# Chapter 12
# Tweedie GLMs


*. . . we cannot know if any statistical technique that we develop is useful unless we use it.*
*Box [5, p. 792]*

## 12.1 Introduction and Overview

This chapter introduces GLMs based on Tweedie EDMs. Tweedie EDMs are distributions that generalize many of the EDMs already seen (the normal, Poisson, gamma and inverse Gaussian distributions are special cases) and include other distributions also. First, Tweedie EDMs are discussed in general (Sect. 12.2), and then two subsets of the Tweedie GLMs which are important are studied: Tweedie EDMs for modelling positive continuous data for which gamma and inverse Gaussian GLMs are special cases (Sect. 12.2.3), then Tweedie EDMs for modelling continuous data with exact zeros (Sect. 12.2.4). We then follow with a description of how to use these Tweedie EDMs to fit Tweedie GLMs (Sect. 12.3).

## 12.2 The Tweedie EDMs

### *12.2.1 Introducing Tweedie Distributions*

Apart from the binomial and negative binomial distributions, the EDMs seen so far in this book have variance functions with similar forms:

- the normal distribution, where $V(\mu) = \mu^0 = 1$ (Chaps. 2 and 3);
- the Poisson distribution, where $V(\mu) = \mu^1$ (Chap. 10);
- the gamma distribution, where $V(\mu) = \mu^2$ (Chap. 11);
- the inverse Gaussian distribution, where $V(\mu) = \mu^3$ (Chap. 11).

These EDMs have power variance functions of the form $V(\mu) = \mu^\xi$, with $\xi = 0, 1, 2, 3$. More generally, any EDM with a variance function $V(\mu) = \mu^\xi$ is called a *Tweedie distribution*, or a *Tweedie* EDM, where $\xi$ can take any real



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_12

**Table 12.1** Features of the Tweedie distributions for various values of the index parameter $\xi$, showing the support $S$ (the permissible values of $y$) and the domain $\Omega$ for $\mu$. The Poisson distribution ($\xi = 1$ and $\phi = 1$) is a special case of the discrete distributions, and the inverse Gaussian distribution ($\xi = 3$) is a special case of positive stable distributions. $\mathbb{R}$ refers to the real line; superscript + means positive real values only; subscript 0 means zero is included in the space (Sect. 12.2.1)

| Tweedie EDM | $\xi$ | $S$ | $\Omega$ | Reference |
|---|---|---|---|---|
| Extreme stable | $\xi < 0$ | $\mathbb{R}$ | $\mathbb{R}^+$ | Not covered |
| Normal | $\xi = 0$ | $\mathbb{R}$ | $\mathbb{R}$ | Chaps. 2 and 3 |
| No EDMs exist | $0 < \xi < 1$ | | | |
| Discrete | $\xi = 1$ | $y = 0, \phi, 2\phi, \ldots$ | $\mathbb{R}^+$ | Chap. 10 for $\phi = 1$ |
| Poisson-gamma | $1 < \xi < 2$ | $\mathbb{R}_0^+$ | $\mathbb{R}^+$ | Sect. 12.2.3 |
| Gamma | $\xi = 2$ | $\mathbb{R}^+$ | $\mathbb{R}^+$ | Chap. 11 |
| Positive stable | $\xi > 2$ | $\mathbb{R}^+$ | $\mathbb{R}^+$ | Sect. 12.2.4 |

value except $0 < \xi < 1$ [25]. $\xi$ is called the *Tweedie index parameter* and is sometimes denoted by $p$. This power-variance relationship has been observed in natural populations for many years [36, 37]. Useful information about the Tweedie distribution appears in Table 5.1 (p. 221).

The four specific cases of Tweedie distributions listed above show that the Tweedie distributions are useful for a variety of data types (Table 12.1). More generally:

- For $\xi \leq 0$, the Tweedie distributions are suitable for modelling continuous data where $-\infty < y < \infty$. The normal distribution ($\xi = 0$) is a special case. When $\xi < 0$, the Tweedie distributions have the unusual feature that data $y$ are defined on the entire real line, but $\mu > 0$. These Tweedie distributions with $\xi < 0$ have no known realistic applications, and so are not considered further.
- For $\xi = 1$ the Tweedie distributions are suitable for modelling discrete data where $y = 0, \phi, 2\phi, 3\phi, \ldots$. When $\phi = 2$, for example, a positive probability exists for $y = 0, 2, 4, \ldots$. The Poisson distribution is a special case when $\phi = 1$.
- For $1 < \xi < 2$, the Tweedie distributions are suitable for modelling positive continuous data with exact zeros. An example is rainfall modelling [12, 31]: when no rain falls, an exact zero is recorded, but when rain *does* fall, the amount is a continuous measurement. Plots of example probability functions are shown in Fig. 12.1. As $\xi \to 1$, the densities show local maxima corresponding to the discrete masses for the corresponding Poisson distribution.
- For $\xi \geq 2$, the Tweedie distributions are suitable for modelling positive continuous data. The gamma ($\xi = 2$) and inverse Gaussian ($\xi = 3$) distributions are special cases (Chap. 11). The distributions become more right skewed as $\xi$ increases (Fig. 12.2).

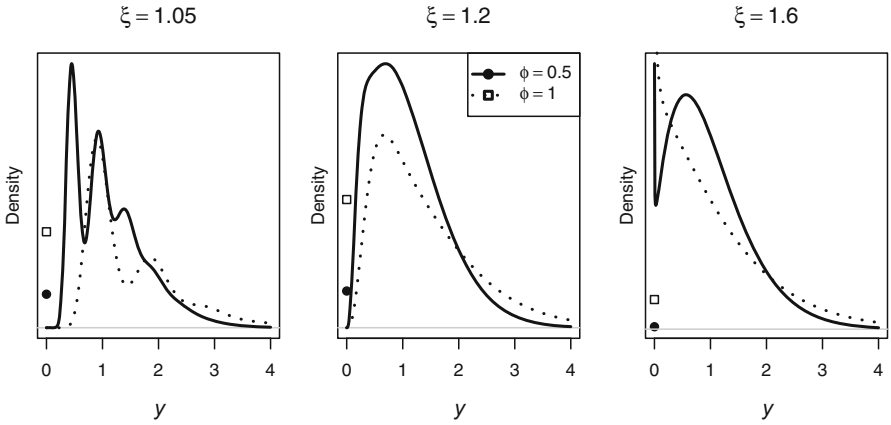


**Fig. 12.1** Examples of Tweedie probability functions with $1 < \xi < 2$ and $\mu = 1$. The solid lines correspond to $\phi = 0.5$ and the dotted lines to $\phi = 1$. The filled dots show the probability of exactly zero when $\phi = 0.5$ and the empty squares show the probability of exactly zero when $\phi = 1$ (Sect. 12.2.1)

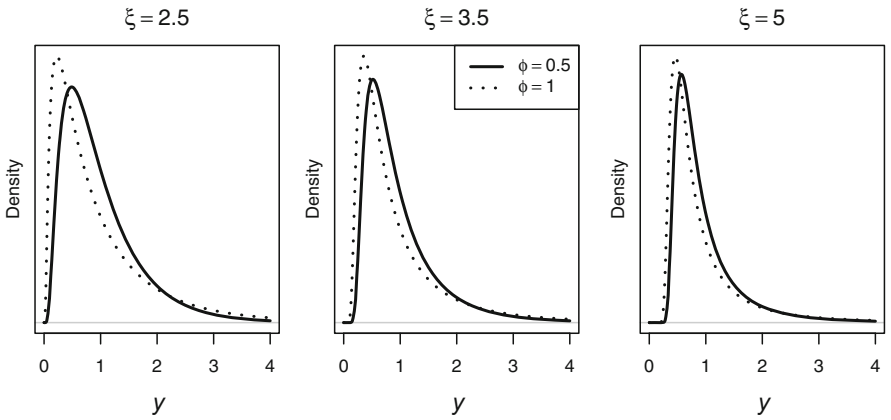


**Fig. 12.2** Examples of Tweedie probability functions with $\xi > 2$ and $\mu = 1$. As $\xi$ gets larger, the distributions become more skewed to the right. The solid lines correspond to $\phi = 0.5$; the dotted lines to $\phi = 1$ (Sect. 12.2.1)

$\xi$ is called the *Tweedie index parameter* for the Tweedie distributions, and specifies the particular distribution in the Tweedie family of distributions. The two cases $1 < \xi < 2$ and $\xi \geq 2$ are considered in this chapter in further detail. (The special cases $\xi = 0, 1, 2, 3$ were considered earlier.)

### 12.2.2 The Structure of Tweedie EDMs

Tweedie distributions are defined as EDMs with variance function $V(\mu) = \mu^\xi$ for some given $\xi$. Using this relationship, $\theta$ and $\kappa(\theta)$ can be determined (following the ideas in Sect. 5.3.6). Setting the arbitrary constants of integration to zero, obtain (Problem 12.1)

$$\theta = \begin{cases} \dfrac{\mu^{1-\xi}}{1-\xi} & \text{for } \xi \neq 1 \\ \log \mu & \text{for } \xi = 1 \end{cases} \quad \text{and} \quad \kappa(\theta) = \begin{cases} \dfrac{\mu^{2-\xi}}{2-\xi} & \text{for } \xi \neq 2 \\ \log \mu & \text{for } \xi = 2 \end{cases}. \tag{12.1}$$

Other parameterizations are obtained by setting the constants of integration to other values. One useful parameterization ensures $\theta$ and $\kappa(\theta)$ are continuous functions of $\xi$ [16] (Problem 12.2). The expressions for $\theta$ and $\kappa(\theta)$ contain $\xi$, so the Tweedie distributions are only EDMs if $\xi$ is known. In practice, the value of $\xi$ is usually estimated (Sect. 12.3.2). If $y$ follows a Tweedie distribution with index parameter $\xi$, mean $\mu$ and dispersion parameter $\phi$, write $y \sim \text{Tw}_\xi(\mu, \phi)$.

Based on these expressions for $\theta$ and $\kappa(\theta)$, the Tweedie probability function may be written in canonical form (5.1). Apart from the special cases identified earlier (the normal, Poisson, gamma and inverse Gaussian distributions), the normalizing constant $a(y, \phi)$ cannot be written in closed form. Consequently, accurate evaluation of the probability function for Tweedie EDMs in general requires numerical methods [15, 16].

The unit deviance is (Problem 12.3)

$$d(y,\mu) = \begin{cases} 2\left\{ \dfrac{\max(y,0)^{2-\xi}}{(1-\xi)(2-\xi)} - \dfrac{y\mu^{1-\xi}}{1-\xi} + \dfrac{\mu^{2-\xi}}{2-\xi} \right\} & \text{for } \xi \neq 1, 2; \\ 2\left\{ y \log \dfrac{y}{\mu} - (y-\mu) \right\} & \text{for } \xi = 1; \\ 2\left( -\log \dfrac{y}{\mu} + \dfrac{y-\mu}{\mu} \right) & \text{for } \xi = 2. \end{cases} \tag{12.2}$$

When $y = 0$, the unit deviance is finite for $\xi \leq 0$ and $1 < \xi < 2$. (Recall $y = 0$ is only admitted for $\xi \leq 0$ and $1 < \xi < 2$; see Table 12.1.)

