

Carolyn J. Anderson  
Jay Verkuilen  
Timothy Johnson

# Applied Generalized Linear Mixed Models: Continuous and Discrete Data

For the Social and Behavioral Sciences

February 10, 2010

Springer





# Contents

|          |  |    |
|----------|--|----|
| <b>1</b> | <b>Introduction</b> .....                        | 1  |
| 1.1      | Clustered Data .....                             | 1  |
| 1.2      | Modeling of Data .....                           | 1  |
| 1.3      | Our Approach .....                               | 1  |
| 1.4      | Other .....                                      | 2  |
| <b>2</b> | <b>Generalized Linear Models</b> .....           | 3  |
| 2.1      | Introduction .....                               | 3  |
| 2.2      | The Three Components of a GLM .....              | 5  |
| 2.2.1    | The Random Component .....                       | 5  |
| 2.2.2    | The Systematic Component .....                   | 15 |
| 2.2.3    | The Link Function .....                          | 16 |
| 2.3      | Examples of GLMs .....                           | 19 |
| 2.3.1    | A Normal Continuous Variable .....               | 20 |
| 2.3.2    | A Skewed Continuous Response Variable .....      | 21 |
| 2.3.3    | A Dichotomous Response Variable .....            | 23 |
| 2.3.4    | A Count Response Variable .....                  | 28 |
| 2.4      | Estimation .....                                 | 31 |
| 2.5      | Assessing Model Goodness-of-Fit to Data .....    | 34 |
| 2.5.1    | Global Measures of Fit .....                     | 35 |
| 2.5.2    | Comparing Models .....                           | 36 |
| 2.5.3    | Local Measures of Fit .....                      | 40 |
| 2.6      | Statistical Inference for Model Parameters ..... | 40 |
| 2.6.1    | Hypothesis Testing .....                         | 40 |
| 2.6.2    | Confidence Intervals .....                       | 45 |
| 2.7      | Summary .....                                    | 47 |
|          | Problems & Exercises .....                       | 48 |
| <b>3</b> | <b>Generalized Linear Mixed Models</b> .....     | 51 |
| 3.1      | Introduction .....                               | 51 |
| 3.2      | Normal Random Variables .....                    | 53 |

|          |   |           |
|----------|---|-----------|
| 3.2.1    | Paired Dependent $t$ -test  | 54        |
| 3.2.2    | Paired $t$ -test as a Random Intercept Model                      | 55        |
| 3.2.3    | Basic Properties of a Random Intercept Model                      | 58        |
| 3.3      | Linear Regression Models with Random Coefficients                 | 63        |
| 3.3.1    | Modeling the Intercept  | 64        |
| 3.3.2    | Modeling the Slope  | 65        |
| 3.4      | Generalized Linear Mixed Models                                   | 72        |
| 3.4.1    | GLMM Formulation of Anorexia Example                              | 72        |
| 3.4.2    | Adding More Predictors: The Normal Case                           | 73        |
| 3.4.3    | Generalized Linear Mixed Models                                   | 75        |
| 3.5      | GLMM for a “Cool” Dichotomous Response Variable                   | 75        |
| 3.5.1    | Fixed versus Random Intercept Models                              | 75        |
| 3.5.2    | Adding Level 2 Predictors   | 77        |
| 3.5.3    | Heterogeneous Variance  | 78        |
| 3.5.4    | The Catch-22 of Multilevel Modeling                               | 79        |
| 3.6      | Cluster-specific, Population Average and Marginal Models          | 80        |
| 3.7      | Estimation  | 82        |
| 3.7.1    | The Normal Case   | 82        |
| 3.7.2    | The Rest  | 83        |
| 3.8      | Summary   | 83        |
|          | Problems & Exercises  | 83        |
| <b>A</b> | <b>The Natural Exponential Dispersion Family of Distributions</b> | <b>85</b> |
| A.0.1    | Likelihood, Score & Information                                   | 88        |
| A.0.2    | Estimation  | 89        |
|          | <b>Index of Data Sets</b>   | <b>91</b> |
|          | <b>References</b>   | <b>93</b> |



## Chapter 2

# Generalized Linear Models

### 2.1 Introduction

Multiple regression and ANOVA dominated statistical analysis of data in the social and behavioral sciences for many years. The recognition that multiple regression and ANOVA are special cases of a more general model, the general linear model, was known for many years by statisticians, but it was not common knowledge to social science researchers until much later<sup>1</sup>. The recognition of the connection between multiple regression and ANOVA by social scientists provided “possibilities for more relevant and therefore more powerful exploitation of research data” [p 426, Cohen, 1968]. As such, the general linear model was a large step forward in the development of regression models.

