MA5771: Applied Generalized Linear Model

Week 5 Instruction Contents

Lesson 5.1 Models for Proportions: Binomial GLMs (Continued)

Related Readings: Sections 9.6, 9.7, 9.8, 9.9, 9.10, and 9.11 in Chapter 9.

Related Readings: Section 8.10 in Chapter 8

Section 5.1.1 Introduction and Overview

In this lesson, we will continue our discussion on the binomial GLM which is the most commonly used GLMs. In **Section 5.1.2**, we illustrate how to use binomial GLMs to estimate median effective dose and use the complementary log-log link in assay analysis. In **Section 5.1.3**, we discuss two problems of binomial GLMs: Wald tests can fail when the fitted value $\hat{\mu}_i$ are close to 1 or 0 and there is no goodness-of-fit for binary response. In **Section 5.1.4**, we discuss the overdispersion problem and introduce the methods based on quasi-likelihood. The lesson is concluded with a case study in **Section 5.1.5.**

Section 5.1.2 Applications of Binomial GLMs

Median Effective Dose, ED50

Binomial GLMs are commonly used to examine the relationship between the dose d of a drug or poison and the proportion y of insects (or plants, or animals) that survive. These models are called dose-response models. Associated with these experiments is the concept of the **median effective dose, ED50**: the dose of poison affecting 50% of the insects. Different fields use different names for similar concepts, such as **median lethal dose (LD50)** or **median lethal concentration (LC50)**. Here, for simplicity, we use ED50 to refer to any of these quantities. The ED50 concept can be applied to other contexts also. By definition, $\mu = 0.5$ at the ED50.

For a binomial GLM using a logit link function, $\eta = \text{logit}(\mu) = 0$ when $\mu = 0.5$. Writing the linear predictor as $\eta = \beta_0 + \beta_1 d$ where d is the dose, then solving for the dose d shows that

$$ED50 = -\frac{\beta_0}{\beta_1}$$

More generally, the dose effective on any proportion ρ of the population, denoted ED(ρ), is estimated by

$$ED(\rho) = \frac{g(\rho) - \beta_0}{\beta_1}$$

where $g()$ is the link function. This is because $\eta = g(\rho) = \beta_0 + \beta_1 d$. By solving this equation and replacing the parameters by their estimates, we can obtain the estimate of $ED(\rho)$ for any given ρ and the link function.

Example 5.1: ED50 of Binomial GLMs with Complementary Log-log Link

We have

$$ED50 = \frac{g(0.5) - \beta_0}{\beta_1}$$

For the complementary log-log link,

$$g(0.5) = \log(-\log(1 - 0.5)) = \log(-\log(0.5)) = \log(\log(2))$$

So

$$ED50 = \frac{\log(\log(2)) - \beta_0}{\beta_1}$$

The estimator of $ED(\rho)$ is $\widehat{ED}(\rho) = \frac{g(\rho) - \hat{\beta}_0}{\hat{\beta}_1}$. It can be difficult to calculate the standard error of $ED(\rho)$ since it is not a linear function of $\hat{\beta}_0$ and $\hat{\beta}_1$. The R function `dose.p()` in the R package **MASS** conveniently returns $\widehat{ED}(\rho)$ and the corresponding estimated standard error. Then the $100(1 - \alpha)\%$ confidence interval of $ED(\rho)$ is:

$$ED(\rho) \pm z_{\frac{\alpha}{2}} * se(ED(\rho))$$

Example 5.2: ED50 of Binomial GLMs

Consider the turbine data again (data set: `turbines`). The ED50 corresponds to the run time for which 50% of turbines would be expected to experience fissures for the commentary log-log link function is: $\widehat{ED}50 = 3993.575$. The corresponding standard error is: 137.8352 (refer R program for more details).

Therefore, the 95% confidence interval of $ED50$ is:

$$\widehat{ED}50 \pm z_{\frac{\alpha}{2}} * se\left(\widehat{ED}50\right) = 3993.575 \pm 1.96 * 137.8352 = (3723.423, 4263.727)$$

Therefore, it is estimated that running the turbines for approximately 3993 hours would produce fissures in about 50% of the turbines. The logit and probit link functions produce similar estimates of ED50.

Exercise 5.1: ED50 of Binomial GLMs

Consider the turbine data again (data set: `turbines`) and the logit link function. Verify that the point estimate, the standard error, and the 95% confidence interval of ED50 are $3926.59, 158.0138$, and $(3616.84, 4236.34)$, respectively.

Complementary Log-Log Link in Assay Analysis

A common problem in biology is to determine the proportion of cells or organisms of interest amongst a much larger population. For example, what is the frequency of adult stem cells in a sample of tissue? Dilution assays are an experimental technique to estimate the frequency of active cells. The idea is to dilute the sample down to the point where some assays yield a positive result (so at least one active cell is present) and some yield a negative result (so no active cells are present). Denote:

- λ: the proportion of active cells in the cell population
- d_i: does, the number of cells of cultures used in assays
- m_i: independent cultures are conducted at dose d_i
- μ_i: the probability of a positive result at dose d_i
- $1 - \mu_i$: the probability of a negative results, or equivalently, the probability of no active cells in the assay

The expected number of active cells in the culture with the dose d_i is λd_i. If the cells behave independently (that is, if there are no community effects amongst the cells), and if the cell dose is controlled simply by dilution, then the actual number of cells in each culture will vary according to a Poisson distribution. A culture will give a negative result only if there are no active cells in the assay. The Poisson probability formula tells us that this occurs with probability

$$\text{probaility of no active cells} = 1 - \mu_i = \exp(-\lambda d_i)$$

This formula can be linearized by taking logarithms of both sides, as

$$\log(1 - \mu_i) = -\lambda d_i$$

or, taking logarithms again,

$$\log(-\log(1 - \mu_i)) = \log(\lambda) + \log(d_i)$$

Therefore, the proportion of active cells can be estimated by fitting a binomial GLM with a complementary log-log link:

$$g(\mu_i) = \log(-\log(1 - \mu_i)) = \beta_0 + \log(d_i)$$

where $\beta_0 = \log(\lambda)$, $\log(d_i)$ is an offset, and $g()$ is the complementary log-log link function.

In principle, a GLM could also have be fitted using a logarithmic link function. However the use of the complementary log-log link is superior, because it leads to a GLM without any constraints on the coefficient β_0.

The point and interval estimates of proportion of active cells (λ) are:

$\hat{\lambda} = \exp(\hat{\beta}_0)$ and $\exp(\hat{\beta}_0 \pm z_{\frac{\alpha}{2}} * se(\hat{\beta}_0))$

We can also obtain the point and interval estimates of $1/\lambda$, which represents the number of cells required on average to obtain one responding cell.

The dilution assay model assumes that a single active cell is sufficient to achieve a positive result, so it is sometimes called the single-hit model. One way to check this model is to fit a slightly larger model in which the offset coefficient is not set to one:

$$g(\mu_i) = \log(-\log(1 - \mu_i)) = \beta_0 + \beta_1 * \log(d_i)$$

The correctness of the single-hit model can then be checked by testing the null hypothesis $H_0: \beta_1 = 1$.

Example 5.3: Assay Analysis

Shackleton et al. showed, for the first time, that a complete mammary milk producing gland could be produced in mice from a single cell. The data (data set: `mammary`) relate to a number of assays in which cells were transplanted into host mice. A positive outcome here consists of seeing a milk gland outgrowth, evidence that the sample of cells included as least one stem cell. The data give the average number of cells in each assay (`N.Cells`), the number of assays at that cell number (`N.Assays`), and the number of assays giving a positive outcome (`N.Outgrowths`). For this data,

(1) Find the point and interval estimates of proportion of active cells (λ).

(2) Test if the single-hit model is sufficient.

Solution: From R program, you can see that:

- The proportion of assays giving a positive outcome (N.Outgrowths/N.Assays) is used as the response.
- The number of assays at that cell number (N.Assays) is used as the weights.
- The logarithm of the average number of cells in each assay (N.Cells) is used as the offset.
- The link function is the complementary log-log link.

From R program, we can get: $\hat{\beta}_0 = -4.164$ and $se(\hat{\beta}_0) = 0.1744$. So the 95% confidence interval of $\hat{\beta}_0$ is

$$\hat{\beta}_0 \pm z_{\frac{\alpha}{2}} * se(\hat{\beta}_0) = -4.164 \pm 1.96 * 0.1744 = (-4.5058, -3.8222)$$

(1) The point and interval estimates of proportion of active cells (λ) are:

$$\hat{\lambda} = \exp(\hat{\beta}_0) = \exp(-4.164) = 0.0155$$

$$(\exp(-4.5058), \exp(-3.8222)) = (0.011, 0.022)$$

The frequency of stem cells is between 0.011 and 0.022.

(2) We can check if the single-hit model is sufficient using the likelihood ratio test. The p-value is 0.421, so there is no evidence of any deviation from the single-hit model.

Section 5.1.3 Wald Tests and Goodness-of-Fit Tests

When Wald Tests Fail

Standard errors and Wald tests experience special difficulties when the fitted values from binomial GLMs are very close to zero or one. Let us use a simple example for illustration. Suppose the logit link function and a single covariate x are used, then

$$\hat{\mu} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}$$

The only way for $\hat{\mu}$ to be zero or one is for $\hat{\beta}_0 + \hat{\beta}_1 x$ to be $\pm\infty$. In either case, $\hat{\beta}_0$ and/or $\hat{\beta}_1$ will approach $-\infty$ or ∞. The phenomenon is the same for other link functions.

When parameter estimates approach $\pm\infty$, the standard errors for those parameters must also approach $\pm\infty$, and Wald test statistics, which are ratios of coefficients to standard errors, become very unreliable. Fortunately, the likelihood ratio and score test usually remain quite serviceable in these situations, even when fitted values are zero or one. This is because the problem of infinite parameters is removable, in principle, by reparameterization, and likelihood ratio and score tests are invariant to reparameterization.

Example 5.4: Failed Wald Tests

A study of the habitats of the noisy miner (a small but aggressive native Australian bird) recorded whether noisy miners were detected in various two hectare transects in buloke woodland patches (data set: `nminer`). We consider fitting a binomial GLM to model the presence of noisy miners in each buloke woodland patch (`Miners`). More specifically, we study whether the presence of noisy miners is impacted by whether or not the number of eucalypts exceeds 15 or not.

The point estimate, standard error, and the p-value for $\hat{\beta}_1$ are 20.41, **3242.45** (**very large**), and 0.995, respectively. The Wald test results indicate that the explanatory variable is not significant. Note the large standard error for the explanatory variable. Compare to the likelihood ratio test results, the p-value is about $1.94 * 10^{-5}$, indicating that the explanatory variable is highly significant.

Despite the Wald test results, a plot of Miners against Eucs15 (**Figure 5.1**) shows an obvious relationship: in woodland patches with more than 15 eucalypts, noisy miners were always observed. The situation is exactly as described in the text, so the Wald test results are not trust worthy. When the number of eucalypts exceeds 15, all woodland patches in the sample have noisy miners, so

$\hat{\mu}_i$ close to 1. This is achieved as $\hat{\beta}_1 \rightarrow \infty$. In this situation, the score or likelihood ratio tests must be used instead of the Wald test.

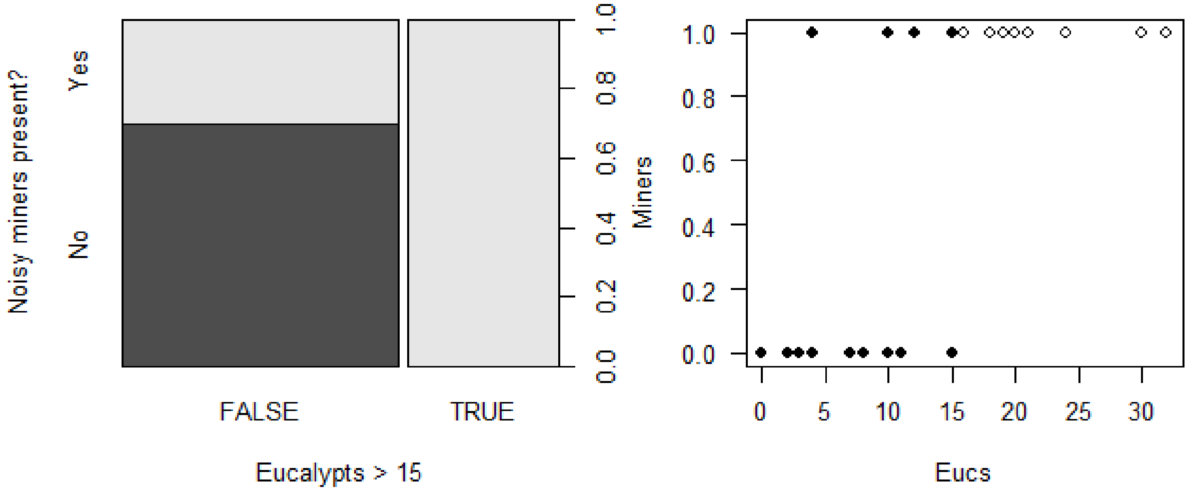

Figure 5.1 The presence of noisy miners. Left panel: the presence of noisy miners as a function of whether 15 eucalypts are observed or not; Right panel: the presence of noisy miners as a function of the number of eucalypts, showing the division at 15 eucalypts.

No Goodness-of-Fit for Binary Responses

When $m_i = 1$for all i, the binomial responses y_1 are all 0 or 1; that is, the data are binary. In this case the residual deviance and Pearson goodness-of- fit statistics are determined entirely by the fitted values. This means that there is no concept of residual variability, and goodness-of-fit tests are not meaningful. For binary data, likelihood ratio tests and score tests should be used, making sure that $p + 1$ is much smaller than the sample size n.

To understand this, the log-likelihood for the sample i can be simplified as:

$$\ell_i = y_i \log(\mu_i) + (1 - y_i)\log(1 - \mu_i) = \begin{cases} \log(\mu_i) & y_i = 1 \\ \log(1 - \mu_i) & y_i = 0 \end{cases}$$

For the maximum model, the maximum of ℓ_i is 0. So the residual deviance is just 2 times the log-likelihood function of the model: $-2\ell_{model}$.

Example 5.5: Binary Data

In the `nminer` example in the previous section, the residual deviance 24.43 is less than the residual degrees of freedom 29. This might be thought to suggest underdispersion, but it has no meaning. The size of the residual deviance is determined only by the sizes of the fitted values, and how far they are from zero and one.

Section 5.1.4 Overdispersion and Quasi-Likelihood

Overdispersion of Binomial GLMs

For the proportion of binomial distribution, the dispersion parameter $\phi = 1$ and $var[y_i] = \frac{\mu_i(1-\mu_i)}{m_i}$. However, in practice the amount of variation in the data can exceed $\mu_i(1 - \mu_i)$, even for ostensibly binomial-like data. This is called **overdispersion**. **Underdispersion** (when actual variance is less than $\mu_i(1 - \mu_i)$) also occurs, but is less common.

Overdispersion has serious consequences for the GLM. It means that standard errors returned by the GLM are underestimated, and tests on the explanatory variables will generally appear to be more significant that warranted by the data, leading to overly complex models.

Overdispersion is detected by conducting a goodness-of-fit test, as described in **Section 3.2.2**. If the residual deviance and Pearson statistics are much greater than the residual degrees of freedom, then there is evidence of lack of fit. Lack of fit may be caused by an inadequate model, for example because important explanatory variables are missing from the model. However, if all relevant or possible explanatory variables are already included in the model, and the data has been checked for outliers that might inflate the residuals, but lack of fit remains, then overdispersion is the alternative interpretation. **Overdispersion means that the binomial model is incorrect in some respect**. Overdispersion can arise from two major causes.

- In some situations, the probabilities μ_i are not constant between observations, even when all the explanatory variables are unchanged. This type of overdispersion can be modelled by a hierarchical model.
- More generally, overdispersion arises when the m_i Bernoulli cases, that make up observation y_i, are positively correlated. For example, positive cases may arrive in clusters rather than as individual cases. This leads to variances

$$var[y_i] = \phi \frac{\mu_i(1-\mu_i)}{m_i}$$

Note that overdispersion cannot arise for binary data with $m_i = 1$.

Example 5.6: Overdispersion of Binomial GLMs

For the seed germination data set `germ`, we can fit a GLM(Binomial; logit) with the types of seeds and the types of root stocks as the explanatory variables. A number of conclusions can be obtained for this model.

(1) We fit a maximal possible explanatory model that include the interaction effects of them.

(2) Pearson statistic: $\frac{P^2}{\phi} = P^2 = \sum_{i=1}^{n} \frac{m_i(y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)} = 31.65$

The p-value of goodness-of-fit test based on Pearson statistic: 0.0167

Deviance: 33.28

The p-value of goodness-of-fit test based on deviance: 0.0104

Both the Pearson statistic and the deviance are bigger than 17, the residual degrees of freedom.

So the overdispersion is clearly presented.

(3) The chi-square approximation to the goodness-of-fit statistics seems good enough. The data includes one observation (number 16) with $m_{16}y_{16} = 0$ and other with $m_6y_6 = 1$, but neither has a large enough residual to be responsible for the apparent overdispersion.

(4) There are no large residuals present that would suggest outliers (**Figure 5.2**).

(5) This a designed experiment, with nearly equal numbers of obser- vations in each combination of the experimental factors Extract and Seeds, so influential observations cannot be an issue.

(6) Having ruled out all alternative explanations, we accept that overdispersion is present. It is understandable. Since seeds are usually planted together in common plots, it is highly possible that they might interact or be affected by common causes; in other words we might well expect seeds to be positively correlated, leading to overdispersion.

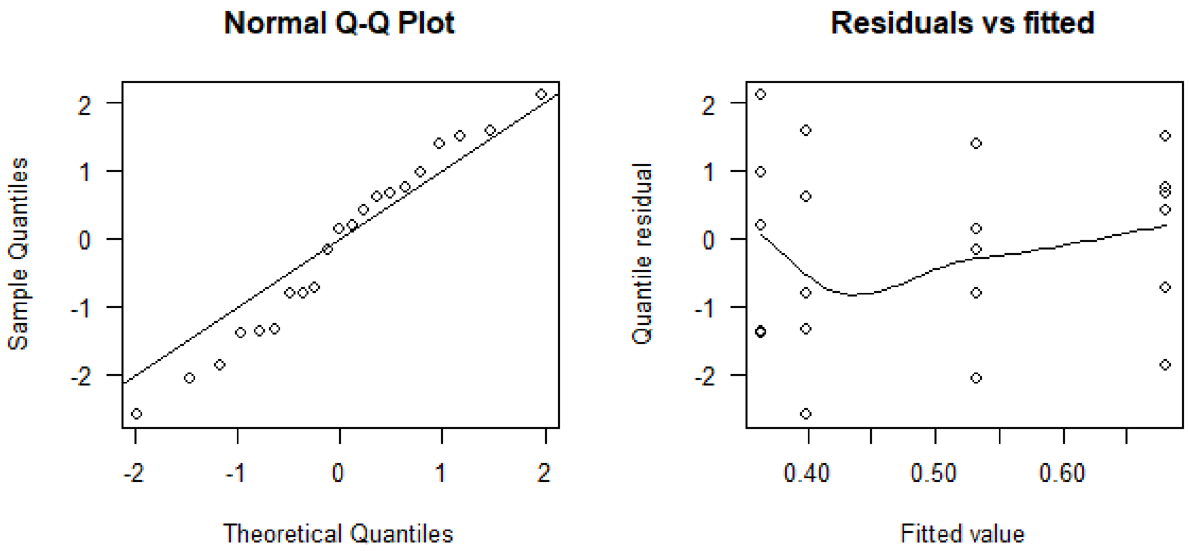

Figure 5.2 Diagnostic plots after fitting a binomial GLM to the seed germination data.

Exercise 5.2: Overdispersion for Binomial GLMs

Machine turbines operate more or less independently, so it seems reasonable to suppose that independence between Bernoulli trials might hold for the turbines data (data set: `turbines`). Verify the following results:

Test	Statistic	Residual DF	*p*-value
Deviance	10.331	9	0.3243
Pearson	9.250	9	0.414

Do we have an overdispersion problem?

Quasi-Likelihood Methods

In rare cases, sometimes the mean-variance relationship for a data set suggests a distribution that is not an EDM. Although 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:

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

where $E[y] = \mu$ and $var[y] = \phi V(\mu)$. This relationship is used in fitting GLMs to find the estimates $\hat{\beta}_j$. The standard errors $se(\hat{\beta}_j)$ 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, we can find some quasi-probability function $\bar{\mathcal{P}}$ exists which satisfies

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

when only the variance function $V()$ is known.

Suppose we have a series of observations y_i, for which we assume $E[y_i] = \mu_i$, and $var[y_i] = \phi V(\mu_i)/w_i$, and $g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}$. Then the **quasi-likelihood function** (more correctly, the quasi-log-likelihood) is defined by

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

The quasi-likelihood behaves like a log-likelihood function, but does not correspond to any probability function. **Table 5.1** compare the log-likelihood function and the quasi-likelihood function.

Table 5.1 Comparison of log-likelihood function and the quasi-likelihood function.

Properties	Log-likelihood	quasi-likelihood	Example Inferences
Probability distribution	Yes	No	
AIC, BIC	Yes	No	
Goodness-of-fit test	Yes	No	
Likelihood ratio test	Yes	No	

Quantile residuals	Yes	No	
Deviance	Yes	Yes	Analysis of deviance
Deviance/Pearson residuals	Yes	Yes	Diagnostic plots
Asymptotic property of $\hat{\beta}_j$	Yes	Yes	Wald type CI

The quasi-likelihood provides a more flexible way to analyze data:

- First, it gives us a way to conduct inference when there is no EDM for a given mean-variance relationship. To specify a quasi-GLM, only the link function and the variance function $V(\mu)$ are needed. For example, we can use $V(\mu) = \exp(\mu)$. The EDMs discussed in our course do not have such mean-variance relationship.
- Second, if $V(\mu)$ is same as the variance function from a genuine EDM, the quasi-GLMs can allow a dispersion parameter that may not be allowed in the genuine EDM. For example, the dispersion in binomial GLMs is 1. While a quasi-GLM with $V(\mu) = \mu(1-\mu)$ allows a dispersion parameter $\phi > 1$ to account the overdispersion of the data.

Therefore, 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.

Quasi-GLMs can be fitted with the R function `glm()` with `family = quasipoisson()`, etc. The R function `quasi()` can be used in `glm()` to fit quasi-GLMs more generally. The link function $g()$ and the variance function $V(\mu)$ are specified with `family = quasi(link =””,variance =””)`.

Quasi-Binomial GLMs

In the presence of overdispersion for a binomial GLM, the quasi-binomial GLM can be used to fit the data. Quasi-binomial GLMs keep the same variance function $V(\mu_i) = \mu_i(1-\mu_i)$ as binomial GLMs, but allow a general positive dispersion ϕ instead of assuming $\phi = 1$. The dispersion parameter is usually estimated by the Pearson estimator. As we have mentioned, although quasi-binomial GLMs do not correspond to any EDM, but the quasi-likelihood theory provides reassurance that the model will still yield consistent estimators provided that the variance function represents the correct mean-variance relationship.

The parameter estimates for binomial and quasi-binomial GLMs are identical (since the estimates $\hat{\beta}_j$ do not depend on the dispersion parameter ϕ), but the standard errors are different. The effect of using the quasi-binomial GLM is to inflate the standard error of the parameter estimates by $\sqrt{\phi}$ so confidence intervals and statistics for testing hypotheses tests will change.

Example 5.7: Quasi-Binomial GLMs

For the seed germination data set `germ`, fit the binomial and quasi-binomial GLMs and compare the results.

Solution: You can find that the R programs for Binomial GLMs and Quasi-Binomial GLMs are also identical. The only difference is that family = binomial() is used in Binomial GLMs while family = quasibinomial() is used in Quasi-Binomial GLMs.

We found the presence of dispersion from **Example 5.6**. The estimated dispersion parameter from the quasi-binomial GLM is $\hat{\phi} = 1.86183$ and $\sqrt{\hat{\phi}} = 1.3645$. **Table 5.2** compares the results of the binomial GLM and quasi-binomial GLM. From the table, we can see that two models have:

(1) Same estimates of β_j but different standard error, test statistic, p-value;

(2) Same residual deviance, deviance residuals, Pearson statistic, and Pearson residuals;

(3) Different test statistic and p-value in the analysis of deviance table.

For the binomial GLM, the dispersion parameter is 1, so the Z test or the chi-square test is used. For the quasi-binomial GLM, the dispersion parameter is unknown and needs to be estimated, so the t-test or F-test is used.

Table 5.2 Comparison of a binomial GLM and a quasi-binomial GLM based on the data set `germ`.

Inference	Inference	Binomial	Quasi-Binomial	Comparison	Comments
Parameter for Extract	$\hat{\beta}_1$	0.5401	0.5401	Same	ϕ dose not affect estimation
	$se(\hat{\beta}_1)$	0.2498	0.3409	Different	$\frac{0.3409}{0.2498} = 1.3645 = \sqrt{\hat{\phi}}$
	z	2.1619	1.5844	Different	Binomial: Z test Quasi: t-test
	p-value	0.0306	0.1315	Different	
Residual DF		17	17	Same	
Log-likelihood		-54.93702	NA	Different	
Deviance		33.28	33.28	Same	So same deviance residuals
Pearson stat		31.65	31.65	Same	So same Pearson residuals
Analysis of deviance `Seeds`	Residual deviance	39.686	39.686	Same	
	Deviance	3.065	3.065	Same	
	Test statistics	3.065	1.6462	Different	Binomial: Chi-square test Quasi: F-test
	p-value	0.080	0.217	Different	

Section 5.1.5 Case Study

An experiment exposed batches of insects to various deposits (in mg) of insecticides (data set: `deposit`). The proportion of insects y_i killed after six days of exposure in each batch of size m_i is potentially a function of the dose of insecticide and the type of insecticide. Since the response is the proportions, the binomial GLM with the probit link function is used.

Exploratory Analysis

There are six doses $(2.00, 2.64, 3.48, 4.59, 6.06,$ and $8.00)$ and three types of insecticide $(A, B,$ and $C)$. A plot of $\Phi^{-1}(y_i)$ against `Deposit` (left panel, **Figure 5.3**) or logarithm of `Deposit` (right panel, **Figure 5.3**) stratified by the type of insecticide. Both plots show insecticides A and B appear to have similar effects, while insecticide C appears different from A and B. The amount of deposit clearly is significant. Second, it shows that the relationship between $\Phi^{-1}(y_i)$ and `Deposit` is not linear. The relationship between $\Phi^{-1}(y_i)$ and logarithm of `Deposit` is more linear but the linearity may be improved by adding a quartic form.

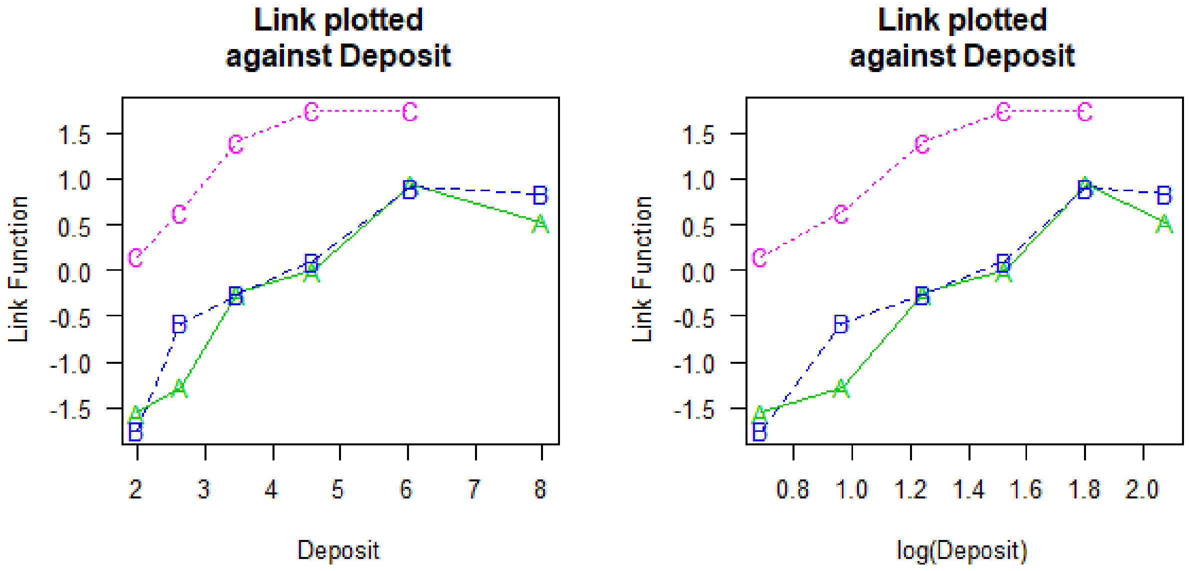

Figure 5.3 A plot of $\Phi^{-1}(y_i)$ against `Deposit` (left panel) or logarithm of `Deposit` (right panel).

Model Selections and Diagnostic

Three models are considered:

- $M1$: `Deposit` as additional explanatory variable
- $M2$: logarithm of `Deposit` as additional explanatory variables
- $M3$: logarithm of `Deposit` and the quadratic logarithm of `Deposit` as additional explanatory variables

$M1$ is considered because this would be an initial model used by most of people. $M2$ is considered because the logarithm of the dose is commonly used in dose-response models. $M3$ is considered because a quartic term could improve the model fitting from **Figure 5.3**. Here the probit link function is used although other link functions such as logit link function can be used (see **Exercise 5.3**).

After the model fitting, we would perform the diagnostic analysis and compare three models.

- **Figure 5.4** shows the data and fitted lines with three different models. Close inspection shows the model $M1$ is inadequate because the fitted for lower and higher doses of insecticides A and B are off. The models $M2$ and $M3$ clearly have the better fit.
- **Figure 5.5** shows the plots of quantile residuals against the constant-information scale of fitted values ($sin^{-1}(\sqrt{\hat{\mu}_i})$, **Table 4.2**) for three models. There are clear patterns for the plot from the model M1 while the plots from the models 2 and 3 are similar.
- **Figure 5.6** shows the plots of the working response against the linear predictors for three models. It is clear that the model M3 has the best linearity.
- The analysis of variance between the models $M2$ and $M3$ returns a p-value 0.00636, indicating that the quadratic model is a statistically significant improvement.
- The AICs and BICs indicate that the model $M3$ has the smallest AIC and BIC thus is ghe best models among these three models.
- **Table 5.3** shows the results based on the deviance and the goodness-of-fit test. Clearly the quadratic model is the best among three models considered here. The model $M1$ is not adequate. There is one observation with $m_i y_i = 2$ and all other $m_i y_i \geq 3$. So we may need to pay some attention about this part.

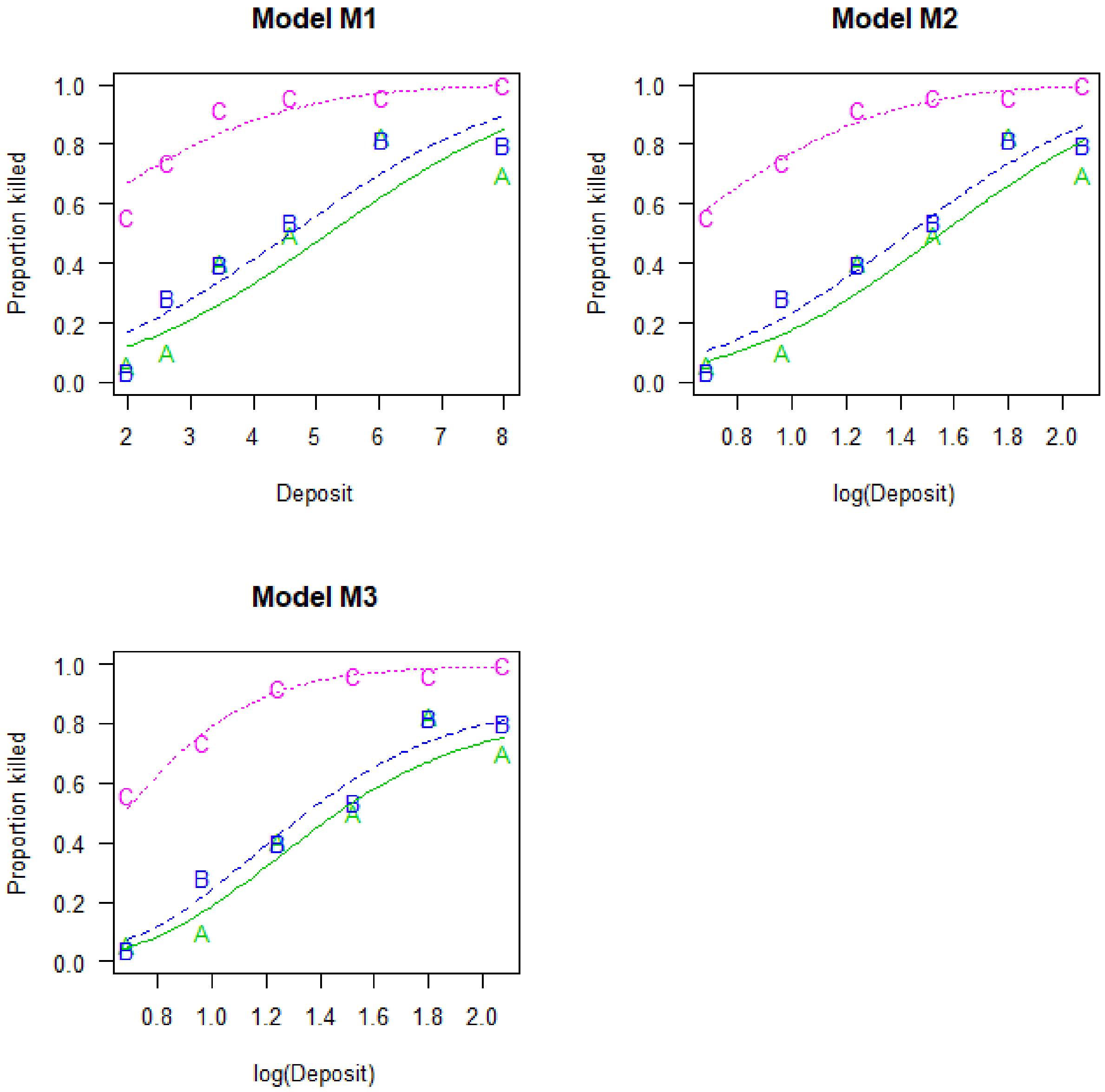

Figure 5.4 The plots of data (in points) and fitted value (in lines) for three binomial GLMs.