The Tweedie probability function can be written in the form of a dispersion model (5.13) also, using the unit deviance (12.2). In this form, the normalizing constant $b(y, \phi)$ cannot be written in closed form, apart from the four special cases. By the saddlepoint approximation, $D(y, \hat{\mu}) \sim \chi^2_{n-p'}$ approximately for a model with $p'$ parameters in the linear predictor. The saddlepoint approximation is adequate if $\phi \leq \min\{y\}^{2-\xi}/3$ for the cases $\xi \geq 1$ considered in this chapter (Prob. 12.4). One consequence of this is that the approximation

is likely to be poor if any $y = 0$ (when $1 < \xi < 2$). Also, recall that $\xi = 3$ corresponds to the inverse Gaussian distribution, for which the saddlepoint approximation is exact.

Of interest is the Tweedie rescaling identity [16]. Writing $\mathcal{P}_\xi(y; \mu, \phi)$ for the probability function of a Tweedie EDM with index parameter $\xi$, then

$$\mathcal{P}_\xi(y; \mu, \phi) = c\mathcal{P}_\xi(cy; c\mu, c^{2-\xi}\phi) \tag{12.3}$$

for all $\xi$, where $y > 0$ and $c > 0$.

### *12.2.3 Tweedie EDMs for Positive Continuous Data*

In most situations, positive continuous responses are adequately modelled using a gamma or inverse Gaussian distribution (Chap. 11). In some circumstances, neither is adequate, especially for severely skewed data. However, all EDMs with variance functions of the form $\mu^\xi$ for $\xi \geq 2$ are suitable for positive continuous data. The gamma ($\xi = 2$) and inverse Gaussian ($\xi = 3$) distributions are just two special cases, and are the only examples of Tweedie EDMs with $\xi \geq 2$ with probability functions that can be written in closed form. One important example corresponds to $V(\mu) = \mu^4$, which is approximately equivalent to using the transformation $1/y$ as the response variable in a linear regression model.

*Example 12.1.* The survival times (in 10 h units) of animals subjected to three types of poison were measured [6] for four different treatments (Table 12.2; data set: `poison`). Four animals were used for each poison–treatment combination (Fig. 12.3, top panels):

```
> data(poison); summary(poison)
  Psn     Trmt        Time
 I  :16   A:12   Min.   :0.1800
 II :16   B:12   1st Qu.:0.3000
 III:16   C:12   Median :0.4000
          D:12   Mean   :0.4794
                 3rd Qu.:0.6225
                 Max.   :1.2400
```

**Table 12.2** Survival times (in 10 h units) for animals under four treatments A, B, C and D, and three poison types I, II and III (Example 12.1)

| Poison I | | | | Poison II | | | | Poison III | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | B | C | D | A | B | C | D | A | B | C | D |
| 0.31 | 0.82 | 0.43 | 0.45 | 0.36 | 0.92 | 0.44 | 0.56 | 0.22 | 0.30 | 0.23 | 0.30 |
| 0.45 | 1.10 | 0.45 | 0.71 | 0.29 | 0.61 | 0.35 | 1.02 | 0.21 | 0.37 | 0.25 | 0.36 |
| 0.46 | 0.88 | 0.63 | 0.66 | 0.40 | 0.49 | 0.31 | 0.71 | 0.18 | 0.38 | 0.24 | 0.31 |
| 0.43 | 0.72 | 0.76 | 0.62 | 0.23 | 1.24 | 0.40 | 0.38 | 0.23 | 0.29 | 0.22 | 0.33 |

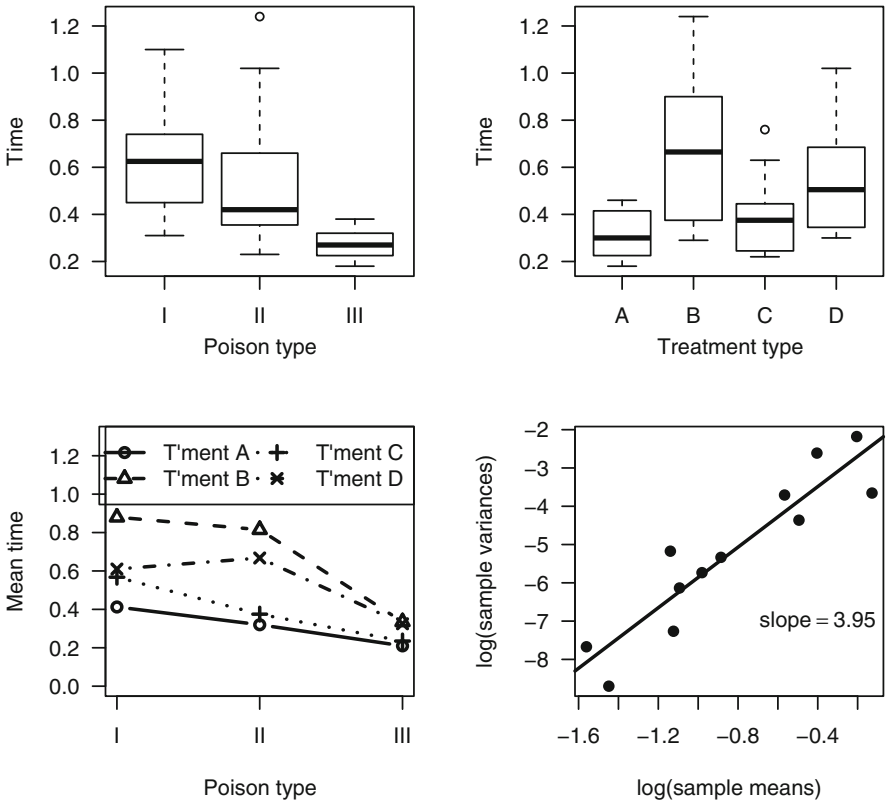


**Fig. 12.3** The poison data. The time to death plotted against poison type (top left panel); the time to death plotted against treatment type (top right panel); the mean of the time to death by poison type and treatment type (bottom left panel); the logarithm of each treatment–poison group variance plotted against the logarithm of the group means (bottom right panel) (Example 12.1)

```
> plot( Time ~ Psn,  xlab="Poison type",    las=1, data=poison )
> plot( Time ~ Trmt, xlab="Treatment type", las=1, data=poison )
> GroupMeans <- tapply(poison$Time, list(poison$Psn, poison$Trmt), "mean")
> matplot( GroupMeans, type="b", xlab="Poison type", ylab="Mean time",
          pch=1:4, col="black", lty=1:4, lwd=2, ylim=c(0, 1.3), axes=FALSE)
> axis(side=1, at=1:3, labels=levels(poison$Psn))
> axis(side=2, las=1); box()
> legend("topright", lwd=2, lty=1:4, ncol=2, pch=1:4,
        legend=c("T'ment A", "T'ment B", "T'ment C", "T'ment D"))
```

Finding the variance and the mean of the four observations in each poison–treatment combination and plotting (Fig. 12.3, bottom right panel) shows that the variance is a function of the mean:

```
>    # Find mean and var of each poison/treatment combination
> mns <- tapply(poison$Time, list(poison$Psn, poison$Trmt), mean)
```

```
> vrs <- tapply(poison$Time, list(poison$Psn, poison$Trmt), var)
>    # Plot
> plot( log(c(vrs)) ~ log(c(mns)), las=1, pch=19,
    xlab="log(sample means)", ylab="log(sample variances)")
> mvline <- lm( log( c(vrs) ) ~ log( c(mns) ) )
> slope <- round( coef( mvline )[2], 2); abline( mvline, lwd=2)
> slope
log(c(mns))
       3.95
```

The slope of this line is 3.95, suggesting a Tweedie EDM with $\xi \approx 4$ may be appropriate. □

### 12.2.4 Tweedie EDMs for Positive Continuous Data with Exact Zeros

Tweedie EDMs with $1 < \xi < 2$ are useful for modelling continuous data with exact zeros. An example of this type of data is insurance claims data [26, 34]. Assume $N$ claims are made in a particular company in a certain time frame, where $N \sim \text{Pois}(\lambda^*)$ where $\lambda^*$ is the Poisson mean number of claims in the time frame. Observe that $N$ could be zero if no claims are made. When $N > 0$, assume the amount of each claim $i = 1, \ldots, N$ is $z_i$, where $z_i$ must be positive. Assume $z_i$ follows a gamma distribution with mean $\mu^*$ and dispersion parameter $\phi^*$, so that $z_i \sim \text{Gam}(\mu^*, \phi^*)$. The *total* insurance payout $y$ is the sum of the $N$ individual claims, such that

$$y = \sum_{i=1}^{N} z_i,$$

where $y = 0$ when $N = 0$. The total claim amount $y$ has a Tweedie distribution with $1 < \xi < 2$. In this interpretation, $y$ is a Poisson sum of gamma distributions, and hence these Tweedie distributions with $1 < \xi < 2$ are sometimes called Poisson–gamma distributions [31], though this term sometimes has another, but related, meaning [17].

*Example 12.2.* The Quilpie rainfall data were considered in Example 4.6 (data set: `quilpie`), where the probability of observing at least 10 mm of total July rainfall was the quantity of interest. In this example, we examine the total July rainfall in Quilpie. Observe that the total monthly July rainfall is continuous, with exact zeros:

```
> library(GLMsData); data(quilpie)
> head(quilpie)
  Year Rain   SOI Phase Exceed y
1 1921 38.4   2.7     2    Yes 1
2 1922  0.0   2.0     5     No 0
```

```
3 1923  0.0 -10.7     3     No 0
4 1924 24.4   6.9     2    Yes 1
5 1925  0.0 -12.5     3     No 0
6 1926  9.1  -1.0     4     No 0
> sum( quilpie$Rain==0 ) # How many months with exactly zero rainfall?
[1] 20
```

For these data, a Tweedie distribution with $1 < \xi < 2$ may be appropriate. The monthly rainfall could be considered as a Poisson sum of rainfall events each July, with each event producing rainfall amounts that follow a gamma distribution. □

The parameters of the fitted Tweedie EDM defined in Sect. 12.2.2, namely $\mu$, $\phi$ and $\xi$, are related to the parameters of the underlying Poisson and gamma distributions by

$$\begin{aligned} \lambda^* &= \frac{\mu^{2-\xi}}{\phi(2-\xi)}; \\ \mu^* &= (2-\xi)\phi\mu^{\xi-1}; \\ \phi^* &= (2-\xi)(\xi-1)\phi^2\mu^{2(\xi-1)}. \end{aligned} \tag{12.4}$$

Tweedie EDMs with $1 < \xi < 2$ are continuous for $y > 0$, but have a positive probability $\pi_0$ at $y = 0$, where [15]

$$\pi_0 = \Pr(y = 0) = \exp(-\lambda^*) = \exp\left\{-\frac{\mu^{2-\xi}}{\phi(2-\xi)}\right\}. \tag{12.5}$$

To compute the MLE of $\pi_0$, the MLEs of $\mu$, $\xi$ and $\phi$ must be used in (12.5) (see the first property of MLEs in Sect. 4.9). The MLEs of $\mu$, $\xi$ and $\phi$ can be computed in R as shown in Sect. 12.3.2.

After computing the MLEs of $\mu$, $\phi$ and $\xi$, the MLEs of $\lambda^*$, $\mu^*$ and $\phi^*$ can be computed using (12.4). These estimates give an approximate interpretation of the model based on the underlying Poisson and gamma models [7, 12, 15], and may sometimes be useful (see Sect. 12.7).

## 12.3 Tweedie GLMs

### *12.3.1 Introduction*

GLMs based on the Tweedie distributions are Tweedie GLMs, specified as GLM(Tweedie, $\xi$; Link function). For both cases considered in this chapter (that is, $\xi > 2$ and $1 < \xi < 2$), we have $\mu > 0$ (Table 12.1). As a result, the usual link function used for Tweedie GLMs is the logarithmic link function. The dispersion parameter $\phi$ is usually estimated using the Pearson estimate

(though the MLE of $\phi$ is necessary for computing the MLE of the probability of exact zeros when $1 < \xi < 2$, as explained in Sect. 12.2.4).

