diagnostic plots (Fig. 8.10) are informative:

```
> library(statmod)             # To find randomized quantile residuals
> qr <- qresid( nm.m1 )
> qqnorm(qr, las=1); qqline(qr)
> plot( qr ~ sqrt(fitted(nm.m1)), las=1 )
> plot( cooks.distance(nm.m1), type="h", las=1 )
> plot( hatvalues(nm.m1), type="h", las=1 )
```

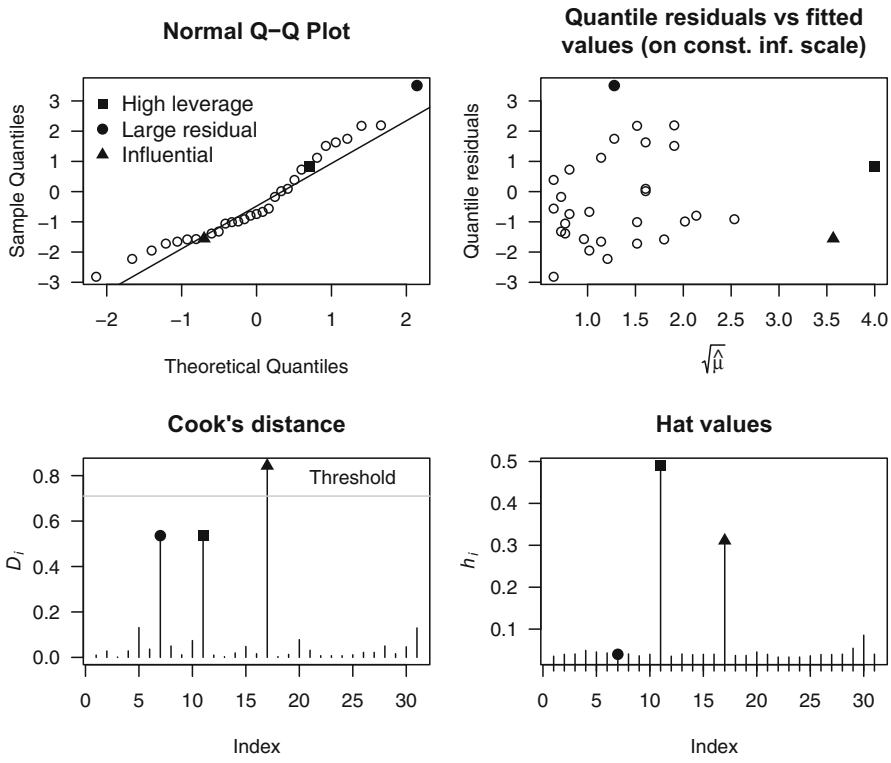

Fig. 8.10 Diagnostic plots for the GLM fitted to the noisy miner data. Top left: Q–Q plot of quantile residuals; top right: quantile residuals against $\sqrt{\hat{\mu}}$ (using the constant-information scale for the Poisson distribution); bottom left: Cook's distance, with the threshold for significance shown; bottom right: the leverages (Sect. 8.12)

We now locate the observations with the largest leverage, the largest absolute quantile residual, and the most influential observation:

```
> maxhat <- which.max( hatvalues(nm.m1) )         # Largest leverage
> maxqr <- which.max( abs(qr) )                   # Largest abs. residual
> maxinfl <- which.max( cooks.distance(nm.m1))    # Most influential
> c( MaxLeverage=maxhat, MaxResid=maxqr, MaxInfluence=maxinfl)
 MaxLeverage.11        MaxResid MaxInfluence.17
             11               7              17
```

Only Observation 17 is influential according to R's criterion (Sect. 3.6.3):

```
> which(influence.measures(nm.m1)$is.inf[,"cook.d"] )
17
17
```

In summary, Observation 11 (plotted with a filled square) has high leverage, but the residual is small and so it is not influential; Observation 7 (plotted with filled circle) has a large residual, but the leverage is small and so it is not