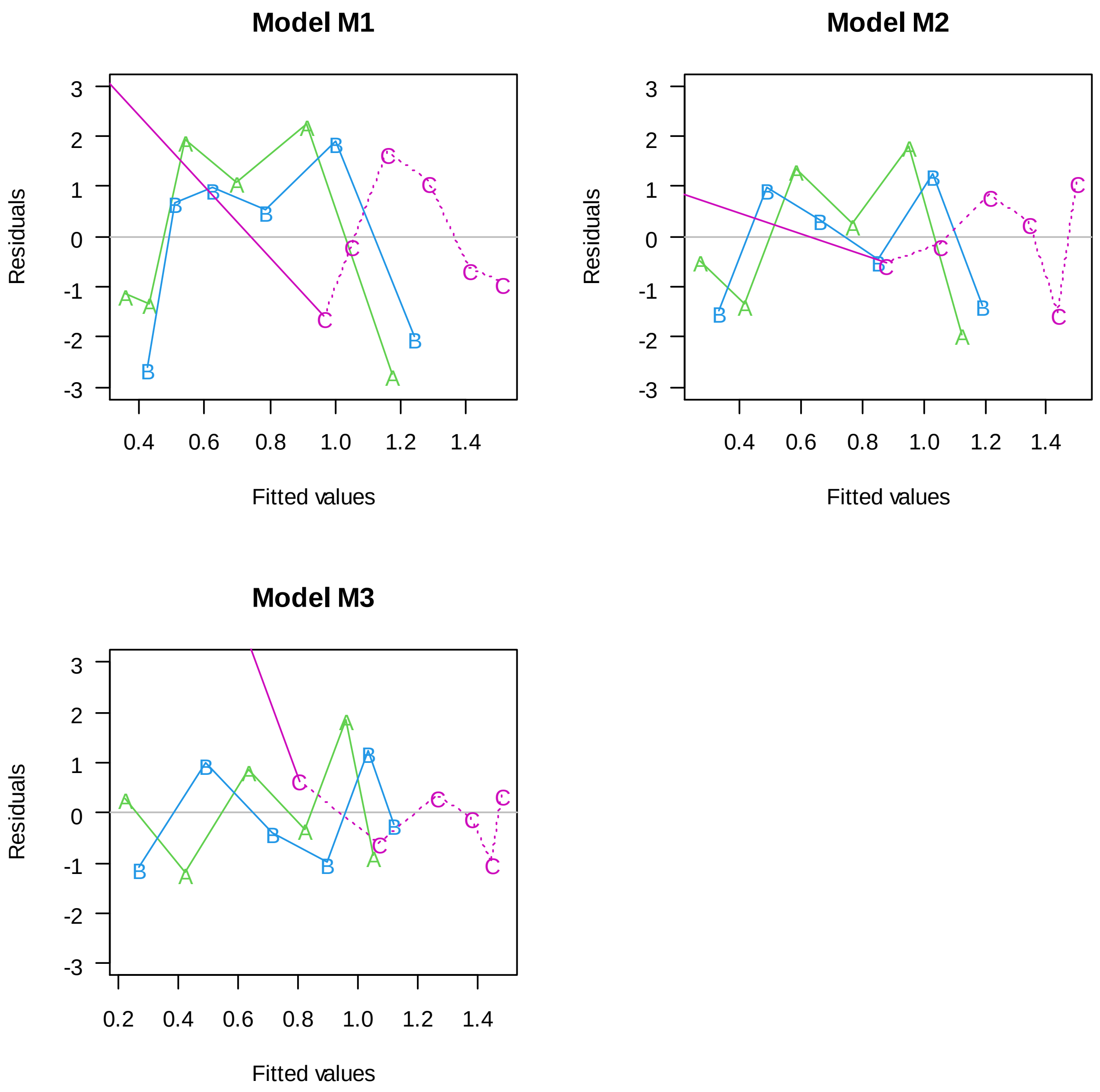

Figure 5.5 The plots of quantile residuals against the constant-information scale of fitted values ($sin^{-1}(\sqrt{\hat{\mu}_i})$, **Table 4.2**) for three models.

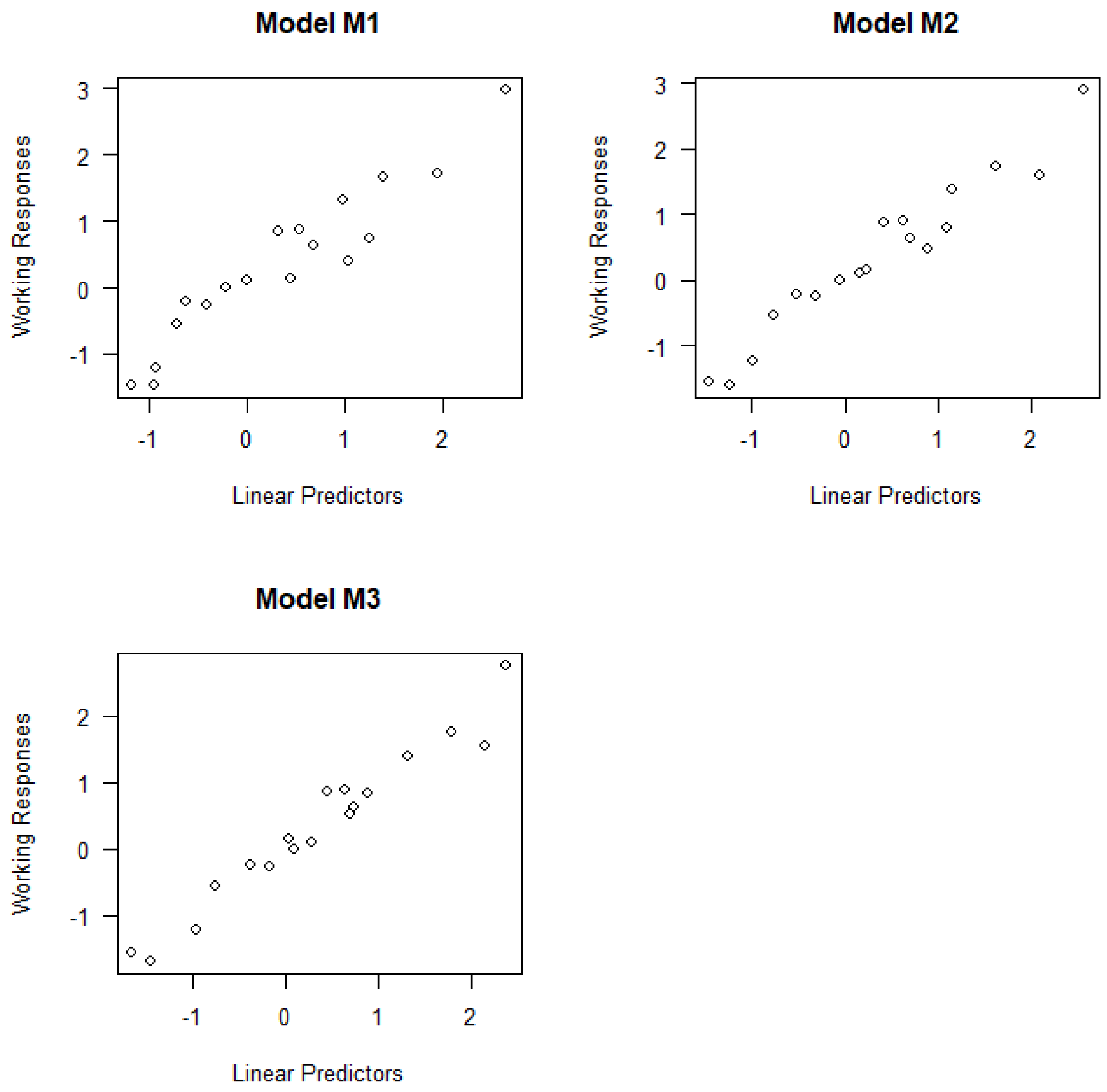

Figure 5.6 The plots of the working response against the linear predictors for three models.

Table 5.3 Deviance and goodness-of-fit test for three model.

Model	Deviance	Residual DF	Deviance/DF	p-value
$\boldsymbol{M1}$	47.39	14	3.39	$1.66 * 10^{-5}$
$\boldsymbol{M2}$	22.35	14	1.60	0.072
$\boldsymbol{M3}$	14.91	13	1.15	0.313

Statistical Inference of Model $M3$

From **Table 5.4**, we can both logarithm and squared logarithm of `Deposit` have the significant effect. There is no evidence to support that insecticide A and B have the different effects. insecticide A and C clearly have the different effects. Further analysis can show that insecticide B and C have the different effects. These are consistent with our observations from **Figure 5.4**. Keep in mind that the coefficients for "Insecticide B" and ""Insecticide C" represents the difference of effects between B and A and between C and C, respectively.

Table 5.4 Coefficients from the model $M3$.

	Estimate	**Standard Error**	**z value**	***p*-vlaue**
Intercept	-3.920	0.524	-7.38	$7.34 * 10^{-14}$
Log(Deposit)	3.775	0.782	4.83	$1.38 * 10^{-6}$
Log(Deposit)^2	-0.750	0.276	1.529	0.00663
Insecticide B	0.195	0.120	1.629	0.1034
Insecticide C	1.701	0.145	11.70	$< 2 * 10^{-16}$

It is also interest to estimate ED50 for each type of insecticide – the dose of insecticide results in 50% insects that are killed. There are three difficulties here: (1) The R function `dose.p()` cannot be used since both $\log(x)$ and $[\log(x)]^2$ are involved. (2) The intercept for insecticide A can be directly obtained from **Table 5.4** but the intercepts for insecticide B and C need to be calculate. (3) We need to solve an equation to find ED50. (4) The standard error for ED50 is difficult to get thus the confidence intervals will not be calculated here.

Since $\beta_0 + \beta_1 \log(x) + \beta_2[\log(x)]^2 = \eta = \Phi^{-1}(\mu)$ and $\Phi^{-1}(0.5) = 0$, the estimated ED50 satisfies this equation:

$$\hat{\beta}_0 + \hat{\beta}_1 \log(x) + \hat{\beta}_2[\log(x)]^2 = 0$$

This quadratic equation have two solutions and we may need to determine which one should be used based on our data. The two solutions are:

$$\log(x) = \frac{-\hat{\beta}_1 \pm \sqrt{(\hat{\beta}_1)^2 - 4\hat{\beta}_2\hat{\beta}_0}}{2\hat{\beta}_2}$$

There are two-way obtain the intercept for each type of intercept. One method is based on the treatment coding and **Table 5.3.** According to the treatment coding:

- Intercept for insecticide A (reference level): -3.920 (intercept in model)
- Intercept for insecticide B: $-3.920 + 0.195 = -3.725$ (intercept + coefficient of B)
- Intercept for insecticide C: $-3.920 + 1.701 = -2.219$ (intercept + coefficient of C)

In the word, the intercept for the reference level is the intercept in the model while the intercept for the other levels is the sum of the intercept in the model and its corresponding coefficient.

The other method to obtain the intercept for each type of intercept is to fit a model without intercept so R is forced to fit a model with separate intercept. You can verify the results same as we have obtained. You can verify that the results are same as we have obtained. Note that "-1" is used on the right side of formula in glm() so a GLM model without the intercept is fitted:

```
Killed/Number ~ logDep + logDep2 + Insecticide "- 1"
```

The following table summarized the estimated ED50 for each type of insecticide. Using $\hat{\beta}_1 = 3.7753$ (the coefficient for the logarithm of `Deposit`), $\hat{\beta}_2 = -0.7501$ (the coefficient for the quadratic logarithm of `Deposit`), and the corresponding $\hat{\beta}_0$ (the coefficients for three types insecticide), we can find that the estimated ED50s for insecticide A, B, and C are 4.324, 3.848, and 1.973, respectively. Insecticides A and B have the similar ED50 and insecticide C has the smallest ED50.

Table 5.5 Estimated ED50 for each type of insecticide based on a probit model.

Insecticide	**Equations**	$\hat{\beta}_0$	**Solution 1**		**Solution 2**		**ED50**
			$\log(x)$	x	$\log(x)$	x	
A	$\hat{\beta}_0 + \hat{\beta}_1 \log(x) +$	-3.919	1.464	4.324	3.569	35.5	4.324
B	$\hat{\beta}_2[\log(x)]^2$=0	-3.725	1.347	3.848	3.686	39.9	3.848
C	$\hat{\beta}_1 = 3.775, \hat{\beta}_2 = -0.7501$	-2.219	0.680	1.973	4.354	77.8	1.973

Conclusions

The probit model with both logarithm and squared logarithm of `Deposit` is adequate (**Table 5.3**) to describe the data. There is no evidence to support that insecticide A and B have the different effects. Insecticide C clearly has the effect different from those from A and B and has the smallest ED50.

Exercise 5.3: Case Study with Logistic Regression

The logistic regression model is used in the textbook. Replicate the results from the textbook and compare the results between the logistic regression and the probit model.

Lesson 5.2 Positive Continuous Data: Gamma and Inverse Gaussian GLMs

Related Readings: Sections 11.1, 11.2, 11.3, 11.4, 11.5, 11.6, 11.7, 11.8, and 11.9 in Chapter 11.

Section 5.2.1 Introduction and Overview

In this lesson, we consider models for positive continuous data. Variables that take positive and continuous values often measure the amount of some physical quantity that is always present. In **Section 5.2.2**, we introduce modeling positive continuous data. In **Section 5.2.3**, we discuss Gamma GLMs, one of two most common GLMs for positive continuous data. In **Section 5.2.4**, we describe another GLMs for positive continuous data, inverse Gaussian GLMs. In **Section 5.2.5**, we discuss the use of link functions and the estimation of the dispersion parameter ϕ. The lesson is concluded with a case study in **Section 5.2.6.**

Section 5.2.2 Modeling Positive Continuous Data

Many applications have response variables which are continuous and positive. Because such variables have the boundary at zero,

- They usually have distributions that are right skew.
- The variance of the response must generally approach zero as the expected value approaches zero. Therefore, positive continuous data tusually shows an increasing mean-variance relationship.

Apart from $V(\mu) = \mu$ (Poisson GLMs), the simplest increasing variance function functions are $V(\mu) = \mu^2$ and $V(\mu) = \mu^3$, which correspond to the Gamma and inverse Gaussian distributions, respectively. For these reasons, GLMs based on the Gamma and inverse Gaussian distributions are useful for modelling positive continuous data.

Example 5.8: Modeling Positive Continuous Data

A series of studies sampled the forest biomass in Eurasia. Part of that data, for small-leaved lime trees, is in the data set `lime`.

A model for the foliage biomass (`Foliage`) y is sought. The mean foliage biomass $\mu = E[y]$ may be related to the surface area, thus proportional to d^2, where d is the diameter of the tree trunk (or `DBH`). In addition, the tree diameter may be related to the age of the tree. However, since diameter measures some physical quantity and is easier to measure precisely, it is expected that the relationship between foliage biomass and `DBH` to be stronger than the relationship between foliage biomass and age. Clearly, the response is always positive. From **Figure 5.7**, the variance in foliage biomass increases as the mean increases, and a relationship exists between foliage biomass and `DBH`, and between foliage biomass and age. The effect of origin is harder to see.

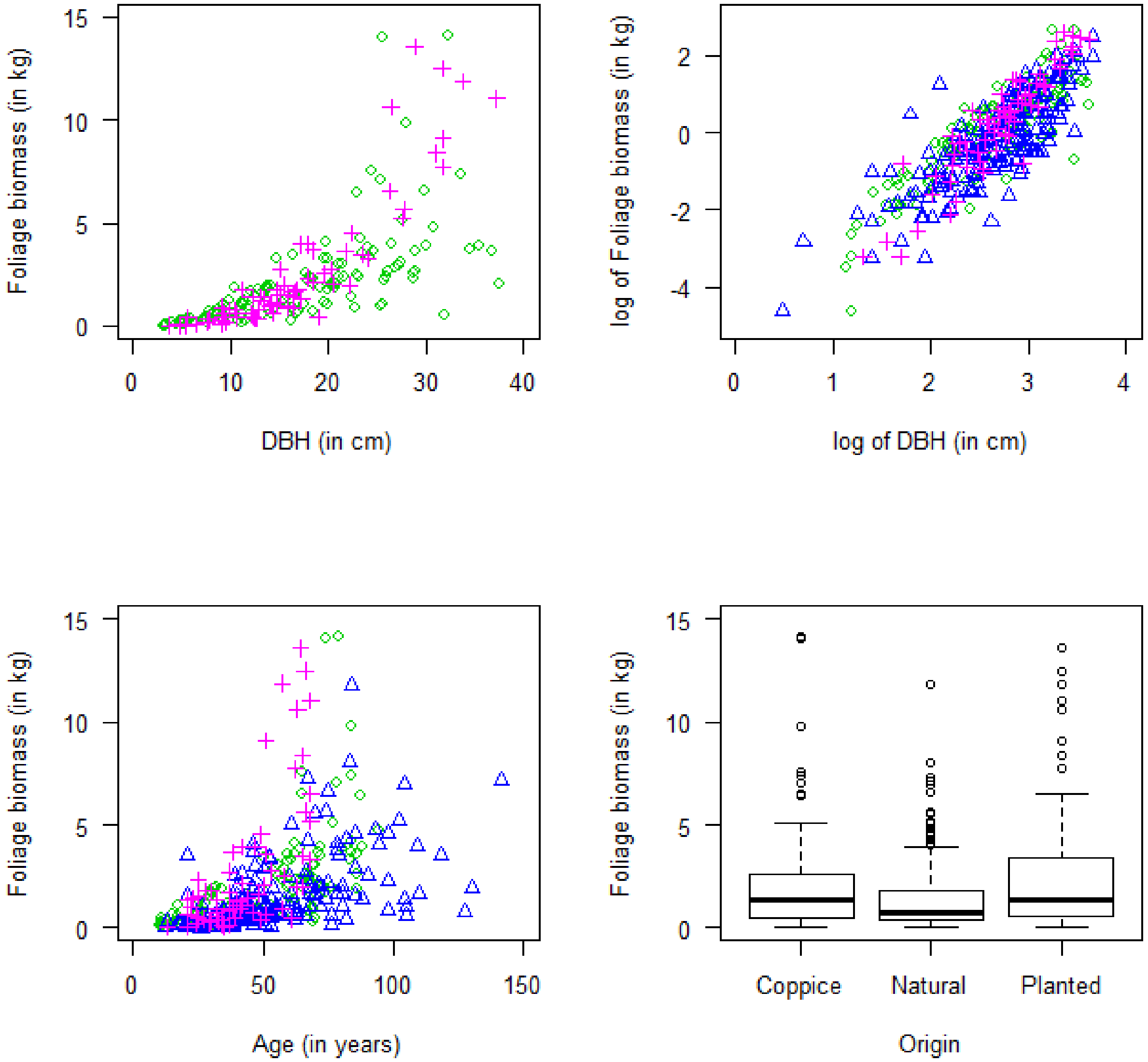

Figure 5.7 The small-leaved lime data. Plots of the foliage biomass (response) against `DBH` (diameter at breast height; top left panel); logarithm of foliage biomass against the logarithm of `DBH` (top right panel); foliage biomass against age (bottom left panel) foliage biomass against origin (bottom right panel).

Section 5.2.3 The Gamma GLMs

Gamma Distribution

The Gamma GMLs are based on the Gamma distribution which has the following probability distribution function:

$$\mathcal{P}(y;\alpha,\beta)=\frac{y^{\alpha-1}\exp(-\frac{y}{\beta})}{\Gamma(\alpha)\beta^{\alpha}}, y>0, \alpha>0, \beta>0$$

where $\Gamma(\alpha)$ is the gamma function (e.g., $\Gamma(n+1)=n!$), α is the shape parameter, and β is the scale parameter. Plots of some example Gamma densities can be found in https://en.wikipedia.org/wiki/Gamma_distribution.

If y has a Gamma distribution with the shape parameter α and the scale parameter β, we have

$$\mu=E[y]=\alpha\beta, var[y]=\alpha\beta^2=\frac{\alpha^2\beta^2}{\alpha}=\frac{1}{\alpha}\mu^2$$

Therefore for the Gamma distribution, the variance function $V(\mu)=\mu^2$ and the dispersion parameter is $\phi=\frac{1}{\alpha}$. The **coefficient of variation** is defined as the ratio of the variance to the mean squared, and is a measure of the relative variation in the data. The coefficient of variation of the Gamma distribution is $\frac{\mu}{\sqrt{\frac{1}{\alpha}\mu^2}}=\sqrt{\alpha}$, which is a constant. Therefore, Gamma GLMs are useful in situations where the coefficient of variation is (approximately) constant.

A special case of the Gamma distribution is $\alpha=1$, the distribution function becomes

$$\mathcal{P}(y;\alpha=1,\beta)=\frac{1}{\beta}\exp(-\frac{y}{\beta})$$

which is the exponential distribution function. Therefore, the exponential distribution is a Gamma distribution with the dispersion parameter $\phi=1$.

From $\mu = \alpha\beta$ and $\phi = \frac{1}{\alpha}$, we can obtain $\alpha = \frac{1}{\phi}$ and $\beta = \frac{\mu}{\alpha} = \phi\mu$. By replacing them in $\mathcal{P}(y;\alpha,\beta)$, we have:

$$\mathcal{P}(y;\mu,\phi) = \frac{y^{\frac{1}{\phi}-1}}{\Gamma\left(\frac{1}{\phi}\right)} \exp\left(-\frac{y}{\phi\mu}\right) \frac{1}{(\phi\mu)^{\frac{1}{\phi}}}$$

$$= \frac{y^{\frac{1}{\phi}-1}}{\Gamma\left(\frac{1}{\phi}\right)} \frac{1}{(\phi)^{\frac{1}{\phi}}} * \exp\left(-\frac{y}{\phi\mu} - \frac{1}{\phi}\log(\mu)\right)$$

$$= a(y;\phi)\exp(\frac{y * \left(-\frac{1}{\mu}\right) - \log(\mu)}{\phi})$$

This is the format of EDMs and the nature parameter will be $-\frac{1}{\mu}$.

Method to Determine if $V(\mu) \approx \mu^2$

Before we can use Gamma GLMS, we would like to see if the variance function $V(\mu) \approx \mu^2$. If $V(\mu) \approx \mu^2$ or more generally, $V(\mu) \approx \mu^c$, then $\log(V(\mu)) \approx c\log(\mu)$, so the following procedure can help us to verify if $V(\mu) \approx \mu^c$:

- **Step 1**: split the data into smaller group according to the explanatory variables
- **Step 2:** calculate the sample variance and sample mean of the response for each group data using the R function `tapply()`.
- **Step 3:** Fit a linear regression with the logarithm of the group variances as the response and the logarithm of the group means as the covariate and obtain the estimate slope, $\hat{\beta}_1$.
- **Step 4:** Choose a number close to $\hat{\beta}_1$as c and use the variance function $V(\mu) \approx \mu^c$.

Note that the above procedures have already been introduced in Lesson 2.2 (please refer to **Example 2.17** and **Quiz Problem 2.4**).

Example 5.9: Variance Function

For the small-leaved lime data (data set: `lime`), the data can be split into smaller groups, and the mean and variance of each group calculated. **Figure 5.8** shows that the variance increases as the mean increases. The slope from the linear regression with the logarithm of the group variances as the response and the logarithm of the group means as the covariate is about 1.7, suggesting the variance function $V(\mu) \approx \mu^2$ is a good approximation. The plot of the group standard deviations against the group means (B in **Figure 5.8**) further confirms that the coefficient of variation is constant.

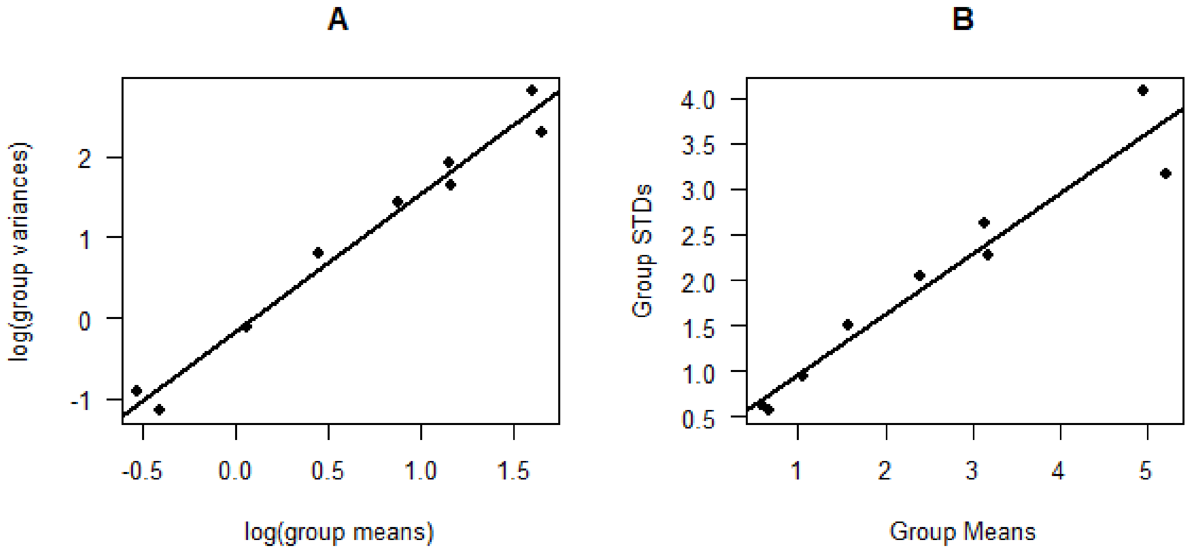

Figure 5.8 The small-leaved limed data: **A** - the logarithm of group variances plotted against the logarithm of the group means; **B** – the group standard deviation plotted against the group means.

Gamma GLMs

Gamma GLMs have the following properties:

- ϕ is almost always unknown and therefore must be estimated, so the t-test and F-test instead of the Z-test and chi-square test are used in the inference.
- One estimate of ϕ is based on the Pearson statistic:

$$\hat{\phi} = \frac{P^2}{n-(p+1)} = \frac{\sum_{i=1}^{n} \frac{w_i(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i^2}}{n-(p+1)}$$

- The residual deviance of Gamma GLMs can be written:

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} w_i * 2 * \left(-\log\left(\frac{y_i}{\hat{\mu}_i}\right) + \frac{y_i-\hat{\mu}_i}{\hat{\mu}_i}\right)$$

The residual deviance $D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) \sim \chi^2_{n-(p+1)}$ approximately when $\phi < \frac{1}{3}$.

- The **canonical link function for the Gamma** distribution is the inverse (or reciprocal) link function $\eta = -\frac{1}{\mu}$. In general, we remove the minus and use $\eta = \frac{1}{\mu}$. Since $-\infty < \eta < \infty$, the link does not guarantee $\mu > 0$ which could cause problems.
- In practice, the **logarithmic link function** is often used. This link should be used when effect of the explanatory variables is suspected to be multiplicative on the mean. This is because for example,
$$\mu = \exp(\beta_0 + \beta_1 x) = \exp(\beta_0) * \exp(\beta_1 x)$$
When the variance is small, Gamma GLMs are similar to Gaussian model with the logarithm of the response.
- The gamma distribution can be used to describe the time between occurrences that follow a Poisson distribution. Please refer the textbook for more details about this topic.

An example of Gamma GLMs will be presented in **Section 5.2.6**.

Section 5.2.4 Inverse Gaussian GLMs

The inverse Gaussian distribution may sometimes be suitable for modelling positive continuous data. The inverse Gaussian has the probability function

$$\mathcal{P}(y; \mu, \phi) = (2\pi y^3 \phi)^{-0.5} \exp\left\{-\frac{1}{2\phi}\frac{(y-\mu)^2}{y\mu^2}\right\}, y > 0, \phi > 0, \mu > 0.$$

Plots of some example inverse Gaussian densities can be found in https://en.wikipedia.org/wiki/Inverse_Gaussian_distribution.

The properties of the inverse Gaussian distribution (or GLMs) include:

- The variance function is $V(\mu) = \mu^3$. Note that the variance function for the Gamma distribution is $V(\mu) = \mu^2$. Therefore, the inverse Gaussian distribution is used when the responses are even more skewed than suggested by the Gamma distribution.
- It becomes more like a “normal (Gaussian)” distribution when $\phi \to 0$. Th name of this distribution can actually be misleading.
- The residual deviance of inverse Gaussian GLMs can be written:

$$D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = \sum\nolimits_{i=1}^{n} w_i \frac{(y_i - \hat{\mu}_i)^2}{y_i \hat{\mu}_i^2}$$

For inverse Gaussian GLMs, the residual deviance $D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})$ has an **exact** $\chi^2_{n-(p+1)}$.

- The ϕ is almost always unknown and therefore must be estimated, so the t-test and F-test instead of the Z-test and chi-square test are used in the inference.
- The inverse Gaussian distribution has an interesting interpretation, connected to **Brownian motion**. You can refer to the textbook for more details if you are interested in this connection.

Again, an example of inverse Gaussian GLMs will be presented in **Section 5.2.6**.

Section 5.2.5 Link Functions and Estimating the Dispersion Parameter

Link Functions

The logarithmic link function is the link function most commonly used for Gamma and inverse Gaussian GLMs, to ensure $\mu > 0$ and for interpretation purposes. For the gamma and inverse Gaussian distributions, R permits the link functions "log", "identity" and "inverse" (the default and canonical link for Gamma GLMs). The link function `link = "1/mu^2"` is also permitted for the inverse Gaussian distribution, and is the default (and canonical) link function. The link function can be determined by the data or the relationship of the response and the explanatory variables in some situations. But in many situations, the choice of the link function can be arbitrary. The diagnostics are important to make sure that the appropriate GLMs and link function are used to describe the data.

Example 5.10: Logarithmic and Inverse Link Functions

For the small-leaved lime data (data set: lime), we have showed that the variance function $V(\mu) \approx \mu^2$ (**Example 5.9**), so a Gamma GLM can be used. However, no turning points or asymptotes are evident for the link function. So we use the logarithm link, the inverse link, and identify link and compare the results from these link functions. The following steps are taken:

(1) Fit the Gamma GLMs with the logarithmic link.

(2) Fit the Gamma GLMs with the inverse link.

(3) Fit the Gamma GLMs with the identity link.

(4) Use diagnostic tools to see if they are appropriate.

Solution: Please refer to the R grogram for some details. Before we fit the model, I would like to point out that:

- First, the model includes the interaction effect between the factor `Origin` and the covariate `DBH`. That means there is an intercept and a slope for `DBH` for each level of `Origin`.
- Second, the response is `Foliage` and the covariate is the logarithm of `DHB` instead of `DHB`. According to **Example 5.8**, we have `Foliage`$\propto$ DHB^2`,`so there is a linear relationship between the logarithm of `Foliage` and the logarithm of `DHB.`In some situations, the Gamma GLMs are equivalent to a linear regression with a logarithm of the response. Therefore, the response is `Foliage` and the covariate is the logarithm of `DHB`.

(1) It is straight to use R to fit the Gamma GLM with the logarithmic link function.

(2) Using the inverse link function produces error messages: `no valid set of coefficients has been found`. This is because the inverse link function does not restrict μ to be positive. In this situation, starting points may be supplied to `glm()` on the scale of the data (using the input `mustart`) or on the scale of the linear predictor (using the input `etastart`) so R can find a valid set of coefficients. Unfortunately, R can not find a valid set of coefficients even the fitted values from the logarithmic link are provided as the starting points. So we do not consider this model further.

(3) Again, R cannot find a valid set of coefficients for the identity link, so this model is not considered either.

(4) From **Figure 5.9**, the model seems appropriate. Other than a few observations, there are no apparent patterns in plot **A**. There is a clear linear relationship in plot **B**. The Q-Q plot looks fine too. Some observations produce large residuals, and the Cook's distance is less than 0.20, so there are no influential observations.

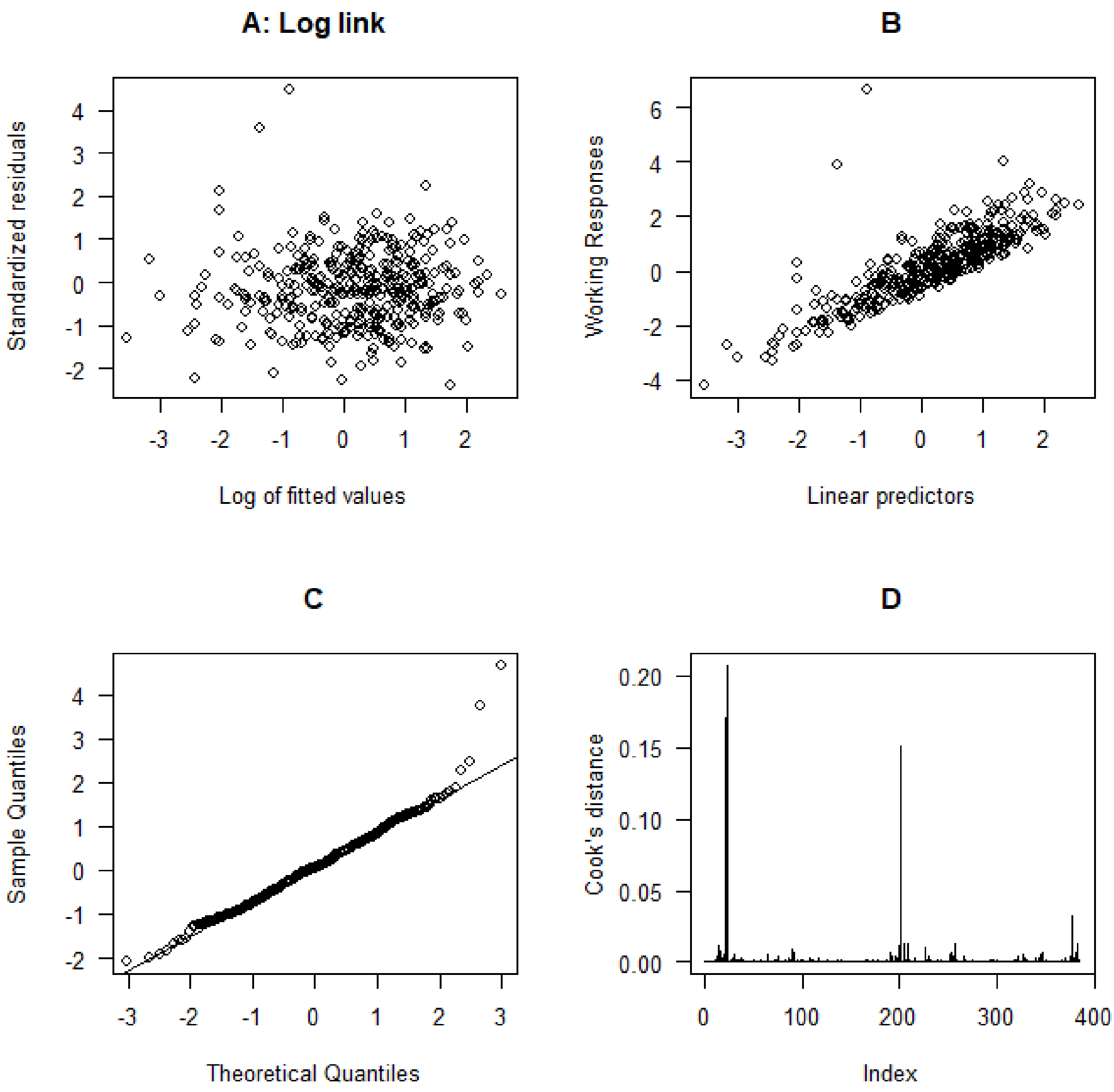

Figure 5.9 Diagnostic plots for a GLM(Gamma; log) based the small-leaved lime data. **A**: standardized residuals against logarithm of fitted values (constant-information scale); **B**: working responses against linear predictors; **C**: Q-Q plot of quantile residuals; and **D**: Cook's distance.

Estimating The Dispersion Parameter

Since $V(\mu_i) = \mu_i^2$ for the Gamma distribution, so the Pearson estimator of ϕ for Gamma GLMs is:

$$\hat{\phi} = \frac{P^2}{n-(p+1)} = \frac{1}{n-(p+1)}\sum_{i=1}^{n} \frac{w_i(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2}$$

where w_i are the prior weights.

Similarly, since $V(\mu_i) = \mu_i^3$ for the inverse Gaussian distribution, so the Pearson estimator of ϕ for inverse Gaussian GLMs is:

$$\hat{\phi} = \frac{P^2}{n-(p+1)} = \frac{1}{n-(p+1)}\sum_{i=1}^{n} \frac{w_i(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^3}$$

Example 5.11: Analysis of Deviance for Gamma GLMs

Using the model lime.log for the small-leaved lime data in **Example 5.10** (data set: `lime`), find $\hat{\phi}$, find the analysis of deviance table and discuss $\hat{\beta}_j$.

Solution: We can obtain the Pearson estimate of the dispersion parameter $\hat{\phi} = 0.5444$. The analysis of deviance table is:

Source	Deviance	Degrees of freedom	Mean Deviance	*F*	*p*-value
`Origin`	19.89	2	9.945	18.27	$< 2 * 10^{-16}$
Log(`DHB`)	328.01	1	328.01	602.54	$< 2 * 10^{-16}$
Interaction	7.89	2	3.945	7.247	0.001
Residual	152.69	379			
Total	508.48	384			

All terms are significant. From $\hat{\beta}_j$ (results not shown), we can conclude: (1) that there is little evidence of a difference between the natural and coppice trees. (2) The coefficient for `DBH` is 1.843, which is close to the expected value 2.