To fit Tweedie GLMs, the particular distribution in the Tweedie family must be specified by defining the value of $\xi$, but usually the value of $\xi$ is unknown and must be estimated before the Tweedie GLM is fitted (Sect. 12.3.2). The correlation between $\hat{\xi}$ and $\hat{\beta}$ is small, so using the estimate $\hat{\xi}$ has only a small effect on inference concerning $\beta$ compared to knowing the true value of $\xi$.

Linear regression models using a Box–Cox transformation of the responses can be viewed as an approximation to the Tweedie GLM with the same underlying mean–variance relationship (Problem 12.7); see Sect. 5.8 (p. 232) and Table 5.2. In terms of inference, the normal approximation to the Box–Cox transformed responses can be quite poor when the responses cover a wide range, especially when the responses include exact zeros or near zeros. As a result, the Tweedie GLM approach can often give superior results.

### *12.3.2 Estimation of the Index Parameter $\xi$*

As noted, fitting a Tweedie GLM requires that the value of the index parameter $\xi$ be known, which identifies the specific Tweedie EDM to use. Since Tweedie distributions are defined as EDMs with $\text{var}[y] = \phi V(\mu) = \phi\mu^{\xi}$, then $\log(\text{var}[y]) = \log\phi + \xi\log\mu$. This shows that a simplistic method for estimating $\xi$ is to divide the data into a small number of groups, and plot the logarithm of the group variances against the logarithm of the group means, as used in Example 12.1 and Example 5.9 (the noisy miner data). However, the estimate of $\xi$ may depend upon how the data are divided.

Note that if exact zeros are present in the data, then $1 < \xi < 2$. However, if the data contains no exact zeros, then $\xi \geq 2$ is common but $1 < \xi < 2$ is still possible. In this situation, one interpretation is that exact zeros are feasible but simply not observed in the given data (Example 12.7).

*Example 12.3.* For the Quilpie rainfall data (data set: `quilpie`), the mean and variance of the monthly July rainfall amounts can be computed within each SOI phase, and the slope computed. An alternative approach is to compute the mean and variance of the rainfall amounts within each decade:

```
>    # Group by SOI Phase
> mn <- with( quilpie, tapply( Rain, Phase, "mean"))
> vr <- with( quilpie, tapply( Rain, Phase, "var"))
> coef( lm( log(vr) ~ log(mn) ) )
(Intercept)     log(mn)
   1.399527    1.553380
>    # Group by Decade
> Decade <- cut( quilpie$Year, breaks=seq(1920, 1990, by=10) )
> mn <- tapply( quilpie$Rain, Decade, "mean")
> vr <- tapply( quilpie$Rain, Decade, "var")
> coef( lm( log(vr) ~ log(mn) ) )
```

```
(Intercept)      log(mn)
  0.2821267    1.9459524
```

The two methods produce different estimates of $\xi$, but both satisfy $1 \leq \xi \leq 2$. □

A more rigorous method for estimating $\xi$, that uses the information in the explanatory variables and is not dependent on the arbitrary dividing of the data, is to compute the maximum likelihood estimator of $\xi$. A convenient way to organize the calculations is via the *profile likelihood* for $\xi$. Various values of $\xi$ are chosen, then the Tweedie GLM is fitted for each value of $\xi$ assuming that $\xi$ is fixed, and the log-likelihood computed at each value of $\xi$. This gives the profile log-likelihood. The value of $\xi$ giving the largest profile log-likelihood is the profile likelihood estimate. A plot of the profile log-likelihood against various values of $\xi$ is often useful.

One difficulty with this method is that the likelihood function for the Tweedie EDMs must be computed, but the probability function for Tweedie EDMs does not have a closed form (Sect. 12.2.2) except in the well-known special cases. However, numerical methods exist for accurately evaluating the Tweedie densities [15, 16], and are used in the R function `tweedie.profile()` (in package **tweedie** [13]) for computing the profile likelihood estimate of $\xi$. The use of `tweedie.profile()` is demonstrated in Example 12.4, and briefly in Example 12.5. Sometimes, estimating $\xi$ using `tweedie.profile()` may be slow, but once the estimate of $\xi$ has been determined fitting the Tweedie GLM using `glm()` is fast (as computing the value of the likelihood is not needed for estimation).

*Example 12.4.* The total monthly July rainfall at Quilpie, considered in Example 12.2 (data set: `quilpie`), is continuous but has exact zeros. Following the conclusion in Sect. 4.12 (p. 202), we consider modelling the total July rainfall as a function of the SOI phase [35]. The SOI phase is clearly of some importance (Fig. 12.4, left panel):

```
> quilpie$Phase <- factor(quilpie$Phase)   # Declare  Phase  as a factor
> plot( Rain ~ Phase, data=quilpie, ylab="Total July rainfall",
       ylim=c(0, 100), las=1)
```

Also observe that the variation is greater for larger average rainfall amounts. A suitable estimate of $\xi$ can be found using `tweedie.profile()`:

```
> library(tweedie)
> out <- tweedie.profile( Rain ~ Phase, do.plot=TRUE, data=quilpie)
```

The profile likelihood plot (Fig. 12.4, right panel) shows the likelihood is computed at a small number of $\xi$ values as filled circles, then a smooth curve is drawn through these points. The horizontal dashed line is the value of the log-likelihood at which the approximate 95% confidence interval for $\xi$ is located, using that, approximately,

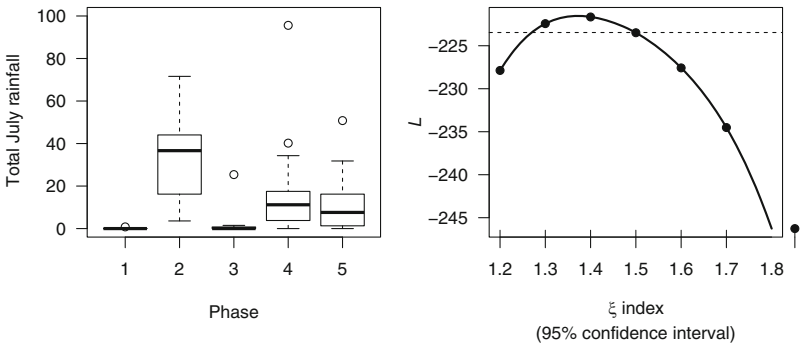


**Fig. 12.4** The total July rainfall at Quilpie plotted against SOI phase (left panel), and the profile likelihood plot for estimating $\xi$ (right panel) (Example 12.4)

$$2\left\{\ell(\hat{\xi}; y; \hat{\phi}, \hat{\mu}) - \ell(\xi; y; \hat{\phi}_\xi, \hat{\mu}_\xi)\right\} \sim \chi^2_1,$$

where $\ell(\xi; y; \hat{\phi}_\xi, \hat{\mu}_\xi)$ is the profile log-likelihood at $\xi$ and $\ell(\hat{\xi}; y; \hat{\phi}, \hat{\mu})$ is the overall maximum.

The output object, named `out` in the above, contains a lot of information (see `names(out)`), including the estimate of $\xi$ (as `xi.max`), the nominal 95% confidence interval for $\xi$ (as `ci`), and the MLE of $\phi$ (as `phi.max`):

```
>    # The index parameter, xi
> xi.est <- out$xi.max
> c( "MLE of xi" = xi.est, "CI for xi" = out$ci )
 MLE of xi CI for xi1 CI for xi2
  1.371429   1.270144   1.499132
>    # Phi
> c("MLE of phi"=out$phi.max)
MLE of phi
  5.558709
```

□

A technical difficulty sometimes arises in estimating $\xi$, which has been observed by many authors [20, 23, 26]. Recall (Sect. 12.2) that the Tweedie distribution with $\xi = 1$ is suitable for modelling discrete data where $y = 0, \phi, 2\phi, 3\phi, \ldots$. If the responses $y$ are rounded to, say, one decimal place, then the log-likelihood may be maximized by setting $\phi = 0.1$ and $\xi = 1$. Likewise, if the data are rounded to zero decimal places, then the log-likelihood may be maximized setting $\phi = 1$ and $\xi = 1$ (Example 12.5). Dunn and Smyth [15] discuss this problem in greater detail. In practice, the profile likelihood plot produced by `tweedie.profile()` should be examined, and values of $\xi$ near 1 should be avoided as necessary.

*Example 12.5.* Consider 100 observations randomly generated from a Tweedie distribution with $\xi = 1.5$, $\mu = 2$ and $\phi = 0.5$.

```
> mu <- 2; phi <- 0.5; xi <- 1.5; n <- 100
> library(tweedie)
> rndm <- rtweedie(n, xi=xi, mu=mu, phi=phi)
```

We then estimate the value of $\xi$ from the original data, and then after rounding to one and to zero decimal places (Fig. 12.5):

```
> xi.vec  <- seq(1.01, 1.75, by=0.05)
> out.est <- tweedie.profile( rndm ~ 1,           xi.vec=xi.vec)
> out.1   <- tweedie.profile( round(rndm, 1) ~ 1, xi.vec=xi.vec)
> out.0   <- tweedie.profile( round(rndm, 0) ~ 1, xi.vec=xi.vec)
```

Now compare the estimates of $\xi$ and $\phi$ for the three cases:

```
> xi.max <- out.est$xi.max
> xi.1 <- out.1$xi.max
> xi.0 <- out.0$xi.max
> compare <- array( dim=c(2, 4))
> colnames(compare) <- c("True", "Estimate", "One d.p.", "Zero d.p.")
> rownames(compare) <- c("xi", "phi")
> compare[1,] <- c(xi, xi.max, xi.1, xi.0)
> compare[2,] <- c(phi, out.est$phi.max, out.1$phi.max, out.0$phi.max)
> round(compare, 3)
    True Estimate One d.p. Zero d.p.
xi   1.5    1.696    1.710     1.010
phi  0.5    0.411    0.407     1.003
```

For these data, rounding to one decimal place only makes a small difference to the log-likelihood, and to the estimate of $\xi$. However, rounding to zero decimal places produces an artificial maximum in the log-likelihood, where $\xi \to 1$ and $\phi \to 1$. □

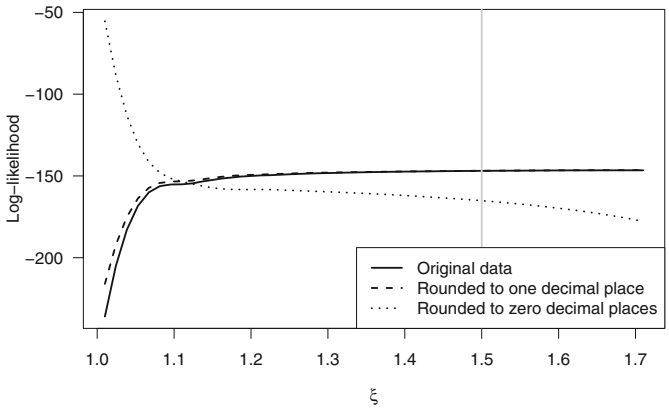


**Fig. 12.5** Estimating $\xi$ for some randomly generated data from a Tweedie distribution with $\xi = 1.5$. The gray vertical line is the true value of $\xi$ (Example 12.5)

### 12.3.3 Fitting Tweedie GLMs

Once an estimate of $\xi$ has been obtained, the Tweedie GLM can be fitted in R using the usual `glm()` function. The Tweedie distributions are denoted in R using `family=tweedie()` in the `glm()` call, after loading the **statmod** package. The call to `family=tweedie()` must specify which Tweedie EDM is to be used (that is, the value of $\xi$), using the input `var.power`; for example, `family=tweedie(var.power=3)` indicates the Tweedie EDM with $V(\mu) = \mu^3$ should be used. The link function is specified using the input `link.power`, where $\eta = \mu^{\texttt{link.power}}$. Usually, `link.power=0` which corresponds to the logarithmic link function. The logarithm link function is the most commonly-used link function with Tweedie GLMs. As usual, the default link function is the canonical link function.