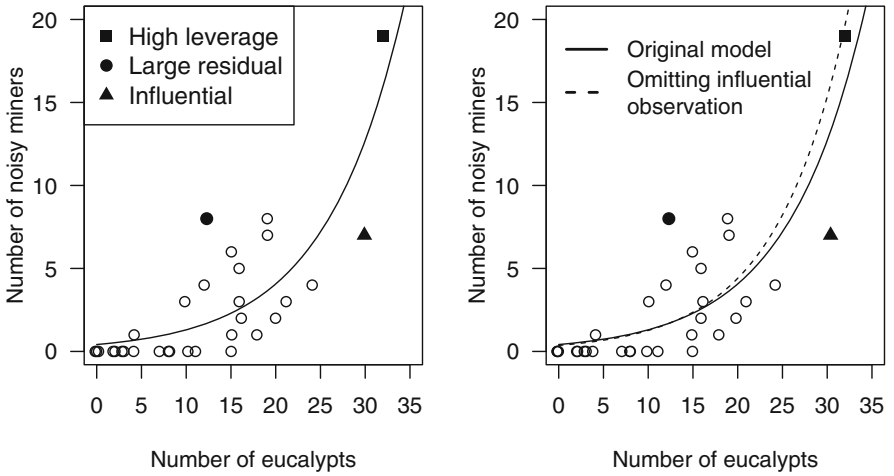

Fig. 8.11 Plots of the noisy miner data: left: the data plotted showing the location of three important observations; right: the data plotted with the fitted models, with and without the influential observation, Observation 17 (Sect. 8.12)

influential; Observation 17 (plotted with a filled triangle) has a reasonably large residual and leverage, and so it is influential.

Observe the changes in the regression coefficients after omitting Observation 17:

```
> nm.m2 <- glm( Minerab ~ Eucs, family=poisson, data=nminer,
           subset=(-maxinfl)) #  A negative index removes the obs.
> c( "Original model"=coef(nm.m1), "Without Infl"=coef(nm.m2))
Original model.(Intercept)        Original model.Eucs
                -0.8762114                  0.1139813
  Without Infl.(Intercept)          Without Infl.Eucs
                -1.0112791                  0.1247156
```

The two fitted models appear slightly different for transects with larger numbers of eucalypts (near Observation 17; Fig. 8.11, right panel):

```
> plot( Minerab ~ jitter(Eucs), data=nminer,
    xlab="Number of eucalypts", ylab="Number of noisy miners")
> newE <- seq( 0, 35, length=100)
> newM1 <- predict( nm.m1, newdata=data.frame(Eucs=newE), type="response")
> newM2 <- predict( nm.m2, newdata=data.frame(Eucs=newE), type="response")
> lines( newM1 ~ newE, lty=1); lines( newM2 ~ newE, lty=2)
```

These results suggest that the two transects with the largest number of eucalypts are important for understanding the data. Overdispersion may be an issue for these data, which we explore in Problem 10.10:

```
> c( deviance(nm.m1), df.residual(nm.m1) )
[1] 63.31798 29.00000
```

8.13 Using R for Diagnostic Analysis of GLMs

Residuals are computed in R for a fitted GLM, say `fit`, using:

- Pearson residuals r_P: `resid(fit, type="pearson")`.
- Deviance residuals r_D: `resid(fit)`, since deviance residuals are the default.
- Quantile residuals r_Q: `qresid(fit)` after loading package **statmod**.
- Partial residuals u_j: `resid(fit, type="partial")`.
- Working residuals e: `resid(fit, type="working")`.
- Response residuals $y - \hat{\mu}$: `resid(fit, type="response")`.
- Standardized deviance residuals r'_D: `rstandard(fit)`.
- Studentized deviance residuals r''_D: approximated using `rstudent(fit)`.

The longer form `residuals(fit)` is equivalent to `resid(fit)`. Each type of residual apart from `type="partial"` returns n values, one for each observation. Using `type="partial"` returns an array with n rows and a column corresponding to each β_j (apart from β_0).

Other useful R commands for diagnostics analysis, used in the same way as for linear regression models, are: `fitted(fit)` for producing fitted values; `hatvalues(fit)` for producing the leverages; `qqnorm()` for producing Q–Q plots of residuals; and `qqline()` for adding reference lines to Q–Q plots.

Measures of influence are computed for GLMs using the same R functions as for linear regression models:

- Cook's distance D: use `cooks.distance(fit)`.
- DFBETAS: use `dfbetas(fit)`.
- DFFITS: use `dffits(fit)`.
- Covariance ratio CR: use `covratio(fit)`.

All these measures of influence, together with the leverages h, are returned using `influence.measures(fit)`. Observations are flagged according to the criteria explained in Sect. 3.6.3 (p. 110).

Fitted GLMs can also `plot()`-ed (Sect. 3.16, p. 146). These commands produce four residual plots by default; see `?plot.lm`.

For remedying problems, the function `poly()` is used to create orthogonal polynomials of covariates, and `bs()` and `ns()` (both in the R package **splines**) for using regression splines in the systematic component.