Exercise 5.4: Analysis of Deviance for Inverse Gaussian GLMs

Fit GLM(Inverse Gaussian; log) for the small-leaved lime data in **Example 5.10** (data set: `lime`), find $\hat{\phi}$, find the analysis of deviance table and discuss $\hat{\beta}_j$. Compare the results with the results from **Example 5.11**.

Solution: We can obtain the Pearson estimate of the dispersion parameter $\hat{\phi} = 1.256$. The analysis of deviance table is:

Source	Deviance	Degrees of freedom	Mean Deviance	F	p-value
`Origin`	10.48	2	5.24	4.172	0.0161
Log(`DHB`)	431.79	1	431.79	343.78	$< 2 * 10^{-16}$
Interaction	24.51	2	12.25	9.759	$7.37 * 10^{-5}$
Residual	406.81	379			
Total	873.60	384			

Note that it may not be a good idea to use the AIC or BIC to compare GLM(Gamma; log) and GLM(inverse Gaussian; log) since AIC or BIC are based on the likelihood function while two models use the different distributions. Based on the deviance, GLM(Gamma; log) is preferred.

Section 5.2.6 Case Studies

Case Study 1: Gamma GLMs

The data set `motorins` from R package **faraway** contains the data on payments (`Payment`) for insurance claims for various areas (`Zone`) of Sweden in 1977. The data is further subdivided by mileage driven (`Kilometers`), the bonus from not having made previous claims (`Bonus`) and the type of car (`Make`). We have information on the number of insured (`Insured`), measured in policy-years, within each of these groups. In this case study, we focus on the data from Zone 1 and the purpose of this study is to evaluate how the claims (`Payment`, y, as the response) are affected by the mileage driven (`Kilometers`), the number of insured (`Insured`), the type of car (`Make`), and the bonus (`Bonus`). The following questions should be answered in order:

(1) What model should be used, a GLM or the linear regression?
(2) What GLM should be used? This depend on the variance function.
(3) Is model appropriate based on the diagnostics?
(4) What conclusions can be obtained?

The response is positive continuous, the linear regression, Gamma GLM, or inverse Gaussian GLM can be used. From the histogram of y (`Payment`) (Plot A of **Figure 5.10**), we can see that the data are heavily right skewed, so a Gamma or inverse Gaussian may be fine. For such data, sometime the linear model also works for transformed response. From the histogram of $\log(y)$

(Plot B of **Figure 5.10**), we can see that the data look more normally distributed. Thus a linear model with $\log(y)$ as the response can also be used.

To determine which GLM should be used, we first to see if the variance function $V(\mu) = \mu^c$. The following analysis is performed: (1) plit the data into smaller groups according to the miles driven (`Kilometres`) and the car type (`Make`); (2) Calculate the sample means and sample variances ; (3) plot the group variances against the group means; (4) plot the logarithm of group variances against the logarithm of group means; (4) fit a linear model. For the plot (Plot D of **Figure 5.10**). We can see that the logarithm of variances increases linearly with the logarithm of means. The estimated slop from the linear regression is 1.94, suggesting the variance function $V(\mu) \approx \mu^2$ is a good approximation.

We fit two models: Gamma GLM with the logarithmic link function and the linear model with $\log(y)$ as the response. Since we expect that the total amount of the claims for a group will be proportionate to the number of insured, it makes sense to treat the logarithm of the number insured as an offset in the model. The variable `Kilometres` is a factor to representing kilomoters per year but is considered as a covariate in the model.

There are important differences between the two models. We can see that mileage class given by Kilometers is statistically significant in the Gamma GLM, but not in the linear model. Some of the coefficients are quite different. For example, we see that for `Make 8`, relative to the reference level of `Make 1`, there are $\exp(0.7504) = 2.1178$ times as much payment when using the Gamma GLM, while the comparable figure for the linear model is $\exp(0.20958) = 1.2332$.

These two models are not nested and have different distributions for the response, which makes direct comparison problematic. The AIC or BIC criterion, which is the maximized likelihood may not be used here since two models use different distribution. From R, we can see that the AIC/BIC form the Gamma GLM is much higher than the AIC/BIC from the linear model. Nevertheless, we note that the null deviance for both models is almost the same (238.97 vs 238.56) while the residual deviance is smaller for the Gamma GLM (155.06 versus 173.53). This improvement relative to the null indicates that the Gamma GLM should be preferred here. Note that purely numerical comparisons such as this are risky and that some attention to residual diagnostics, scientific context and interpretation is necessary.

Now we focus our analysis on the Gamma GLM. From **Figure 5.11**, the model may still need to be improved. There is a clear linear relationship in plot B. The Q-Q plot looks fine too. Some observations produce large residuals, and the Cook's distance is less than 0.25, so there are no influential observations. There are some problems from plot A: the variances of standardized deviance residuals decrease with larger fitted values. One reason is that some important variables may be missing. This may be also due to the smaller number of observations in that range.

For the Gamma GLM, the residual deviance and degrees of freedom are 155.06 and 284, respectively. The p-value obtained from χ^2_{284} is 1.000, indicating the model is sufficient.

For conclusions, we focus on the analysis of variance (**Table 5.6**). The deviance and F-test statistic are obtained by comparing the full model and the model without that variable ("Type III" deviance). From the table, we can conclude that all three variables are statistically significant.

Table 5.6 The analysis of deviance for a Gamma GLM with the `motorins` data.

Source	**Deviance**	**Degrees of freedom**	**Mean Deviance**	F	p**-value**
`Kilometres`	8.41	1	8.41	15.13	$1.25 * 10^{-4}$
`Make`	29.83	8	3.73	6.708	$4.75 * 10^{-8}$
`Bonus`	45.84	1	45.84	8.245	$7.37 * 10^{-5}$
Residual	155.06	284			

We may also make predictions. For `Make = "1", Kilometres = 1, Bonus = 1, Insured = 100`, we have $\hat{\eta} = 11.052$ and $se(\hat{\eta}) = 0.154$. The point and 95% interval estimate of the response are:

$$\hat{\mu} = \exp(11.052) = 63061$$

95% CI: $\exp\left(\hat{\eta} \pm t_{\frac{\alpha}{2},284} * se(\hat{\eta})\right) = \exp(11.052 \pm 1.9684 * 0.154) = (46571, 85390)$

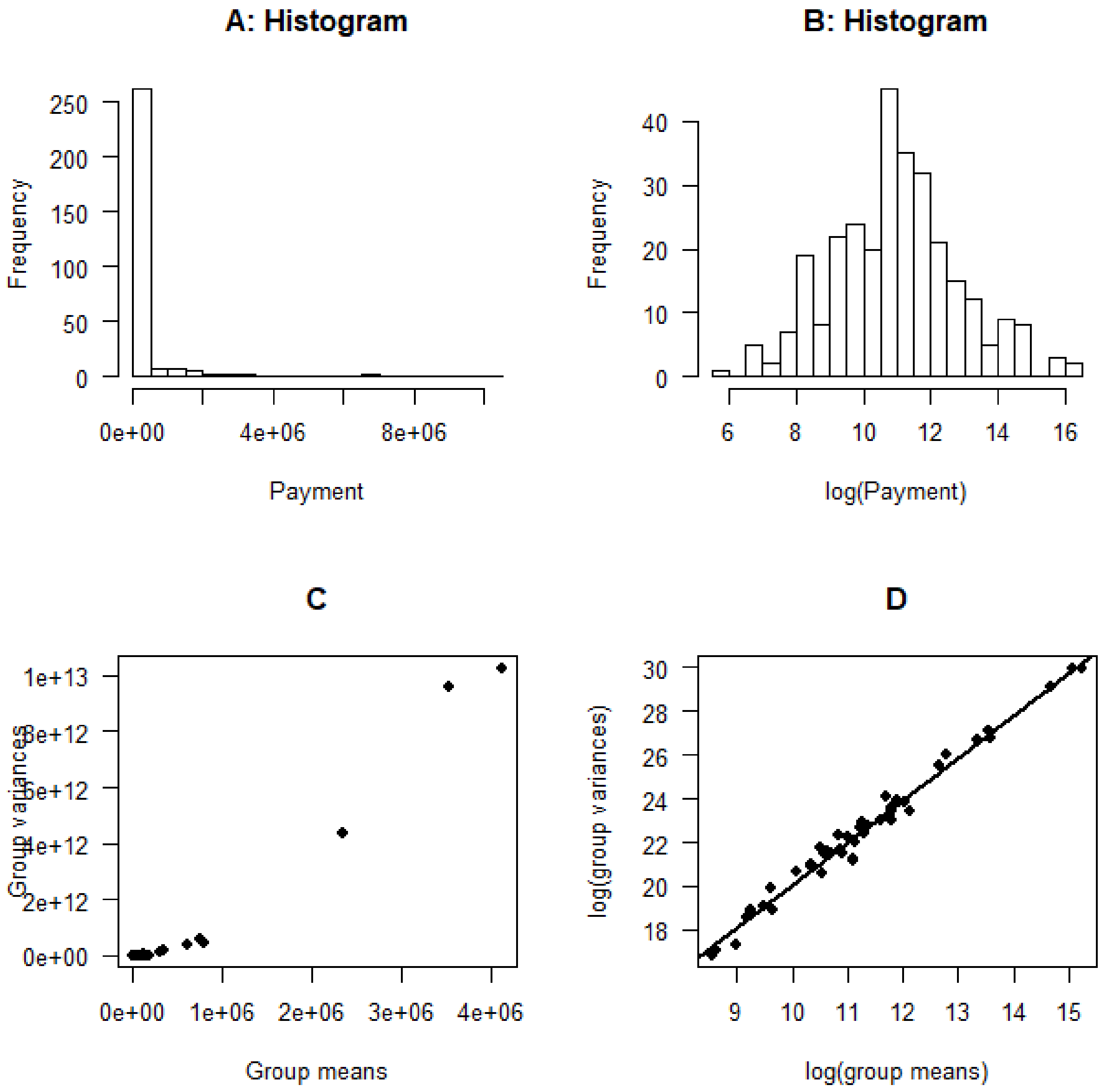

Figure 5.10 Exploratory plots from `motions` Zone 1 data: **A** - histogram of the response; **B** - histogram of the logarithm of the response; **C** – plot of the group variances against of groups means; and **D** - plot of the logarithm of group variances against of the logarithm of groups means.

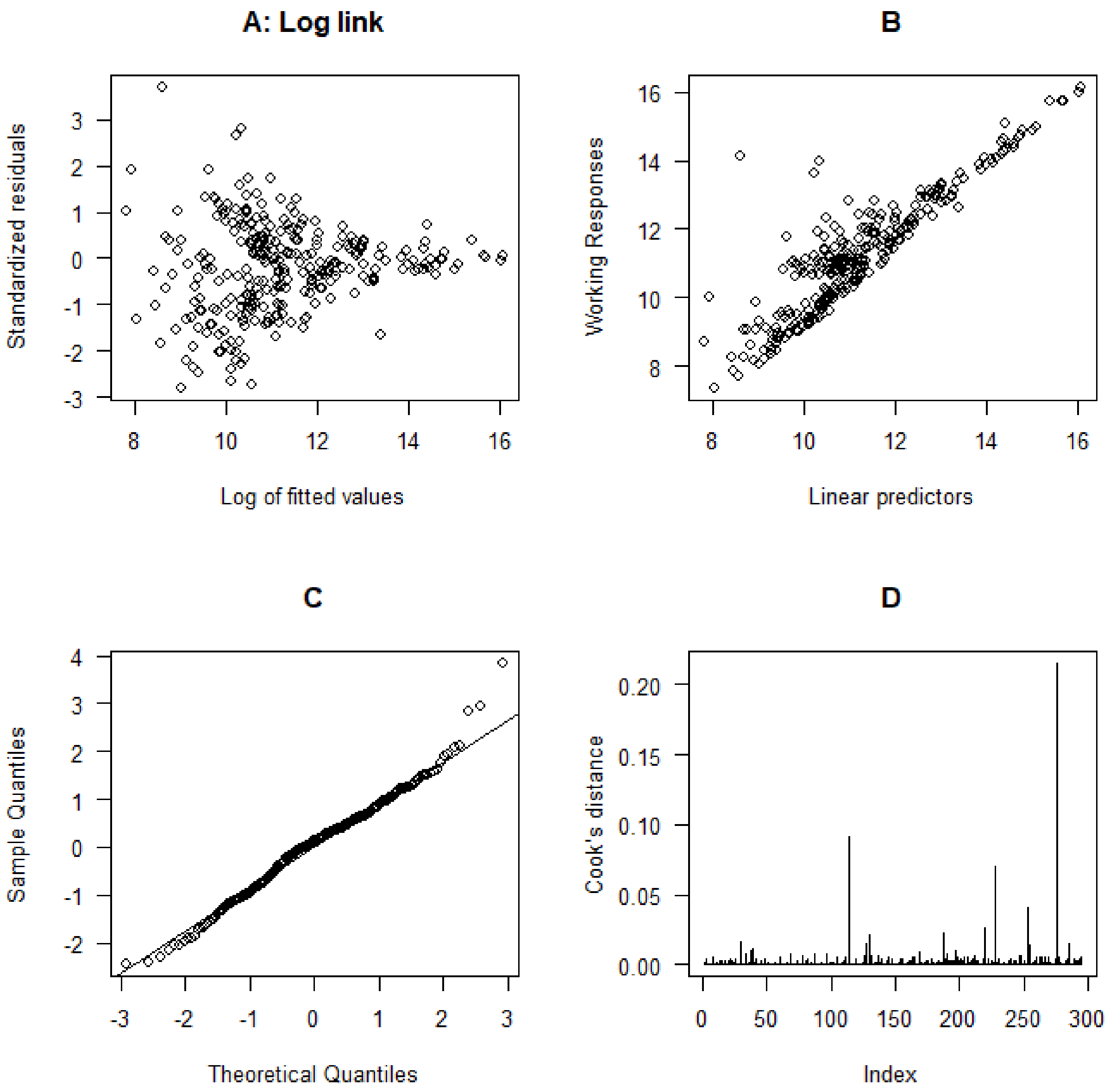

Figure 5.11 Diagnostic plots for a GLM(Gamma; log) based the `motorins` Zone 1 data. **A**: standardized residuals against logarithm of fitted values (constant-information scale); **B**: working responses against linear predictors; **C**: Q-Q plot of quantile residuals; and **D**: Cook's distance.

Exercise 5.5: Gamma GLM and Linear Model

For `motorins` Zone 7 data, perform the analysis that are similar to the analysis from Case Study 1.

Case Study 2: Inverse Gaussian GLMs

In a study of sheets of building materials, the permeability of three sheets was measured on three different machines over nine days, for a total of 81 sheets, all of equal thickness. Each measurement is an average permeability of eight random pieces cut from each of the 81 sheets (data set: `perm`). The inverse Gaussian model may be appropriate: particles move at random according to Brownian motion through the building material assuming uniform material, drifting across the sheet. Boxplots (Plots A and B of **Figure 5.12**) show that the variance increases with the mean, and shows one large observation that is a potential outlier. Further analysis (Plots C and D of **Figure 5.12**) show that the logarithm of variances and logarithm of means have a linear relationship with an estimated slope 2.2. Because the inverse Gaussian distribution has a sensible interpretation for these data, we adopt the inverse Gaussian model. We also select the logarithmic link function, when the parameters are interpreted as having a multiplicative effect on the response. The model is fitted with the main effects of two factors: days and machines.

The analysis of deviance shows that `Day` is not significant after the adjustment of `Mach`: the deviance, F-test statistic, and the p-value for Day are 0.069, 1.56, and 0.153, respectively. So we omit Day from the model. Recall the deviance has an exact distribution for the inverse Gaussian distribution, so these results do not rely on small-dispersion or large-sample asymptotics.

Machine A is used as the reference level, the estimated coefficients for Machine B and C are -0.63898 and -0.17286, respectively. Therefore, the model suggests the permeability measurements on Machine B are, on average, $\exp(-0.63898) = 0.5278$ times those for Machine A. Likewise, the permeability measurements on Machine C are, on average, $\exp(0.1729) = 0.8413$ times those for Machine A. The p-values suggest Machine C is very similar to Machine A, but Machine B is different.

We can now examine the fitted model to determine if the large observation identified in **Figure 5.12** is an outlier, and if it is influential. No residuals appear too large. The largest Cook's distance is 0.35, so no observations are influential either.

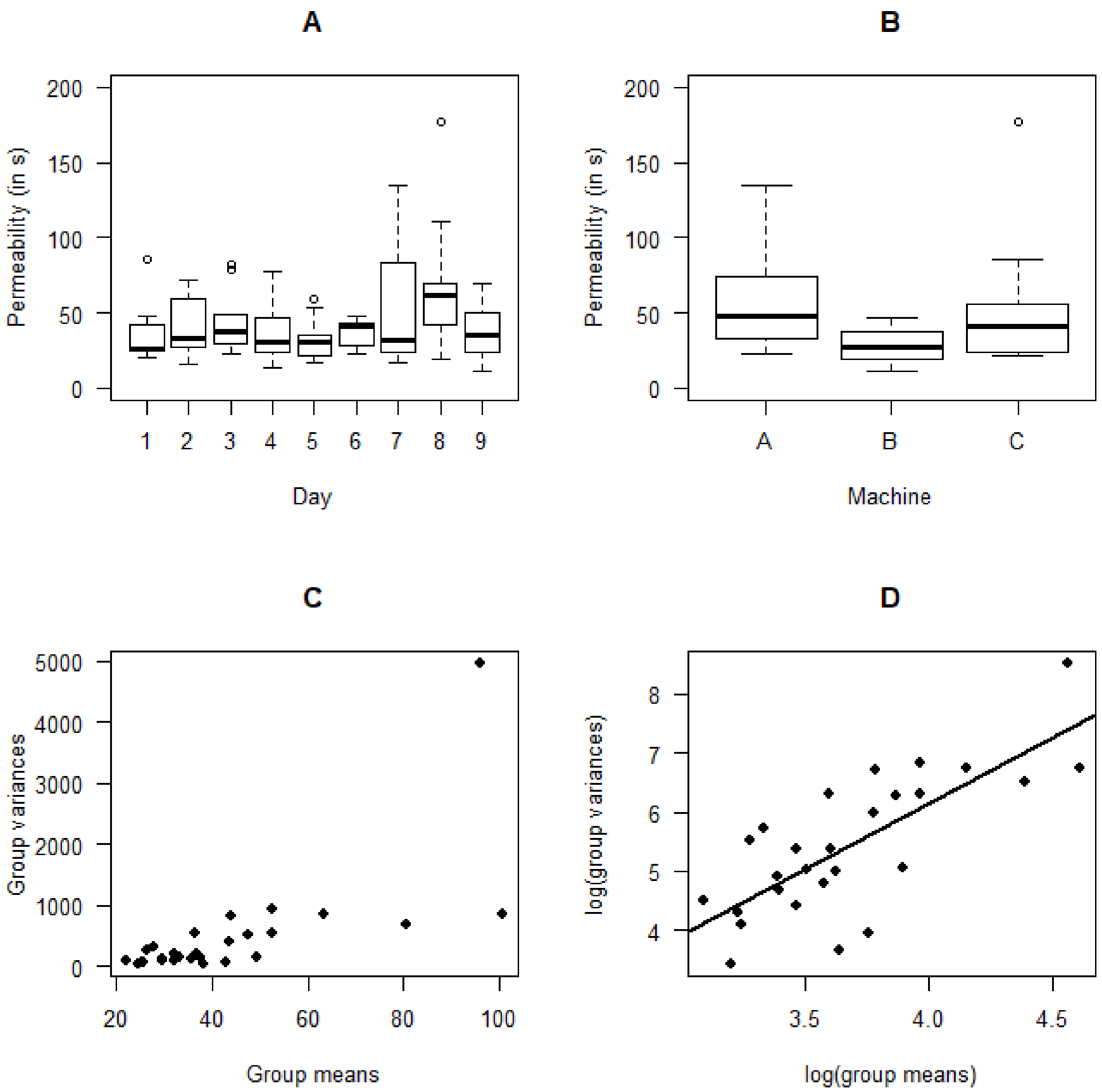

Figure 5.12 Exploratory plots from `perm` data set. **A** – boxplots of the response (permeability) stratified by days; **B** - boxplots of the response (permeability) stratified by machines; **C** – plot of the group variances against of groups means; and **D** - plot of the logarithm of group variances against of the logarithm of groups means.

Table 5.2 Comparison of a binomial GLM and a quasi-binomial GLM based on the data set `germ`.

Inference	Inference	Binomial	Quasi-Binomial	Comparison	Comments
Parameter for Extract	$\hat{\beta}_1$	0.5401	0.5401	Same	ϕ dose not affect estimation
	$se(\hat{\beta}_1)$	0.2498	0.3409	Different	$\frac{0.3409}{0.2498} = 1.3645 = \sqrt{\hat{\phi}}$
	z	2.1619	1.5844	Different	Binomial: Z test Quasi: t-test
	p-value	0.0306	0.1315	Different	
Residual DF		17	17	Same	
Log-likelihood		-54.93702	*NA*	Different	
Deviance		33.28	33.28	Same	So same deviance residuals
Pearson stat		31.65	31.65	Same	So same Pearson residuals
Analysis of deviance `Seeds`	Residual deviance	39.686	39.686	Same	
	Deviance	3.065	3.065	Same	
	Test statistics	3.065	1.6462	Different	Binomial: Chi-square test Quasi: F-test
	p-value	0.080	0.217	Different	

MA5771: Applied Generalized Linear Model

Week 6 At-a-Glance

Title: Model for Counts: Poisson and Negative Binomial GLMs

Overview

In this week’s lesson, we will discuss several GLMs, including Poisson GLMs, negative binomial GLMs, and quasi-Poisson GLMs, to analyze count data. We will focus on Poisson GLMs for three types of count data: models with covariates, models for rates, and models for contingency tables. We will discuss the overdispersion problem in Poisson GLMs and then introduce two models to tackle the overdispersion problem: negative binomial GLMs and quasi-Poisson models as alternative models.

Learning Objectives

When you complete this module, you should be able to:

1. Use appropriate Poisson GLMs to analyze count data, including counts with covariate, rates, and counts that can be organized as tables.
2. Use appropriate Poisson GLMs to analyze contingency tables, including two-dimensional tables, three-dimensional tables, high-order tables, and tables with structural zeros.
3. Detect the overdispersion in Poisson GLMs and tackle the overdispersion problem with either negative binomial GLMs and/or quasi-Poisson models.
4. Choose a suitable GLM among several Poisson, negative binomial, and quasi-Poisson GLMs.

Instruction Content

See others file for details.

Exam

NA

Quiz

See other files for details.

Homework

See other files for details.

MA5771: Applied Generalized Linear Model

Week 6 Instruction Contents

Lesson 6.1 Models for Counts: Poisson GLMs

Related Readings: Sections 10.1, 10.2, 10.3, and 10.4 in Chapter 10.

Section 6.1.1 Introduction and Overview

Data in the form of counts arise often in practice. Examples include: the number of cases of leukemia reported per year in a certain jurisdiction; the number of flaws per meter of electrical cable. In this lesson, we focus on Poisson GLMs for counts when the events being counted are independent, where there is no clear upper limit for the number of events that can occur, or where the upper limit is very much greater than any of the actual counts. In **Section 6.1.2**, we list important information about the Poisson distribution and Poisson GLMs. We then focus on describing the models for two types of count data: models for rates (**Section 6.1.3**) and models for counts organized in tables. Different GLMs are describe for two-dimensional tables (**Section 6.1.4**), three-dimensional tables (**Section 6.1.5**), high-order tables (**Section 6.1.6**), and tables with structural zeros (**Section 6.1.7**).

Section 6.1.2 Summary of Poisson GLMs

The distribution most often used for modelling counts is the Poisson distribution, which has the probability function

$$\mathcal{P}(y;\mu) = \frac{\exp(-\mu)\mu^y}{y!}$$

for $y = 0, 1, 2, \ldots$, with expected counts $E[y] = \mu > 0$. The Poisson distribution has already been established as an EDM, and a Poisson GLM proposed for the noisy miner data in previous lessons. Useful information about the Poisson distribution and Poisson GLMs is listed below:

(1) The Poisson distribution function can be written as:

$$\mathcal{P}(y;\mu) = \frac{\exp(-\mu)\mu^y}{y!} = \frac{1}{y!}\exp(y\log(\mu) - \mu)$$

So the nature parameter is $\log(\mu)$, the dispersion parameter is $\phi = 1$, and the canonical link function is the logarithmic link: $g(\mu) = \log(\mu)$.

(2) Since $\phi = 1$, the inference of Poisson GLMs is based on the z-test and χ^2 distribution.

(3) The variance function is the identity function: $V(\mu) = \mu$.

(4) The residual deviance of Poisson GLMs is:

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = \sum\nolimits_{i=1}^{n} w_i * 2\left\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right\}$$

where w_i are the prior weights. When $y_i = 0$, since $lim_{y_i \to 0} y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) = 0$, the corresponding deviance becomes $w_i * 2\{0 - (0 - \hat{\mu}_i)\} = 2w_i\hat{\mu}_i$. $D(\boldsymbol{y};\hat{\boldsymbol{\mu}})$ is approximate $\chi^2_{n-(p+1)}$ and such approximation is adequate if $\min(y_i) \geq 3$.

(5) The most common link function used for Poisson GLMs is the canonical link function – the logarithmic link, which ensures $\mu > 0$ and enables the regression parameters to be interpreted as having multiplicative effects.

Using the logarithmic link function, the general form of a Poisson GLM is

$$\begin{cases} y_i \sim \text{Poisson}(\mu_i) \\ \log(\mu_i) = \beta_0 + \sum\nolimits_{j=1}^{p} \beta_j x_{ji} \end{cases}$$

The systematic component of Poisson GLMs can be written as

$$\mu_i = \exp\left(\beta_0 + \sum\nolimits_{j=1}^{p} \beta_j x_{ji}\right)$$

$$= \exp(\beta_0) * \exp(\beta_1 x_{1i}) * \cdots * \exp(\beta_p x_{pi})$$

This shows that the impact of each explanatory variable is multiplicative. Increasing x_{ji} by one increases μ_i by factor of $\exp(\beta_j)$. If $\beta_j = 0$ then $\exp(\beta_j) = 1$ and μ_i is not related to x_{ji}. If $\beta_j > 0$, then $\exp(\beta_j) > 1$ so μ_i increases if x_{ji} increases; ff $\beta_j < 0$, then $\exp(\beta_j) < 1$ so μ_i decreases if x_{ji} increases.

(6) Sometimes, the link functions "identity" ($\eta = g(\mu) = \mu$) or "sqrt" ($\eta = g(\mu) = \sqrt{\mu}$) are used with Poisson GLMs.

(7) A Poisson GLM is denoted GLM(Poisson; link), and is specified in R using `family = poisson()` in the R function `glm()`.

When the explanatory variables are all qualitative (that is, factors), the data can be summarized as a **contingency table** and the model is often called a **log-linear model** (**Section 6.1.4**). If any of the explanatory variables are quantitative (that is, covariates), the model is often called a **Poisson regression model**. Since Poisson regression has been discussed as a case study earlier (**Section 4.1.5**), we do not consider Poisson regression models further but only present a case study using Poisson GLMs.

When the linear predictor includes a constant (intercept) term (as is almost always the case), and the log-link function is used, the residual deviance can be simplified to

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = \sum\nolimits_{i=1}^{n} w_i * 2\left\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right)\right\}$$

that is, the second term in the unit residual deviance can be dropped as it sums to zero. This identity will be used later to clarify the analysis of contingency tables. The proof of this identity is given in **Example 6.1.** You can skip this example but need to understand the conclusion of it.

Example 6.1 A Property of Residual Deviance for Poisson GLM with log Link

Show that if $\log(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}$ in a Poisson GLM with the weight w_i, then

$$\sum\nolimits_{i=1}^{n} w_i(y_i - \hat{\mu}_i) = 0$$

Proof: Note that $\log(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}$ and $\mu_i = \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji})$

$$\mathcal{P}(y_i; \mu_i, w_i) = \frac{1}{y_i!}\exp(\frac{y_i \log(\mu_i) - \mu_i}{1/w_i})$$

$$\ell_i = \log\big(\mathcal{P}(y_i; \mu_i)\big) = -\log(y_i!) + \frac{y_i \log(\mu_i) - \mu_i}{1/w_i}$$

$$= -\log(y_i!) + \frac{y_i(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}) - \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji})}{1/w_i}$$

So the log-likelihood function is

$$\ell = \sum_{i=1}^{n} \left\{ -\log(y_i!) + \frac{y_i(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}) - \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji})}{1/w_i} \right\}$$

We can obtain that the score for β_0:

$$\frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{n} \left\{ \frac{y_i - \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}\right)}{1/w_i} \right\}$$

Note that the maximum likelihood estimator of $\hat{\boldsymbol{\beta}}$ satisfies

$$0 = \sum_{i=1}^{n} \left\{ \frac{y_i - \exp\left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ji}\right)}{1/w_i} \right\} = \sum_{i=1}^{n} \frac{y_i - \hat{\mu}_i}{1/w_i} = \sum_{i=1}^{n} w_i(y_i - \hat{\mu}_i)$$

Section 6.1.3 Modeling Rates

We describe the Poisson GLMs for the counts that are number of independent events:

- The **maximum number** of events is **known but large**; that is, there is an upper bound for each count response, but the upper bound is very large.
- The maximum number of events is usually representative of some populations.
- The size of each population needs to be specified to make comparisons meaningful.
- The response can be usefully viewed as a **rate** rather than just as a count.
- In principle, rates can be treated as proportions, and analyzed using binomial GLMs, but Poisson GLMs are more convenient when the populations are large and the rates are relatively small, less than 1% say. This is because the binomial distribution can be approximated by the Poisson distribution when the rates are small.
- In Poisson GLMs, the logarithm of the population size should be considered as an offset in the model.

For example, consider comparing the number of people with a certain disease in various cities. The number of cases in each city may be useful information for planning purposes. However, quoting just the number of people with the disease in each city is an unfair comparison, as some cities have a far larger population than others. Comparing the number of people with the disease per unit of population (for example, per thousand people) is a fairer comparison. That is, the disease rate is often more suitable for modelling than the actual number of people with the disease.

Before we look at an example, let us look at why the logarithm of the population size should be used as an offset in Poisson GLMs. Define y_i as the count for the ith population and T_i as the size of corresponding population. The rate per unit of population is $\frac{y_i}{T_i}$, and the expected rate is

$$E\left[\frac{y_i}{T_i}\right] = \frac{\mu_i}{T_i}$$

where μ_i possibly depends on the explanatory variables, and T_i is known. Using a logarithmic link function, the suggested systematic component is

$\log\left(\frac{\mu_i}{T_i}\right) = \eta_i$, which is equivalent to $\log(\mu_i) = \log(T_i) + \eta_i$

Therefore the model suggested for cancer rates is

$$y_i \sim \text{Poisson}(\mu_i)$$

$$\log(\mu_i) = \log(T_i) + \eta_i$$

You can see that $\log(T_i)$ should be an offset in the systematic part. **Keep in mind that in order to model rate with the count data and the Poisson GLM in R, you still need to use the count (y_i) not the rate ($\frac{y_i}{T_i}$) as the response and the logarithm of population size ($\log(T_i)$) as the offset. Please refer to the R code for Example 6.2 to find out how to correctly use R to fit a Poisson GLM to model the rate.**

Example 6.2: Poisson GLMs for Rates

As a numerical example, consider the number of incidents of lung cancer from 1968 to 1971 in four Danish cities (data set: `danishlc`), recorded by age groups. Fit a Poisson GLM and draw your conclusions.

Solution: The following steps are performed in order: (1) exploratory analysis; (2) model building; (3) checking approximation; (3) model diagnostics; and (5) conclusions.

Exploratory Analysis. For this data, it is more informative to model the rate – the number of lung cancer cases per unit of population than to model the number of cases. For example, the number of cases of lung cancer in each age group (11, 11, 11, 0, 11, and 10) is remarkably similar for Fredericia. However, the rates differ substantially. A plot of the cancer rates against city and age

(**Figure 6.1**) suggests the lung cancer rate may change with age. In addition, the rates are between 0.127% and 0.222% so a Poisson GLM instead of a Binomial GLM is used here.

Model Building. This is to determine which GLMs, link function, and explanatory variables should be used.

Note that here Age is created as an order factor. The default coding for an ordered factor in R is called "polynomial contrasts", which are not appropriate here (the ordered categories are not equally spaced) and are hard to interpret anyway. To instruct R to use the familiar treatment coding for ordered factors, use the R function `options()`:

```
options(contrasts = c("contr.treatment", "contr.treatment"))
```

The first input tells R to use treatment coding for unordered factors (which is the default), and the second to use treatment coding for ordered factors (rather than the default ”`contr.poly`”). You can use the R function `model.matrix()` to see how they differ using the treatment coding and polynomial contrast.

In addition, you can see **that `Cases` is used as the response and the logarithm of `Pop` is used as the offset in the R function `glm()` although the Poison GLM is used to model the rate**.

After a Poisson GLM with the interaction is fitted (keep in mind that the logarithm of population size is an offset in the model), we can obtain the analysis of deviance table:

Source	**Deviance**	**Degrees of freedom**	**Chi-square**	***p*-value**
`City`	3.393	3	3.393	0.335
`Age`	103.068	5	103.068	$< 2 * 10^{-16}$
Interaction	23.447	15	23.447	0.075
Residual	0	0		
Total	129.908	23		

First, the residual deviance is 0 and has 0 degrees of freedom. Second, that both the city and interaction effects between the city and age are not significant so the final model only include the age as the explanatory variable.

Now, we look at an alternative Poisson GLM which considered `Age` as quantitative (since the categories are not equally spaced) and use the lower class boundary of each class. The lower boundary is preferred since the final class only has a lower boundary. In addition, **Figure 6.1** may suggest a possible quadratic relationship. The AIC for the four models and the analysis of deviance of nested models are listed in **Table 6.1**. We can see that there is no sufficient evidence to suggest that the interaction model is different from the model with `Age` only. The quadratic model is an improvement over the linear model. The quadratic model has the smallest AIC among four models compared but its AIC is similar to the AIC of the model with `Age` only.

Checking approximations. The saddlepoint approximation is suitable for Poisson distributions when $y_i \geq 3$ for all observations. For this data, only one observation is less than 3. So we may need to be cautious when we use the goodness-of-fit tests.

Model Diagnostics. The goodness-of-fit tests show that both the model with Age only and the quadratic model are reasonably adequate (**Table 6.1**).

Note that diagnostic plots (**Figure 6.2**), the quantile residuals are used. Also, the constant-information scale of $\hat{\mu}_i$ ($\sqrt{\hat{\mu}_i}$ for Poisson GLMs) are used. The diagnostic plots suggest that both models are reasonable models, though we prefer the model with the age as the factor, since the quadratic model appears to show three observations with high influence relative to the other observations, and is a simpler model.

Conclusions. For the model with the age as the factor, one interesting question is to see how the rates differ according to the age groups. **Table 6.2** lists the estimated differences of two adjacent age groups and **simultaneous 95% confidence intervals** of β_j. The results are obtained by the R function `glht()` in R package **multcomp** and `confint()`.

Since we need to calculate 5 confidence intervals, the **simultaneous** $100(1-\alpha)\%$ confidence intervals are calculated. The confidence level for a single confidence interval is based on the probability that the random interval will be "correct" (meaning that the random interval will contain the true value of the contrast or function). It is shown below that when several confidence intervals are calculated, the probability that they are all simultaneously correct can be alarmingly small. The wider confidence intervals are needed to guarantee that the **overall confidence level** is at least $1-\alpha$. The overall confidence level is the probability that all parameters are simultaneously contained in the corresponding confidence interval. The same principle is also applied to situations where multiple hypothesis tests are conducted.