Once the model has been fitted, quantile residuals [14] are recommended for diagnostic analysis, especially when $1 < \xi < 2$ when exact zeros may be present. Using more than one set of quantile residuals is recommended, due to the randomization used at $y = 0$ (Sect. 8.3.4.2).

*Example 12.6.* For the Quilpie rainfall data (data set: `quilpie`), the estimate of $\xi$ found in Example 12.4 is $\xi \approx 1.37$. To fit this model in R:

```
> xi.est <- round(xi.est, 2); xi.est
[1] 1.37
> m.quilpie <- glm( Rain ~ Phase, data=quilpie,
                    family=tweedie(var.power=xi.est, link.power=0) )
> printCoefmat(coef(summary(m.quilpie)))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.1691     1.9560 -1.1089 0.271682
Phase2        5.6923     1.9678  2.8927 0.005239 **
Phase3        3.5153     2.0600  1.7064 0.092854 .
Phase4        5.0269     1.9729  2.5480 0.013287 *
Phase5        4.6468     1.9734  2.3547 0.021665 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can compare the Pearson, deviance and quantile residuals (Fig. 12.6):

```
> dres  <-  resid(m.quilpie)    # The default residual
> pres  <-  resid(m.quilpie, type="pearson")
> qres1 <- qresid(m.quilpie)   # Quantile resids, replication 1
> qres2 <- qresid(m.quilpie)   # Quantile resids, replication 2
> qqnorm(dres,  main="Deviance residuals", las=1); qqline(dres)
> qqnorm(pres,  main="Pearson residuals",  las=1); qqline(pres)
> qqnorm(qres1, main="Quantile residuals (set 1)", las=1); qqline(qres1)
> qqnorm(qres2, main="Quantile residuals (set 2)", las=1); qqline(qres2)
```

Compare the Q–Q plot of the deviance, Pearson and quantile residuals (Fig. 12.6): the exact zeros appear as bands in the bottom left corner when using the deviance residuals. When the data contain a large number of exact zeros, this feature makes the plots of the deviance residuals hard to read.

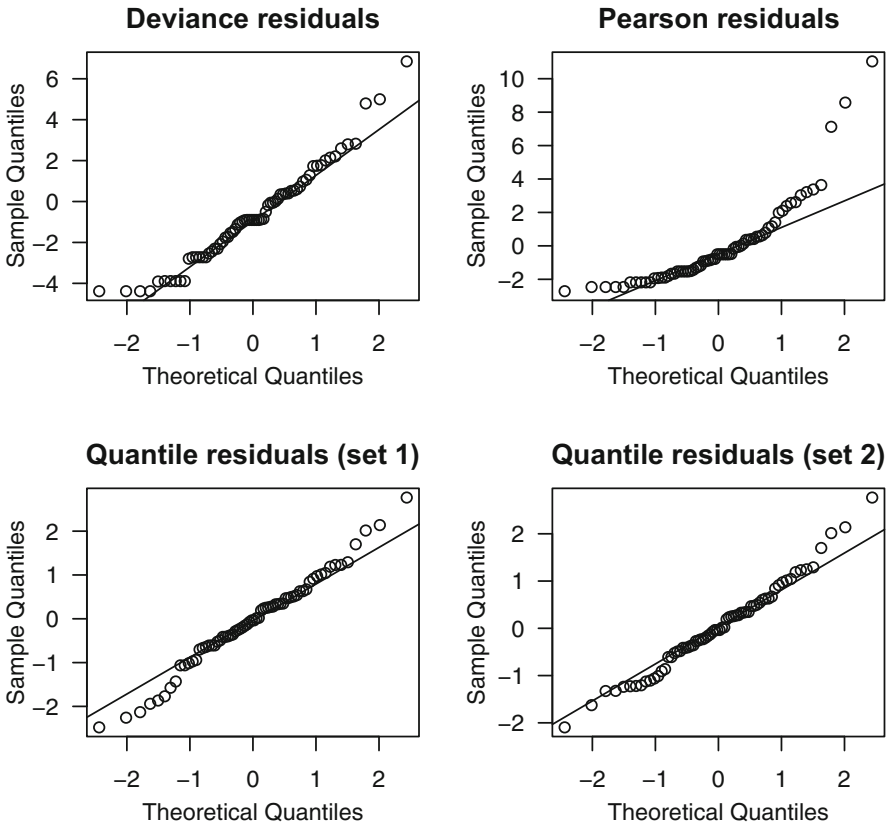


**Fig. 12.6** Q–Q plots for the Pearson, deviance and quantile residuals for the Tweedie GLM fitted to the Quilpie rainfall data. Two realization of the quantile residuals are shown (Example 12.6)

The quantile residuals use a small amount of randomization (Sect. 8.3.4.2) to remove these bands. The Q–Q plot of the quantile residuals for these data suggest the model is adequate. Q–Q plots of the other residuals make it difficult to draw definitive conclusions. For this reason, the use of quantile residuals is strongly recommended for use with Tweedie GLMs with $1 < \xi < 2$.

Other model diagnostics (Fig. 12.7) also suggest the model is reasonable:

```
> plot( qres1 ~ fitted(m.quilpie), las=1,
      xlab="Fitted values", ylab="Quantile residuals" )
> plot( cooks.distance(m.quilpie), type="h", las=1,
      ylab="Cook's distance, D")
> plot( qresid(m.quilpie) ~ factor(quilpie$Phase), las=1,
      xlab="Phase", ylab="Quantile residuals" )
```

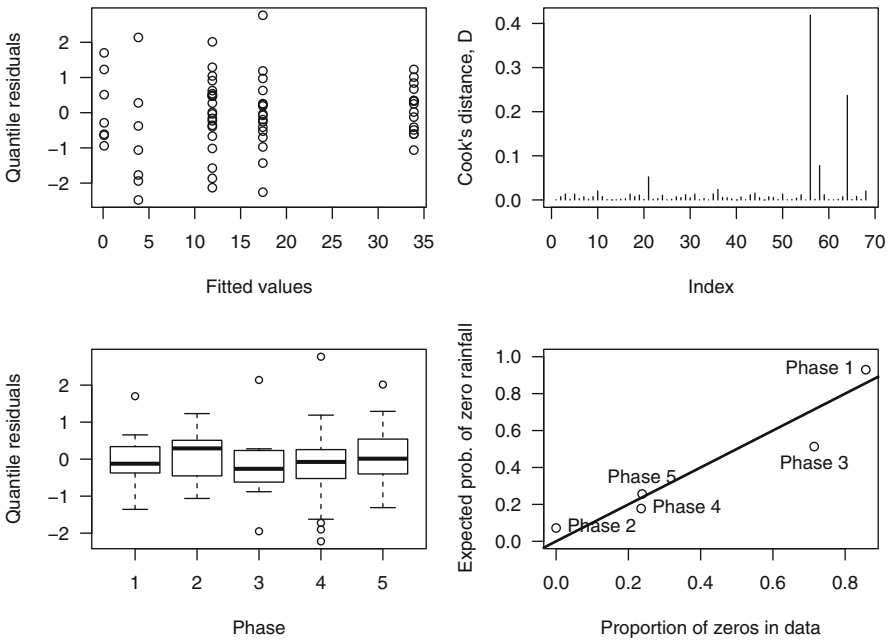


**Fig. 12.7** The diagnostics for the Tweedie GLM fitted to the Quilpie rainfall data (Examples 12.6 and 12.7)

No observations are identified as influential using Cook's distance, though DFFITS identifies one observation as influential and CV identifies eight:

```
> q.inf <- influence.measures(m.quilpie)
> colSums(q.inf$is.inf)
  dfb.1_ dfb.Phs2 dfb.Phs3 dfb.Phs4 dfb.Phs5    dffit    cov.r   cook.d
       0        0        0        0        0        1        8        0
     hat
       0
```

□

As shown in Sect. 12.2.4, Tweedie GLMs with $1 < \xi < 2$ can be developed as a Poisson sum of gamma distributions. A fitted GLM can be interpreted on this basis too.

*Example 12.7.* For the Quilpie rainfall data (data set: `quilpie`), the predicted number of zero-rainfall months $\hat{\pi}_0$ for each SOI phase can be compared to the actual proportion of months in the data with zero rainfall for each SOI phase.

To find the MLE of $\pi_0$ using (12.5), the MLE of $\phi$ must be used, which was conveniently returned by `tweedie.profile()` as `phi.max` (Example 12.4). The plot of the expected probability of a zero against the proportion of zeros in the data for each SOI phase is shown in Fig. 12.7 (bottom right panel):

```
>    # Modelled probability of P(Y=0)
> new.phase <- factor( c(1, 2, 3, 4, 5) )
> mu.phase <- predict(m.quilpie, newdata=data.frame(Phase=new.phase),
                      type="response")
> names(mu.phase) <- paste("Phase", 1:5)
> mu.phase
   Phase 1    Phase 2    Phase 3    Phase 4    Phase 5
 0.1142857 33.8937500  3.8428573 17.4235294 11.9142857
> phi.mle <- out$phi.max
> pi0 <- exp( -mu.phase^(2 - xi.est) / (phi.mle * (2 - xi.est) ) )
>    #
>    # Observed probability of P(Y=0)
> prop0 <- tapply(quilpie$Rain, quilpie$Phase,
                  function(x){sum(x==0)/length(x)})
> #
> plot( pi0 ~ prop0, xlab="Proportion of zeros in data", ylim=c(0, 1),
       ylab="Expected prob. of zero rainfall", las=1 )
> abline(0, 1, lwd=2)  # The line of equality
> text(prop0, pi0,     # Adds labels to the points
    labels=paste("Phase", levels(quilpie$Phase)),
    pos=c(2, 4, 1, 4, 3))  # These position the labels; see  ?text
```

The proportion of months with zero rainfall are predicted with reasonable accuracy. The Tweedie GLM seems a useful model for the total July rainfall in Quilpie.

As suggested in Sect. 12.2.4 (p. 463), the estimated parameters of the GLM can be used to interpret the underlying Poisson and gamma distributions. To do so, use the `tweedie.convert()` function in package **tweedie**:

```
> out <- tweedie.convert(xi=xi.est, mu=mu.phase, phi=phi.mle)
> downscale <- rbind("Poisson mean"     = out$poisson.lambda,
                     "Gamma mean"       = out$gamma.mean,
                     "Gamma dispersion" = out$gamma.phi)
> colnames(downscale) <- paste("Phase", 1:5)
> downscale
                    Phase 1   Phase 2    Phase 3   Phase 4    Phase 5
Poisson mean     0.07281493  2.628215  0.6668339  1.728229  1.3602174
Gamma mean       0.16582834  1.362530  0.6088689  1.065178  0.9254371
Gamma dispersion 1.44678583 97.673944 19.5044793 59.694036 45.0588947
```

In the context of rainfall modelling, this interpretation in terms of $\lambda^*$, $\mu^*$ and $\phi^*$ is a form of *statistical downscaling* [11]. The estimates of the Poisson mean $\lambda^*$ show the mean number of rainfall events in July when the SOI is in each phase, and the estimates of the gamma mean $\mu^*$ give the mean amount of rainfall in each rainfall event for each SOI phase. For Phase 2 the model predicts a mean of 2.628 rainfall events occur in July, with a mean of 1.363 mm in each. The mean monthly July rainfall predicted by the model agrees with the observed mean rainfall in the data:

```
> tapply( quilpie$Rain, quilpie$Phase, "mean")  # Mean rainfall from data
        1         2         3         4         5
 0.1142857 33.8937500  3.8428571 17.4235294 11.9142857
```

```
> mu.phase                                        # Mean rainfall from model
   Phase 1    Phase 2    Phase 3    Phase 4    Phase 5
 0.1142857 33.8937500  3.8428573 17.4235294 11.9142857
```

(Note that the boxplots in Fig. 12.4 show the *median* rainfall, not the *mean*.) The estimates of $\mu^*$ and $\phi^*$ are the mean and dispersion parameters for the gamma distribution fitted to the total July rainfall amount for each SOI phase.

Notice that $1 < \xi < 2$ since exact zeros are present in the data. However, exact zeros are not present in every SOI Phase:

```
> tapply(quilpie$Rain, quilpie$Phase, "min")
  1   2   3   4   5
0.0 3.6 0.0 0.0 0.0
```

In other words, even though no months with exactly zero rainfall were observed during Phase 2, the Tweedie GLM assigns a (small) probability that such an event could occur:

```
> round(out$p0, 2)
[1] 0.93 0.07 0.51 0.18 0.26
```

□

## 12.4 Case Studies

### *12.4.1 Case Study 1*

A study of performance degradation of electrical insulation from accelerated tests [28, 29, 32] measured the dialetric breakdown strength (in kilovolts) for eight time periods (in weeks) and four temperatures (in degrees Celsius). Four measurements are given for each time–temperature combination (data set: `breakdown`), and the study can be considered as a $8 \times 4$ factorial experiment.

```
> data(breakdown)
> breakdown$Time <- factor(breakdown$Time)
> breakdown$Temperature <- factor(breakdown$Temperature)
> summary(breakdown)
    Strength          Time    Temperature
 Min.   : 1.00   1      :16   180:32
 1st Qu.:10.00   2      :16   225:32
 Median :12.00   4      :16   250:32
 Mean   :11.24   8      :16   275:32
 3rd Qu.:13.53   16     :16
 Max.   :18.50   32     :16
                 (Other):32
```

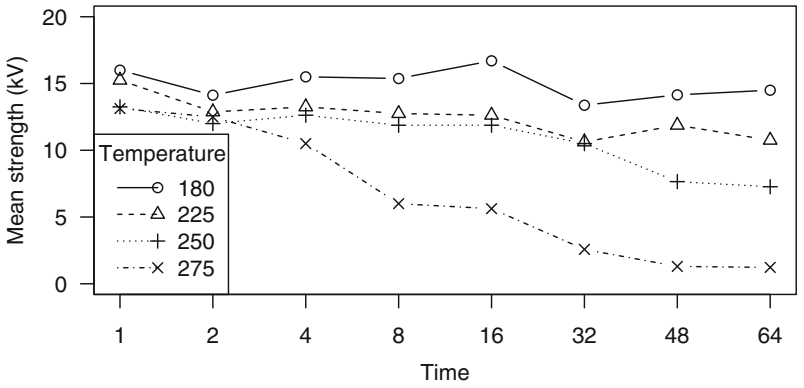


**Fig. 12.8** A plot of the dialetric breakdown data (Sect. 12.4.1)

A plot of the data (Fig. 12.8) may suggest that a temperature of 275°C is different than the rest:

```
> bd.means <- with(breakdown,
                   tapply(Strength, list(Time, Temperature), "mean"))
> matplot( bd.means, type="b", col="black",
           pch=1:4, lty=1:4, las=1, ylim=c(0, 20),
           xlab="Time", ylab="Mean strength (kV)", axes=FALSE)
> axis(side=1, at=1:8, labels=levels(breakdown$Time))
> axis(side=2, las=2); box()
> legend("bottomleft", pch=1:4, lty=1:4, merge=FALSE,
    legend=levels(breakdown$Temperature), title="Temperature" )
```

The plot also seems to show that the variance increases as `Time` increases. To consider fitting a Tweedie GLM to the data, we use `tweedie.profile()` to find an estimate of $\xi$:

```
> bd.xi <- tweedie.profile(Strength~Time*Temperature, data=breakdown,
           do.plot=TRUE, xi.vec=seq(1.2, 2, length=11))
> bd.m <- glm( Strength~factor(Time) * factor(Temperature), data=breakdown,
           family=tweedie(link.power=0, var.power=bd.xi$xi.max))
> anova(bd.m, test="F")
```

Notice that $1 < \xi < 2$ even though all breakdown strengths are positive:

```
> bd.xi$xi.max
[1] 1.591837
```

The Q–Q plot (Fig. 12.9, right panel) suggests no major problems with the model:

```
> qqnorm( resid(bd.m), las=1 ); qqline( resid(bd.m) )
```

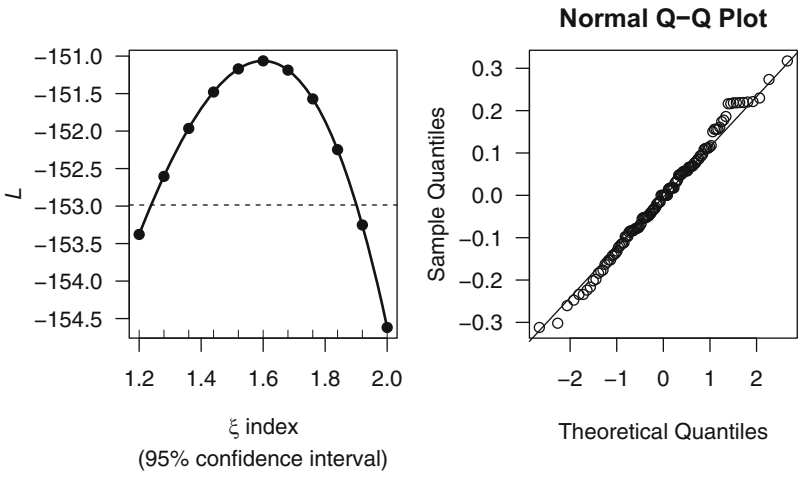


**Fig. 12.9** The profile-likelihood plot (left panel) and Q–Q plot of quantile residuals (right panel) for the dialetric breakdown data (Sect. 12.4.1)

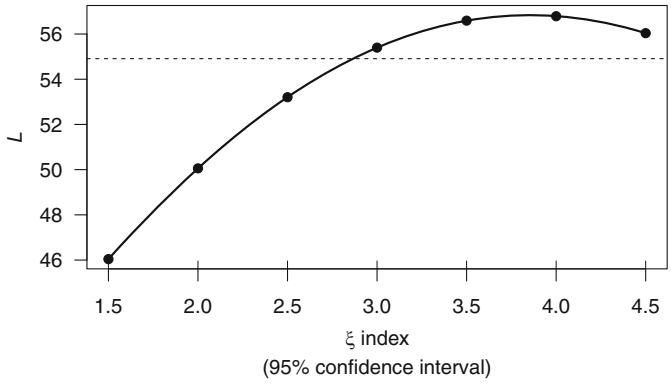


**Fig. 12.10** The profile likelihood plot for estimating the value of the Tweedie index parameter $\xi$ for the poison data (Sect. 12.4.2)

## *12.4.2 Case Study 2*

Consider the survival times data first introduced in Example 12.1, where a Tweedie EDM with $\xi \approx 4$ was suggested for modelling the data (data set: `poison`). To find the appropriate Tweedie EDM for modelling the data more formally, initially determine an estimate of $\xi$ using the profile likelihood (Fig. 12.10), using the R function `tweedie.profile()` from the package **tweedie**:

```
> data(poison)
> library(tweedie)   # To provide  tweedie.profile()
```

```
> pn.profile <- tweedie.profile( Time ~ Trmt * Psn, data=poison,
    do.plot=TRUE)
.......Done.
> c("xi: MLE"=pn.profile$xi.max, "xi: CI"=pn.profile$ci)
 xi: MLE  xi: CI1  xi: CI2
3.826531 2.866799       NA
```

These results suggest that fitting a Tweedie GLM using $\hat{\xi} = 4$ is not unreasonable:

```
> library(statmod)  # To provide the  tweedie()  family
> poison.m1 <- glm( Time ~ Trmt * Psn, data=poison,
    family=tweedie(link.power=0, var.power=4))
> anova( poison.m1, test="F")
         Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                        47     62.239
Trmt      3   19.620        44     42.619 32.7270 2.189e-10 ***
Psn       2   32.221        42     10.398 80.6195 5.053e-14 ***
Trmt:Psn  6    2.198        36      8.199  1.8334      0.12
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interaction is not significant. The fitted model without the interaction term is:

```
> poison.m2 <- update( poison.m1, . ~ Trmt + Psn )
> summary(poison.m2)
Call:
glm(formula = Time ~ Trmt + Psn, family = tweedie(link.power = 0,
    var.power = 4), data = poison)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.29925  -0.32135  -0.03321   0.20951   0.94121

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.82828    0.07938 -10.435 3.10e-13 ***
TrmtB        0.61792    0.08812   7.012 1.40e-08 ***
TrmtC        0.15104    0.06414   2.355   0.0233 *
TrmtD        0.49832    0.08053   6.188 2.13e-07 ***
PsnII       -0.22622    0.09295  -2.434   0.0193 *
PsnIII      -0.77091    0.08007  -9.628 3.43e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Tweedie family taken to be 0.2656028)

    Null deviance: 62.239  on 47  degrees of freedom
Residual deviance: 10.398  on 42  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 8
```

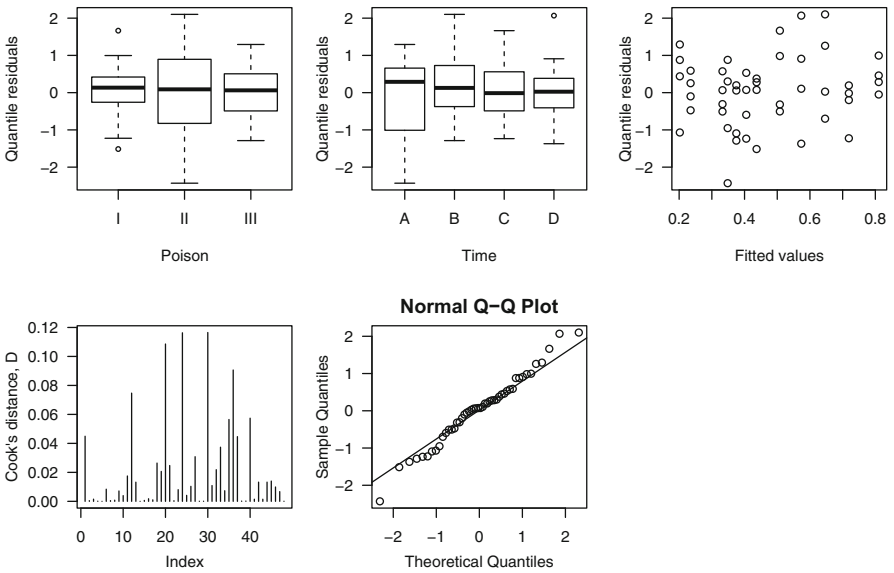


**Fig. 12.11** The diagnostics for the final model `poison.m2` fitted to the poison data (Sect. 12.4.2)

Notice the AIC is not computed by default, because the necessary numerical computations may be time consuming. However, the AIC can be computed explicitly using the function `AICtweedie()` in package **tweedie**, suggesting the non-interaction model is preferred:

```
> c("With int"     = AICtweedie(poison.m1),
    "Without int." = AICtweedie(poison.m2))
    With int Without int.
   -87.57423    -88.32050
```

The diagnostic plots suggest model `poison.m2` is adequate (Fig. 12.11), though the residuals for Poison 2 are more variable than for other poisons:

```
> plot( qresid(poison.m2) ~ poison$Psn, las=1,
       xlab="Poison", ylab="Quantile residuals" )
> plot( qresid(poison.m2) ~ poison$Trmt, las=1,
    xlab="Time", ylab="Quantile residuals" )
> plot( qresid(poison.m2) ~ fitted(poison.m2), las=1,
       xlab="Fitted values", ylab="Quantile residuals" )
> plot( cooks.distance(poison.m2), type="h", las=1,
       ylab="Cook's distance, D")
> qqnorm( qr<-qresid(poison.m2), las=1 ); qqline(qr)
```

The final model is GLM(Tweedie, $\xi = 4$; log):

$$\begin{cases} y \sim \mathrm{Tw}_{\xi=4}(\hat{\mu}, \bar{\phi} = 0.2656) & \text{(random)} \\ \log \mathrm{E}[y] = \log \hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 & \text{(systematic)} \end{cases}$$

where the $x_j$ represent dummy variables for the treatment type ($j = 1, 2, 3$) and poison type ($j = 4, 5$). Observe the Pearson estimate of $\phi$ is given in the output of `summary(poisson.m2)` as $\bar{\phi} = 0.2656$.