Fit quasi-GLMs in R using the `glm()` function, but using specific `family` functions:

- `quasibinomial()` is used to fit quasi-binomial models. The default link function is the `"logit"` link function as for binomial GLMs. `"probit"`, `"cloglog"` (complementary log-log), `"cauchit"` and `"log"` links are also permitted, as for binomial GLMs (Sect. 9.8).
- `quasipoisson()` is used to fit quasi-Poisson models. The default link function is the `"log"` link function as for Poisson GLMs. `"identity"` and `"sqrt"` links are also permitted, as for Poisson GLMs (Sect. 10.5.3).

- `quasi()` is used to fit quasi-models more generally. Because this function is very general, any of the link functions provided by R are permitted (but may not all be sensible): `"identity"` (the default), `"logit"`, `"probit"`, `"cloglog"`, `"cauchit"`, `"log"`, `"identity"`, `"sqrt"` and `"1/mu^2"` are all permitted. Additional link functions can be defined using the `power()` function; for example, `link=power(lambda=1/3)` uses a link function of the form $\mu^{1/3} = \eta$. Using `lambda=0` is equivalent to using the logarithmic link function.
  To fit the quasi-models, the variance structure must also be defined, using for example, `family = quasi(link="log", variance="mu")`, which uses the variance function $V(\mu) = \mu$. The possible variance structures permitted for the `variance` are:
  - `"constant"`, the default, for which $V(\mu)$ is constant;
  - `"mu(1-mu)"` for which $V(\mu) = \mu(1 - \mu)$;
  - `"mu"` for which $V(\mu) = \mu$;
  - `"mu^2"` for which $V(\mu) = \mu^2$;
  - `"mu^3"` for which $V(\mu) = \mu^3$.

  Other variance functions can also be specified by writing appropriate R functions, but are rarely required and require extra effort and so are not discussed further.

The AIC is not shown in the model `summary()` for quasi-models, since the AIC is not defined for quasi-models. `summary()`, `anova()` and `glm.scoretest()` work as usual for quasi-models.

8.14 Summary

Chapter 8 discusses methods for identifying possible violations of assumptions in GLMs, and then remedying or ameliorating these problems.

The assumptions for GLMs are, in order of importance (Sect. 8.2):

- Lack of outliers: The model is appropriate for all observations.
- Link function: The correct link function $g()$ is used.
- Linearity: All important explanatory variables are included, and each explanatory variable is included in the linear predictor on the correct scale.
- Variance function: The correct variance function $V(\mu)$ is used.
- Dispersion: The dispersion parameter ϕ is constant.
- Independence: The responses y_i are independent of each other.
- Distribution: The responses y_i come from the specified EDM.

The main tool for diagnostic analysis is residuals. Pearson, deviance and quantile residuals can be used for GLMs (Sect. 8.3). Quantile residuals are highly recommended for discrete EDMs. Standardized or Studentized residuals are preferred as they have approximately constant variance (Sect. 8.6).

For GLMs, the leverages are the diagonal elements of the hat matrix $\mathrm{H} = \mathrm{W}^{1/2}\mathrm{X}(\mathrm{X}^T\mathrm{W}\mathrm{X})^{-1}\mathrm{X}^T\mathrm{W}^{1/2}$ (Sect. 8.4.2).

A strategy for diagnostic analysis of GLMs is (Sects. 8.7 and 8.9):

- Check for independence of the responses (Sect. 8.7.2). If the residuals show non-independence, use other methods.
- Plot residuals against $\hat{\mu}$ and residuals against each x_j (Sect. 8.7.3). If the variation is not constant, an incorrect EDM may have been used.
  If a trend exists, the systematic component may need changing: change the link function, add extra explanatory variables, or transform a covariates.
- To further examine the link function, plot z against $\hat{\eta}$ (Sect. 8.7.3).
- To determine if the source of the non-linearity is that covariate x_j is included on the incorrect scale, plot u_j against x_j (called a *component plus residual plot* or a *partial residual plot*) (Sect. 8.7.3).
- The choice of distribution can be checked using a Q–Q plot of quantile residuals (Sect. 8.7.4).

Outliers can be identified using Studentized residuals (Sect. 8.8). Outliers and influential observations also may be remedied by changes made to the model (Sect. 8.8). Influential observations can be identified using Cook's distance, DFFITS, DFBETAS or CR (Sect. 8.8).

Quasi-likelihood may be used when a suitable EDM cannot be identified, but information about the mean and variance is available (Sect. 8.10).

Collinearity occurs when at least some of the covariates are highly correlated with each other, implying they measure almost the same information (Sect. 8.11).

Problems

Selected solutions begin on p. 539.