Various methods have been developed to ensure that the overall confidence level is not too small and the overall significance level is not too high. In this course, we introduce a simple method called **Bonferroni method** to calculate the simultaneous $100(1-\alpha)\%$ confidence intervals and the p-values when there are multiple tests. Let m be the number of confidence intervals (or tests), then the simultaneous $100(1-\alpha)\%$ confidence intervals for a parameter is:

$$point\ estimate \pm z_{\frac{\alpha}{2m}} * standard\ error$$

Comparing with the formula $100(1-\alpha)\%$ confidence intervals, the critical value is $z_{\frac{\alpha}{2m}}$ instead of $z_{\frac{\alpha}{2}}$ is used. This corresponds to construct a $100(1-\alpha^*)\%$ confidence intervals, where $\alpha^* = \frac{\alpha}{m}$. The simultaneous $100(1-\alpha)\%$ confidence intervals can be obtained by specifying the appropriate critical value $z_{\frac{\alpha}{2m}}$ using `confint(, calpha = )` (Please refer to R program for more details).

Similarly, the Bonferroni adjusted p-value is:

$$m * 2 * \Pr(Z > |z|) = m * original\, p - \text{value}$$

Let us illustrate how we can use the Bonferroni method to calculate the simultaneous 95% confidence interval of β_1.

Without the Bonferroni method, the critical value is: $z_{0.025} = 1.96$. There are 5 confidence intervals, so the new critical value is: $z_{\frac{0.05}{2*5}} = z_{0.005} = 2.5758$. Therefore the 95% confidence interval for β_1 is:

- Without adjustment: $1.082 \pm 1.96 * 0.2481 = (0.596, 1.568)$
- With Bonferroni method: $1.082 \pm 2.5758 * 0.2481 = (0.443, 1.721)$

Table 6.2 lists the simultaneous 95% confidence interval of difference of coefficients between two adjacent age groups. Note that the age group $44-54$ is the refence group. We can see only the first simultaneous 95% confidence interval does not contain 0, meaning the lung cancer rates are significantly different between these two groups. Since $\exp(\hat{\beta}_1) = \exp(1.082) = 2.95$, the cancer rate in the age group $55 - 59$ is about 2.95 times the cancer rate in the age group $44 - 54$.

We do not know if the cancer rates between two non-adjacent age groups are significantly different. Such conclusions should be based on the simultaneous confidence intervals of appropriate linear combinations of β_j. For example, we need to obtain the simultaneous confidence intervals of $\beta_5 - \beta_3$ if we would like to know if the cancer rates are different between the age group > 74 and the age group $64 - 69$.

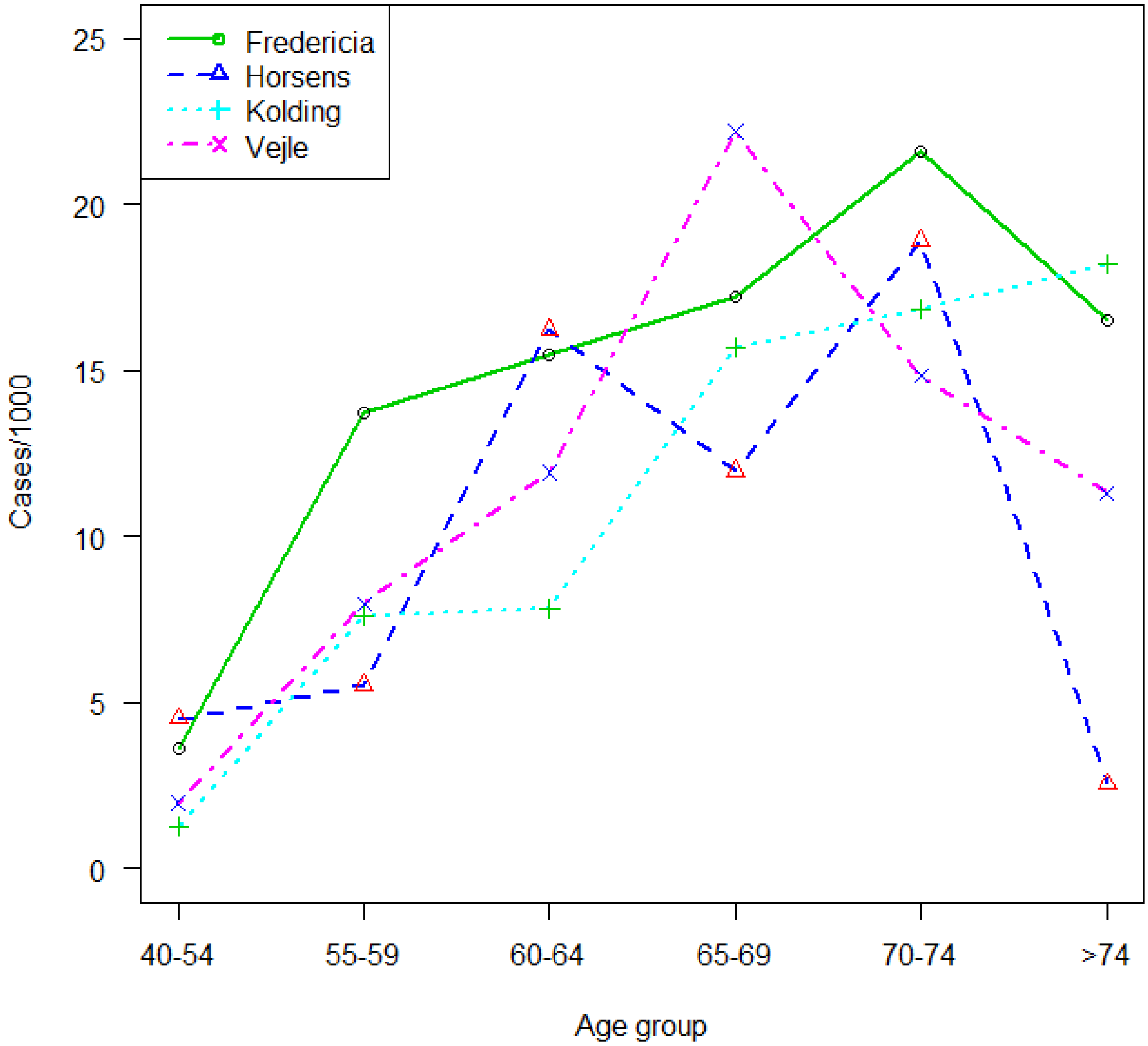

Figure 6.1 The Danish lung cancer rates for various age groups in different cities.

Table 6.1 The AIC and the analysis of deviance.

Model	**Model**	**Deviance**	**Deviance DF**	p**-value**	**AIC**	χ^2	p**-value**
1	With Interaction	0	0		144.39	28.31	0.0575
2	Factor `Age`	28.31	18	0.058	136.69		
3	Numerical `Age`	49.0	22	$8 * 10^{-4}$	149.36	16.468	$4.95 * 10^{-5}$
4	Quadratic	32.50	21	0.052	134.89		

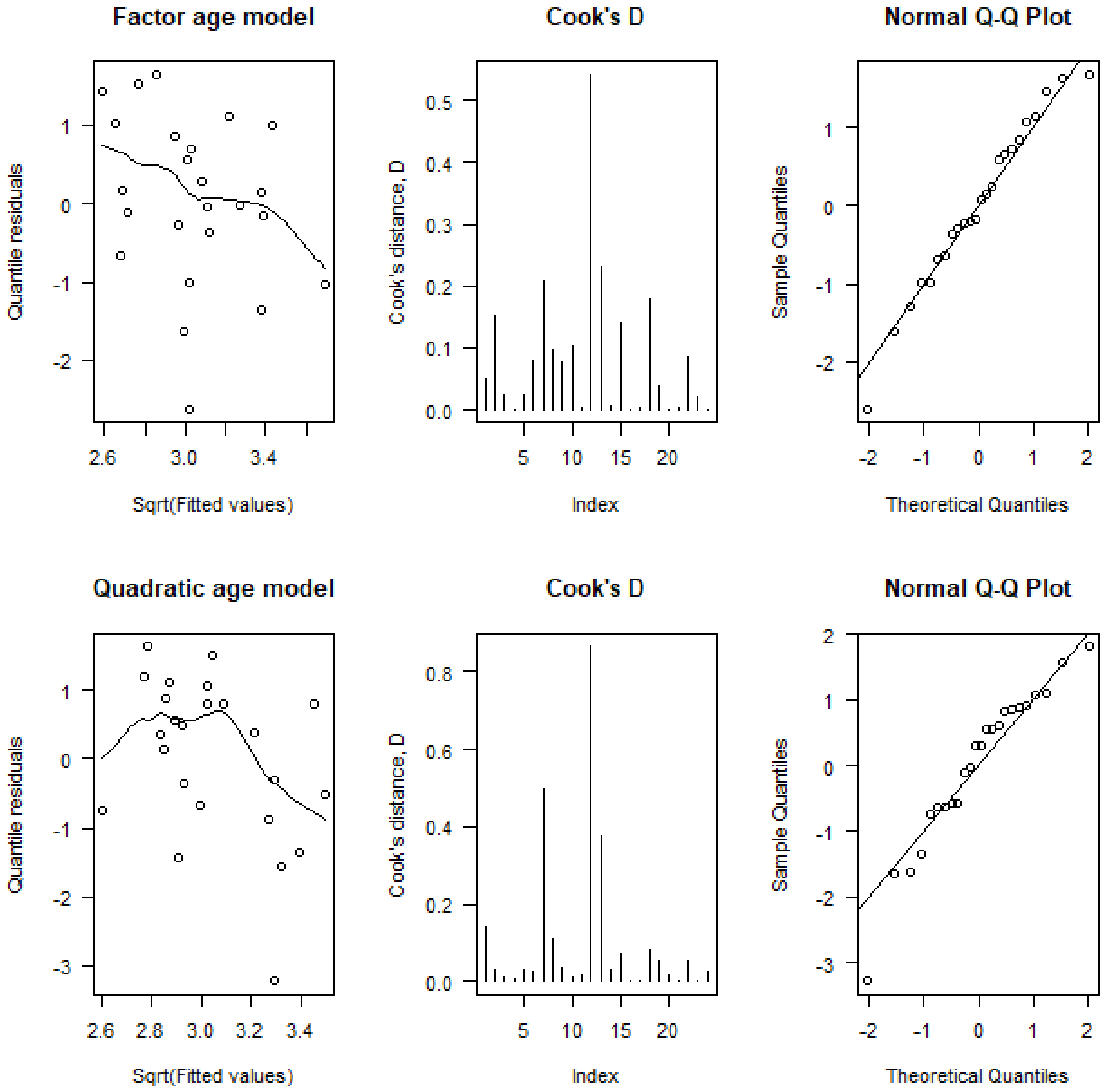

Figure 6.2 Diagnostic plots for two models fitted model to the Danish lung cancer data. **Top panels**: treating age as a factor; **bottom panels**: fitting a quadratic in age. The quantile residuals are used in all plots.

Table 6.2 The point and interval estimates of linear combinations of parameters from the Poisson GLM Model in **Example 6.2**.

Age Group 1	Age Group 2	Parameter	Point Estimate	Standard Error	95% CI	95% CI (Bonferroni)
$\mathbf{55-59}$	$\mathbf{40-54}$	β_1	1.082	0.2481	$(0.596, 1.569)$	$(0.443, 1.721)$
$\mathbf{60-64}$	$\mathbf{55-59}$	$\beta_2-\beta_1$	0.491	0.2335	$(-0.038, 0.877)$	$(-0.182, 1.021)$
$\mathbf{65-69}$	$\mathbf{60-64}$	$\beta_3-\beta_2$	0.249	0.2133	$(-0.169, 0.666)$	$(-0.301, 0.798)$
$\mathbf{70-74}$	$\mathbf{65-69}$	$\beta_4-\beta_3$	0.097	0.2173	$(-0.329,0.523\ )$	$(-0.463, 0.657)$
$\mathbf{>74}$	$\mathbf{70-74}$	$\beta_6-\beta_3$	-0.439	0.2393	$(-0.908, 0.030)$	$(-1.055, 0.177)$

Exercise 6.1: Poisson GLMs for Rates

Based on Results in **Table 6.2**, verify the following results:

Age Group 1	Age Group 2	Parameter	Point Estimate	Standard Error	$\boldsymbol{p}$-value	$\boldsymbol{p}$-value (Bonferroni)
$\mathbf{55-59}$	$\mathbf{40-54}$	β_1	1.0823	0.2481	$1.29*10^{-5}$	$6.43*10^{-5}$
$\mathbf{60-64}$	$\mathbf{55-59}$	$\beta_2-\beta_1$	0.4913	0.2335	0.0725	0.362
$\mathbf{65-69}$	$\mathbf{60-64}$	$\beta_3-\beta_2$	0.2486	0.2133	0.2437	1.000
$\mathbf{70-74}$	$\mathbf{65-69}$	$\beta_4-\beta_3$	0.09769	0.2173	0.6555	1.000
$\mathbf{>74}$	$\mathbf{70-74}$	$\beta_6-\beta_3$	-0.4389	0.2393	0.0666	0.333

Solution: Let us verify the p-value for testing $H_0\colon \beta_3-\beta_2=0$. The z-test statistic is:

$$z=\frac{0.2486}{0.2133}=1.1655$$

$$p-\text{value}=2*\Pr(Z>|z|)=2*\Pr(Z>1.16555)=0.2438$$

There are 5 tests, so the Bonferroni adjusted p-value is: $0.2438*5=1.219$. Since the p-value is a probability so 1 is used here.

The point and interval estimates, the standard errors, and the p-values of **Table 6.2** and the table in **Exercise 6.1** can be obtained by the R functions `glht()` and `summary()`.

Section 6.1.4 Contingency Tables: Log-Linear Models

Count data commonly appear in tables, called **contingency tables**, where the observations are cross-classified according to the levels of the classifying factors. We start with two cross-classifying factors (two-dimensional tables) then extend to three cross-classifying factors (three-dimensional tables) and then extend to higher-order tables.

Two Dimensional Tables: Systematic Component

The simplest contingency table is a two-way (or two-dimensional) table, with factors A and B. If factor A has I levels and factor B has J levels, the contingency table has size $I \times J$. In general, the entries in an $I \times J$ table are defined as shown in **Table 6.3**, where y_{ij} refers to the observed count in row i and column j for $i = 1, 2, \ldots I$ and $j = 1, 2, \ldots J$.

Table 6.3 The general $I \times J$ contingency table. The cell count y_{ij} corresponds to level i of A and level j of B.

		Factor *B*				
		Column 1	**Column 1**	$\cdots$	**Column *J***	**Row Total**
Factor *A*	**Row 1**	y_{11}	y_{12}	$\cdots$	y_{1J}	$m_{1\cdot}$
	Row 2	y_{21}	y_{22}	$\cdots$	y_{2J}	$m_{2\cdot}$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	Row *I*	y_{I1}	y_{I2}	$\cdots$	y_{IJ}	$m_{I\cdot}$
	Column Total	$m_{\cdot 1}$	$m_{\cdot 1}$	$\cdots$	$m_{\cdot 1}$	m

Write μ_{ij} for the expected count in cell (i, j). For convenience, also define π_{ij} as the expected probability that an observation is in cell (i, j), where $\mu_{ij} = m\pi_{ij}$, and m is the total number of observations. We write $m_{i\cdot}$ to mean the sum of counts in row i over all columns, and $m_{\cdot j}$ to mean the sum of counts in column j over all rows. The use of the dot · in this context means to sum over all the elements of the index that the dot replaces. For example, $m_{\cdot 2} = m_{12} + m_{22} + \cdots + m_{I2}$.

If factors A and B are independent, then $\pi_{ij} = \pi_{i\cdot} * \pi_{\cdot j}$ is true. Writing

$$\mu_{ij} = m\pi_{ij} = m\pi_{i\cdot}\pi_{\cdot j}$$

take logarithms to obtain

$$\log(\mu_{ij}) = \log(m) + \log(\pi_{i\cdot}) + \log(\pi_{\cdot j})$$

for the systematic component. This systematic component may be re-expressed using dummy variables, since the probabilities $\pi_{i\cdot}$ depend on which unique row the observation is in, and the probabilities $\pi_{\cdot j}$ depends on which unique column the observation is in.

Example 6.3: Two Dimensional Tables

To demonstrate and fix ideas, first consider the smallest possible table of counts: a 2×2 table. The data in **Table 6.4** were collected between December 1996 and January 1997, and comprise a two-dimensional (or two-way) table of counts collating the attitude of Australians to genetically modified (GM) foods (factor A) according to their income (factor B). The purpose of this example is to generate a data object from Table 6.4 so the data object can be analyzed by Poisson GLMs in R.

Table 6.4 The attitude of Australians to genetically modified foods (factor A) according to income (factor B) (**Example 10.2**). Note that x_1 and x_2 correspond to the dummy coding for each factor, respectively.

	High income ($\boldsymbol{x_2 = 0}$)	**Low income** ($\boldsymbol{x_2 = 1}$)	**Total**
For GM foods ($\boldsymbol{x_1 = 0}$)	263	258	521
Against GM foods ($\boldsymbol{x_1 = 1}$)	151	222	373
Total	414	480	894

Solution: We create a data frame that contains three columns: response, factor A, and factor B. Then the data frame can be analyzed by R and a table similarly with **Table 6.4** can be generated.

Two-Dimensional Tables: Random Component

The random component of Poisson GLMs for contingency tables depends on sampling schemes. A table of counts may arise from several possible sampling schemes, each suggesting a different probability model. Three possible scenarios are:

- The m observations are allocated to factors A and B as the observations randomly arrive; neither row nor column totals are fixed.
    - **Example**: A hospital records the gender and race of patients who visit the hospital for a time period. m is not known in advance. The number of patients within each gender or race are random.

- A fixed total number of m observations are cross-classified by the factors A and B.
    - **Example**: A hospital records the gender and race of first 300 patients who visit the hospital for a time period. Here m is fixed and known before the experiment starts.
- The row totals are fixed, and observations allocated to factor B within each level of A. (Alternatively, the column total are fixed, and observations allocated to factor A within each level of B.)
    - **Example**: A hospital records the race of first 150 female and 150 male patients who visit the hospital for a time period. The number of female and male patients are fixed and known in advance.

No Marginal Total Are Fixed: Random Component

In this situation, if the total number of individuals observed (the grand total in the table) can be viewed as Poisson distributed, and if the individuals give responses independently of one another, then each of the counts in the table must follow a Poisson distribution. The log-likelihood function for the $I \times J$ table is

$$\ell(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(-\mu_{ij} + y_{ij} \log(\mu_{ij})\right)$$

ignoring the terms not involving the parameters μ_{ij}. The residual deviance is

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{I} \sum_{j=1}^{J} 2\left\{y_{ij} \log\left(\frac{y_{ij}}{\hat{\mu}_{ij}}\right)\right\}$$

omitting the term $y_{ij} - \hat{\mu}_{ij}$, which always sums to zero if the log-linear predictor contains the constant term (**Example 6.1**).

Example 6.4: Log-Linear Models for Contingency Tables

The GM data (**Example 6.3** and **Table 6.4**) are collated from survey forms completed by customers randomly arriving at a large shopping center over 1 week. For this data, no marginal total is fixed; no limits exist on how large the counts can be (apart from the city population, which is much larger than the counts in the table). We would like to answer several questions:

(1) Fit a Poisson GLM with main effects only and discuss the significance of each factor.
(2) For Poisson GLM with main effects only, discuss why such model is not our main interest.

(3) Fit a Poisson GLM with the interaction effects.

Solution:

Fit a Poisson GLM with main effects only.

A Poisson GLM with main effects only can be fitted:

```
glm(Counts ~ Att + Inc, family = poisson, data = gm)
```

Recall the logarithmic link function is the default in R for Poisson GLMs. This model fits a log-linear model, and hence assumes that attitude and income are independent. Both `Att` and `Inc` are statistically significant in the order they are fitted according to the analysis of deviance.

A Poisson GLM with main effects only is not our main interest

Based on the coefficients, the model has the systematic component

$$\log(\hat{\mu}_{ij}) = 5.486 - 0.3342 * x_1 + 0.1479 * x_2$$

where x_1 and x_2 are either 0 or 1 and defined in **Table 6.4**. This systematic component is the usual regression model representation of the systematic component, where dummy variables are explicitly used for the rows and columns. Since each cell of the table belongs to just one row and one column, the dummy variables are often zero for any given cell.

Log-linear models are often easier to interpret when converted back to the scale of the fitted values. In particular, $\exp(\hat{\beta}_0)$ gives the fitted expected count for the first cell in the table, while similar expressions for the other parameters give the relative increase in counts for one level of a factor over the first. By unlogging, the systematic component becomes

$$\hat{\mu}_{ij} = \exp(5.486 - 0.3342 * x_1 + 0.1479 * x_2)$$

$$= \exp(5.486) * \exp(-0.3342 * x_1) * \exp(0.1479 * x_2)$$

$$= 241.3 * 0.7159^{x_1} * 1.159^{x_2}$$

Compare the values of $\hat{\mu}_{ij}$ when $x_2 = 1$ to the values for when $x_2 = 0$:

- When $x_2 = 0$ (High income): $\hat{\mu}_{i1} = 241.3 * 0.7159^{x_1}$
- When $x_2 = 1$ (Low income): $\hat{\mu}_{i2} = 241.3 * 0.7159^{x_1} * 1.159$

Under this model, the fitted values for $\hat{\mu}_{i2}$ are always 1.159 times the fitted values for $\hat{\mu}_{i2}$, for either value of x_1. From **Table 6.4**, the ratio of the number of low income respondents and the number of high income respondents (column marginal totals) is $\frac{480}{414} = 1.159$. This value is exactly the factor $\exp(0.1479) = 1.159$, which is no coincidence. **This demonstrates an important feature of the main effects terms in log-linear models: the main effect terms in the model simply model the marginal totals**. These marginal totals are usually not of main interest.

A Poisson GLM with interaction effects

The purpose of the `GM` study, for example, is to determine the relationship between income and attitudes towards genetically modified foods, not to estimate the proportion of Australians with high incomes. That is, the real interest lies with the interaction term in the model:

```
gm.int <- glm(Counts ~ Att * Inc, family = poisson, data = gm)
```

Several conclusions can be obtained:

- Notice that after fitting the interaction term, both the residual deviance and the residual degrees of freedom are 0, so the fit is perfect. This indicates that the number of coefficients in the model is the same as the number of entries in the table. This means that the 2×2 table cannot be summarized by a smaller set of coefficients.
- The analysis of deviance table shows the interaction term is necessary in the model. Thus, the data suggest an association between income levels and attitude towards genetically modified foods.

We can examine the percentage of low and high income respondents who are `For` and `Against` genetically modified foods by income level using the R function `prop.table()`. The table shows that 63.5% high income Australians are in favor of genetically modified foods while only 53.8% low income Australians are in favor of genetically modified foods. Thus, the data shows that high income Australians are more likely to be in favor of genetically modified foods than low income Australians.

The Grand Total Is Fixed: Random Component

Another scenario that may have produced the data in **Table 6.4** assumes a fixed number of 894 people were sampled. For example, the researchers may have decided to survey 894 people in total, and then classify each respondent as Low or High income, and also classify each respondent as For or Against gm foods. While the counts are free to vary within the table, the counts have the restriction that their sum is capped at 894. However, the Poisson distribution has no upper limits on y by definition. Instead, the **multinomial distribution** is appropriate. For a 2×2 table, the probability function for the multinomial distribution is

$$\begin{aligned}\mathcal{P}&(y_{11}, y_{12}, y_{21}, y_{22}; \mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}; m) \\ &= \frac{m!}{y_{11}!\, y_{12}!\, y_{21}!\, y_{22}!} \left(\frac{\mu_{11}}{m}\right)^{y_{11}} \left(\frac{\mu_{12}}{m}\right)^{y_{12}} \left(\frac{\mu_{21}}{m}\right)^{y_{21}} \left(\frac{\mu_{22}}{m}\right)^{y_{22}} \\ &= \frac{m!}{\prod_{i=1}^{2}\prod_{j=1}^{2} y_{ij}!} \prod\nolimits_{i=1}^{2} \prod\nolimits_{j=1}^{2} \left(\frac{\mu_{ij}}{m}\right)^{y_{ij}}\end{aligned}$$

The grand total m is retained in the distribution functions and other related function to indicate the grand total m is fixed.

For a $I \times J$ table, the probability function for the multinomial distribution is

$$\mathcal{P}(\boldsymbol{y}; \boldsymbol{\mu}; m) = \frac{m!}{\prod_{i=1}^{I}\prod_{j=1}^{J} y_{ij}!} \prod\nolimits_{i=1}^{I} \prod\nolimits_{j=1}^{J} \left(\frac{\mu_{ij}}{m}\right)^{y_{ij}}$$

The multinomial distribution is a generalization of binomial distribution. For example, it models the probability of counts for each side of a k-sided die rolled n times. For n independent trials each of which leads to a success for exactly one of k categories, with each category having a given fixed success probability, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories. When $k = 2$ and $n > 1$, it becomes the binomial distribution. When $k = 2$ and $n = 1$, it becomes a Bernoulli distribution.

Ignoring terms not involving μ_{ij}, the log-likelihood function of a $I \times J$ table (modeled by a multinomial distribution) is:

$$\ell(\boldsymbol{\mu}; \boldsymbol{y}, m) = \sum\nolimits_{i=1}^{I} \sum\nolimits_{j=1}^{J} y_{ij} \log(\mu_{ij})$$

and the residual deviance is

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}, m) = \sum\nolimits_{i=1}^{I} \sum\nolimits_{j=1}^{J} 2 \left\{ y_{ij} \log\left(\frac{y_{ij}}{\hat{\mu}_{ij}}\right)\right\}$$

after ignoring terms not involving $\hat{\mu}_{ij}$. Estimating μ_{ij} by maximizing the log-likelihood for the multinomial distribution requires the extra condition $\sum_i \sum_j \hat{\mu}_{ij} = m$ to ensure that the grand total is fixed at $\sum_i \sum_j y_{ij} = m$ as required by the sampling scheme.

We can compare the Poisson and multinomial distributions:

- **Distribution functions**: the Poisson distribution has an additional term, $-\mu_{ij}$ which is extra condition to ensure the grand total is fixed. The second term is identical.
- **Residual deviances**: are identical for Poisson and multinomial distributions, after ignoring terms not involving μ_{ij}.

These similarities for the multinomial and Poisson distributions have one fortunate implication: even though the multinomial distribution is the appropriate probability model, a Poisson GLM can be used to model the data under appropriate conditions. When the grand total is fixed, the appropriate condition is that the constant term β_0 must appear in the linear predictor, because this ensures $\sum_i \sum_j \hat{\mu}_{ij} = \sum_i \sum_j y_{ij} = m$ by **Example 6.1**. The effect of including the constant term in the model is that all inferences are conditional on the grand total. The Poisson model, conditioning on the grand total, is equivalent to a multinomial model. Thus, a Poisson model is still an appropriate model for the randomness, provided the constant term is in the model. In summary, for two-dimensional contingency tables:

- If no marginal totals are fixed: Poisson GLMs with or without the constant term (β_0) can be used but in general the constant term will be in the model.
- If the grand total is fixed: Poisson GLMs with the constant term (β_0) can be used.

The Column (or Row) Totals are Fixed

A third scenario that may have produced the data in **Table 6.4** assumes that the column (or row) totals are fixed. For example, the researchers may have decided to survey 480 low income people and 414 high income people, then record their attitudes towards gm foods. In this case, the totals in each column are fixed and the counts again have restrictions.

If the column totals are fixed, a multinomial distribution applies separately within each column of the table, because the numbers in each column are fixed and not random. Assuming the counts in each column are independent, the probability function of a $I \times J$ table is the product of multinormal distribution for each column:

$$\mathcal{P}(\boldsymbol{y};\boldsymbol{\mu};m_{\cdot 1},\cdots,m_{\cdot J}) = \prod_{j=1}^{J}\left\{\frac{m_{\cdot j}}{\prod_{i=1}^{I} y_{ij}!}\prod_{i=1}^{I}\left(\frac{\mu_{ij}}{m_{\cdot j}}\right)^{y_{ij}}\right\}$$

$\frac{m_{\cdot j}}{\prod_{i=1}^{I} y_{ij}!}\prod_{i=1}^{I}\left(\frac{\mu_{ij}}{m_{\cdot j}}\right)^{y_{ij}}$ is the distribution function of jth column and it is a multinomial distribution function with a column total $m_{\cdot j}$. Again, $m_{\cdot 1},\cdots,m_{\cdot J}$ are retained in the distribution function and other related functions to indicate these columns totals are fixed.

The log-likelihood function is

$$\ell(\boldsymbol{\mu};\boldsymbol{y},m_{\cdot 1},\cdots,m_{\cdot J}) = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}\log(\mu_{ij})$$

when terms not involving the parameters μ_{ij} are ignored. To solve for the parameters μ_{ij}, the extra J constraints

$$\sum_{i=1}^{I} \mu_{ij} = m_{\cdot j}(j = 1,2,\cdots,J)$$

must also be added to ensure both column totals are fixed.

Again, notice the similarity between this log-likelihood and the log-likelihood for the Poisson distribution. The residual deviances are exactly the same, after ignoring terms not involving $\hat{\mu}_{ij}$. This means the Poisson distribution can be used to model the data, provided the coefficients corresponding to the column totals appear in the linear predictor, since this ensures

$$\sum_{i=1}^{I} \hat{\mu}_{ij} = \sum_{i=1}^{I} y_{ij} = m_{\cdot j}(j = 1,2,\cdots,J)$$

Similarly, if the column totals or both row totals and column totals are fixed, a Poisson GLM is appropriate if the coefficients corresponding to the column totals and/or row totals are in the model.

Section 6.1.5 Three-Dimensional Tables

Three-dimensional tables cross-classify subjects according to three factors, say A, B and C. If the factors have I, J and K levels respectively, the table is an $I \times J \times K$ table. Similar with two-dimensional tables, we use the following notations:

- y_{ijk}: the observed count in row i $(i = 1, \cdots, I)$ and column $j(j = 1, \cdots, J)$ for group $k(k = 1, \cdots, K)$
- μ_{ijk}: the expected count in cell (i, j, k)
- m: the total number of counts
- $\pi_{ijk} = \frac{\mu_{ijk}}{m}$: the expected probability that an observation is in cell (i, j, k).
- $m_{ij\cdot}, m_{i\cdot k}, m_{\cdot jk}$: the sum over one of factors given the levels of the other two factors
- $m_{i\cdot\cdot}, m_{\cdot j\cdot}, m_{\cdot\cdot k}$: the sum over two factors given the level of the third factor

The meaning of the main effect terms in a Poisson GLM has been discussed for two-dimensional tables: the main effect terms model the marginal totals. Scientific interest focuses on the interactions between the factors. The model with main-effects only acts as the base model for contingency tables against which interaction models are compared. In a three-dimensional table, three two-factor interactions are possible, as well as a three-factor interaction term. Different interpretations exist depending on which interaction terms appear in the final model. These interpretations are considered here. We now introduce the example data to be used.

Example 6.5: Three-dimensional Tables

The example data in this section (data set: `kstones`; **Table 6.5**) comes from a study of treatments for kidney stones, comparing the success rates of various methods for small and large kidney stones.

Table 6.5 The kidney stone data. The success rates of two methods are given by size; `S` means a success, and `F` means a Failure in the table.

	Small stones			**Large stones**					
	S	**F**	**Total**	**S**	**F**	**Total**	**Total S**	**Total F**	**Total**
Method A	81	6	87	192	71	263	273	77	350
Method B	234	36	270	55	25	80	289	61	350
Total	315	42	357	247	96	343	562	138	700

Such data is generally presented in a table similar to **Table 6.5**. To fit a Poisson GLM for it in R, it is still organized as data frame with one column as the response variable and three columns as three factors.

In this section, we treat the method as factor A, the kidney stone size as factor B, and the outcome (success or failure) as factor C. Note that 350 patients were selected for use with each method. Since this marginal total is fixed, the corresponding main effect term `Method` must appear in the Poisson GLM. **The Poisson GLM with all three main effect terms ensures all the marginal totals from the original table are retained, but the parameters themselves are of little interest**.

Mutual Independence

If A, B and C are independent, then $\pi_{ijk} = \pi_{i\cdot\cdot} * \pi_{\cdot j\cdot} * \pi_{\cdot\cdot k}$ so that, on a log-scale,

$$\log(\mu_{ijk}) = \log(m * \pi_{ijk}) = \log(m) + \log(\pi_{i\cdot\cdot}) + \log(\pi_{\cdot j\cdot}) + \log(\pi_{\cdot\cdot k})$$

This is called **mutual independence**. As seen for the two-dimensional tables,

- Including the main effect terms effectively ensures the marginal totals are preserved.
- If the mutual independence model is appropriate, then the table may be understood from just the marginal totals.
- For the kidney stone data, the mutual independence model states that the success or failure is independent of the method used, and independent of the size of the kidney stones, and that the method used is also independent of the size of the kidney stone.
- Adopting this model assumes the data can be understood for each variable separately. In other words,
    - equal proportions of patients are in each method ($350/700 = 50\%$);
    - $138/700 = 19.7\%$ of all treatments fail
    - $343/700 = 49.0\%$ of patients have large kidney stones.

In this section, we will fit the models then comment and compare the models after all the models are fitted.

Partial Independence

Suppose A and B are not independent, but both are independent of C; then

$$\pi_{ijk} = \pi_{ij\cdot} * \pi_{\cdot\cdot k}$$

$$\log(\mu_{ijk}) = \log(m * \pi_{ijk}) = \log(m) + \log(\pi_{ij\cdot}) + \log(\pi_{\cdot\cdot k})$$

Since A and B are not independent, $\pi_{ij\cdot} \neq \pi_{i\cdot\cdot} * \pi_{\cdot j\cdot}$. To ensure that the marginal totals are preserved, the main effects are also included in the model). This means that the model should be:

$$\log(\mu_{ijk}) = \log(m) + \log(\pi_{i\cdot\cdot}) + \log(\pi_{\cdot j\cdot}) + \log(\pi_{ij\cdot}) + \log(\pi_{\cdot\cdot k})$$

This systematic component has one two-factor interaction AB . This is called **partial independence** (or **joint independence**). If a partial independence model is appropriate, then the two-way tables for each level of C are multiples of each other, apart from randomness. The data can be understood by combining the tables over C. For the kidney stone data, we can fit all three models that have one of the two-factor interactions.

Conditional Independence

Suppose that A and B are independent of each other when considered separately for each level of C. Then the probabilities π_{ijk} are **independent conditional** on the level of k, when

$$\pi_{ij|k} = \pi_{i\cdot|k} * \pi_{\cdot j|k}$$

Each conditional probability can be written in terms of marginal totals according to the definition of conditional probability:

$$\pi_{ij|k} = \frac{\pi_{ijk}}{\pi_{\cdot\cdot k}}, \pi_{i\cdot|k} = \frac{\pi_{i\cdot k}}{\pi_{\cdot\cdot k}}, \pi_{\cdot j|k} = \frac{\pi_{\cdot jk}}{\pi_{\cdot\cdot k}}$$

so we have

$$\frac{\pi_{ijk}}{\pi_{\cdot\cdot k}} = \frac{\pi_{i\cdot k}}{\pi_{\cdot\cdot k}} * \frac{\pi_{\cdot jk}}{\pi_{\cdot\cdot k}}$$

Which can further be simplified as

$$\pi_{ijk} = \frac{\pi_{i\cdot k} * \pi_{\cdot jk}}{\pi_{\cdot\cdot k}}$$

In other words,

$$\log(\mu_{ijk}) = \log(m) + \log(\pi_{i\cdot k}) + \log(\pi_{\cdot jk}) - \log(\pi_{\cdot\cdot k})$$

To ensure the marginal totals are preserved, use the model

$$\log(\mu_{ijk}) = \log(m) + \log(\pi_{i\cdot\cdot}) + \log(\pi_{\cdot j\cdot}) + \log(\pi_{\cdot\cdot k}) + \log(\pi_{i\cdot k}) + \log(\pi_{\cdot jk})$$

which includes the main effects. The systematic component has the two two-factor interactions `AC` and `BC`. This is called **conditional independence**.