These data have also been analysed [6] using the Box–Cox transformation $\lambda = -1$, corresponding to $y^* = 1/y$. This transformation is the variance-stabilizing transformation approximating the Tweedie GLM with $\xi = 4$ (Table 5.2).

## 12.5 Using R to Fit Tweedie GLMs

Fitting Tweedie GLMs require extra R libraries to be installed (Sect. A.2.5):

- The **tweedie** package [13] is useful for estimating the appropriate value of $\xi$ for a given data set using the function `tweedie.profile()`.
- The **statmod** package [33] is essential for fitting Tweedie GLMs, providing the `tweedie()` GLM family function. It also provides the function `qresid()` for computing quantile residuals, whose use is strongly recommended with Tweedie GLMs.

The `tweedie.profile()` function fixes the value of $\xi$ and fits the Tweedie GLM, then computes the log-likelihood. After doing so for various values of $\xi$, the profile likelihood estimate of $\xi$ is the value producing the largest value of the log-likelihood. The function may be slow for very large data sets.

The use of `tweedie.profile()` requires a `formula` for specifying the systematic component in the same form as used for `glm()`. Other important inputs are:

- `xi.vec`: The vector of $\xi$-values to consider. By default, if the response contains zeros then `xi.vec = seq(1.2, 1.8, by=0.1)`, and if the response does not contain zeros then `xi.vec = seq(1.5, 5, by=0.5)`. The likelihood function is smoothed by default (unless `do.smooth=FALSE`) through the likelihood values computed at these values of $\xi$ given in `xi.vec`.
- `do.plot`: Indicates whether to produce a plot of the log-likelihood against $\xi$, called a *profile likelihood plot*. Producing the plot is recommended to ensure the function has worked correctly and to ensure the problem identified in Sect. 12.3.2 has not occurred. If the plot is not smooth, the `method` may need to be changed. The log-likelihood is evaluated numerically at the values of $\xi$ in `xi.vec`, and these evaluations shown with a filled circle in the profile likelihood plot if `do.plot=TRUE` (by default, `do.plot=FALSE`). An interpolation spline is drawn if `do.smooth=TRUE` (the default).
- `method`: The method used for numerically computing the log-likelihood. Occasionally the method needs to be changed explicitly to avoid difficulties (errors messages may appear; the log-likelihood may be computed as $\pm\infty$ (shown as `Inf` or `-Inf` in R); or the plot of the log-likelihood against

$\xi$ is not smooth). The options include `method = "series"`, `method = "inversion"`, or `method = "interpolation"`. The series method [15] often works well when the inversion method fails [16]. The interpolation method uses either the series or an interpolation of the inversion method results, so is often faster but may produce discontinuities in the profile likelihood plot when the computations change regimes.

- `do.ci`: Produces a nominal 95% confidence interval for the MLE of $\xi$ when `do.ci=TRUE` (which is the default).

The function `tweedie.profile()` returns numerous quantities, the most useful of which are:

- `xi.max`: The profile likelihood estimate of $\xi$.
- `phi.max`: The MLE of $\phi$.
- `ci`: The limits of the approximate 95% confidence interval for $\xi$ (returned if `do.ci=TRUE`, which is the default).

See `?tweedie.profile` for further information.

After installing the **statmod** package, specify a Tweedie GLM in R using `glm(formula, family=tweedie(var.power, link.power))`, where the value of $\xi$ is `var.power`, and `link.power` specifies the link function in the form $\mu^{\texttt{link.power}} = \eta$. Most commonly, `link.power` is zero, specifying the logarithmic link function. (The default link function is the canonical link function; Problem 12.5.) The AIC is not computed and shown in the model `summary()`, because the computations may be slow. If necessary, the AIC can be computed directly using `AICtweedie()` in package **tweedie**.

## 12.6 Summary

Chapter 12 focuses on fitting Tweedie GLMs to two types of data: Tweedie GLMs for positive continuous data, and Tweedie GLMs for positive continuous data with exact zeros.

The Tweedie distributions are EDMs with the variance function $V(\mu) = \mu^{\xi}$, for $\xi \notin (0, 1)$ (Sect. 12.2). Special cases of Tweedie distributions previously studied are the normal ($\xi = 0$), Poisson ($\xi = 1$ and $\phi = 1$), gamma ($\xi = 2$) and inverse Gaussian ($\xi = 3$) distributions (Sect. 12.2).

The unit deviance is given in (12.2). The residual deviance $D(y, \hat{\mu})$ is suitably described by a $\chi^2_{n-p'}$ distribution if $\phi \leq y^{2-\xi}/3$, but is exact when $\xi = 3$ (the inverse Gaussian distribution) (Sect. 12.2.2).

For $\xi \geq 2$, the Tweedie distributions, and hence Tweedie GLMs, are appropriate for positive continuous data. For $1 < \xi < 2$, the Tweedie distributions, and hence Tweedie GLMs, are appropriate for positive continuous data with exact zeros (Sect. 12.2).

The value of $\xi$ is estimated using the `tweedie.profile()` function from the R package **tweedie** (Sect. 12.3).

## Problems

Selected solutions begin on p. 547.

**12.1.** Deduce the expressions for $\theta$ and $\kappa(\theta)$ for the Tweedie EDMs, as given in (12.1) (p. 460), using that $V(\mu) = \mu^{\xi}$. Set the arbitrary constants of integration to zero. (HINT: Follow the approach in Sect. 5.3.6, p. 217.)

**12.2.** In Problem 12.1, expressions for $\theta$ and $\kappa(\theta)$ were found by setting the arbitrary constants of integration to zero. In this problem we consider an alternative parameterization [15].

1. By appropriately choosing the constants of integration, show that alternative expressions for $\theta$ and $\kappa(\theta)$ can be written as

$$\theta = \begin{cases} \dfrac{\mu^{1-\xi} - 1}{1-\xi} & \text{for } \xi \neq 1 \\ \log \mu & \text{for } \xi = 1 \end{cases} \quad \text{and} \quad \kappa(\theta) = \begin{cases} \dfrac{\mu^{2-\xi} - 1}{2-\xi} & \text{for } \xi \neq 2 \\ \log \mu & \text{for } \xi = 2 \end{cases} \tag{12.6}$$

2. Show that $\theta$ is continuous in $\xi$. (HINT: Use that $\lim_{\alpha \to 0}(x^{\alpha} - 1)/\alpha \to \log x$.)
3. Likewise, show that $\kappa(\theta)$ is continuous in $\xi$.

**12.3.** Deduce the unit deviance for the Tweedie EDMs given in (12.2) (p. 460).

**12.4.** Using the guideline presented in Sect. 5.4.5 (p. 226), show that the residual deviance $D(y, \hat{\mu})$ is likely to follow a $\chi^2_{n-p'}$ distribution when $\phi \leq y^{2-\xi}/3$ when $\xi \geq 1$. Hence show that the saddlepoint approximation is likely to be poor for continuous data with exact zeros.

**12.5.** Deduce the canonical link function for the Tweedie EDMs.

**12.6.** Consider the rescaling identity in (12.3).

1. Using this identity, deduce the Tweedie EDM for which the value of $\phi$ does not change when a change of measurement units (say, from grams to kilograms) is applied to the data $y$.
2. Using this identity, deduce the Tweedie EDM for which value of $\phi$ increases by the same factor as that used for a change of measurement units in the data $y$.
3. What does the identity reveal about the case of the inverse Gaussian distribution in the case of a change in measurement units in $y$?
4. Show that the probability function for any Tweedie EDM $\mathcal{P}_{\xi}(y; \mu, \phi)$ can be computed by an evaluation at $\mu = 1$ (that is, $\mathcal{P}_{\xi}(y^*; 1, \phi^*)$), by finding the appropriately-redefined values of $y^*$ and $\phi^*$.

**12.7.** Consider the Box–Cox transformation (Sect. 3.9, p. 116).

1. Show that the Box–Cox transformation for any $\lambda$ approximates fitting a GLM based on a EDM with variance function $V(\mu) = \mu^{2(1-\lambda)}$ if $\mu > 0$. (Use a Taylor series of the transformation expanded about the mean $\mu$, as in Sect. 5.8.)
2. No Tweedie EDMs exist when $0 < \xi < 1$. Use this result to show no equivalent power-variance GLM exists for the Box–Cox transformations corresponding to $0.5 < \lambda < 1$.

**12.8.** A study of monthly rainfall in Australia [22] fitted Tweedie GLMs to a number of different rainfall stations using $\hat{\xi} = 1.6$. For Bidyadanga monthly rainfall from 1912 to 2007, the fitted systematic component was

$$\log \hat{\mu}_m = 2.903 + 1.908 \sin(2\pi m/12) + 0.724 \cos(2\pi m/12),$$

where $m = 1, 2, \ldots 12$ corresponds to the month of the year (for example, February corresponds to $m = 2$). The standard errors for the parameter estimates are (respectively) 0.066, 0.090 and 0.085, and the MLE of $\phi$ is 8.33.

1. Compute the Wald statistic for testing if each regression parameter is zero.
2. Plot the value of $\hat{\mu}_m$ against $m$ for $m = 1, \ldots, 12$ for Bidyadanga.
3. Plot the predicted value of $\pi_0$ against $m$ for $m = 1, \ldots, 12$ for Bidyadanga.

**12.9.** A study [10] of the walking habits of adults living in south-east Queensland, Australia, compared different types of Statistical Areas classified by their *walk score* [9] as 'Highly walkable', 'Somewhat walkable', 'Car-dependent' or 'Very car-dependent' (Table 12.3). The Tweedie GLM was fitted using $\hat{\xi} = 1.5$.

1. Explain the differences between the predicted mean walking times in both sections of the table. Why are the predicted means all larger for the second model ('walking adults')?
2. A Tweedie GLM was fitted for 'All adults' and a gamma GLM for 'Walking adults'. Explain why these models may have been chosen.
3. The deviance from the fitted Tweedie GLM was 5976.08 on 1242 degrees of freedom. Use this information to find an estimate of $\phi$.
4. Using the Tweedie GLM, find an estimate of the proportion of all adults who did no walking in each of the four types of walkability descriptions, and comment. Why are these values not the MLEs of the $\pi_0$?