8.1. Consider the Poisson distribution.

1. For $y = 0$, show that the smallest possible value of r_P is $-\sqrt{w\hat{\mu}}$.
2. For $y = 0$, show that the smallest possible value of r_D is $-\sqrt{2w\hat{\mu}}$.
3. For $y = 0$, what is the smallest value r_Q can take? Explain.
4. Comment on the normality of the residuals in light of the above results.

8.2. Show that the Pearson residuals for a gamma EDM cannot be less than $r_P = -1/\sqrt{w}$, but have no theoretical upper limit. Use these results to comment on the approximate normality of Pearson residuals for gamma EDMs. What range of values can be taken by deviance and quantile residuals?

8.3. Consider the binomial distribution.

1. Determine the deviance residuals for the binomial distribution.
2. In the extreme case $m = 1$, show that r_D will either take the value $\sqrt{2\log(1-\hat{\mu})}$ or $-\sqrt{2\log\hat{\mu}}$.

8.4. Use the R function `rpois()` to generate 1000 random numbers, say `y`, from a Poisson distribution with mean 1. Fit a Poisson GLM using the systematic component `y~1`. Then, plot the Q–Q plot of the residuals from this model using the Pearson, deviance and quantile residuals, and comment on the Q–Q plots produced using the different types of residuals. (Remember to generate more than one set of quantile residuals due to the added randomness.)

8.5. Consider the situation where the observations y come from distributions with known mean μ and known ϕ. Show that the Pearson residuals have mean zero and variance ϕ for any EDM.

8.6. The standardized deviance residual $r'_{D,i}$ is approximately the reduction in the residual deviance when Observation i is omitted from the data. Demonstrate this in R using the `trees` data as follows.

- Fit the model `cherry.m1` (Sect. 8.3.1). Compute the residual deviance, the Pearson estimate of ϕ, and the standardized deviance residuals from this model.
- Omit Observation 1 from `trees`, and refit the model. Call this model `cherry.omit1`.
- Compute the difference between the residual deviance for the full model `cherry.m1` and for model `cherry.omit1`. Show that this differences divided by the Pearson estimate of ϕ is approximately the standardized deviance residuals squared.

Repeat the above process for every observation i. At each iteration, call this model `cherry.omiti`. Then, compute the difference between the deviance for the full model `cherry.lm` and for model `cherry.omiti`. Show that these differences divided by ϕ are approximately the standardized residuals squared.

8.7. Consider the exponential distribution (4.37) defined for $y > 0$.

1. When $\mu = 3.5$ and $y = 1.5$, compute the Pearson, deviance and quantile residuals when the weights are all one.
2. When $\mu = 3.5$ and $y = 3.5$, compute the Pearson, deviance and quantile residuals when the weights are all one.
3. Comment on what the above shows.

8.8. Consider a transformation $A(y)$ of a response variable y.

1. Expand $A(y)$ about μ using the first two terms of the Taylor series to show that $A(y) - A(\mu) \approx A'(\mu)(y - \mu)$.

2. Using the previous result, compute the variance of both sides to show that
$$r_A = \frac{A(y) - A(\mu)}{A'(\mu)\sqrt{V(\mu)}},$$
called the Anscombe residual [10, 12], has a variance of ϕ approximately.
3. For GLMs, $A(t) = \int V(t)^{-1/3}(t)\,\mathrm{d}t$, where $V(\mu)$ is the variance function. Hence show that the Anscombe residuals for the Poisson distribution are
$$r_A = \frac{3(y^{2/3} - \mu^{2/3})}{2\mu^{1/6}}.$$
4. Compute the Anscombe residuals for the gamma and inverse Gaussian distributions.

8.9. Suppose a situation implies a variance function of the form $V(\mu) = \mu^2(1-\mu)^2$, where $0 < \mu < 1$ (for example, see [10, §9.2.4]). This variance function does not correspond to any known EDM.

1. Deduce the quasi-likelihood.
2. Deduce the unit deviance.

8.10. A study [16] counted the number of birds from four different species of seabirds in ten different quadrats in the Anadyr Strait (off the Alaskan coast) during summer, 1998 (Table 8.2; data set: `seabirds`). Because the responses are counts, a Poisson GLM may be appropriate.

1. Fit the Poisson GLM with a logarithmic link function, using the systematic component `Count ~ Species + factor(Quadrat)`.
2. Using the guidelines in Sect. 7.5 to determine when the Pearson and deviance residuals are expected to be adequate or poor.
3. Using this model, plot the deviance residuals against the fitted values, and also against the fitted values transformed to the constant-information scale. Using the plots, determine if the model is adequate.
4. Using the same model, plot the quantile residuals against the fitted values, and also against the fitted values transformed to the constant-information scale. Using the plots, determine if the model is adequate.
5. Comparing the plots based on the deviance and quantile residuals, which type of residual is easier to interpret?