If a **conditional independence model** is appropriate, then each two-way table for each level of C considered separately shows independence between A and B. The data can be understood by creating separate tables involving factors A and B, one for each level of C.

Uniform Association

Consider the case where all three two-factor interactions are present but the three-factor interaction ABC only is absent. This means that each two-factor interaction is unaffected by the level of the third factor. No interpretation in terms of independence or through the marginal totals is possible. Such model can be written:

$$\log(\mu_{ijk}) = \log(m) + \log(\pi_{i\cdot\cdot}) + \log(\pi_{\cdot j\cdot}) + \log(\pi_{\cdot\cdot k}) + \log(\pi_{ij\cdot}) + \log(\pi_{i\cdot k}) + \log(\pi_{\cdot jk})$$

which contains all two-way interactions. This is called **uniform association**. If the uniform association model is appropriate, then the data can be understood by examining all three individual two-way tables. Uniform association is simple enough to define from a mathematical point of view, but is often difficult to interpret from a scientific point of view.

The Saturated Model

If all interaction terms are necessary in the linear predictor, the model is the saturated model:

$$\log(\mu_{ijk}) = \log(m) + \log(\pi_{i\cdot\cdot}) + \log(\pi_{\cdot j\cdot}) + \log(\pi_{\cdot\cdot k}) +$$

$$\log(\pi_{ij\cdot}) + \log(\pi_{i\cdot k}) + \log(\pi_{\cdot jk}) + \log(\pi_{ijk})$$

which includes all interactions. The model has zero residual deviance and zero residual degrees of freedom. In other words, the model produces a perfect fit. This means that there are as many parameter estimates as there are cells in the table, and so the data cannot be summarized using a smaller set of coefficients. If the saturated model is appropriate, then the data cannot be presented in a simpler form than giving the original $I \times J \times K$ table.

Before we discuss different models using the data set `kstones`, we summarize the properties of different models for $I \times J \times K$ tables in **Table 6.6**.

Table 6.6 Comparison of different models for $I \times J \times K$ tables.

Model	Main Effects	Two-way Interaction	Three-factor interaction	Equivalent Contingency Tables
Mutual Independence (Marginal Model)	Yes	None	None	Three one-dimensional tables
Partial Independence	Yes	One	None	One two-dimensional table plus one dimensional table
Conditional Independence	Yes	Two	None	Two-dimensional tables for each level of a factor
Uniform Association	Yes	Three	None	Three two-dimensional tables
Saturated Model	Yes	Three	Yes	One three-dimensional table

Table 6.7 Models used in comparison from the data set `kstones`.

Model	Name	Model
Mutual Independence (Marginal Model)	`ks.mutind`	`Method + Size + Outcome`
Partial Independence	`ks.SM`	`Method + Size + Outcome + Method:Size`
	`ks.SO`	`Method + Size + Outcome + Size:Outcome`
	`ks.OM`	`Method + Size + Outcome + Method:Size`
Conditional Independence	`ks.noMo`	`Size *(Method + Outcome)`
	`ks.noOS`	`Method *(Size + Outcome)`
	`ks.noMS`	`Outcome *(Method + Size)`
Uniform Association	`ks.no3`	`Method * Size * Outcome – Size:Method:OutCome`
Saturated Model	`ks.all`	`Method * Size * Outcome`

Comparison of Models

For the kidney stone data and 9 Poisson GLMs listed in **Table 6.7**, we have the following conclusions:

(1) The saddlepoint approximation is sufficiently accurate since $\min(y_i) \geq 3$. This means that goodness-of-fit tests can be used to examine and compare the models.

(2) The mutual independence model, `ks.mutind`, is not appropriate, since its residual deviance 234.44 is much larger than the residual degrees of freedom 4 and the p-value of the goodness-of-fit test is 0.

(3) For the same reasons, the partial independence models, `ks.SM`, `ks.SO`, and `ks.OM`, are not appropriate either.

(4) For the same reasons, two conditional independence models, `ks.noMs` and `ks.noOS` are not appropriate.

(5) The conditional independence model `ks.noMO` appears the simplest suitable model. This implies that the data are best understood by creating separate tables for large and small kidney stones, but small and large kidney stones data should not be combined.

(6) More complicate model, uniform association model (`ks.no3`) and the saturated model (`ks.all`) are fine but the model `ks.noMO` should be used since it is the simplest suitable model.

Simpson's Paradox

Understanding which interaction terms are necessary in a log-linear model has important implications for condensing the tabular data. If a table is collapsed over a factor incorrectly, incorrect and misleading conclusions may be reached. An extreme example of this is **Simpson's paradox**.

Example 6.6: Simpson's Paradox

Find which method (method A or method B) is preferred using (1) separated 2×2 tables of `Method` and `Outcome` for large stones and small stones; and (2) the 2×2 table of `Method` and `Outcome`.

Solution: The tables can obtained with the R function `xtabs()`, `prop.table()`, etc. Please refer to R program for more details. **Table 6.8** presents the results.

Table 6.8 2×2 tables of `Method` and `Outcome`. The number in the parentheses is the proportion of failures or successes for each method.

	Large Stones		**Small Stones**		**All Stones**	
	Failure	**Success**	**Failure**	**Success**	**Failure**	**Success**
Method A	71 (27%)	192 (73%)	6 (7%)	81 (93%)	77 (22%)	273 (78%)
Method B	25 (31%)	55 (69%)	36 (13%)	234 (87%)	17 (17%)	289 (83%)

The most suitable model appears to be model `ks.noMO`. This model has two two-factor interactions, indicating conditional independence between `Outcome` and `Method`, depending on the size of the kidney stones.

(1) The dependence on `Size` means that the data must be stratified by kidney stone size for the correct relationship between `Method` and `Outcome` to be seen. From **Table 6.8**, we can see that the success rates of the method A are about 73% and 95% for the large stones and the small stones, respectively. Both of them are greater than the corresponding success rates of the method B which are about 69% and 87% for the large stones and the small stones, respectively. Clearly, the method A is better than the method B.

(2) Combining the data over the size of stones, and considering a single combined two-way table of `Method` and `Outcome` (and hence ignoring the size), is an incorrect summary. From **Table 6.8**, the success rate for the method A is about 78%, and for the method B the success rate is about 83%, so the method B would be preferred.

In this example, incorrectly collapsing the table over $Size$ has completely changed the conclusion. This is called **Simpson's paradox**, which is a result of incorrectly collapsing a table. This is because:

- The method A uses a larger number of large stones (263 large stones versus 87 small stone).
- For the larger stones, the success rates for both methods are lower.
- So the method A reports a higher number of total failures when the two groups are combined.
- The method B uses a larger number of small stones (270 small stones versus 80 large stones).
- For the smaller stones, the success rates for both methods are higher.
- So the method B reports a smaller number of total failures.

An extreme situation is that the method A only uses the large stone while the method B only uses the small stones. The combined data (although there is not meaningful to combine the data in such situation) will show that the method B is better than the method A.

Section 6.1.6 Higher-Order Tables

Extending these ideas to situations with more than three factors is easy in practice using R, though interpreting the final models is often difficult.

Example 6.7: Four-Dimensional Tables

A study of seriously emotionally disturbed (SED) and learning disabled (LD) adolescents reported their depression levels (**Table 6.9**; data set: dyouth). The data are counts classified by four factors: Age (using 12-14 as the reference group), Group (either LD or SED), Gender and level of Depression (either low L or high H). We would like to find a suitable model and interpret the findings from the model.

Table 6.9 Depression levels in youth.

		Depression low L		**Depression high H**	
Age	**Group**	**Males**	**Females**	**Males**	**Females**
12-14	**LD**	79	34	18	14
	SED	11	5	5	8
15-16	**LD**	63	26	10	11
	SED	32	15	3	7
17-18	**LD**	36	16	13	1
	SED	36	12	5	2

Solution: First, since none of the totals were fixed beforehand and are free to vary randomly, no variables need to be included.

Second, there 4 main effects, $\binom{4}{2} = 6$ two-factor interactions, $\binom{4}{3} = 4$ three-factor interactions, and 1 four-factor interactions. Therefore, there are large number of possible Poisson GLMs can be used.

Third, we will use a model that includes all four main effects, two-factor interactions and three-factor interactions of Age, Depression, and Gender, and two-factor interactions between Age and Group.

Fourth, the analysis of deviance table is shown in **Table 6.10**. Overall, the model shows an association between depression and age and gender (the p-value is 0.01557). Since the two-factor interaction between Depression and Group are not included as a significant interaction in the model, no difference in depression rates between the two groups once the demographic variables have been taken into account.

Table 6.10 The analysis of deviance table for the data dyouth.

Source	**Df**	**Deviance**	$\chi^2 -$ **stat**	p**-value**	
Age	2	11.963	11.963	0.002525	**

Depression	1	168.375	168.375	$< 2.2 * 10^{-16}$	***
Gender	1	58.369	58.369	$2.17 * 10^{-14}$	***
Group	1	69.104	69.104	$< 2.2 * 10^{-16}$	***
Age:Depression	2	3.616	3.616	0.164	
Age:Gender	2	3.631	3.631	0.163	
Depression:Gender	1	7.229	7.229	0.0072	**
Age:Group	2	27.090	27.090	$1.31 * 10^{-6}$	***
Age:Depression:Gender	2	8.325	8.325	0.0156	*
Residual	9	10.35			
Total	23	368.05			

Fifth, the three-way interaction shows that the relationship between age and depression is different for males and females. Given the fitted model, collapsing the table into a simpler table would be misleading. The proportion tables (**Table 6.11**) show that the rate of high depression decreases with age for girls, especially for 17 years and older, whereas for males the rate of high depression decreases at age 15–16 then increases again for 17–18. This difference in pattern explains the three-way interaction detected by the analysis of deviance table.

Table 6.11 2 × 2 tables of `Age` and `Depression` stratified by the gender. The number in the parentheses is the proportion of `Depression H` or `L` with each age group.

	Males		**Female**	
Age	**Depression H**	**Depression L**	**Depression H**	**Depression L**
12-14	23 (20%)	93(80%)	22 (36%)	39 (64%)
15-16	13 (12%)	95 (88%)	18 (31%)	41 (69%)
17-18	18 (20%)	72 (80%)	3 (10%)	28 (90%)

Sixth, the model also finds a significant interaction between `Age` and `Group`, meaning simply that the `SED` and `LD` groups contain different proportions of the age groups. This is not particularly of interest, but it is important to keep the `Age:Group` term in the model, so that the tests for interactions involving Depression should adjust for these demographic proportions.

Exercise 6.2: Model Selection

For the depressed youth data (data set: `dyouth`), perform the following analysis. A Poisson GLM with the logarithmic link function should be used.

(1) Show that the four-factor interaction is not significant.

(2) Show that only one three-factor interaction is significant in the model.

(3) Then show that four two-factor interactions are needed in the model (some because they are significant, some because of the marginality principle).

(4) Show that the model is adequate by examining the model diagnostics.

Solution: Refer to R codes and output for more details.

(1) The chi-square test statistic, degrees of freedom, and the p-value from the analysis of deviance table between two nested models, the model with the four-factor interaction and the model without the four-factor interaction are: 0.064, 2, and 0.725, respectively.

(2) From the "type I" analysis of deviance table from the Poisson GLM without the four-factor interaction, we can see that the only significant three-factor interaction is `Age:Depression:Gender` with a p-value 0.0155. Two other three-way interactions, `Age:Depression:Group` and `Depression:Gender:Group`, have the p-values 0.063 and 0.085, respectively.

The p-values from the "Type I" analysis of deviance table depend on the order of terms in the model. So if possible, the "Type III" analysis of deviance table is preferred. Again, from the "Type III" analysis of deviance table, only the three-factor interaction is `Age:Depression:Gender` is significant.

(3) The model used here include all main effects, two-factor interactions, and the three-factor interaction is `Age:Depression:Gender`. Since `Age:Depression:Gender` is in the model, the two-factor interactions between `Age`, `Depression`, and `Gender` should be included in the final model even if they are not significant. Therefore, we only need to look at the other three two-factor interactions involving `Group`. From the "Type III" analysis of deviance table, the two-factor interaction, `Age:Group` (p-value$= 9.8 * 10^{-7}$) is significant. and `Depression:Group` (p-value$= 0.722$) and `Gender:Group` (p-value$= 0.433$) are not significant.

(4) The final model has main effects, two-factor and three-factor interactions between `Age`, `Depression`, and `Gender,` and the two-way interaction `Age:Group`. You can perform the diagnostic analysis to verify that the final model is adequate.

Section 6.1.7 Structural Zeros in Contingency Tables

Contingency tables may contain cells with zero counts. Depending on the reason for a zero count, different approaches must be taken when modelling. There are two types of zeros.

- **Sampling zeros** or **random zeros** appear by chance, simply because no observations occurred in that category due to low frequency and/or small sample size.
    - Larger samples may produce non-zero counts in those cells.
    - Cells with such zero counts can modelled like the other counts in the data.
    - However, computing fitted values for cells with zero counts is sensible.
    - In addition, the presence of the zeros means the saddlepoint approximation is likely to be very poor.
    - As a result, levels of one or more factors may be combined to increase the minimum count. For example, "Strongly agree" and "Agree" may be combined sensibly into a single "Agreement" category.
- **Structural zeros** appear because the outcome is impossible.
    - For example, in a cross-tabulation of gender and surgical procedures, the cell corresponding to male hysterectomies must contain a zero count.
    - Producing fitted values for cells with structure zeros makes no sense.
    - Structural zeros require special attention since computing expected counts for impossible events is nonsense.
    - As a result, cells containing structural zeros must be removed from the data before analysis.

Example 6.8: Structural Zeros in Tables

The types of cancer diagnosed in Western Australia in 1996 were recorded for males and females (**Table 6.12**; data set: `wacancer`) to ascertain whether the number of cancers differs between genders. How should we analyze such data with a Poisson GLM?

Table 6.12 The number of cancers diagnosed by gender in Western Australia during 1996.

	Cancer Type						
Gender	**Prostate**	**Breast**	**Colorectal**	**Lung**	**Melanoma**	**Cervix**	**Other**
Males	923	0	511	472	362	0	1406
Females	0	875	355	211	282	77	1082

Solution: Three cells have zeros recorded. Two of these three cells are **structural zeros** since they are impossible - females cannot have prostate cancer, and males cannot have cervical cancer. Breast cancer is a possible, but very rare, disease among men (about 100 times as many cases in females compared to males, in the USA). The zero for male breast cancer is technically a **sampling zero**. Since breast cancer is already known to be a rare disease for males, the analysis should focus on gender differences for other types of cancers, such as colorectal, lung, melanoma and other cancers.

We first fit a GLM using all data. Then we remove the data for the breast cancer and two structural zeros and use the processed data to fit a Poisson GLM. Note that the degrees of freedom for the variables are different for the two models from the analysis of deviance tables (**Table 6.13**). For both models, the interaction term is very significant, so the number of people diagnosed with the different types of cancers differs according to gender, even after eliminating prostate, breast and cervical cancer, which are obviously gender-linked. However, note that the degrees of freedom are different for the two models.

Although the conclusions from these two models are similar, the data with removed zeros should be used.

Table 6.13 The analysis of deviance tables for the data set `wacancer`.

	All Data			**Data with Removed Zeros**		
Source	**DF**	**Deviance**	***p*-value**	**DF**	**Deviance**	***p*-value**
`Cancer`	6	3281.5	$< 2.2 * 10^{-16}$	5	2591.47	$< 2.2 * 10^{-16}$
`Gender`	1	95.9	$< 2.2 * 10^{-16}$	1	144.74	$< 2.2 * 10^{-16}$
`Cancer:Gender`	6	2686.2	$< 2.2 * 10^{-16}$	3	38.11	$2.68 * 10^{-8}$
Residual	0	0		0	0	
Total	13	6063.7		9	2774.32	

Lesson 6.2 Models for Counts: Negative Binomial GLMs

Related Readings: Sections 10.5, 10.6, 10.7, and 10.8 in Chapter 10.

Section 6.2.1 Introduction and Overview

In this lesson, we describe another type of GLMs for modeling counts: negative binomial GLMs. We first discuss the overdispersion problem in Poisson GLMs (**Section 6.2.2**). We then introduce two models to tackle the overdispersion problem: negative binomial GLMs (**Section 6.2.3**) and quasi-Poisson models (**Section 6.2.4**) as alternative models. The lesson is concluded with a case study in **Section 6.2.5**.

Section 6.2.2 Overdispersion for Poisson GLMs

In **Section 5.1.4**, we discussed the overdispersion problem for binomial GLMs. The overdispersion problem also exists for Poisson GLMs. For a Poisson distribution, the dispersion parameter $\phi = 1$ and the variance function $V(\mu) = \mu = var[y]$. However, in practice the apparent variance of the data often exceeds μ. This is called **overdispersion**. **Underdispersion** also occurs, but is less common.

Overdispersion arises either because the mean μ retains some innate variability, even when all the explanatory variables are fixed, or because the events that are being counted are positively correlated. Overdispersion typically arises because the events being counted arise in clusters or are mutually supporting in some way. This causes the underlying events to be positively correlated, and overdispersion of the counts is the result.

The presence of overdispersion have the following consequences if it is not considered:

- The parameter estimates $\hat{\beta}_j$ may or may not be affected, depending on the nature of the overdispersion.
- The standard errors $se(\hat{\beta}_j)$ are underestimated, meaning that the estimated standard errors are less than the true standard errors.

- Consequently, tests on the explanatory variables will generally appear to be more significant that warranted by the data, meaning that the tests have the inflated type I error rate.
- Similarly, confidence intervals for the parameters will be narrower than warranted by the data.

Overdispersion is detected by conducting a goodness-of-fit test. If the residual deviance and Pearson goodness-of-fit statistics are much larger than the residual degrees of freedom, then either the fitted model is inadequate or the data are overdispersed. If lack of fit remains even after fitting the maximal possible explanatory model, and after eliminating any outliers, then overdispersion is the alternative explanation.

When the counts are very small, so asymptotic approximations to the residual deviance and Pearson statistics are suspect, then overdispersion may be difficult to judge. However the goodness-of-fit statistics are more likely to be underestimated than overestimated in small count situations, so large goodness-of-fit statistics should generally be taken to indicate lack of fit.

Example 6.9: Absence of Overdispersion

For the final model fitted to the kidney stone data (see **Table 6.7**), the residual deviance, the residual degrees of freedom, and the p-value from the goodness-of-fit test based on the residual deviance are 3.45, 2, and 0.178, respectively. The Pearson statistic and the p-value of Pearson goodness-of-fit test are 3.18 and 0.204, respectively. Note $\min(y_i) = 6$ so the approximations are fine. A goodness-of-fit test does not reject the hypothesis that the model is adequate. So there is no overdispersion problem here.

Example 6.10: Presence of Overdispersion

In an experiment to assess viral activity, pock marks were counted at various dilutions of the viral medium (data set: `pock`). We use the logarithm to base 2 of `Dilution` as a covariate, since the dilution levels are in increasing powers of 2 suggesting this was factored into the design. We fit GLM(Poisson; log) to see if the model is adequate.

Solution: There are several parts for the data analysis: exploratory analysis, model fitting, diagnostic, and conclusions.

Exploratory analysis. The plot of the pock counts against $\log_2(\text{Dilution})$ shows a definite relationship between the variables (**Figure 6.3**, left panel).

To verify that the variance function $V(\mu) \approx \mu$, we can plot the logarithm of group variances against the logarithm of group means. From the right panel of **Figure 6.3**, we can see that the logarithm of variances increases linearly with increasing logarithm of means.

We can see that the sample variances are much higher than the sample means for each group, indicating the presence of overdispersion. This is not unexpected since intuitively, pock marks are more likely to appear in clusters rather than independently. Not only are the variances greater than the means, but their ratio increases with the mean as well. The slope of the trend in the right panel of **Figure 6.3** is 1.44. This suggests a variance function approximately of the form $V(\mu) \approx \mu^{1.5}$. The mean–variance relationship here is in some sense intermediate between that for the Poisson ($V(\mu) = \mu$) and Gamma ($V(\mu) = \mu^2$) distributions.

Modeling fitting: GLM(Poisson; log) is fitted.

Diagnostic: The residual deviance, the residual degrees of freedom, and the p-value from the goodness-of-fit test based on the residual deviance are 290.44, 46, and 0, respectively. This suggests the substantial lack of fit. Possible reasons of lack of fit include missing important explanatory variables, the presence of the overdispersion, etc. The saddlepoint approximation is satisfactory here as $\min(y_i) = 5$. In addition, the deviance and Pearson goodness-of-fit statistics (not shown here) are nearly identical. There are two ways to model the overdispersion: negative binomial GLMs and quasi-Poisson models.

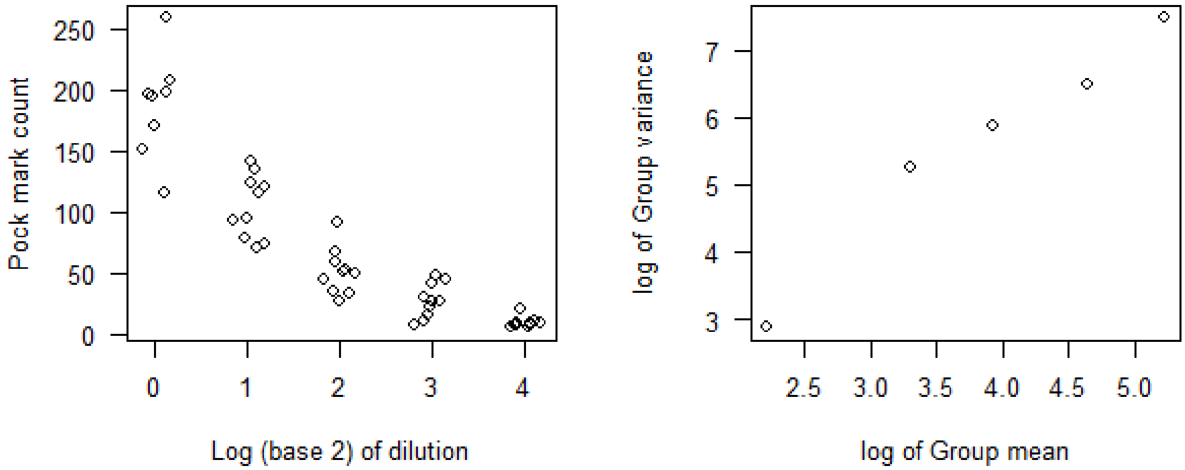

Figure 6.3 The pock data. **Left panel**, the counts against the logarithm of dilution; **right panel**, the logarithm of the group variances against the logarithm of the group means

Section 6.2.3 Negative Binomial GLMs

Hierarchical Models

One way to model overdispersion is through a hierarchical model. Instead of assuming $y_i \sim Poisson(\mu_i)$, we can add a second layer of variability by allowing μ_i itself to be a random variable. Suppose instead that

$y_i|\lambda_i \sim Poisson(\lambda_i)$ and $\lambda_i \sim G(\mu_i, \psi)$

where $G(\mu_i, \psi)$ denotes a distribution with mean μ_i and coefficient of variation ψ. It can be shown, under the hierarchical model, that

$$E[y_i] = \mu_i, var[y_i] = \mu_i + \psi\mu_i^2$$

so the variance contains an overdisperion term $\psi\mu_i^2$. The larger ψ, the greater the overdispersion.

A popular choice is to assume that the mixing distribution G is a gamma distribution. The coefficient of variation of a gamma distribution is its dispersion parameter, so the second layer of the hierarchical model becomes $\lambda_i \sim Gamma(\mu_i, \psi)$. With this assumption, We can show that y_i follows a **negative binomial distribution** with probability function

$$\mathcal{P}(y_i\,; \mu_i, k) = \frac{\Gamma(y_i + k)}{\Gamma(y_i + 1)\Gamma(k)} \left(\frac{\mu_i}{\mu_i + k}\right)^{y_i} \left(1 - \frac{\mu_i}{\mu_i + k}\right)^{k}$$

where $k = \frac{1}{\psi}$ and $\Gamma()$ is the gamma function.

Negative Binomial Distribution and GLMs

Negative binomial distribution and corresponding GLMs have the following properties (without proof):

- For **any fixed $\boldsymbol{k}$**, the negative binomial distribution is an EDM.
- The dispersion parameter for the negative binomial distribution is $\phi = 1$. Therefore, the standard normal distribution and the chi-square distribution instead of the t-distribution and the F-distribution are used in the statistical inference.
- $E[y_i] = \mu_i$ and $var[y_i] = \mu_i + \frac{\mu_i^2}{k}$
- The negative binomial distribution with fixe k can used in GLMs to model the count data.
- In practice, k is rarely known and so negative binomial glms are usually used with an estimated value for k.
- In R, the function `glm.nb()` from package **MASS** can be used in place of `glm()` to fit the model.
- The R function `glm.nb()` undertakes maximum likelihood estimation for both k and the GLM coefficients β_j simultaneously.
- The estimation of k introduces an extra layer of uncertainty into a negative binomial GLM. However the maximum likelihood estimator $\hat{k}$ of k is uncorre- lated with the $\hat{\beta}_j$, according to the usual asymptotical approximations. Hence the glm fit tends to be relatively stable with respect to estimation of k.
- Negative binomial GLMs give larger standard errors than the corresponding Poisson GLMs, depending on the size of $k = \frac{1}{\psi}$.
- The coefficient estimates $\hat{\beta}_j$ from a negative binomial GLM may be similar (but not identical) to those produced from the corresponding Poisson GLM.

- The default link function for `glm.nb()` is the logarithmic link function. Indeed the log-link is almost always used with negative binomial GLMs to ensure $\mu_i > 0$ for any value of the linear predictor.
- The function `glm.nb()` also allows the “`sqrt`” and “`identity`” link function.
- As usual, the quantile residuals are strongly recommended for negative binomial GLMs.

In summary, the negative binomial GLMs can be used to model counts data in the presence of overdispersion.

Example 6.11: Negative Binomial GLMs

The pock data shows overdispersion (data set: `pock`). We fit a negative binomial GLM, estimating k using the function `glm.nb()` in package **MASS** (note that `glm.nb()` uses `theta` to denote k).

Solution: After fitting the model, we can see that $k \approx 10$, the negative binomial model is using the variance function $V(\mu) = \mu + \frac{\mu^2}{10}$. The coefficient of variation of the mixing distribution ($\psi = 1/k$) is estimated to be about 10%, a reasonable level for replicate to replicate variation.

Do not use the R function `glm.convert()` as suggested by the textbook because the results can be confusing.

The comparison between the Poisson GLM and the negative binomial GML is summarized in **Table 6.14** and **Figure 6.4**.

- The parameter estimates are similar while the standard errors are quite different.
- The negative binomial GLM has a much smaller residual deviance than the Poisson GLM.
- The goodness-of-fit test from negative binomial GLM has a p-value 0.288, so the hypothesis that the model is adequate cannot be rejected.
- The diagnostic plots (bottom panel, **Figure 6.4**) suggest the negative binomial model is adequate. No observations are particularly influential.
- For Poisson GLM, there are a few influential observations. Two Cook’s distances are greater than 1.0.

In summary, the negative GLM is preferred over the Poisson model for the `pock` data.

Table 6.14 The results from a Poisson GLM and a negative binomial GLM for the data set `pock`.

Model	$\hat{\beta}_0$	$s(\hat{\beta}_0)$	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	Residual Deviance	Residual DF	p-value from goodness-of-fit
Poisson	5.268	0.0226	−0.681	0.0154	290.44	46	0
NB	5.334	0.0879	−0.725	0.0389	50.86	46	0.288

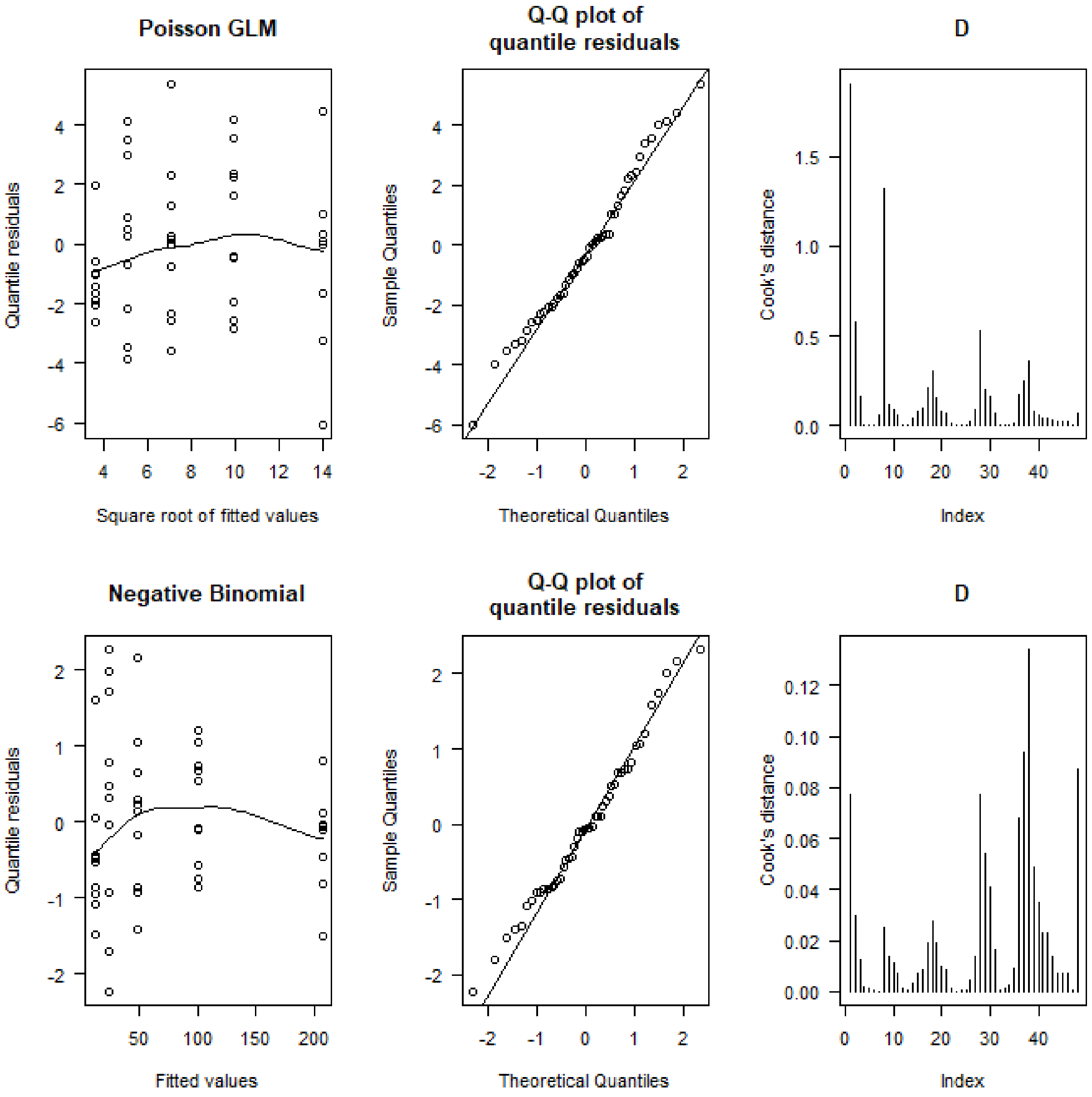

Figure 6.4 Diagnostic plots from fitting the Poisson GLM (top panel) and the negative binomial model bottom panels) to the `pock` data.

Exercise 6.3: Diagnostic Plots of a Negative Binomial Model

Fit GLM(Poisson; log) and GLM(negative binomial; log) and verify **Figure 6.4**. Note that the quantile residuals are used in all plots.

Section 6.2.4 Quasi-Poisson Models

The simplest to use, and therefore most commonly used, approach to overdispersed counts are **quasi-Poisson models**. Quasi-Poisson models keep the Poisson variance function $V(\mu) = \mu$ but simply allow a general positive dispersion parameter ϕ, so that $var[y_i] = \phi\mu_i$. Here $\phi > 1$ corresponds to overdispersion. This approach can be motivated in the same way as were quasi-binomial models (**Section 5.1.4**).

When $\phi \neq 1$, there is no EDM with this variance function that gives positive probability to integer values of y_i. As outlined in **Table 5.1**, the likelihood function, AIC, BIC, and quantile residuals, etc., cannot be calculated/defined. Nevertheless, the quasi-likelihood methods of **Section 5.1.4**, still apply, so quasi-Poisson GLMs yield consistent estimators and consistent standard errors for the β_j. The residual deviance can be calculated too.

Since ϕ does not affect the parameter estimation, the coefficient estimates from a quasi-Poisson GLM are identical to those from the corresponding Poisson GLM, but the standard errors are inflated by a factor of $\sqrt{\phi}$. Confidence intervals and statistics for testing hypotheses tests will change for the same reason.

Note that quasi-Poisson and the negative binomial model both produce overdispersion relative to the Poisson distribution but they assume different mean-variance relationships. Quasi-Poisson models assume a linear variance function ($V(\mu) = \mu$) whereas negative binomial models use a quadratic variance function ($V\ (\mu) = \mu + \frac{\mu^2}{k}$).

Example 6.12: Quasi-Poisson Models

Fit a quasi-Poisson model to the `pock` data.

Solution: After a quasi-Poisson model fitted to the data, we can obtain the following results:

- The estimated dispersion parameter $\hat{\phi} = 6.34$.

- The residual deviance and the residual degrees of freedom are 290.44 and 46, respectively. They are identical to those from the Poisson GLM. This is expected because they use the same variance function $V(\mu) = \mu$. Since the dispersion parameter ϕ is unknown in the quasi-likelihood model, the goodness-of-fit test cannot be performed.
- From The diagnostic plots (**Figure 6.5**) suggest the quasi-Poisson model is broadly adequate, and no observations are particularly influential. Note that the standardized deviance residuals are used since the quantile residuals cannot be defined from the quasi-Poisson model.
- The estimates $(\hat{\beta}_j)$ from the quasi-Poisson and Poisson models are identical.
- Comparing the standard errors from the quasi-Poisson model to the standard errors produced from the Poisson GLM, the standard errors in the quasi-Poisson model are scaled by $\sqrt{\hat{\phi}}$.
- The confidence intervals of parameters from the quasi-Poisson model are based on the t-distribution since the dispersion parameter is unknown and must be estimated.
- The fitted models say that the expected number of pock marks decreased by a factor of about $\exp(-0.7) \approx 0.5$ for every 2-fold dilution. In other words, the expected number of pock marks is directly proportional to the concentration of the viral medium.