**12.10.** A study of polythene use by cosmetic companies in the UK [19] hypothesized a relationship with company turnover (Table 12.4; data set: `polythene`). Consider two Tweedie GLMs models for the data, both using a logarithmic link function for the systematic component: the first using `Polythene~Turnover`, and the second using `Polythene~log(Turnover)`.

1. Find estimates of $\xi$ for each model.

**Table 12.3** Predicted mean number of minutes of walking per day in four types of regions, adjusted for work status, household car ownership and driver's license status (Problem 12.9)

| | All adults | | Walking adults | |
|---|---|---|---|---|
| | $n$ | Predicted mean | $n$ | Predicted mean |
| Highly walkable | 214 | 7.5 | 155 | 25.5 |
| Somewhat walkable | 407 | 4.7 | 255 | 25.4 |
| Car-dependent | 441 | 2.9 | 254 | 21.2 |
| Very car-dependent | 187 | 2.5 | 90 | 18.3 |

**Table 12.4** The company turnover and polythene use for 23 cosmetic companies in the UK (to preserve confidentiality, the data were scaled) (Problem 12.10)

| Polythene use (in tonnes) | Turnover (in £00 000) | Polythene use (in tonnes) | Turnover (in £00 000) | Polythene use (in tonnes) | Turnover (in £00 000) |
|---|---|---|---|---|---|
| 0.04 | 0.02 | 31.50 | 9.85 | 587.83 | 83.94 |
| 1.60 | 0.23 | 472.50 | 21.13 | 1068.92 | 106.13 |
| 0.00 | 3.17 | 0.00 | 24.40 | 676.20 | 156.01 |
| 0.00 | 3.46 | 94.50 | 30.18 | 1056.30 | 206.43 |
| 3.78 | 3.55 | 55.94 | 40.13 | 1503.60 | 240.51 |
| 29.40 | 4.62 | 266.53 | 68.40 | 1438.50 | 240.93 |
| 8.00 | 5.71 | 252.53 | 70.88 | 2547.30 | 371.68 |
| 95.13 | 7.77 | | | 4298.70 | 391.33 |

2. Fit the GLMs to the data, and interpret the models.
3. On two separate plots of polythene use against turnover, plot the systematic components of both models, including the 95% confidence interval for the fitted lines. Comment on the models.
4. Compute the AIC for both models, and comment.
5. Produce the appropriate diagnostic plots for both models.
6. Deduce a suitable model for the data.

**12.11.** Consider the permeability of building material data given in Table 11.2 (data set: `perm`). In Sect. 11.7 (p. 440), the positive continuous response was modelled using an inverse Gaussian GLM for interpretation reasons. Jørgensen [24] also considers a gamma ($\xi = 2$) GLM for the data.

1. Determine an estimate of $\xi$ using `tweedie.profile()`. What EDM is suggested?
2. Fit a suitable Tweedie GLM ensuring an appropriate diagnostic analysis.

**12.12.** A study of human energy expenditure measured the energy expenditure $y$ of 104 females over a 24-h period (Table 12.5; data set: `energy`), and also recorded their fat-tissue mass $x_1$ and non-fat tissue $x_2$ mass [18, 24]. A model for the energy expenditure is $\mathrm{E}[y] = \beta_1 x_1 + \beta_2 x_2$, assuming the

**Table 12.5** The energy expenditure and mass of 104 females (units not given). Only the first six observations are shown (Problem 12.12)

| Energy expenditure | Mass of fat tissue | Mass of non-fat tissue |
|---|---|---|
| 60.08 | 17.31 | 43.22 |
| 60.08 | 34.09 | 43.74 |
| 63.69 | 33.03 | 48.72 |
| 64.36 | 9.14 | 50.96 |
| 65.37 | 30.73 | 48.67 |
| 66.05 | 20.74 | 65.31 |
| ⋮ | ⋮ | ⋮ |

energy expenditure for each tissue type is homogenous. Since the total mass is $M = x_1 + x_2$, divide by $M$ and rewrite as $\mathrm{E}[\bar{y}] = \beta_2 + (\beta_1 - \beta_2)\bar{x}$, where $\bar{y} = y/M$ is the energy expenditure per unit mass, and $\bar{x} = x_1/M$ is the proportion of fat-tissue mass.

1. Plot $\bar{y}$ against $\bar{x}$ and confirm the approximate linear relationship between the variables.
2. Use `tweedie.profile()` to estimate $\xi$ for the data. Which Tweedie EDMs is appropriate?
3. Find a suitable GLM for the data, ensuring a diagnostic analysis.

**12.13.** The data described in Table 12.6 (data set: `motorins1`) concern third party motor insurance claims in Sweden for the year 1977 [1, 21, 32]. The description of the data states that Swedish motor insurance companies "apply identical risk arguments to classify customers, and thus their portfolios and their claims statistics can be combined" [1, p. 413]. The data set contains 315 observations representing one of the zones in the country (covering Stockholm, Göteborg, and Malmö with surroundings).

For the remainder of the analysis, consider payments in millions of Kroner. Policies are categorized by kilometres of travel (five categories), the no-claim bonus (seven categories) and make of car (nine categories), for a total of 315 categories. Of these, 20 contain exactly zero claims, so the total payout in those categories is exactly zero; in other categories, the total payout can be consider continuous. Find an appropriate model for the data. (HINT: You will need to change the range of $\xi$ values considered by `tweedie.profile()` using the `xi.vec` input.)

Using your fitted model, interpret the model using the parameters of the underlying Poisson and gamma distributions. (HINT: See (12.4), p. 464.)

**12.14.** The total monthly August rainfall for Emerald (located in Queensland, north eastern Australia) from 1889 to 2002 is shown in Table 12.7 (data set: `emeraldaug`) with the monthly average southern oscillation index (SOI). Negative values of the SOI often indicate El Niño episodes, which are often associated with reduced rainfall in eastern and northern Australia [27].

**Table 12.6** A description of the variables used in the Swedish insurance claims data set (Problem 12.13)

| Variable | Description |
|---|---|
| `Kilometres`: | Kilometres travelled per year:<br>1: Less than 1000<br>2: 1000–15,000<br>3: 15,000–20,000<br>4: 20,000–25,000<br>5: More than 25,000 |
| `Bonus`: | No claims bonus; the number of years since last claim, plus one |
| `Make`: | 1–8 represent eight different common car models. All other models are combined in class 9 |
| `Insured`: | Number of insured in policy-years |
| `Claims`: | Number of claims |
| `Payment`: | Total value of payments in Skr (Swedish Kroner) |

**Table 12.7** The total monthly rainfall in August from 1889–2002 in Emerald, Australia, plus the monthly average SOI and corresponding SOI phases. The first five observations are shown (Problem 12.14)

| Year | Rain (in mm) | SOI | SOI phase |
|---|---|---|---|
| 1889 | 15.4 | 2.1 | 5 |
| 1890 | 47.5 | −3.1 | 5 |
| 1891 | 45.7 | −8.9 | 5 |
| 1892 | 0.0 | 5.9 | 2 |
| 1893 | 108.7 | 7.8 | 2 |
| ⋮ | ⋮ | ⋮ | ⋮ |

1. Argue that the Poisson–gamma models are appropriate for monthly rainfall data, along the lines of the argument in Sect. 12.2.4 (p. 463).
2. Perform a hypothesis test to address the relationship between rainfall and SOI given earlier in the question to see if it applies at Emerald: "Negative values of the SOI... are often associated with reduced rainfall in eastern and northern Australia."
3. Fit an appropriate EDM for modelling the total monthly August rainfall in Emerald from the SOI.
4. Compute the 95% confidence interval for the SOI parameter, and determine the practical importance of SOI for August rainfall in Emerald.
5. Fit an appropriate EDM for modelling the total monthly August rainfall in Emerald from the SOI phases.
6. Interpret the fitted model using SOI phases, using the parameters of the underlying Poisson and gamma distributions. (HINT: See (12.4), p. 464.)

**Table 12.8** Data from 194 trawls in the South East Fisheries ecosystem regarding the catch of tiger flathead. Distance is measured north to south on the 100 m depth contour (Problem 12.15)

| Longitude of trawl | Latitude of trawl | Depth (in m) | Distance (in m) | Swept area (in ha) | Number of tiger flathead | Biomass of tiger flathead (in kg) |
|---|---|---|---|---|---|---|
| 149.06 | −37.81 | −33 | 91 | 4.72260 | 1 | 0.02 |
| 149.08 | −37.83 | −47 | 90 | 5.00040 | 0 | 0.00 |
| 149.11 | −37.87 | −74 | 89 | 6.11160 | 153 | 30.70 |
| 149.22 | −38.02 | −117 | 88 | 5.83380 | 15 | 7.77 |
| 149.27 | −38.19 | −212 | 88 | 3.04222 | 0 | 0.00 |
| 150.29 | −37.41 | −168 | 48 | 6.11160 | 25 | 6.90 |
| 150.19 | −37.33 | −113 | 48 | 5.83380 | 53 | 15.30 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

**12.15.** A study on the South East Fisheries ecosystem near Australia [4] collected data about the number of fish caught from fish trawl surveys. One analysis of these data [17] studied the number of tiger flathead (Table 12.8; data set: `flathead`).

1. The data record the number of flathead caught per trawl plus the total biomass of the flathead caught. Propose a mechanism for the total biomass that leads to the Tweedie GLM as a possible model (similar to that used in Sect. 12.2.4).
2. The paper that analysed the data [17] fits a Poisson GLM to model the number of tiger flathead caught. The paper states

   > ... the dependence on covariates, if any, is specified using orthogonal polynomials in the linear predictor. The dependency on depth used a second order polynomial and the dependency on along-coast used a third order polynomial... The log of the area swept variable was included as an offset (p. 542).

   Explain why area is used as an offset.
3. Based on the information above, fit an appropriate Poisson GLM for modelling the *number* of tiger flathead caught (using `Depth` and `Distance` as covariates, in the manner discussed in the quote above). Show that this model has large overdispersion, and hence fit a quasi-Poisson model. Propose a reason why overdispersion is observed.
4. Based on the above information, plot the logarithm of biomass against the depth and distance, and comment on the relationships.
5. The paper that analysed the biomass data [17] stated that

   > There is no reason to include an extra spatial dimension... as it would be highly confounded with depth (p. 541).

   Determine if any such correlation exists between depth, and the latitude and longitude.

**Table 12.9** Feeding rates (in feeds per hour) of chestnut-crowed babblers (Problem 12.16)

| Feeding rate | Observation time (h) | Sex | Chick age (days) | Non-breeding birds ages | Brood size |
|---|---|---|---|---|---|
| 0.000 | 11.09 | M | 1 | Adult | 3 |
| 0.000 | 11.16 | M | 2 | Adult | 4 |
| 0.000 | 12.81 | M | 3 | Adult | 1 |
| 0.238 | 12.59 | M | 4 | Adult | 1 |
| 1.316 | 12.16 | M | 5 | Adult | 1 |
| 1.041 | 11.53 | M | 6 | Adult | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 0.321 | 6.22 | F | 19 | Adult | 3 |
| 0.000 | 6.22 | M | 19 | Yearling | 3 |

6. The paper that analysed the biomass data [17] used a Tweedie GLM (using `Depth` and `Distance` as covariates, in the manner discussed in the quote above). Based on the above information, fit a suitable Tweedie GLM, and assess the model using diagnostics.
7. Compare the Q–Q plot of the deviance and quantile residuals from the Tweedie GLM, and comment.