8.11. Children were asked to build towers as high as they could out of cubical and cylindrical blocks [8, 14]. The number of blocks used and the time taken were recorded (data set: `blocks`). In this problem, only consider the number of blocks used y and the age of the child x.

In Problem 6.10, a GLM was fitted for these data. Perform a diagnostic analysis, and determine if the model is suitable.

Table 8.2 The number of each species of seabird counted in ten quadrats in the Anadyr Strait during summer, 1998 (Problem 8.10)

	Quadrat									
Species	1	2	3	4	5	6	7	8	9	10
Murre	0	0	0	1	1	0	0	1	1	3
Crested auklet	0	0	0	2	3	1	5	0	1	5
Least auklet	1	2	0	0	0	0	1	3	2	3
Puffin	1	0	1	1	0	0	3	1	1	0

8.12. Nambe Mills, Santa Fe, New Mexico [3, 15], is a tableware manufacturer. After casting, items produced by Nambe Mills are shaped, ground, buffed, and polished. In 1989, as an aid to rationalizing production of its 100 products, the company recorded the total grinding and polishing times and the diameter of each item (Table 5.3; data set: `nambeware`).

In Problem 6.11, a GLM was fitted to these data. Perform a diagnostic analysis, and determine if the model is suitable.

8.13. In Problem 3.24 (p. 157), a linear regression model was fitted to artificial data (data set: `triangle`), generated so that $\mu = \sqrt{x_1^2 + x_2^2}$; that is, x_1 and x_2 are the lengths of the sides of a right-angled triangle, and $\mathrm{E}[y] = \mu$ is the length of the hypotenuse (where some randomness has been added).

1. Based on the true relationships between the variables, write down the corresponding systematic component for fitting a GLM for modelling the hypotenuse. What link function is necessary?
2. Fit an appropriate GLM to the data, using the normal and gamma distributions to model the randomness. Which GLM is preferred?

References

[1] Box, G.E.P.: Science and statistics. Journal of the American Statistical Association **71**, 791–799 (1976)
[2] Breslow, N.E., Clayton, D.G.: Approximate inference in generalized linear mixed models. Journal of the American Statistical Association **88**(421), 9–25 (1993)
[3] Data Desk: Data and story library (DASL) (2017). URL http://dasl.datadesk.com
[4] Diggle, P.J., Tawn, J.A., Moyeed, R.A.: Model-based geostatistics. Applied Statistics **47**(3), 299–350 (1998)
[5] Dunn, P.K., Smyth, G.K.: Randomized quantile residuals. Journal of Computational and Graphical Statistics **5**(3), 236–244 (1996)

[6] Gotway, C.A., Stroup, W.W.: A generalized linear model approach to spatial data analysis and prediction. Journal of Agricultural, Biological and Environmental Statistics **2**(2), 157–178 (1997)
[7] Hardin, J.W., Hilbe, J.M.: Generalized Estimating Equations. Chapman and Hall/CRC, Boca Raton (2012)
[8] Johnson, B., Courtney, D.M.: Tower building. Child Development **2**(2), 161–162 (1931)
[9] Maron, M.: Threshold effect of eucalypt density on an aggressive avian competitor. Biological Conservation **136**, 100–107 (2007)
[10] McCullagh, P., Nelder, J.A.: Generalized Linear Models, second edn. Chapman and Hall, London (1989)
[11] McCulloch, C.E.: Generalized linear mixed models. Institute of Mathematical Statistics (2003)
[12] Pierce, D.A., Shafer, D.W.: Residuals in generalized linear models. Journal of the American Statistical Association **81**, 977–986 (1986)
[13] Pregibon, D.: Goodness of link tests for generalized linear models. Applied Statistics **29**(1), 15–24 (1980)
[14] Singer, J.D., Willett, J.B.: Improving the teaching of applied statistics: Putting the data back into data analysis. The American Statistician **44**(3), 223–230 (1990)
[15] Smyth, G.K.: Australasian data and story library (OzDASL) (2011). URL http://www.statsci.org/data
[16] Solow, A.R., Smith, W.: Detecting in a heterogenous community sampled by quadrats. Biometrics **47**(1), 311–317 (1991)
[17] Weisberg, S.: Applied Linear Regression. John Wiley and Sons, New York (1985)
[18] Williams, D.A.: Generalized linear models diagnostics using the deviance and single-case deletions. Applied Statistics **36**(2), 181–191 (1987)