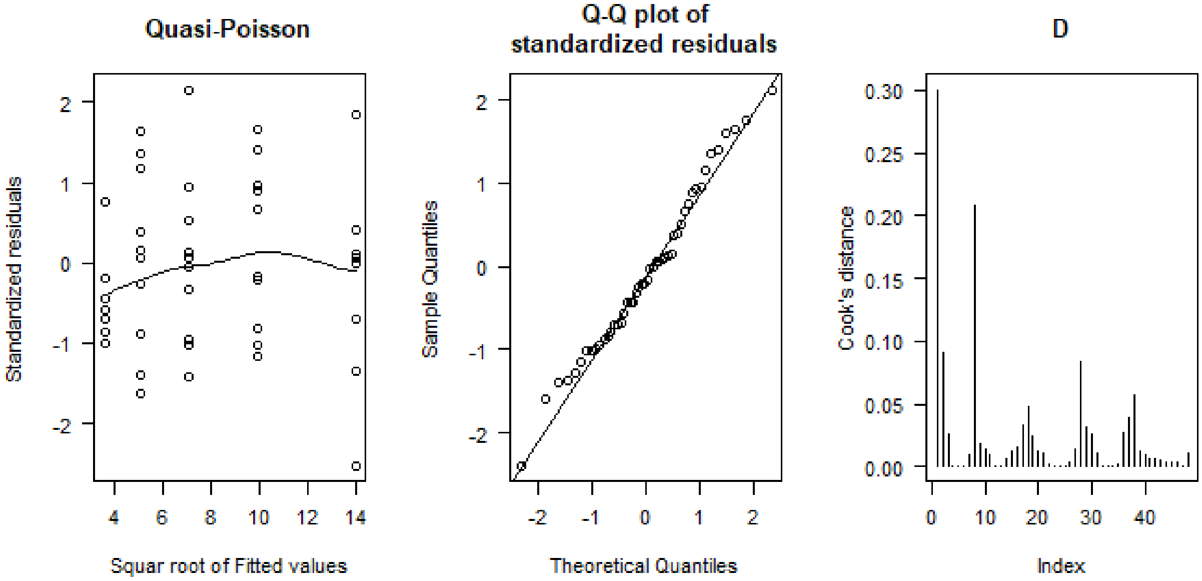

Figure 6.5 Diagnostic plots from fitting the quasi-Poisson GLM to the `pock` data. Note that the quantile residuals are not defined for the quasi-Poisson model, so the standardized deviance residuals are used.

Exercise 6.4: Confidence Interval of μ

For three model discussed on **Examples 6.10, 6.11,** and **6.12**, find the predicted and 99% confidence interval of mean number of pock marks when `Dilution`$= 2$.

Solution: Since $x = 2$ and $\log_2(x) = 1$, so $\hat{\mu} = \exp(\hat{\beta}_0 + \hat{\beta}_1)$. To obtained the 99% confidence interval of μ,

- Obtain $\hat{\beta}_0 + \hat{\beta}_1$ and $se(\hat{\beta}_0 + \hat{\beta}_1)$ from R function predict.
- Find 99% confidence interval of $\beta_0 + \beta_1$:
$$(c_l, c_h) = \hat{\beta}_0 + \hat{\beta}_1 \pm critical\ value * se(\hat{\beta}_0 + \hat{\beta}_1)$$
- Find 99% confidence of μ: $(\exp(c_l), \exp(c_h))$

Note that for the Poisson GLM and Negative Binomial GLM, the critical value is based on the standard normal distribution while for the quasi-Poisson model, the critical value is based on the t-distribution. The results are summarized in Table **6.15**. It can be seen that the point estimates $\hat{\mu}$ from three models are similar. The Poisson GLM produces the narrowest 99% confidence interval of μ while the negative binomial and quasi-Poisson models produce similar confidence intervals. Note the Poisson GLM is not adequate due to the presence of overdispersion, so the confidence intervals from the Poisson GLM shouldn't be used.

Table 6.15 Predicted and 99% confidence interval of mean number of pock marks when `Dilution`$= 2$.

Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0 + \hat{\beta}_1$	$se(\hat{\beta}_0 + \hat{\beta}_1)$
Poisson	5.27693	-0.68094	4.58699	0.017142
Negative binomial	5.33284	-0.72460	4.60825	0.061016
Quasi-Poisson	5.26793	-0.68094	4.58699	0.043159
Method	$z_{0.005}$ or $t_{0.005,46}$	99% CI of $\beta_0 + \beta_1$	$\hat{\mu} = \exp(\hat{\beta}_0 + \hat{\beta}_1)$	99% CI of μ
Poisson	2.57583	$(4.5843, 4.6312)$	98.198	$(93.958, 102.633)$
Negative binomial	2.57583	$(4.4511, 4.7654)$	100.308	$(85.720, 117.380)$
Quasi-Poisson	2.68701	$(4.4710, 4.7030)$	98.198	$(87.447,110.274)$

Section 6.2.5 Case Study

In a study of nesting female horseshoe crabs, each with an attached male, the number of other nearby male crabs (called satellites) were counted (data set: `hcrabs`). The color of the female, the condition of her spine, her carapace width, and her weight were also recorded. The purpose of the study is to understand the factors that attract satellite crabs. Are they more attracted to larger females? Does the condition or color of the female play a role?

As we have discussed before, the data can be analyzed in several steps: theoretical consideration, exploratory analysis, model building and selection, diagnostics, and conclusions.

Theoretical consideration. In this step, the possible models that should be considered can be determined based on the study and/or data. Clearly the response is the number satellite crabs attracted by the female crab and is the counts data. Therefore the Poisson GLMs (or quasi-Poisson GLMs, negative binomial GLMs) should be used.

Crabs tend to congregate and interact with one another, rather than behaving independently, hence we should expect overdispersion a `priori` relative to Poisson for the counts of satellite crabs. Therefore, quasi-Poisson GLMs or negative binomial GLMs would be more appropriate than Poisson GLMs.

Exploratory analysis. Plots and/or tables can be used to (1) summary statistics of response and the explanatory variables; (2) assess the relationship between the response and the explanatory variables; (3) determine if we need to transform the explanatory variables. Specifically, the following analysis are conducted:

- Color is on a continuum from light to dark, and spine condition counts the number of intact sides, so both of them are defined as ordered factors.
- Plotting the response (`Sat`) or its logarithm against each of the explanatory variables shows trends for more satellite crabs to congregate around females that are larger (in weight and width), are lighter in color, and have no spinal damage (**Figure 6.6**).
    - The R function `jitter()` is used to avoid overplotting by adding a small amount of noise to the response.
    - log(`Sat`+1) is used to avoid taking logarithm of zero.

  - Plots on the log-scale are preferable because the values of `Wt` and `Width` are distributed more symmetrically on the log-scale, and because the relationships between them and `Sat` are more likely to be relative rather than additive.
- Plots between the explanatory variables shows that they are inter-related (**Figure 6.7**):
  - `Wt` is the most obvious overall summary of the size of each female.
  - The lighter-colored females are also typically heavier.
  - The females with no spine damage are also typically heavier.
  - The relationships observed between `Sat` and `Col` and `Spine` might be explained by the relationship between `Sat` and `Wt`.
  - `Wt` should be proportional to the volume of each female, hence should be approximately proportional to Width^3, if the females are all the same shape. Indeed, log(`Wt`) is nearly linearly related to log(`Width`) with an estimated slope 2.56, which is close to 3.

Model Building. We fit a quasi-Poisson GLM with the logarithmic link function. The explanatory variables are the color, the spine condition, the logarithm of weight, and the logarithm of width.

- Since the color and the spine condition are ordered factors, use the R function:

```
options(contrasts = c("contr.treatment", "contr.treatment"))
```

  to ensure that the treatment coding instead of polynomial contrast is used. Note the different coding schemes for the dummy variable can change the meaning and the interpretation of coefficients in the model but will not change the likelihood function, residual deviances, etc.
- Pearson statistic and the residual degrees of freedom are 528 and 165 respectively.
- The Pearson estimate of the dispersion, $\hat{\phi} = \frac{528}{165} = 3.2$, so our expectation of overdispersion seems confirmed.
- Based on the “type I” analysis of deviance table, log(`Wt`) is significant, other explanatory variables, including log(`Width`), the color, and the spine condition are not significant after adjusting for log(`Wt`).

Therefore we fit a quasi-Poisson GLM with the logarithmic link function. The only explanatory variable is logarithm of weight. This model is named “`cr.m2`” in R and this section.

Model diagnostic. The diagnostic plots suggest that the model cr.m2 is a reasonable model (**Figure 6.8**).

- The constant-information scale of the fitted values is used. For the quasi-Poisson GLM, they are square root of the fitted values.
- Quantile residuals are not defined for the quasi-Poisson GLM. The standardized deviance residuals are used.
- No observation is identified as influential using Cook's distance.

Note that nearly half of the values of the response Sat are 0 or 1, which may suggest a problem for the distribution of the residual deviance and the evaluation of overdispersion. How should we assess if the saddlepoint approximation is fine here? Here, we introduce a **parametric bootstrap** method to evaluate if the residual deviance has an approximate χ^2 distribution. We use the model cr.m2 as an illustration.

- **Step 1**: Fit the model and obtain the predicted $\hat{\mu}_i$.
- **Step 2**: Generate n random numbers according to the Poisson distribution with the mean $\hat{\mu}_i$.
- **Step 3**: Use the random numbers from Step 2 as the response and the explanatory variables in the original data to fit the same Poisson GLM and calculate the corresponding residual deviance.
- **Step 4**: Repeat Steps 2 and 3 N times and obtain N residual deviances.
- **Step 5**: Compare the empirical distribution of N residual deviances from Step 4 and the theoretical distribution if the approximation is fine. Again this can be done with a Q-Q plot. For the model cr.m2, the the theoretical distribution χ^2_{171}.

We generate 200 random residual deviances using model cr.m2 and compare them with χ^2_{171} using the Q-Q plot (**Figure 6.9**). Almost all points are in a straight line, indicating that even nearly half of the values of the response Sat are 0 or 1, the approximate distribution of residual deviances from model cr.m2 is χ^2_{171}.

Conclusions. The fitted systematic component of model cr.m2 is

$$\log(\mu) = -12.57 + 1.744 * \log(W)$$

or equivalently

$$\mu = 0.000003483 * W^{1.744}$$

where W is the weight of the crabs in grams. The quasi-Poisson model indicates that heavier crabs have more satellites on average. If the regression coefficient for $, \log(W)$ is 1, then the expected number of satellite crabs would be directly proportional to the weight of the female. The number of satellites seems to increase just a little faster than this.

It is tempting to speculate on the biological implications. It might well be possible for a male crab to sense the overall weight of the female crab by smell or other chemical senses, because the amount of chemical emitted by a female should be proportional to her size, whereas width, color or spine damage would need vision. The results perhaps suggest that the crabs do not use vision as their primary sense.

Exercise 6.5: Negative Binomial GLM for Crab Data

As we have mentioned at the beginning, the negative binomial GLM can be used to analyze the data. Here, we fit a negative binomial GLM and compare it with the quasi-Poisson model. As an exercise, please perform the following analysis:

(1) Fit GLM(negative binomial; log) with `Sat` as the response and log(`Wt`) as the only explanatory variable.

(2) Find the estimate of k in the negative binomial distribution ($\hat{k} = 0.9580$).

(3) Check if the model is adequate with appropriate diagnostic plots.

(4) Show the residual deviance has an approximate χ^2_{171}.

(5) Compare estimated parameters and their standard errors from the negative binomial GLM and the quasi-Poisson GLM.

(6) Reproduce Fig. 10.9 from the textbook based on the R codes in the textbook. The differences between the two models becomes apparent for heavier crabs, for both the systematic components and the random components.

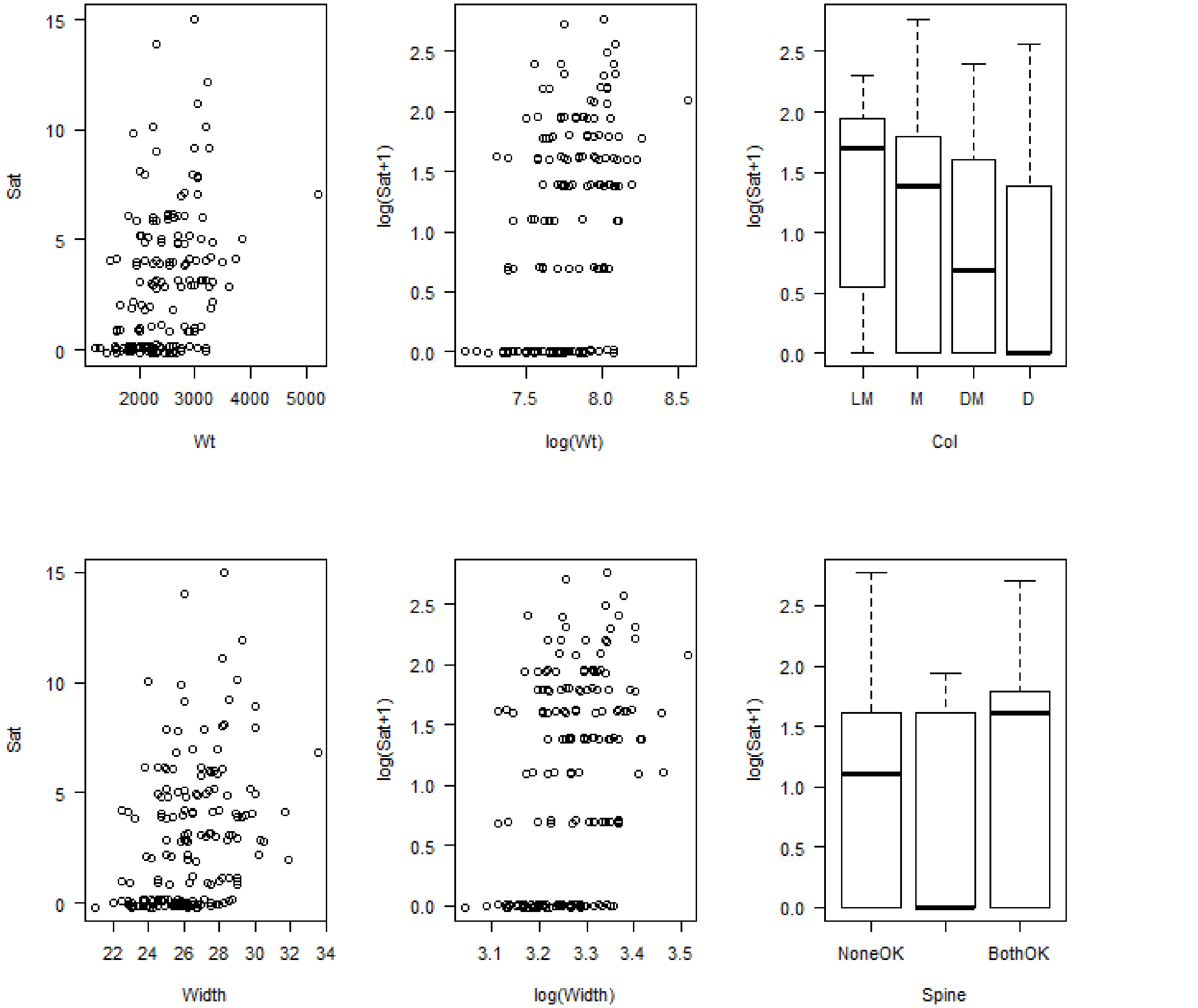

Figure 6.6 The number of satellites on each female horseshoe crab plotted against the weight, color, width and spine condition.

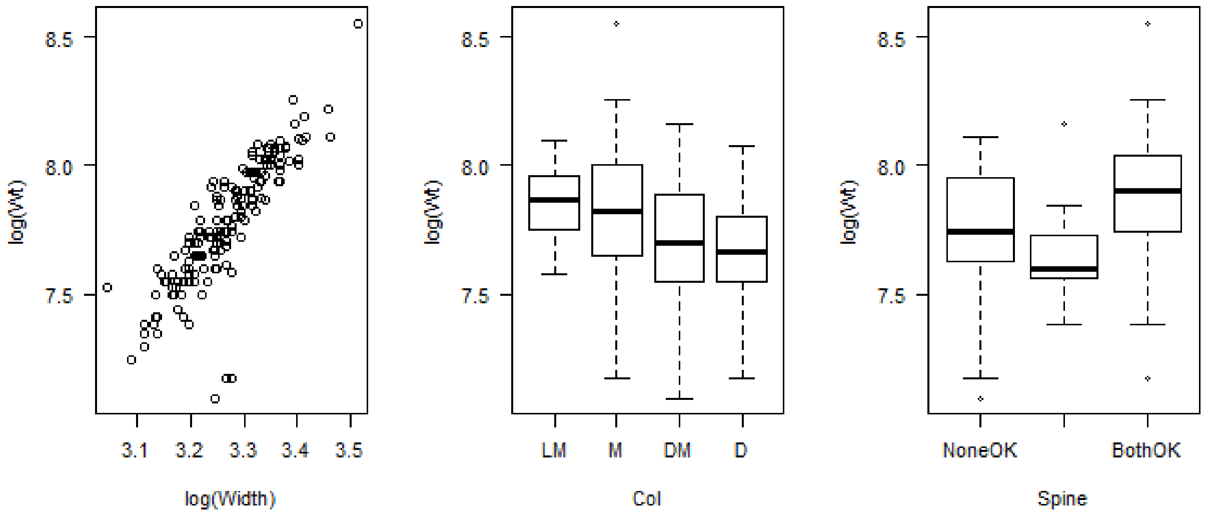

Figure 6.7 Weight of each female horseshoe crab plotted against width, color and spine condition.

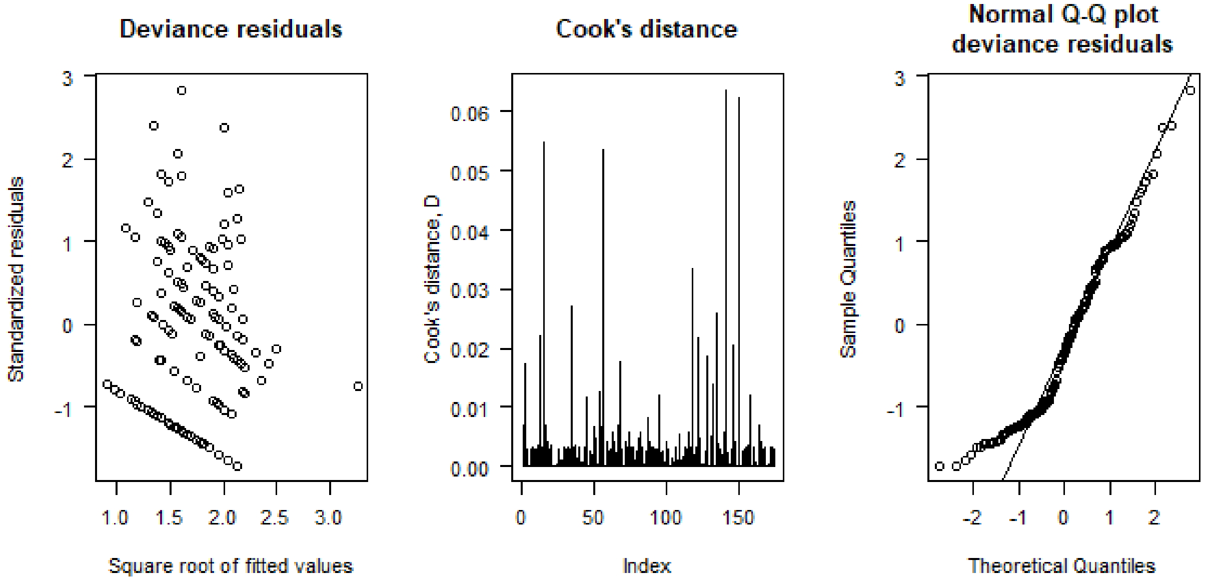

Figure 6.8 Diagnostic plots for the quasi-Poisson model `cr.m2`. The deviance residuals against fitted values (left panel); Cook's distance (center panel); a Q–Q plot of the quantile residuals (right panel).

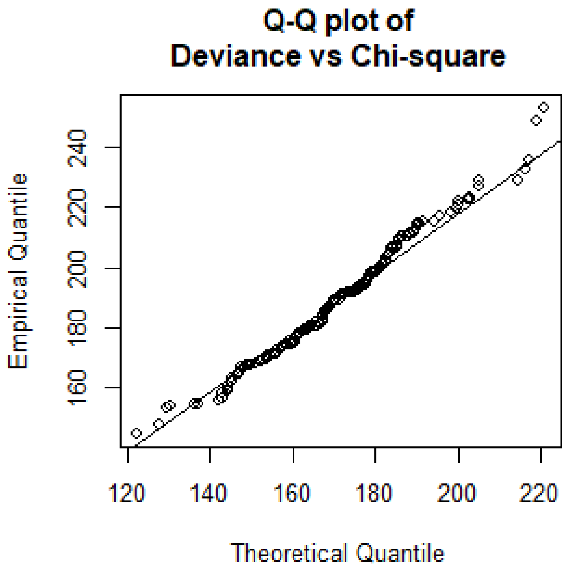

Figure 6.9 Q-Q plot of random residual deviances from model cr.m2 against χ^2_{171}.

Table 6.1 The AIC and the analysis of deviance.

Model	**Model**	**Deviance**	**Deviance DF**	p**-value**	**AIC**	χ^2	p**-value**
1	With Interaction	0	0		144.39	28.31	0.0575
2	Factor Age	28.31	18	0.058	136.69		
3	Numerical Age	49.0	22	$8 * 10^{-4}$	149.36	16.468	$4.95 * 10^{-5}$
4	Quadratic	32.50	21	0.052	134.89		

Table 6.2 The point and interval estimates of linear combinations of parameters from the Poisson GLM Model in **Example 6.2**.

Age **Group 1**	**Age** **Group 2**	**Parameter**	**Point** **Estimate**	**Standard** **Error**	**95% CI**	**95% CI** **(Bonferroni)**
$\mathbf{55-59}$	$\mathbf{40-54}$	β_1	1.082	0.2481	$(0.596, 1.569)$	$(0.443, 1.721)$
$\mathbf{60-64}$	$\mathbf{55-59}$	$\beta_2-\beta_1$	0.491	0.2335	$(-0.038, 0.877)$	$(-0.182, 1.021)$
$\mathbf{65-69}$	$\mathbf{60-64}$	$\beta_3-\beta_2$	0.249	0.2133	$(-0.169, 0.666)$	$(-0.301, 0.798)$
$\mathbf{70-74}$	$\mathbf{65-69}$	$\beta_4-\beta_3$	0.097	0.2173	$(-0.329,0.523\ )$	$(-0.463, 0.657)$
$>\mathbf{74}$	$\mathbf{70-74}$	$\beta_6-\beta_3$	-0.439	0.2393	$(-0.908, 0.030)$	$(-1.055, 0.177)$

Table 6.3 The general $I \times J$ contingency table. The cell count y_{ij} corresponds to level i of A and level j of B.

		Factor *B*				
		Column 1	**Column 1**	$\cdots$	**Column *J***	**Row Total**
Factor *A*	**Row 1**	y_{11}	y_{12}	$\cdots$	y_{1J}	$m_{1\cdot}$
	Row 2	y_{21}	y_{22}	$\cdots$	y_{2J}	$m_{2\cdot}$
	…	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	Row *I*	y_{I1}	y_{I2}	$\cdots$	y_{IJ}	$m_{I\cdot}$
	Column Total	$m_{\cdot 1}$	$m_{\cdot 1}$	$\cdots$	$m_{\cdot 1}$	m

MA5771: Applied Generalized Linear Model

Week 7 At-a-Glance

Title: Tweedie GLMs and Nominal/Ordinal Logistic Models

Overview

In this week, we will focus on models for several types of data. These models include Tweedie GLMs for positive continuous data, Tweedie GLMs for positive continuous data with exact zeros, nominal logistic regression for nominal responses, and proportional odds model for ordinal data. We will discuss Tweedie EDMs in general, including the definition and structure of Tweedie distributions, two special Tweedie EDMs, and a profile likelihood method to estimate the Tweedie index parameter. We will describe the multinominal distribution and it can be used to model nominal and ordinal responses.

Learning Objectives

When you complete this module, you should be able to:

1. Use R to estimate the index parameter in a Tweedie GLM.
2. Fit an appropriate Tweedie GLM for a given data and interpret the model accordingly.
3. Use nominal logistic regression to analyze nominal responses
4. Use proportional odds model to analyze ordinal responses

Instruction Content

See others file for details.

Exam

NA

Quiz

N/A.

Homework

See other files for details.

MA5771: Applied Generalized Linear Model

Week 7 Instruction Contents

Lesson 7.1 Tweedie GLMs

Related Readings: Sections 12.1, 12.2, 12.3, 12.4, 12.5, and 12.6 from Chapter 12.

Section 7.1.1 Introduction and Overview

In this lesson, we introduce 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. In **Section 7.1.2**, we discuss Tweedie GLMs in general, including the definition and structure of Tweedie distributions, two special Tweedie EDMs to model positive continuous data, and positive continuous data with exact zeros. In **Section 7.1.3**, we describe a profile likelihood method to estimate the Tweedie index parameter and R functions to fit Tweedie GLMs. This lesson is concluded with two case studies in **Section 7.1.4**.

Section 7.1.2 The Tweedie EDMs

Introduction of Tweedie Distributions

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

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

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 ξ can take any real value except $0 < \xi < 1$. ξ is called the **Tweedie index**

parameter. This power-variance relationship has been observed in natural populations for many years. Useful information about the Tweedie distribution appears in **Table 7.1**.

Table 7.1 Features of the Tweedie distributions for various values of the index parameter ξ, showing the support S (the permissible values of y) and the domain Ω for μ. The Poisson distribution ($\xi = 1$ and $\varphi = 1$) is a special case of the discrete distributions, and the inverse Gaussian distribution ($\xi = 3$) is a special case of positive stable distributions. R refers to the real line; superscript $+$ means positive real values only; subscript 0 means zero is included in the space.

Tweedie EDM	ξ	S	Ω	**Covered?**
Extreme Stable	$\xi < 0$	R	R^+	Not covered
Normal	$\xi = 0$	R	R	Yes. Week 1
No EDMs exist	$0 < \xi < 1$			
Discrete	$\xi = 1$	$y = 0, \varphi, 2\varphi, \cdots$	R^+	Yes. Week 6
Poisson-Gamma	$1 < \xi < 2$	R_0^+	R^+	Yes. Week 7
Gamma	$\xi = 2$	R^+	R^+	Yes. Week 5
Positive Stable	$\xi > 2$	R^+	R^+	Yes. Week 7

The four specific cases of Tweedie distributions listed above show that the Tweedie distributions are useful for a variety of data types (**Table 7.1**). 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 $y = 0, \varphi, 2\varphi, 3\varphi, \cdots$. When $\varphi = 2$, for example, a positive probability exists for $y = 0, 2, 4, \cdots$. The Poisson distribution is a special case when $\varphi = 1$.
- For $1 < \xi < 2$, the Tweedie distributions are suitable for modelling positive continuous data with exact zeros. An example is rainfall modelling: 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 **Figure 7.1**. 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. The distributions become more right skewed as ξ increases (**Figure 7.2**).

ξ 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 $\xi > 2$ are considered in **Lesson 7.1** in further detail. (The special cases $\xi = 0, 1, 2, 3$ were considered earlier.)

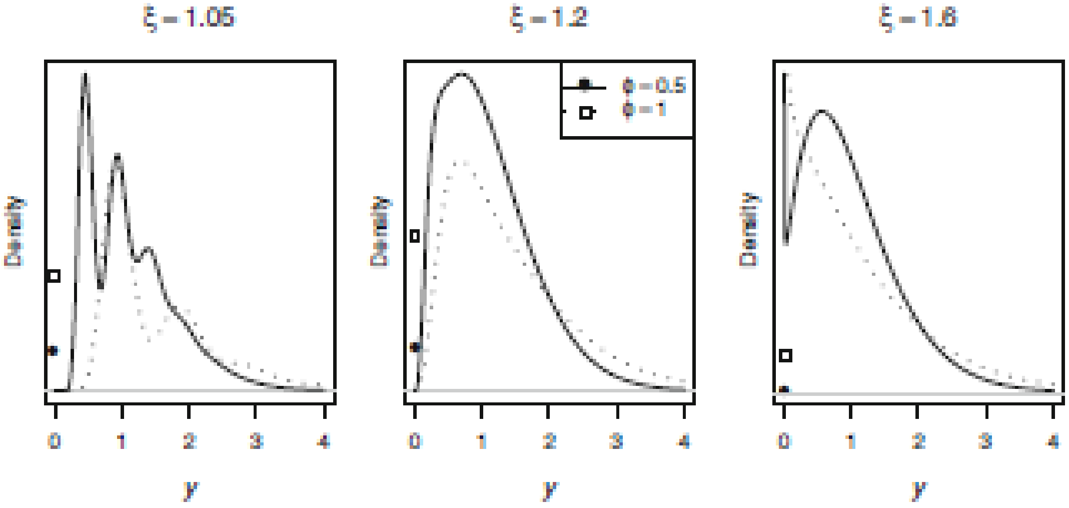

Figure 7.1 Examples of Tweedie probability functions with $1 < \xi < 2$ and $\mu = 1$. The solid lines correspond to $\varphi = 0.5$ and the dotted lines to $\varphi = 1$. The filled dots show the probability of exactly zero when $\varphi = 0.5$ and the empty squares show the probability of exactly zero when $\varphi = 1$.

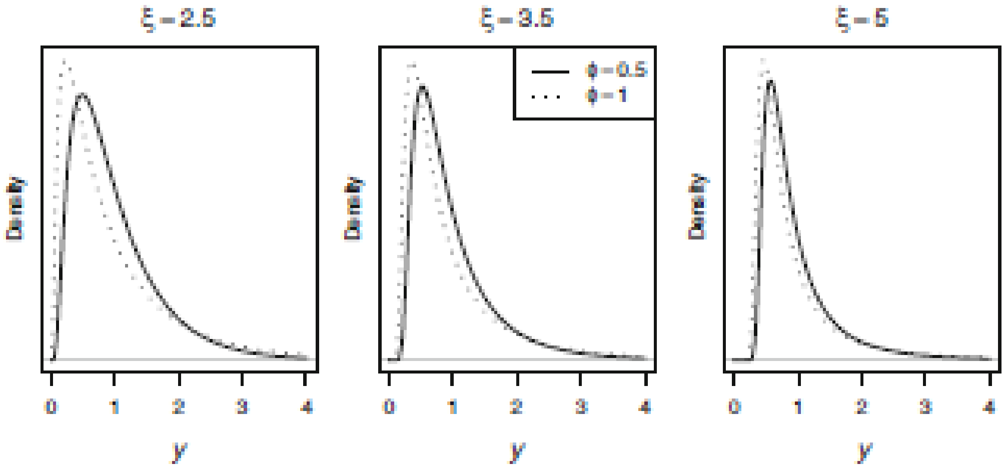

Figure 7.2 Examples of Tweedie probability functions with $\xi > 2$ and $\mu = 1$. As ξ gets larger, the distributions become more skewed to the right. The solid lines correspond to $\varphi = 0.5$; the dotted lines to $\varphi = 1$.

The Structure of Tweedie EDMs

The canonical probability distribution function of an EDM can be written as:

$$\mathcal{P}(y;\theta,\phi) = a(y,\phi)\exp\{\frac{y\theta - \kappa(\theta)}{\phi}\}$$

Tweedie distributions are defined as EDMs with variance function $V(\mu) = \mu^{\xi}$ for some given ξ. Using this relationship, θ and $\kappa(\theta)$ can be determined.

- There are different methods to obtain θ and $\kappa(\theta)$.
- Note that θ and $\kappa(\theta)$ are functions of $\mu = E[y]$ and ξ so the Tweedie distributions are only EDMs if ξ is known.
- In practice, the value of ξ is usually estimated.
- If y follows a Tweedie distribution with index parameter ξ, mean μ and dispersion parameter ϕ, write $y \sim Tw_{\xi}(\mu,\phi)$.
- 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.
- The approximate distribution of residual deviances is adequate when $\phi \leq \frac{[\min\{y_i\}]^{2-\xi}}{3}$.
    - For example for the Poisson distribution, the condition becomes $1 \leq \frac{[\min\{y_i\}]^{2-1}}{3}$, which is equivalent to $\min\{y_i\} \geq 3$, as we have discussed before.

Tweedie EDMs for Positive Continuous Data

In most situations, positive continuous responses are adequately modelled using a Gamma or inverse Gaussian distribution. In some circumstances, neither is adequate, especially for severely skewed data. However, all EDMs with variance functions of the form $V(\mu) = \mu^{\xi}$ 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 7.1: Tweedie EDMs for Positive Continuous Data

The survival times (in 10-hour units) of animals subjected to three types of poison were measured for four different treatments (data set: $poison$). Four animals were used for each poison-treatment combination. We need to find an appropriate GLM of this data.

Solution: We first look at the relationship between the survival time and the types of poison and different treatments. The box plots of the survival time against the types of poison and treatments (plots A and B of **Figure 7.3**) clearly show that the survival varies according to different types of poison and treatments. The lines from the interaction plot (plot C of **Figure 7.3**) are parallel, indicating the little interaction effects between types of poison and treatments.

The data is positive continuous, so a Tweedie EDM with $\xi \geq 2$ can be used. To find an appropriate ξ, we use the following steps:

(1) Calculate the sample means and variances for each combination of poison and treatment.
(2) Plot the logarithm of group variances against the logarithm of group means.
(3) Fit a linear regression for the logarithm of group variances and the logarithm of group means and estimate the slope.

It can be seen that the logarithm of group variances increases linearly with the logarithm of group means (plot D of **Figure 7.3**). Th estimated slope from the linear regression is 3.95, suggesting a Tweedie EDM with $\xi \approx 4$ may be appropriate.

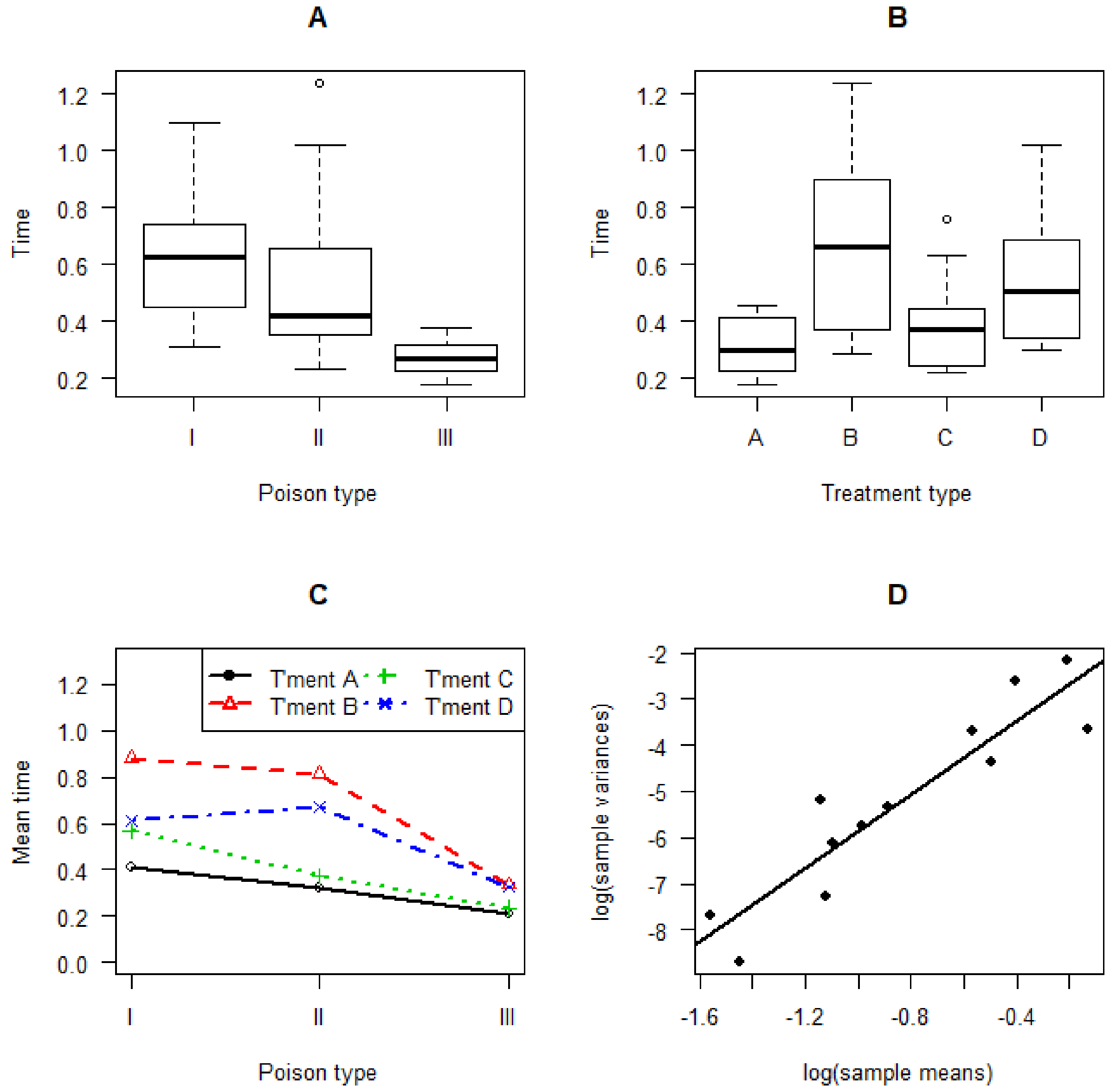

Figure 7.3 The poison data. The time to death plotted against poison type (plot A); the time to death plotted against treatment type (plot B); the mean of the time to death by poison type and treatment type (plot C); the logarithm of each treatment-poison group variance plotted against the logarithm of the group means (plot D).