**12.16.** Chestnut-crowned babblers are medium-sized Australian birds that live in social groups. A study of their feeding habits [8] recorded, among other things, the rates at which they fed, in feeds per hour (Table 12.9; data set: `babblers`). About 18% of the feeding rates are exact zeros. Fit a Tweedie GLM to the data to model the feeding rates.

**12.17.** A study comparing two different types of toothbrushes [2, 30] measured the plaque index for females and males before and after brushing (Table 12.10; data set: `toothbrush`). Smaller values mean cleaner teeth. The 26 subjects all used both toothbrushes. One subject received the same plaque index before and after brushing.

Assuming the plaque index cannot become worse after brushing, fit an appropriate GLM to the data for modelling the difference (Before − After), and deduce if the toothbrushes appear to differ in their teeth-cleaning ability, and if this seems related to the sex of the subject.

**12.18.** An experiment [3] to quantify the effect of ketamine (an anaesthetic) measured the amount of sleep (in min) for 30 guinea pigs, using five different doses (Table 12.11; data set: `gpsleep`).

1. Explain what the exact zeros mean.
2. Plot the data, and show that the variance increases with the mean.
3. Plot the logarithm of the group variances against the logarithm of the group means, where the groups are defined by the doses. Show this implies $\xi \approx 1$.

**Table 12.10** The plaque index before and after brushing for two types of toothbrushes; smaller values indicate cleaner teeth (Problem 12.17)

| Conventional brush | | | | Hugger (new) brush | | | |
|---|---|---|---|---|---|---|---|
| Females | | Males | | Females | | Males | |
| Before | After | Before | After | Before | After | Before | After |
| 1.20 | 0.75 | 3.35 | 1.58 | 2.18 | 0.43 | 0.90 | 0.15 |
| 1.43 | 0.55 | 1.50 | 0.20 | 2.05 | 0.08 | 0.58 | 0.10 |
| 0.68 | 0.08 | 4.08 | 1.88 | 1.05 | 0.18 | 2.50 | 0.33 |
| 1.45 | 0.75 | 3.15 | 2.00 | 1.95 | 0.78 | 2.25 | 0.33 |
| 0.50 | 0.05 | 0.90 | 0.25 | 0.28 | 0.03 | 1.53 | 0.53 |
| 2.75 | 1.60 | 1.78 | 0.18 | 2.63 | 0.23 | 1.43 | 0.43 |
| 1.25 | 0.65 | 3.50 | 0.85 | 1.50 | 0.20 | 3.48 | 0.65 |
| 0.40 | 0.13 | 2.50 | 1.15 | 0.45 | 0.00 | 1.80 | 0.20 |
| 1.18 | 0.83 | 2.18 | 0.93 | 0.70 | 0.05 | 1.50 | 0.25 |
| 1.43 | 0.58 | 2.68 | 1.05 | 1.30 | 0.30 | 2.55 | 0.15 |
| 0.45 | 0.38 | 2.73 | 0.85 | 1.25 | 0.33 | 1.30 | 0.05 |
| 1.60 | 0.63 | 3.43 | 0.88 | 0.18 | 0.00 | 2.65 | 0.25 |
| 0.25 | 0.25 | | | 3.30 | 0.90 | | |
| 2.98 | 1.03 | | | 1.40 | 0.24 | | |

**Table 12.11** Amount of sleep (in min) for 30 guinea pigs after receiving intravenous doses of ketamine (Problem 12.18)

| 0.60 mg/kg | | 1.04 mg/kg | | 1.44 mg/kg | | 2.00 mg/kg | | 2.75 mg/kg | |
|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 3.60 | 5.59 | 7.67 | 0.00 | 1.71 |
| 0.00 | 0.00 | 2.85 | 5.92 | 8.32 | 8.50 | 9.40 | 9.77 | 11.15 | 11.89 |
| 3.99 | 4.78 | 7.36 | 10.43 | 12.73 | 13.20 | 10.92 | 24.80 | 14.48 | 14.75 |

4. Using `tweedie.profile()`, show that $\hat{\xi} = 1.1$. (HINT: Try using `xi.vec = (1.02, 1.4, by=0.02)` to ensure you obtain a good estimate of $\xi$.)
5. Show that a quadratic Tweedie GLM in `Dose` is significantly better than the Tweedie GLM linear is `Dose`.
6. Also consider the linear and quadratic Tweedie GLM using `log(Dose)` in place of `Dose`.
7. Also consider a Tweedie GLM using a natural cubic spline, with `knots= quantile(Dose, c(0.33, 0.67))`.
8. Plot all five systematic component on a plot of the data, and comment.
9. Use the AIC to determine a model from the five considered, and show the quadratic model in `Dose` is the preferred model.

## References


[1] Andrew, D.F., Herzberg, A.M.: Data: A Collection of Problems from Many Fields for the Student and Research Worker. Springer, New York (1985)

[2] Aoki, R., Achcar, J.A., Bolfarine, H., Singer, J.M.: Bayesian analysis of null-intercept errors-in-variables regression for pretest/post-test data. Journal of Applied Statistics **31**(1), 3–12 (2003)

[3] Bailey, R.C., Summe, J.P., Hommer, L.D., McCracken, L.E.: A model for the analysis of the anesthetic response. Biometrics **34**(2), 223–232 (1978)

[4] Bax, N.J., Williams, A.: Habitat and fisheries production in the South East fishery ecosystem. Final Report 1994/040, Fisheries Research and Development Corporation (2000)

[5] Box, G.E.P.: Science and statistics. Journal of the American Statistical Association **71**, 791–799 (1976)

[6] Box, G.E.P., Cox, D.R.: An analysis of transformations (with discussion). Journal of the Royal Statistical Society, Series B **26**, 211–252 (1964)

[7] Brown, J.E., Dunn, P.K.: Comparisons of Tobit, linear, and Poisson-gamma regression models: an application of time use data. Sociological Methods & Research **40**(3), 511–535 (2011)

[8] Browning, L.E., Patrick, S.C., Rollins, L.A., Griffith, S.C., Russell, A.F.: Kin selection, not group augmentation, predicts helping in an obligate cooperatively breeding bird. Proceedings of the Royal Society B **279**, 3861–3869 (2012)

[9] Carr, L.J., Dunsiger, S.I., Marcus, B.H.: Validation of Walk Score for estimating access to walkable amenities. British Journal of Sports Medicine **45**(14), 1144–1148 (2011)

[10] Cole, R., Dunn, P., Hunter, I., Owen, N., Sugiyama, T.: Walk score and Australian adults' home-based walking for transport. Health & Place **35**, 60–65 (2015)

[11] Connolly, R.D., Schirmer, J., Dunn, P.K.: A daily rainfall disaggregation model. Agricultural and Forest Meteorology **92**(2), 105–117 (1998)

[12] Dunn, P.K.: Precipitation occurrence and amount can be modelled simultaneously. International Journal of Climatology **24**, 1231–1239 (2004)

[13] Dunn, P.K.: tweedie: Tweedie exponential family models (2017). URL https://CRAN.R-project.org/package=tweedie. R package version 2.3.0

[14] Dunn, P.K., Smyth, G.K.: Randomized quantile residuals. Journal of Computational and Graphical Statistics **5**(3), 236–244 (1996)

[15] Dunn, P.K., Smyth, G.K.: Series evaluation of Tweedie exponential dispersion models. Statistics and Computing **15**(4), 267–280 (2005)

[16] Dunn, P.K., Smyth, G.K.: Evaluation of Tweedie exponential dispersion models using Fourier inversion. Statistics and Computing **18**(1), 73–86 (2008)
[17] Foster, S.D., Bravington, M.V.: A Poisson–gamma model for analysis of ecological data. Environmental and Ecological Statistics **20**(4), 533–552 (2013)
[18] Garby, L., Garrow, J.S., Jørgensen, B., Lammert, O., Madsen, K., Sørensen, P., Webster, J.: Relation between energy expenditure and body composition in man: Specific energy expenditure in *vivo* of fat and fat-free tissue. European Journal of Clinical Nutrition **42**(4), 301–305 (1988)
[19] Gilchrist, R.: Regression models for data with a non-zero probability of a zero response. Communications in Statistics—Theory and Methods **29**, 1987–2003 (2000)
[20] Gilchrist, R., Drinkwater, D.: Fitting Tweedie models to data with probability of zero responses. In: H. Friedl, A. Berghold, G. Kauermann (eds.) Statistical Modelling: Proceedings of the 14th International Workshop on Statistical Modelling, pp. 207–214. International Workshop on Statistical Modelling, Gräz (1999)
[21] Hallin, M., François Ingenbleek, J.: The Swedish automobile portfolio in 1997. Scandinavian Actuarial Journal pp. 49–64 (1983)
[22] Hasan, M.M., Dunn, P.K.: A simple Poisson–gamma model for modelling rainfall occurrence and amount simultaneously. Agricultural and Forest Meteorology **150**, 1319–1330 (2010)
[23] Jørgensen, B.: Exponential dispersion models (with discussion). Journal of the Royal Statistical Society, Series B **49**, 127–162 (1987)
[24] Jørgensen, B.: Exponential dispersion models and extensions: A review. International Statistical Review **60**(1), 5–20 (1992)
[25] Jørgensen, B.: The Theory of Dispersion Models. Monographs on Statistics and Applied Probability. Chapman and Hall, London (1997)
[26] Jørgensen, B., de Souza, M.C.P.: Fitting Tweedie's compound Poisson model to insurance claims data. Scandinavian Actuarial Journal **1**, 69–93 (1994)
[27] McBride, J.L., Nicholls, N.: Seasonal relationships between Australian rainfall and the southern oscillation. Monthly Weather Review **111**(10), 1998–2004 (1983)
[28] National Institute of Standards and Technology: Statistical reference datasets (2016). URL http://www.itl.nist.gov/div898/strd
[29] Nelson, W.: Analysis of performance-degradation data from accelerated tests. IEEE Transactions on Reliability **30**(2), 149–155 (1981)
[30] Singer, J.M., Andrade, D.F.: Regression models for the analysis of pretest/posttest data. Biometrics **53**, 729–725 (1997)

[31] Smyth, G.K.: Regression analysis of quantity data with exact zeros. In: Proceedings of the Second Australia–Japan Workshop on Stochastic Models in Engineering, Technology and Management, pp. 572–580. Technology Management Centre, University of Queensland, Brisbane (1996)

[32] Smyth, G.K.: Australasian data and story library (OzDASL) (2011). URL http://www.statsci.org/data

[33] Smyth, G.K.: statmod: Statistical Modeling (2017). URL https://CRAN.R-project.org/package=statmod. R package version 1.4.30. With contributions from Yifang Hu, Peter Dunn, Belinda Phipson and Yunshun Chen.

[34] Smyth, G.K., Jørgensen, B.: Fitting Tweedie's compound Poisson model to insurance claims data: Dispersion modelling. In: Proceedings of the 52nd Session of the International Statistical Institute. Helsinki, Finland (1999). Paper Meeting 68: Statistics and Insurance

[35] Stone, R.C., Auliciems, A.: SOI phase relationships with rainfall in eastern Australia. International Journal of Climatology **12**, 625–636 (1992)

[36] Taylor, L.R.: Aggregation, variance and the mean. Nature **189**, 732–735 (1961)

[37] Tweedie, M.C.K.: The regression of the sample variance on the sample mean. Journal of the London Mathematical Society **21**, 22–28 (1946)