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:

- Assume N claims are made in a particular company in a certain time frame.
- $N \sim Poisson(\lambda^*)$ where λ^* is the Poisson mean number of claims in the time frame.
- $N = 0$ when no claims are made.
- When $N > 0$, assume the amount of each claim $i = 1, \dots, N$ is z_i, where z_i must be positive.
- Assume z_i follows a gamma distribution with mean μ^* and dispersion parameter ϕ^*, so that $z_i \sim Gamma(\mu^*, \phi^*)$.
- 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.

Example 7.2: Tweedie EDMs for Positive Continuous Data with Exact Zeros

The total July rainfall (in millimeters) at Quilpie, Australia, has been recorded (data set: `quilpie`), together with the value of the monthly mean southern oscillation index (`SOI`). The `SOI` is the standardized difference between the air pressures at Darwin and Tahiti, and is known to have relationships with rainfall in parts of Australia. Some Australian farmers may delay planting crops until a certain amount of rain has fallen (a "rain threshold") within a given time frame (a "rain window"). In this example, we examine the total July rainfall in Quilpie. Observe that the total monthly July rainfall is continuous, with exact zeros. There are 20 months with 0 rainfall.

For this 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 in the corresponding Tweedie distribution are μ, ϕ, and ξ, which can be estimated using R. Several quantities related to underlying Poisson and Gamma distributions can be calculated. Specifically,

- The Poisson mean number, λ^* is

$$\lambda^* = \frac{\mu^{2-\xi}}{\phi(2-\xi)}$$

- The mean of Gamma distribution, μ^* is:

$$\mu^* = (2-\xi)\phi\mu^{\xi-1}$$

- The dispersion parameter of Gamma distribution, ϕ^* is:

$$\phi^* = (2-\xi)(\xi-1)\phi^2\mu^{2(\xi-1)}$$

- Tweedie EDMs with $1 < \xi < 2$ are continuous for $y > 0$, but have a positive probability π_0 at $y = 0$,

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

Once the MLEs of μ, ϕ, and ξ are obtained, the MLEs of λ^*, μ^*, ϕ^* and π_0 can be computed using the above formulas. These estimates give an approximate interpretation of the model based on the underlying Poisson and gamma models, and may sometimes be useful.

Section 7.1.3 Tweedie GLMs

Introduction

GLMs based on the Tweedie distributions are Tweedie GLMs, specified as GLM(Tweedie, ξ; Link function). Here we list a few things about Tweedie GLMs.

- For both cases considered in this lesson (that is, $\xi > 2$ and $1 < \xi < 2$), we have $\mu > 0$ (**Table 7.1**). As a result, the usual link function used for Tweedie GLMs is the logarithmic link function.
- The dispersion parameter ϕ is usually estimated using the Pearson estimate. In some situations, the MLE of ϕ is needed.
- To fit Tweedie GLMs, ξ must be known. Usually the value of ξ is unknown and must be estimated before the Tweedie GLM is fitted.
- The correlation between $\hat{\xi}$ and $\hat{\boldsymbol{\beta}}$ is small, so using the estimate $\hat{\xi}$ has only a small effect on inference concerning β compared to knowing the true value of ξ.

- 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.
- 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.

Estimation of the Index Parameter ξ

For a Tweedie EDM, we have

$$var[y] = \phi V(\mu) = \phi\mu^{\xi} = \phi(E(y))^{\xi}$$

Taking the logarithm at both sides, we obtain:

$$\log(var[y]) = \log(\phi) + \xi\log(E(y))$$

This shows that a simplistic method for estimating ξ 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 7.1** and before. However,

- The estimate of ξ may depend upon how the data are divided.
- If exact zeros are present in the data, then $1 < \xi < 2$.
- 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 7.3: Estimation of the Index Parameter

Estimate the Tweedie index parameter ξ from the Quilpie rainfall data.

Solution: The sample variances and means are calculated with each `SOI` phase and then the slope is estimated to be 1.55.

An alternative approach is to compute the mean and variance of the rainfall amounts within each decade. The estimated slope is 1.95.

The two methods produce quite different estimates of ξ, but both satisfy $1 < \xi < 2$.

A more rigorous method for estimating ξ, 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 ξ. A convenient way to organize the calculations is via the **profile likelihood** for ξ. The following steps can be taken:

- **Step** 1: for a fixed value of ξ, fit the Tweedie GLM.
- **Step** 2: compute the maximum log-likelihood function of GLM.
- **Step** 3: choose another ξ and repeat Steps 1 and 2.
- **Step** 4: The log-likelihoods for different values of ξ is called the **profile log-likelihoods**. The value of ξ giving the largest profile log-likelihood is used as an estimate of ξ and called the **profile likelihood estimate**.

For the method, a plot of the profile log-likelihood against various values of ξ is often useful. The R function `tweedie.profile()` (in R package **tweedie**) can be conveniently used to compute the profile likelihood estimate of ξ. In practice, the profile likelihood plot produced by `tweedie.profile()` should be examined, and values of ξ near 1 should be avoided as necessary.

Example 7.4: Profile Likelihood Estimate of Index Parameter

The total monthly July rainfall at Quilpie, considered in **Examples 7.2** and **7.3** (data set: `quilpie`), is continuous but has exact zeros. We consider modelling the total July rainfall as a function of the `SOI Phase`. Find the profile likelihood estimate of index parameter.

Solution: From the box plot of the rainfall against `SOI Phase` (**Figure 7.4**), we can observe that the variation is greater for larger average rainfall amounts.

The profile likelihood plot (**Figure 7.4**, right panel) shows the likelihood is computed at a small number of ξ 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 ξ is located.

The output object, named `out` in the above, contains a lot of information (see `names(out)`), including the estimate of ξ (as `xi.max`), the nominal 95% confidence interval for ξ (as `ci`), and the MLE of the dispersion parameter ϕ (as `phi.max`). For this example, we have

- Profile likelihood estimate of ξ: $\hat{\xi} = 1.37$
- 95% CI of ξ: $(1.27, 1.50)$
- MLE of dispersion parameter ϕ: $\hat{\phi} = 5.56$

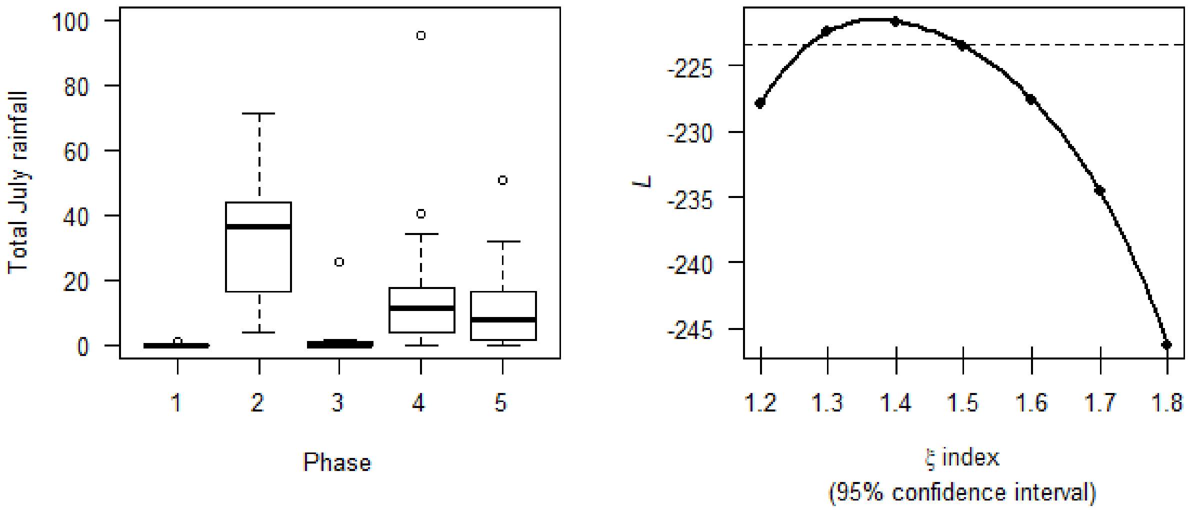

Figure 7.4 The total July rainfall at Quilpie plotted against `SOI Phase` (left panel), and the profile likelihood plot for estimating ξ (right panel).

Fitting Tweedie GLMs

Once an estimate of ξ has been obtained, the Tweedie GLM can be fitted in R using the usual `glm()` function.

- The **statmod** package must be loaded first.
- The Tweedie distributions are denoted in r using `family = tweedie()` in the `glm()` call.
- The call to `family = tweedie()` must specify which the value of ξ, 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`. Usually, `link.power = 0` which corresponds to the logarithmic link function, which is most commonly used link function with Tweedie GLMs.
- Once the model has been fitted, quantile residuals 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 $y = 0$.

Example 7.5: Tweedie GLM

For the Quilpie rainfall data (data set: `quilpie`), the estimate of ξ found in **Example 7.4** is $\xi \approx 1.3714$. We focus on the diagnostic plots for this example.

We first compare the Pearson, deviance and quantile residuals with Q-Q plots (**Figure 7.5**).

- We can that 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.
- The quantile residuals use a small amount of randomization 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 Tweedie GLMs with $1 < \xi < 2$.

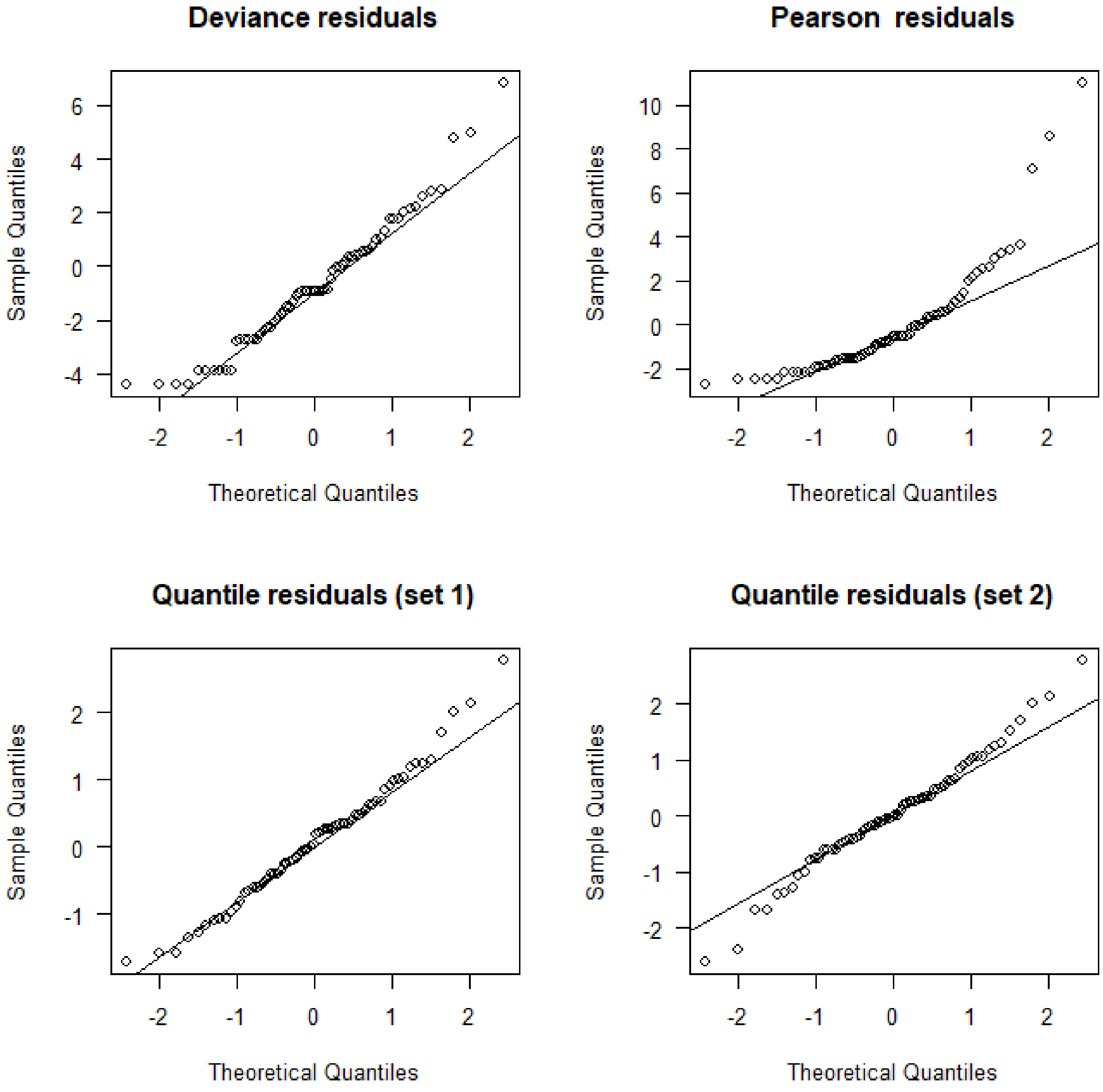

Figure 7.5 Q–Q plots for the Pearson, deviance and quantile residuals for the Tweedie GLM fitted to the Quilpie rainfall data. Two realization of the quantile residuals are shown.

Exercise 7.1: Diagnostic of Tweedie GLMs

Replicate Fig. 12.7 in the textbook to confirm that: (1) The diagnostic plots suggest the model is reasonable; (2) No observations are identified as influential using Cook's distance.

Solution: Please refer to R program for the details. Note your plots may not be identical to Fig. 12.7 in the textbook due to the randomization of the quantile residuals.

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 7.6: Poisson sum of Gamma Distributions

For the Quilpie rainfall data (data set: `quilpie`) and the Tweedie GLM fitted in **Example 7.5**, find the corresponding parameter in Poisson and Gamma distributions, and the predicted number of zero-rainfall months $\hat{\pi}_0$ for each `SOI Phase`.

Solution: We need to obtain the MLE of ϕ, ξ and μ for each `SOI Phase`.

- $\hat{\xi} = 1.3714$: obtained from `tweedie.profile()` (**Example 7.4**)
- $\hat{\phi} = 5.5587$: obtained from `tweedie.profile()` (**Example 7.4**)
- $\hat{\phi} = 6.806$: Pearson estimate from `glm()`, should not be used here
- $\hat{\mu} = [0.1143, 33.8938, 3.8428, 17.4235, 11.9143]$: obtained from `predict()`

Since $\hat{\pi}_0 = \exp\left\{-\frac{\hat{\mu}^{2-\hat{\xi}}}{\hat{\phi}(2-\hat{\xi})}\right\}$, we can obtain

$$\hat{\pi}_0 = [0.9294, 0.0727, 0.5132, 0.1782, 0.2570]$$

For example, $\exp\left\{-\frac{0.1142^{2-1.3714}}{5.5587*(2-1.3714)}\right\} = 0.9294$.

Figure 7.6 shows the plot of the expected probability of a zero against the proportion of zeros in the data for each `SOI Phase`. 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. Notice that $1 < \xi < 2$ since exact zeros are present in the data. Note that no months with exactly zero rainfall were observed during Phase 2, the Tweedie GLM assigns a (small) probability (0.0727) that such an event could occur.

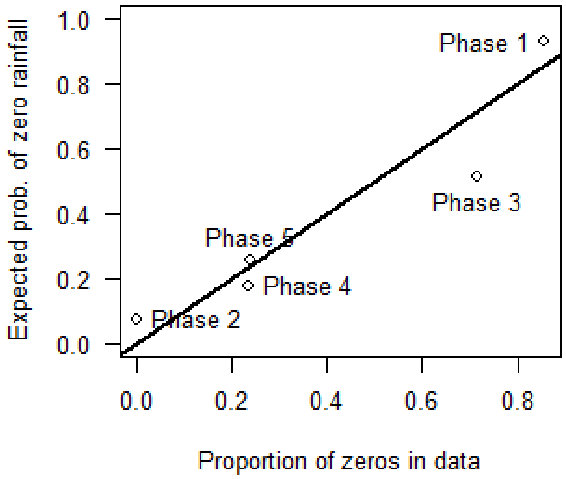

Figure 7.6 plot of the expected probability of a zero against the proportion of zeros in the data for each `SOI Phase`.

The estimated parameters of the Tweedie GLM can be used to interpret the underlying Poisson and Gamma distributions. To do so, use the `tweedie.convert()` function in package **tweedie** (**Table 7.2**). In the context of rainfall modelling, this interpretation in terms of λ^*, μ^* and ϕ^* is a form of **statistical downscaling**. The estimates of the Poisson mean λ^* show the mean number of rainfall events in July when the SOI is in each phase, and the estimates of the gamma mean μ^* give the mean amount of rainfall in each rainfall event for each SOI phase. For Phase 2 the model predicts a mean of 2.621 rainfall events occur in July, with a mean of 1.375 mm in each. The mean monthly July rainfall predicted by the model agrees with the observed mean rainfall in the data.

Table 7.2 The MLE of parameters in the Poisson and Gamma distributions.

		Estimates				
Parameter	**Parameter**	**Phase 1**	**Phase 2**	**Phase 3**	**Phase 4**	**Phase 5**
Poisson Mean	λ^*	0.0732	2.6210	0.6671	1.7251	1.3585
Gamma Mean	μ^*	1.5611	12.9318	5.7608	10.1000	8.7702
Gamma Dispersion	ϕ^*	0.5909	0.5909	0.5909	0.5909	0.5909

Exercise 7.2: Parameters in Poisson and Gamma Distributions

Verify **Table 7.2** based on:

$$\lambda^* = \frac{\mu^{2-\xi}}{\phi(2-\xi)}, \mu^* = (2-\xi)\phi\mu^{\xi-1}, \phi^* = (2-\xi)(\xi-1)\phi^2\mu^{2(\xi-1)}$$

$$\hat{\xi} = 1.3714, \hat{\phi} = 5.5587, \hat{\mu} = [0.1143, 33.8938, 3.8428, 17.4235, 11.9143]$$

Section 7.1.4 Case Studies

Case Study 1

A study of performance degradation of electrical insulation from accelerated tests 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 × 4 factorial experiment. The following analysis can be done by the order.

Exploratory analysis. A plot of the data (**Figure 7.7**) may suggest that a temperature of 275◦C is different than the rest. The plot also seems to show that the variance increases as `Time` increases.

Estimate index parameter ξ. To consider fitting a Tweedie GLM to the data, we use `tweedie.profile()` to find an estimate of ξ : $\hat{\xi} = 1.592$. The main effects and interaction effects of `Temperature` and `Time`. The profile likelihood is plotted (**Figure 7.8**, left panel). Note that $1 < \hat{\xi} < 2$ even though all breakdown strengths are positive.

Tweedie GLM. GLM(Tweedie, $\hat{\xi} = 1.592$; log) is fitted. The Q-Q plot (**Figure 7.8**, right panel) suggests no major problems with the model. The analysis of deviance table shows that both main effects and the interaction effects are significant.

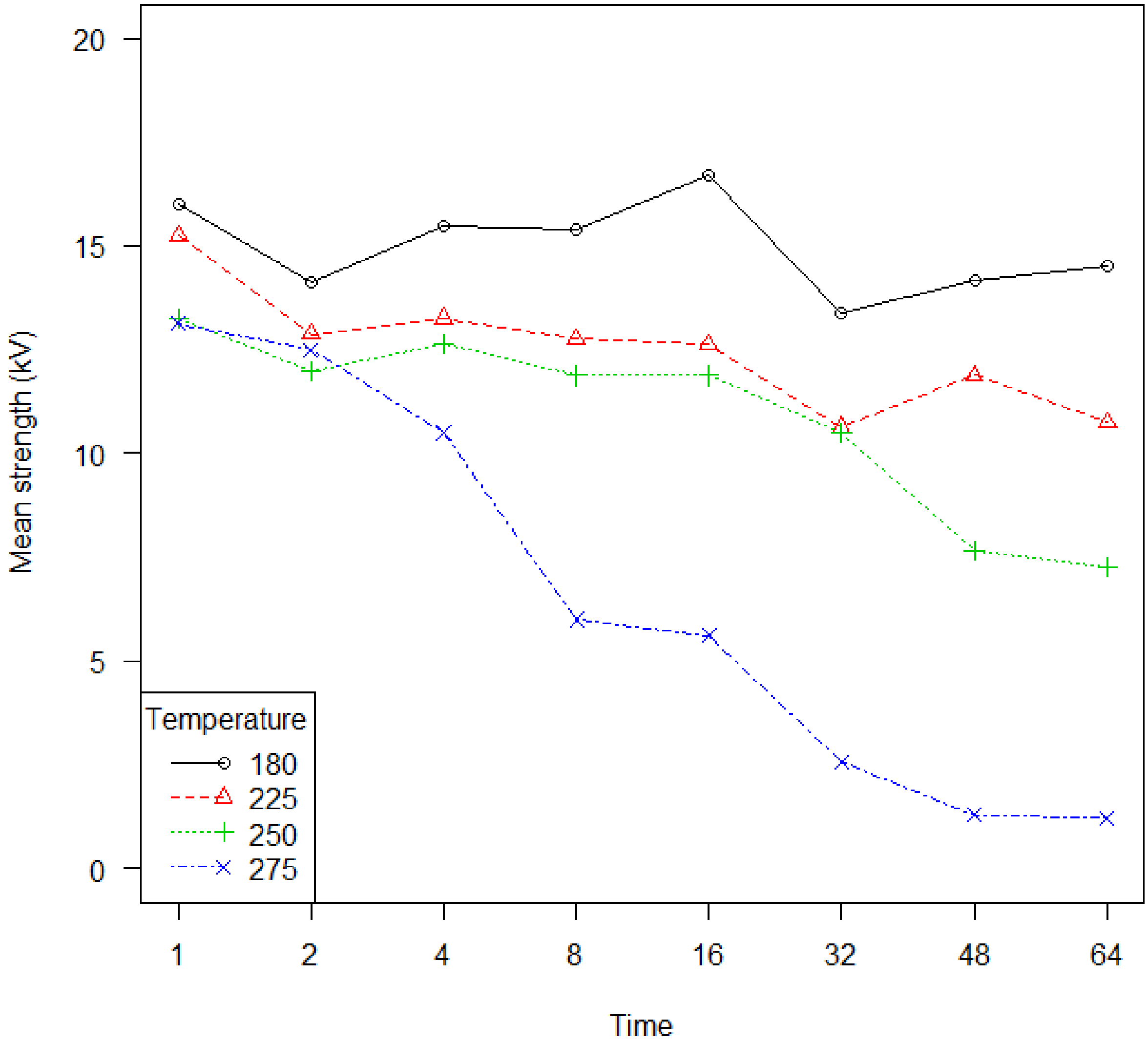

Figure 7.7 A plot of the dialetric breakdown data.

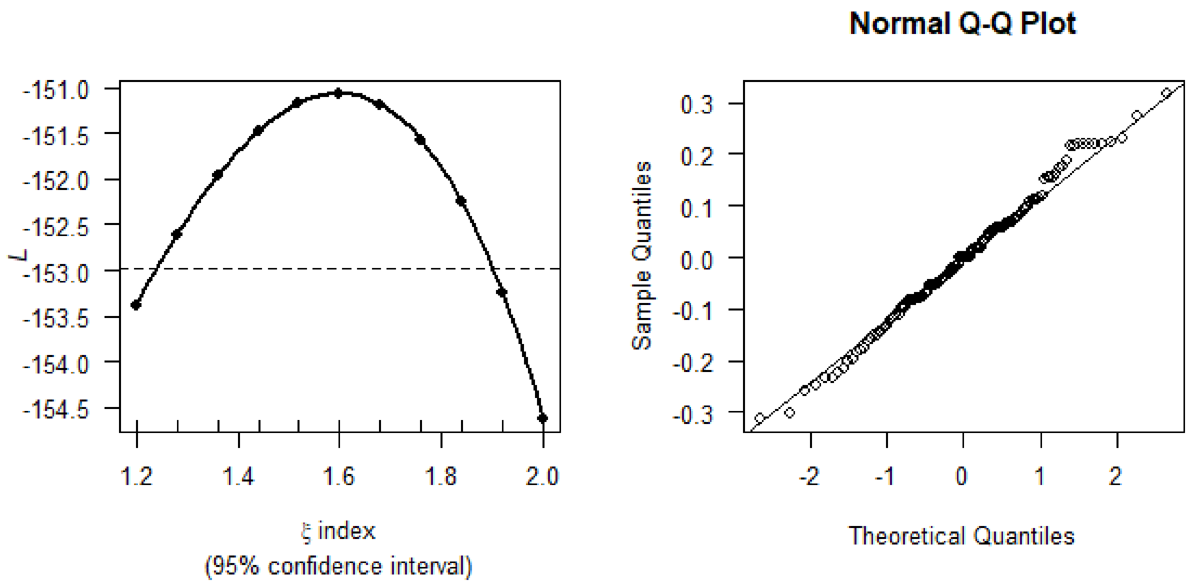

Figure 7.8 The profile-likelihood plot (left panel) and Q–Q plot of quantile residuals (right panel) for the dialetric breakdown data.

Case Study 2

Consider the survival times data first introduced in **Example 7.1**, where a Tweedie EDM with $\xi = 4$ was suggested for modelling the data (data set: `poison`). To find the appropriate Tweedie GLM for modelling the data more formally, initially determine an estimate of ξ using the profile likelihood (**Figure 7.9**). Using the R function `tweedie.profile()` from the package **tweedie**, we estimate $\hat{\xi} = 3.83$ and the 95% confidence interval of ξ is $(2.87, 4.88)$. Therefore, we fit a Tweedie model with $\xi = 4$.

The first model includes the main effects and the interaction effects of the types of poison and treatments. The analysis of deviance table shows that the interaction is not significant. The final model only contains the main effects.

The diagnostic plots suggest model poison.m2 is adequate (**Figure 7.10**), though the residuals for Poison 2 are more variable than for other poisons.

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

- $y \sim \mathrm{Tw}_{\xi=4}(\hat{\mu}, \hat{\phi} = 0.2656)$ (random)

- $\log(E[y]) = \log(\hat{\mu}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5$ (systematic)

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 the dispersion parameter ϕ is given in the output of `summary()` as $\hat{\phi} = 0.2656$.

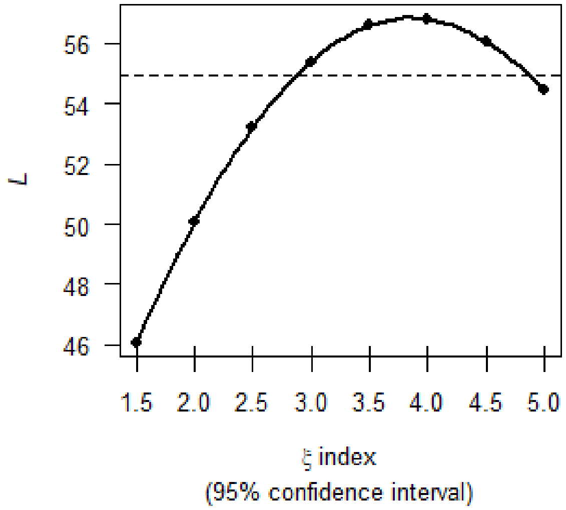

Figure 7.9 The profile likelihood plot for estimating the value of the Tweedie index parameter ξ for the `poison` data.

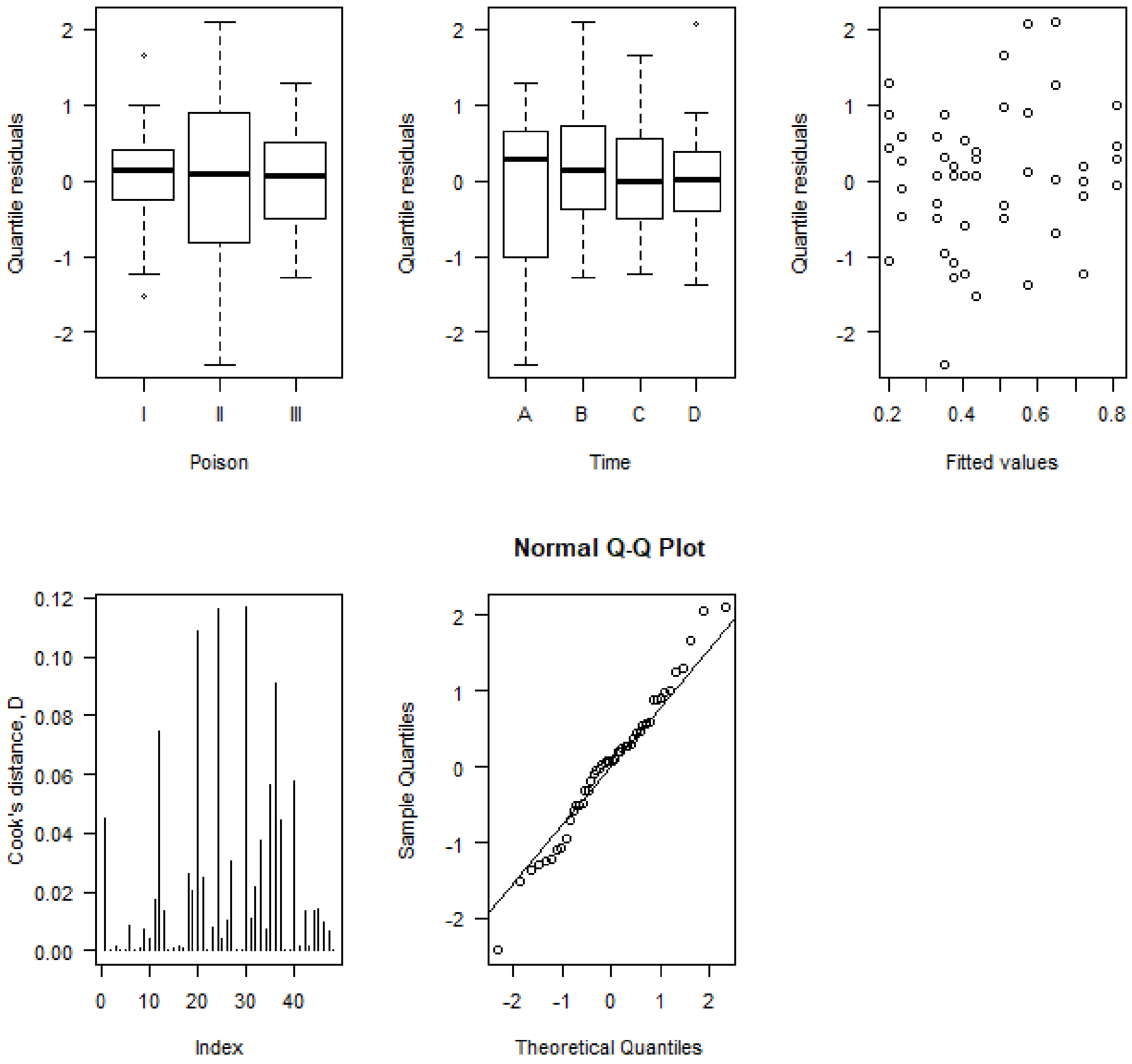

Figure 7.10 The diagnostics for the final Tweedie GLM model fitted to the poison data.

Lesson 7.2 Nominal and Ordinal Logistic Regression

Related Readings: Chapter 8 from "introduction to Generalized Linear Models" by Annette J. Dobson and Adrian G. Barnett.

Section 7.1.1 Introduction and Overview

In this lesson, we will describe methods to model nominal responses and ordinal responses. We will discuss nominal logistic regressions for nominal responses (**Section 7.2.2**) and present a case study using nominal logistic regression (**Section 7.2.3**). We will then introduce several ordinal logistic regressions but focus on proportional odds model (ordinal logistic regression) (**Section 7.2.4**).

Section 7.2.2 Nominal Logistic Regression

Introduction

If the response variable is categorical, with more than two categories, then there are two options for generalized linear models. The first option is to model the frequencies or counts for the covariate patterns as the response variables with Poisson distributions. This approach is also called log-linear modelling, is covered in **Lesson 6.1**. The second approach relies on generalizations of logistic regression from dichotomous responses, described in **Lessons 4.2** and **5.1**. Such approach is called nominal or ordinal logistic regression.

Both log-linear models and nominal or ordinal logistic regression can analyze data that can be organized in contingencies tables. There are also some differences between log-linear models and nominal or ordinal logistic regressions.

- For nominal or ordinal logistic regression, one of the measured or observed categorical variables is regarded as the response, and all other variables are regarded explanatory variables. The quantitative variables can be used as the explanatory variables.
- For log-linear models, all the variables are treated alike. The quantitative variables cannot be used as the explanatory variables.

- Nominal or ordinal logistic regression yield odds ratio estimates which are relatively easy to interpret if there are no interactions (or only fairly simple interactions).
- Log-linear models are good for testing hypotheses about complex interactions, but the parameter estimates are less easily interpreted.

Therefore, the choice between log-linear model and nominal or ordinal logistic regression in a particular situation depends on several factors.

- If one variable is clearly a "response" (for example, the outcome of a prospective study), then nominal or ordinal logistic regression should be used.
- If several variables have the same status (as may be the situation in a cross-sectional study), then log-linear models should be used.
- Additionally, the choice may depend on how the results are to be presented and interpreted.

Multinomial Distribution

The binomial distribution can be used to describe the number of an outcome y of a total number m when there are only two possible categories. The multinomial distribution is a generalization of the binomial distribution. It can be used to describe the number of an outcome y of a total number m when there are more than two possible categories. Although we have described the multinomial distribution in **Lesson 6.1**, we provide more details of this distribution.

Consider a random variable Y with K categories. Let $\mu_1, \cdots, \mu_K$ denote the respective probabilities (or expected proportion), with $\mu_1 + \cdots + \mu_K = 1$. If there are m independent observations of Y which result in y_1 outcomes in category 1, y_2 outcomes in category 2, and so on, then $y_1 + \cdots + y_K = m$ and the distribution of $y_1, \cdots, y_K$ is

$$\mathcal{P}(y_1, \cdots, y_K; \mu_1, \cdots, \mu_K, m) = \frac{m!}{y_1! \cdots y_K!} \mu_1^{y_1} \cdots \mu_K^{y_K}$$

The properties of the multinomial distribution include:

- When $K = 2$, then $\mu_2 = 1 - \mu_1$, $y_2 = m - y_1$ and

  $$\mathcal{P}(y_1, y_2; \mu_1, \mu_2, m) = \frac{m!}{y_1!\, y_2!} \mu_1^{y_1} \mu_2^{y_2} = \binom{m}{y_1} \mu_1^{y_1} (1 - \mu_1)^{m - y_1}$$

  which is the binomial distribution.

- In general, the multinomial distribution is not an EDM. However, each of $y_k(k = 1, \cdots . K)$ has a binomial distribution with parameter μ_k and m thus is an EDM. Therefore, $E[y_k] = \mu_k$ and $var[y_k] = \mu_k(1 - \mu_k)$.
- There is a relationship between the multinomial distribution and the Poisson distribution. Such relationship ensures that generalized linear model is appropriate.

Now we can write out the likelihood function. Let $y_1, \cdots, y_n$ be n observations from a response variable with K categories. Each y_i will be a number of outcomes from a specific category k_i. When $K = 2$, we can set k_i be one of two categories. Depending on how the data is collected, the log-likelihood function can be different. But after ignoring items without μ, the log-likelihood function is:

$$\ell(y_1, \cdots, y_n; \boldsymbol{\mu}) = \sum_{i=1}^{n} y_i * \log(\mu_{k_i,i})$$

where $\mu_{k_i,i}$ depends on k_i (the category of y_i), the corresponding values of explanatory variables, and the link function. This formula is difficult to understand but we can use the following example as an illustration.

Example 7.7: Car preferences

In a study of motor vehicle safety, men and women driving small, medium and large cars were interviewed about vehicle safety and their preferences for cars, and various measurements were made of how close they sat to the steering wheel. There were 50 subjects in each of the six categories (two sexes and three car sizes). They were asked to rate how important various features were to them when they were buying a car. The data in **Table 7.3** and the file `car.csv` shows the ratings for air conditioning and power steering, according to the sex and age of the subject (the categories “not important” and “of little importance” have been combined). For this data, we would like to write out its log-likelihood function.

Table 7.3 Importance of air conditioning and power steering in cars (row percent- ages in brackets).

		Response			
		No or Little		**Very**	
Sex	**Age**	**Important**	**Important**	**Important**	**Total**
Women	$18-23$	26 (58%)	12 (27%)	7 (16%)	45
	$24-40$	9 (20%)	21 (47%)	15 (33%)	45

	> 40	5 (8%)	14 (23%)	41 (68%)	60
Men	$18-23$	40 (62%)	17 (26%)	8 (12%)	65
	$24-40$	17 (39%)	15 (34%)	12 (27%)	44
	> 40	8 (20%)	13 (37%)	18 (44%)	41
Total		105	94	101	300

Solution: The authors stated that there were 50 subjects in each of the six categories (two sexes and three car sizes). It seems this statement is not true from **Table 7.3**. Nonetheless, let us assume that the row totals are fixed so the three numbers of each row follow a multinomial distribution. Let us consider 2nd row (Women aged $24-40$). The log-likelihood function for these three observations $y_4 = 9$, $y_5 = 21$, and $y_6 = 15$ are

$$\ell = \log\left(\frac{45}{9!\,21!\,15!}\right) + 9 * \log(\mu_{1,4}) + 21 * \log(\mu_{2,5}) + 15 * \log(\mu_{3,6})$$

Here, $\log\left(\frac{45}{9!21!15!}\right)$ is an item without parameters. $\mu_{1,4}$ refers to the mean of y_4 since it is the number of "No or little improve" which is the 1st category of the response variable. $\mu_{1,4}$ depends on the link function and the corresponding values of explanatory variable. If we only consider the main effects of the gender and the age group and use the treatment coding with women aged $18-23$ as the reference levels, we have

$$g(\mu_{1,4}) = \eta_4 = \beta_{0,1} + \beta_{24-40,1}$$

where $\beta_{0,1}$ is the intercept for the 1st category of the response variable and $\beta_{24-40,1}$ is the coefficient of age group $24-40$ for the 1st category of the response variable. Similarly, we can have:

$$g(\mu_{2,5}) = \eta_5 = \beta_{0,2} + \beta_{24-40,2}$$

$$g(\mu_{3,6}) = \eta_6 = \beta_{0,3} + \beta_{24-40,3}$$

Note that for each category of the response variable, we have a set of coefficients in the linear predictor. Because $\mu_{1,4} + \mu_{2,5} + \mu_{3,6} = 1$, we actually only need two sets of coefficients.

As another example, for the last three observations $y_{16} = 8$, $y_{17} = 15$, and $y_{18} = 44$ (Men aged > 40), their log-likelihood function is:

$$\mathscr{b} = \log\left(\frac{41}{8!\,15!\,18!}\right) + 8 * \log(\mu_{1,16}) + 15 * \log(\mu_{2,17}) + 18 * \log(\mu_{3,18})$$

And we can have:

$$g(\mu_{1,16}) = \eta_{16} = \beta_{0,1} + \beta_{\text{Men},1} + \beta_{>40,1}$$

$$g(\mu_{2,17}) = \eta_{17} = \beta_{0,2} + \beta_{\text{Men},2} + \beta_{>40,2}$$

$$g(\mu_{3,16}) = \eta_{16} = \beta_{0,3} + \beta_{\text{Men},3} + \beta_{>40,3}$$

For this data, the set of parameters could be:

$$\beta_{0,1}, \beta_{0,2}, \beta_{0,3}, \beta_{\text{Men},1}, \beta_{\text{Men},2}, \beta_{\text{Men},3}, \beta_{24-40,1}, \beta_{24-40,2}, \beta_{24-40,3}, \beta_{>40,1}, \beta_{>40,2}, \beta_{>40,3}$$

In addition, you can see that the log-likelihood function of all data is:

$$\ell(y_1, \cdots, y_n; \boldsymbol{\mu}) = \sum_{i=1}^{n} y_i * \log(\mu_{k_i,i})$$

after ignoring items without model parameters.

Nominal Logistic Regression: Systematic Component

Nominal logistic regression models are used when there is no natural order among the response categories. For the response variable with two categories, the logistic regression (or the binomial GLM with the logit link function) has the following systematic component:

$$\text{logit}(\mu_2) = \log\left(\frac{\mu_2}{1-\mu_2}\right) = \log\left(\frac{\mu_2}{\mu_1}\right) = \eta$$

In the logistic regression, the linear predictor, η, can be defined by the following two steps:

(1) Choose one of categories as the reference level. For example, choose the category 1 as the reference category.
(2) The linear component η, is defined as the logarithmic ratio of the expect portion of the other category and the expect portion of the reference category.

Such link function can be easily extended to the response variable with more than two categories. For a response variable with K possible categories, one category is arbitrarily chosen as the reference category. Suppose that the first category is the reference category. Then the logits for the other categories are defined by

$$\text{logit}(\mu_k) = \log\left(\frac{\mu_k}{\mu_1}\right) = \eta_k$$

Each linear predictor η_k is a linear function of parameters $\beta_{0k}, \beta_{1k}, \cdots, \beta_{pk}$. So the model has a total $(K-1)*(p+1)$ parameters. When $K = 2$, this model reduces to the logistic regression.

From $\text{logit}(\mu_k) = \log\left(\frac{\mu_k}{\mu_1}\right) = \eta_k$, we can obtain

$$\mu_k = \mu_1 \exp(\eta_k)\,, k = 2, \cdots, K.$$

From

$$1 = \mu_1 + \mu_2 + \cdots + \mu_K = \mu_1 + \mu_1 * \exp(\eta_2) + \cdots + \mu_1 * \exp(\eta_K)$$

$$= \mu_1(1 + \exp(\eta_2) + \cdots + \exp(\eta_K))$$

we have

$$\mu_1 = \frac{1}{1 + \exp(\eta_2) + \cdots + \exp(\eta_K)}$$

So $\mu_1, \mu_2, \cdots, \mu_K$ are function of parameters $\beta_{0k}, \beta_{1k}, \cdots, \beta_{pk}$. Then the maximum likelihood estimators of parameters $\beta_{0k}, \beta_{1k}, \cdots, \beta_{pk}$ can be obtained. Again, these concepts and formulas can be difficult to understand, so we use the following example as an illustration.

Example 7.8: Nominal Logistic Regression

For the car data (**Table 7.3**) in **Example 7.7**, find the log-likelihood function of nominal logistic regression in terms of parameters $\beta_{0k}, \beta_{1k}, \cdots, \beta_{pk}$.

Solution: We still use the data from the 2nd row and 6th row as the examples.

The log-likelihood function for these three observations $y_4 = 9$, $y_5 = 21$, and $y_6 = 15$ are

$$\ell = \log\left(\frac{45}{9!\,21!\,15!}\right) + 9 * \log(\mu_{1,4}) + 21 * \log(\mu_{2,5}) + 15 * \log(\mu_{3,6})$$

We need to re-write $\mu_{1,4}$, $\mu_{2,5}$, and $\mu_{3,6}$ as the function of parameters $\beta_{0k}, \beta_{1k}, \cdots, \beta_{pk}$. In this example, $p = 3$ and

$$\mu_{1,4} = \frac{1}{1 + \exp(\beta_{0,2} + \beta_{24-40,2}) + \exp(\beta_{0,3} + \beta_{24-40,3})}$$

$$\mu_{2,5} = \mu_{1,4} * \exp(\beta_{0,2} + \beta_{24-40,2})$$

$$\mu_{3,6} = \mu_{1,4} * \exp(\beta_{0,3} + \beta_{24-40,3})$$

Exercise 7.3: Nominal Logistic Regression

For the car data (**Table 7.3**) in **Example 7.7**, the log-likelihood function of the last three observations $y_{16} = 8$, $y_{17} = 15$, and $y_{18} = 44$ (Men aged > 40), is:

$$\ell = \log\left(\frac{41}{8!\,15!\,18!}\right) + 8 * \log(\mu_{1,16}) + 15 * \log(\mu_{2,17}) + 18 * \log(\mu_{3,18})$$

Show that:

$$\mu_{1,16} = \frac{1}{1 + \exp(\beta_{0,2} + \beta_{\mathrm{Men},2} + \beta_{>40,2}) + \exp(\beta_{0,3} + \beta_{\mathrm{Men},3} + \beta_{>40,3})}$$

$$\mu_{2,17} = \mu_{1,16} * \exp(\beta_{0,2} + \beta_{\mathrm{Men},2} + \beta_{>40,2})$$

$$\mu_{3,18} = \mu_{1,16} * \exp(\beta_{0,3} + \beta_{\mathrm{Men},3} + \beta_{>40,3})$$

Nominal Logistic Regression: $\boldsymbol{\beta}$ and Odds Ratio

Often it is easier to interpret the effects of explanatory factors in terms of odds ratios than the parameters β. For simplicity, consider a response variable with K categories and a binary explanatory variable x which denotes whether an "exposure" factor is present ($x = 1$) or absent ($x = 0$). The odds ratio for exposure for response k ($k = 2, \cdots, K$) relative to the reference category $k = 1$:

$$OR_k = \frac{\mu_{kp}}{\mu_{ka}} / \frac{\mu_{1p}}{\mu_{1a}}$$

where μ_{kp} and μ_{ka} denote the probabilities of response category k ($k = 1, \cdots, K$) according to whether exposure is present or absent, respectively. For the model

$$\log\left(\frac{\mu_k}{\mu_1}\right) = \beta_{0k} + \beta_{1k}x, k = 2, \cdots, K$$

The logarithm of odds are

- $\log\left(\frac{\mu_{ka}}{\mu_{1a}}\right) = \beta_{0k}$, when $x = 0$, indicating the exposure is absent, and

- $\log\left(\frac{\mu_{kp}}{\mu_{1p}}\right) = \beta_{0k} + \beta_{1k}$, when $x = 1$, indicating the exposure is present

Therefore, the logarithm of the odds ratio can be written as

$$\log(OR_k) = \log\left(\frac{\mu_{kp}}{\mu_{ka}}\right) - \log\left(\frac{\mu_{1p}}{\mu_{1a}}\right) = \beta_{0k} + \beta_{1k} - \beta_{0k} = \beta_{1k}$$

Hence $OR_k = \exp(\beta_{1k})$:

- The estimated odds ratio is $OR_k = \exp(\hat{\beta}_{1k})$.
- The $100(1-\alpha)\%$ confidence interval of odds ratio OR_k is

$$\exp(\hat{\beta}_{1k} \pm z_{\frac{\alpha}{2}} * se(\hat{\beta}_{1k}))$$

- If $\beta_{1k} = 0$, $OR_k = 1$ which corresponds to the exposure factor having no effect.

Fit Nominal Logistic Regression with R

The nominal logistic regression can be fitted with the R function `multinom()` for package **nnet**. The program uses the optimization method from the neural net trainer in **nnet** to compute he maximum likelihood, but there is no deeper connection to neural networks. Once the model parameters are estimated ($\hat{\beta}_{0k}, \hat{\beta}_{1k}, \cdots, \hat{\beta}_{pk}, k = 2, \cdots, K$), theoretically we can calculate different types of residuals, the deviance, the Pearson statistic, etc. However, due to the implementation of `multinom()`, many functions that work for GLM don't work as expected. Here we list what can be calculated with a nominal logistic regression from `multinom()`. Suppose that the fitted model is named as `fit` in R:

- `multinom()`: The response is a factor with two or more levels. y_i (the counts in the table) are supplied as the weights in the call.
- `summary(fit)`: provides some basic information of fitted model.
  - `summary(fit)$fitted.values`: estimated probabilities for the observation. For each observation, K probabilities are estimated.
  - `summary(fit)$residuals`: the difference of observed probability and the estimated probabilities. Such residuals cannot be used in diagnostic plots.
  - `summary(fit)$weights`: the prior weights used in the model fitting
  - `summary(fit)$edf`: the number of parameters used in the model

  - `summary(fit)$coefficients`: the estimated parameters $\hat{\beta}_{0k}, \hat{\beta}_{1k}, \cdots, \hat{\beta}_{pk}$.
  - `summary(fit)$standard.errors`: the standard errors of $\hat{\beta}_{0k}, \hat{\beta}_{1k}, \cdots, \hat{\beta}_{pk}$
  - `summary(fit)$deviance`: incorrect deviance so it should not be used
  - `summary(fit)$AIC`: AIC of the model, can be used to compare two non-nested models.
- `logLik(fit)`: the log-likelihood function after ignoring the constant
- `fitted(fit)`: same as `summary(fit)$fitted.values`
- `predict(fit)`: same as `summary(fit)$fitted.values`
- `resid(fit)` or `residuals(fit)`: same as `summary(fit)$residuals`
- `deviance(fit)`: same as `summary(fit)$deviance`, so it should not be used
- `confint(fit)`: confidence intervals of coefficients
- `anova()`: compare two nested models
- `AIC(fit)`: AIC of the model

Section 7.2.3 Case Study of Logistic: Car Preferences

We would like to fit a nominal logistic regression using the car data (**Table 7.3**). For these data the response, importance of air conditioning and power steering, is rated on an ordinal scale but for the purpose of this case study the order is ignored and the 3-point scale is treated as nominal. The category "no or little" importance is chosen as the reference category. We only consider the main effects of the gender and the age group and use the treatment coding with women aged $18-23$ as the reference levels. Age is also ordinal, but initially we will regard it as nominal. The analysis can be divided into several steps: (1) exploratory analysis; (2) modeling fitting and selection; (3) model diagnostic; (4) conclusions.

Exploratory analysis. The proportions of responses in each category by age and sex are shown in **Figure 7.11**. It is clear that the importance of air-conditioning and power steering increased with age. Also men considered these features less important than women did. The plot does not suggest a strong interaction between the age group and the gender since the lines in both plots have similar patterns.

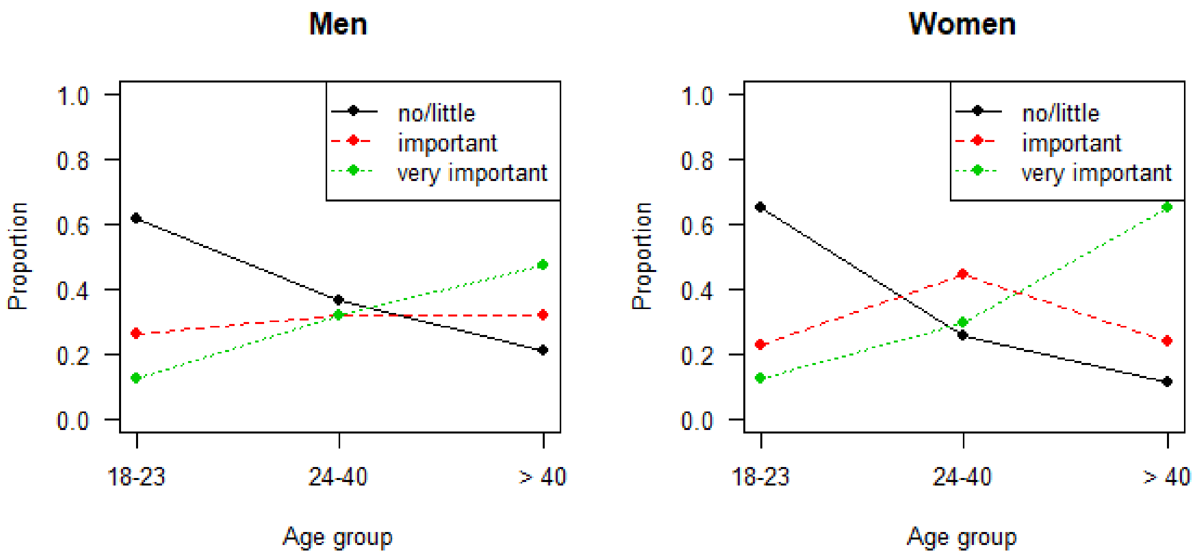

Figure 7.11 Preferences for air conditioning and power steering: percentages of responses in each category by age and sex.

Model Selection. From **Figure 7.11**, a model with the main effects is preferred. We fit three nominal logistic regression models:

- model 0: the model with intercept – response ~ 1
- model 1: the model with main effects – response ~ age + gender
- model 2: the model with main effects – response ~ age + gender + age:gender

Note that the model 2 is also the maximum model here. It has 6 parameters (a constant and coefficients for sex, two age categories and two age by sex interactions) for "important vs no/little important" and 6 parameters "very important vs no/little important", giving a total of 12 parameters. The model 0 has 2 parameters and model 1 has 8 parameters. The model 0 can be considered as the minimum model. We can use the likelihood ratio test to compare these three nested models. The likelihood test statistic for the model 1 and model 2 is $2 * (-288.3 - (-290.35)) = 3.94$. The degrees of freedom associated with this deviance are 12 − 8 = 4 because the maximal model has 12 parameters and the fitted model has 8 parameters. The corresponding p-value is 0.414. The likelihood test statistic, the associated degrees of freedom, and the p-value for the model 0 and the model 1 are 77.84, 6, and ≈ 0. Three conclusions can be obtained: (1) The model 1 is a significant improvement over the model 1, showing the overall importance of the explanatory variables; (2)

The model 1 and model 2 are not significantly different. (3) Since the model 2 is the maximum model, thus the likelihood ratio test between is also a goodness-of-fit test, indicating that the model 1 provides a good description of the data.

Model diagnostic. Such analysis should be still based on the residuals. Unfortunately, the residuals provided from R are not useful, so we skip this part of analysis.

Conclusions. We first look at the estimated coefficients and odds ratios (**Table 7.4**).

Table 7.4 Results of fitting the nominal logistic regression model (8.11) to the car preference data.

Parameter $\boldsymbol{\beta}$	**Estimate** $\widehat{\boldsymbol{\beta}}$	**Standard Error**	**Parameter 95% CI**	**Odds ratio** $(\exp(\widehat{\boldsymbol{\beta}}))$	**Odds ratio 95% CI**
			important vs. no/little important		
Intercept	-0.5908	0.2840			
Men	-0.3881	0.3005	$(-0.9771, 0.2009)$	0.6783	$(0.3753, 1.2225)$
$\mathbf{24-40}$	1.1283	0.3416	$(0.4587, 1.7979)$	3.0903	$(1.5819, 6.0368)$
$\mathbf{>40}$	1.5877	0.4029	$(0.7980, 2.3773)$	4.8925	$(2.2211, 10.7766)$
			very important vs. no/little important		
Intercept	-1.0391	0.3305			
Men	-0.8130	0.3210	$(-1.4422, -0.1837)$	0.4435	$(0.2364, 0.8321)$
$\mathbf{24-40}$	1.4781	0.4009	$(0.6923, 2.2639)$	4.3846	$(1.9983, 9.6206)$
$\mathbf{>40}$	2.9168	0.4229	$(2.0878, 3.7457)$	18.4813	$(8.0674, 42.3378)$

From the Wald statistics and the odds ratios and the confidence intervals, it is clear that the importance of air-conditioning and power steering increased significantly with age. Also men considered these features less important than women did, although the statistical significance of this finding is dubious (especially considering the small frequencies in some cells).

Probabilities can be estimated too. For example, consider the preference of women aged $24-40$ age, we have

$$\hat{\eta}_2 = \log\left(\frac{\hat{\mu}_2}{\hat{\mu}_1}\right) = -0.591 + 1.128 = 0.537$$

$$\hat{\eta}_3 = \log\left(\frac{\hat{\mu}_3}{\hat{\mu}_1}\right) = -1.039 + 1.478 = 0.439$$

According to the formula derived before, we have:

$$\hat{\mu}_1 = \frac{1}{1+\exp(\hat{\eta}_2)+\cdots+\exp(\hat{\eta}_K)} = \frac{1}{1+\exp(0.537)+\cdots+\exp(0.439)} = 0.234$$

$$\hat{\mu}_2 = \hat{\mu}_1 * \exp(\hat{\eta}_2) = 0.234 * \exp(0.537) = 0.402$$

$$\hat{\mu}_2 = \hat{\mu}_1 * \exp(\hat{\eta}_3) = 0.234 * \exp(0.439) = 0.364$$

These estimated probabilities can be multiplied by the total frequency for each sex by age group to obtain the "expected" frequencies or fitted values. The residuals and log-likelihood can further be calculated (**Table 7.5**).

Table 7.5 Results of fitting the nominal logistic regression model to the car preference data.

Sex	Age	Response	Observed y_i	Probability $\hat{\mu}$	Fitted value	Log-likelihood $y_i \log(\hat{\mu})$
Women	$18-23$	no/little	26	0.5242	23.59	-16.7929
		important	12	0.2903	13.07	-14.8402
		very	7	0.1865	8.35	-11.7947
	$24-40$	no/little	9	0.2346	10.56	-13.0494
		important	21	0.4015	18.07	-19.1620
		very	15	0.3639	16.37	-15.1637
	>40	no/little	5	0.0975	5.85	-11.6355
		important	14	0.2644	15.87	-18.6228
		very	41	0.6380	38.28	-18.4263
Men	$18-23$	no/little	40	0.6525	42.41	-17.0792
		important	17	0.2451	15.93	-23.9005
		very	8	0.1024	6.65	-18.2326
	$24-40$	no/little	17	0.3510	15.44	-17.7988
		important	15	0.4075	17.93	-13.4647
		very	12	0.2415	10.63	-17.0517
	>40	no/little	8	0.1743	7.15	-13.9870
		important	15	0.3204	13.13	-17.0751
		very	18	0.5054	20.72	-12.2842
Total			300		300	-290.35

Exercise 7.4

Based on the results from **Table 7.4**, verify three rows corresponding to "men" and "> 40" of **Table 7.5**.

Section 7.2.4 Ordinal Logistic Regression

Introduction

If there is an obvious natural order among the response categories, then this can be taken into account in the model specification. The example on car preferences provides an illustration as the study participants rated the importance of air conditioning and power steering in four categories from “not important” to “very important.” Ordinal responses like this are common in areas such as market research, opinion polls and fields such as psychiatry where “soft” measures are common.

Consider a random variable Y with K ordered categories. Let $\mu_1, \cdots, \mu_K$ denote the respective probabilities (or expected proportion), with $\mu_1 + \cdots + \mu_K = 1$. If there are m independent observations of Y which result in y_1 outcomes in category 1, y_2 outcomes in category 2, and so on, then $y_1 + \cdots + y_K = m$ and the distribution of $y_1, \cdots, y_K$ is still a multinomial distribution. However, we need to consider the order when we model $\mu_1, \cdots, \mu_K$.

In some situations there may, conceptually, be a continuous variable U, such as severity of disease, which is difficult to measure. It is assessed by some crude method that amounts to identifying “cut points,” C_k, for the latent variable so that, for example, patients with small values are classified as having “no disease,” those with larger values of U are classified as having “mild disease” or “moderate disease” and those with high values are classified as having “severe disease”. The cutpoints $C_1, \cdots, C_{K-1}$ define K ordinal categories with associated probabilities $\mu_1, \cdots, \mu_K$ (with $\mu_1 + \cdots + \mu_K = 1$).

Not all ordinal variables can be thought of in this way because the underlying process may have many components, as in the car preference example. Nevertheless, the idea is helpful for interpreting the results from statistical models. For ordinal categories, there are several different commonly used models including cumulative logit model, proportional odds model, adjacent categories logit model, and continuation ratio logit model. The proportional odds model is most commonly used model and it is our focus in the course.

Cumulative Logit Model and Proportional Odds Model

For the nominal logistic regression, the linear predictor is the logarithm of odds for kth category versus the reference category:

$$\text{logit}(\mu_k) = \log\left(\frac{\mu_k}{\mu_1}\right) = \eta_k, k = 2, \cdots, K.$$

For the **cumulative logit model**, the linear predictor is the logarithm of cumulative odds:

$$\text{logit}\big(P(Y \leq k)\big) = \log\left(\frac{P(Y \leq k)}{P(Y > k)}\right) = \log\left(\frac{\mu_1 + \cdots + \mu_k}{\mu_{k+1} + \cdots + \mu_K}\right) = \eta_k, k = 1, \cdots, K - 1.$$

where

$$\frac{P(Y \leq k)}{P(Y > k)} = \frac{\mu_1 + \cdots + \mu_k}{\mu_{k+1} + \cdots + \mu_K}, k = 1, \cdots, K - 1.$$

is called the **cumulative odds** for the kth category.

The linear predictor η_k can be written as:

$$\eta_k = \beta_{0,k} + \beta_{1,k}x_1 + \cdots + \beta_{p,k}x_p$$

Similar as the nominal logistic regression, there are $(K - 1) * (p + 1)$ parameters in the model.

If the linear predictor in the cumulative logit model has an intercept term $\beta_{0,k}$ which depends on the category k, but the other explanatory variables do not depend on k, then the model is

$$\log\left(\frac{\mu_1 + \cdots + \mu_k}{\mu_{k+1} + \cdots + \mu_K}\right) = \eta_k = \beta_{0,k} + \beta_1 x_1 + \cdots + \beta_p x_p$$

This is called the **proportional odds model** (or **ordinal logistic regression** model). It is based on the assumption that the effects of the covariates are the same for all categories on the logarithmic scale. There are $K - 1 + p$ parameter in the proportional odds model. As for the nominal logistic regression model, the odds ratio associated with an increase of one unit in an explanatory variable x_j is $\exp(\beta_j)$, where $j = 1, \cdots, p$.

The proportional odds model is fitted using the R function `polr()` in R package **MASS**. Here, we need to empathize the parameterization used in `polr()`. The following formula is used `polr()`:

$$\log\left(\frac{\mu_1 + \cdots + \mu_k}{\mu_{k+1} + \cdots + \mu_K}\right) = \eta_k = \beta_{0,k} - \beta_1 x_1 - \cdots - \beta_p x_p$$

In other words, "$-\beta_i(i = 1,2, \cdots, p)$" instead of "$\beta_i(i = 1,2, \cdots, p)$" used in `polr()`. We skip the technical details of the model but use an example for the illustration.

Case Study of Proportional Odds Model: Car Preference

The response variable for the car preference data is, of course, ordinal. The proportional odds model is fitted and the results are shown in **Table 7.6**.

Table 7.6 Results of proportional odds ordinal regression model for the car preference data using `polr()` in R package **MASS**.

Parameter $\boldsymbol{\beta}$	**Estimate** $\widehat{\boldsymbol{\beta}}$	**Standard Error**	**Parameter 95% CI**	**Odds ratio** (**exp**$(\widehat{\boldsymbol{\beta}})$)	**Odds ratio 95% CI**
β_{01}	0.0435	0.2323			
β_{02}	1.6550	0.2556			
men	-0.5762	0.2262	$(-1.0214, -0.1338)$	0.5620	$(0.3601, 0.8747)$
$24-40$	1.1471	0.2776	$(0.6073, 1.6970)$	3.1490	$(1.8355, 5.4677)$
> 40	2.2325	0.2915	$(1.6702, 2.8142)$	9.3228	$(5.3132, 16.6798)$

The parameter estimates for the proportional odds model are all quite similar to those from the nominal logistic regression model. The log-likelihood for the nominal logistic regression and the proportional odds model are -290.35 and -290.65 which are almost identical. The estimated probabilities are also similar. For example, consider the preference of women aged $24-40$ age, we have

$$\hat{\eta}_1 = \log\left(\frac{\hat{\mu}_1}{\hat{\mu}_2 + \hat{\mu}_3}\right) = 0.044 - 1.147 = -1.103$$

$$\hat{\eta}_2 = \log\left(\frac{\hat{\mu}_1 + \hat{\mu}_2}{\hat{\mu}_3}\right) = 1.655 - 1.147 = 0.508$$

Note that -1.147 instead of 1.147 is used due to the parametrization implemented in R function `polr()`.

From the first equation, we have

$$\log\left(\frac{\hat{\mu}_1}{\hat{\mu}_2 + \hat{\mu}_3}\right) = \log\left(\frac{\hat{\mu}_1}{1 - \hat{\mu}_1}\right) = -1.103$$

So we can get

$$\hat{\mu}_1 = \frac{\exp(-1.103)}{1 + \exp(-1.103)} = 0.249$$

From the second equation, we have

$$\log\left(\frac{\hat{\mu}_1 + \hat{\mu}_2}{\hat{\mu}_3}\right) = \log\left(\frac{1 - \hat{\mu}_3}{\hat{\mu}_3}\right) = -\log\left(\frac{\hat{\mu}_3}{1 - \hat{\mu}_3}\right) = 0.508$$

So we can get

$$\hat{\mu}_3 = \frac{\exp(-0.508)}{1+\exp(-0.508)} = 0.376$$

With $\hat{\mu}_1 + \hat{\mu}_2 + \hat{\mu}_3 = 1$, we have

$$\hat{\mu}_1 = 0.249, \hat{\mu}_2 = 0.375, \hat{\mu}_3 = 0.376$$

These probabilities are similar with those from **Table 7.5**.

In summary, the proportional odds logistic model for ordinal data and the nominal logistic model produce similar results.

General Comments

Although the models described in this lesson are developed from the logistic regression model for binary data, other link functions such as the probit or complementary log-log functions can also be used. Logits and probits are appropriate if the distribution is symmetric but the complementary log-log link may be better if the distribution is very skewed.

If there is doubt about the order of the categories, then nominal logistic regression will usually be a more appropriate model than any of the models based on assumptions that the response categories are ordinal. Although the resulting model will have more parameters and hence fewer degrees of freedom and less statistical power, it may give results very similar to the ordinal models (as in the car preference example).

To determine if the proportional odds model or the cumulative logit model should be used, the likelihood ratio test and other test have been developed. However, such methods are not implemented in R. We will not discuss this topic in our course.

Table 7.1 Features of the Tweedie distributions for various values of the index parameter ξ, showing the support S (the permissible values of y) and the domain Ω for μ. The Poisson distribution ($\xi = 1$ and $\varphi = 1$) is a special case of the discrete distributions, and the inverse Gaussian distribution ($\xi = 3$) is a special case of positive stable distributions. R refers to the real line; superscript $+$ means positive real values only; subscript 0 means zero is included in the space.

Tweedie EDM	$\boldsymbol{\xi}$	$\boldsymbol{S}$	$\boldsymbol{\Omega}$	**Covered?**
Extreme Stable	$\xi < 0$	R	R^+	Not covered
Normal	$\xi = 0$	R	R	Yes. Week 1
No EDMs exist	$0 < \xi < 1$			
Discrete	$\xi = 1$	$y = 0, \varphi, 2\varphi, \cdots$	R^+	Yes. Week 6
Poisson-Gamma	$1 < \xi < 2$	R_0^+	R^+	Yes. Week 7
Gamma	$\xi = 2$	R^+	R^+	Yes. Week 5
Positive Stable	$\xi > 2$	R^+	R^+	Yes. Week 7

Table 7.2 The MLE of parameters in the Poisson and Gamma distributions.

		Estimates				
Parameter	**Parameter**	**Phase 1**	**Phase 2**	**Phase 3**	**Phase 4**	**Phase 5**
Poisson Mean	λ^*	0.0732	2.6210	0.6671	1.7251	1.3585
Gamma Mean	μ^*	1.5611	12.9318	5.7608	10.1000	8.7702
Gamma Dispersion	ϕ^*	0.5909	0.5909	0.5909	0.5909	0.5909

Table 7.3 Importance of air conditioning and power steering in cars (row percent- ages in brackets).

		Response			
Sex	**Age**	**No or Little Important**	**Important**	**Very Important**	**Total**
Women	18 − 23	26 (58%)	12 (27%)	7 (16%)	45
	24 − 40	9 (20%)	21 (47%)	15 (33%)	45
	> 40	5 (8%)	14 (23%)	41 (68%)	60
Men	18 − 23	40 (62%)	17 (26%)	8 (12%)	65
	24 − 40	17 (39%)	15 (34%)	12 (27%)	44
	> 40	8 (20%)	13 (37%)	18 (44%)	41
Total		105	94	101	300

Table 7.4 Results of fitting the nominal logistic regression model (8.11) to the car preference data.

Parameter $\boldsymbol{\beta}$	**Estimate** $\widehat{\boldsymbol{\beta}}$	**Standard Error**	**Parameter 95% CI**	**Odds ratio (exp**$(\widehat{\boldsymbol{\beta}})$**)**	**Odds ratio 95% CI**
important vs. no/little important					
Intercept	-0.5908	0.2840			
Men	-0.3881	0.3005	$(-0.9771, 0.2009)$	0.6783	$(0.3753, 1.2225)$
$\mathbf{24-40}$	1.1283	0.3416	$(0.4587, 1.7979)$	3.0903	$(1.5819, 6.0368)$
$> \mathbf{40}$	1.5877	0.4029	$(0.7980, 2.3773)$	4.8925	$(2.2211, 10.7766)$
very important vs. no/little important					
Intercept	-1.0391	0.3305			
Men	-0.8130	0.3210	$(-1.4422, -0.1837)$	0.4435	$(0.2364, 0.8321)$
$\mathbf{24-40}$	1.4781	0.4009	$(0.6923, 2.2639)$	4.3846	$(1.9983, 9.6206)$
$> \mathbf{40}$	2.9168	0.4229	$(2.0878, 3.7457)$	18.4813	$(8.0674, 42.3378)$

Table 7.5 Results of fitting the nominal logistic regression model to the car preference data.

Sex	Age	Response	Observed	Probability	Fitted	Log-likelihood
			y_i	$\hat{\mu}$	value	$y_i \log(\hat{\mu})$
Women	$18-23$	no/little	26	0.5242	23.59	−16.7929
		important	12	0.2903	13.07	−14.8402
		very	7	0.1865	8.35	−11.7947
	$24-40$	no/little	9	0.2346	10.56	−13.0494
		important	21	0.4015	18.07	−19.1620
		very	15	0.3639	16.37	−15.1637
	>40	no/little	5	0.0975	5.85	−11.6355
		important	14	0.2644	15.87	−18.6228
		very	41	0.6380	38.28	−18.4263
Men	$18-23$	no/little	40	0.6525	42.41	−17.0792
		important	17	0.2451	15.93	−23.9005
		very	8	0.1024	6.65	−18.2326
	$24-40$	no/little	17	0.3510	15.44	−17.7988
		important	15	0.4075	17.93	−13.4647
		very	12	0.2415	10.63	−17.0517
	>40	no/little	8	0.1743	7.15	−13.9870
		important	15	0.3204	13.13	−17.0751
		very	18	0.5054	20.72	−12.2842
Total			300		300	−290.35

Table 7.6 Results of proportional odds ordinal regression model for the car preference data using `polr()` in R package **MASS**.

Parameter $\boldsymbol{\beta}$	**Estimate** $\widehat{\boldsymbol{\beta}}$	**Standard Error**	**Parameter 95% CI**	**Odds ratio (exp**$(\widehat{\boldsymbol{\beta}})$**)**	**Odds ratio 95% CI**
β_{01}	0.0435	0.2323			
β_{02}	1.6550	0.2556			
men	-0.5762	0.2262	$(-1.0214, -0.1338)$	0.5620	$(0.3601, 0.8747)$
$24 - 40$	1.1471	0.2776	$(0.6073, 1.6970)$	3.1490	$(1.8355, 5.4677)$
> 40	2.2325	0.2915	$(1.6702, 2.8142)$	9.3228	$(5.3132, 16.6798)$