In the general linear model framework, response variables are assumed to be normally distributed, have constant variance over the values of the predictor variables, and equal linear functions of predictor or explanatory variables. Transformations of data were developed as ways to force data into a normal linear regression model; however, this is no longer necessary nor optimal. Generalized linear models (GLM) go beyond the general linear model by allowing for non-normally distributed response variables, heteroscedasticity, and non-linear relationships between the mean of the response variable and the predictor or explanatory variables.

First introduced by Nelder & Wedderburn (1972), GLMs provide a unifying framework that encompasses many seemingly disparate models. Special cases of GLMs include not only linear regression and ANOVA, but also logistic regression, probit models, Poisson regression, log-linear models, and many more. An additional advantage of the GLM framework is that there is a common computational method for fitting the models to data. The implementation of this method in software programs opened up the ability of researchers to design models to fit their data and to

---

<sup>1</sup> Fisher (1928) was one of the first (if not the first) to realized the connection between multiple regression and ANOVA (see also Fisher (1934)). The relationship was fully described in paper by Wishart (1934). The general linear model representation of ANOVA can also be found in Scheffe (1959)’s text on ANOVA. A classical reference in the social sciences is Cohen (1968).

fit a wide variety of models, including those not previously proposed in the literature. Many software packages are now available for fitting GLMs to data, including SAS (SAS Institute, 2003), S-Plus (Insightful Corporation, 2007), R (R Core Team, 2006), Stata (–reference–) and others.

In the GLM framework, models are constructed to fit the type of data and problem at hand. Three major decisions must be made. The first is the *random component* that consists of choosing a probability distribution for the response variable. The distribution can be any member from the *natural exponential dispersion family* distributions or the *exponential family*, for short<sup>2</sup> Special cases of this family of distributions include the normal, binomial, Poisson, gamma, and others. The second component of a GLM is the *systematic component* or *linear predictor* that consists of a linear combination of predictor or explanatory variables. Lastly, a *link function* must be chosen that maps the mean of the response variable onto the linear predictor.

As an example, consider research on cognition and aging by Stine-Morrow, Miller, Gagne & Hertzog (2008). In their research, they measure the time it takes an elderly individual to read words presented on a computer screen. Reaction times are non-negative continuous variables that tend to have positively skewed distributions. A common strategy is to use normal linear regression by trying to find a transformation of reaction times such that they are normally distributed with equal variances and linearly related to the predictor variables. Rather than using a normal distribution, a positively skewed distribution with values that are positive real numbers can be selected. The systematic component of the model can potentially equal any real number, but the link function can be chosen to ensure that the predicted means are in the permissible range (i.e., non-negative real numbers). In regression, we model means conditional on explanatory or predictor variables. The link is not applied to the data, but to the expected value or mean of the response. As a result, the choice of a distribution for the responses is based on the nature of the response variable without regard to what transformation (of means) is chosen.

Although GLMs do not take into account clustering or nesting of observations into larger units (e.g., repeated measures on an individual, students within peer groups, children within families), GLMs are an ideal starting point for our modeling approach. GLMs include models for response variables that are continuous or metric variables and those that are discrete. In the subsequent chapters, the GLM approach is extended to include random effects as a way to deal with dependency between observations created by grouping, clustering or nesting of observations into larger units.

In Section 2.2, the three components of a GLM are discussed in detail. In Sections 2.3, examples for continuous variables and discrete variables are presented to illustrate how GLMs are formed, as well as introduce some of the data sets that will be re-analyzed later in the book. In Section 2.4, a general overview of estimation is given that also includes problems and solutions sometimes encountered when fit-

---

<sup>2</sup> The *natural exponential dispersion* family and the *natural exponential family* of distributions are often used interchangeably. The former include distributions characterized by a single parameters (i.e., location or mean); whereas the latter is a more general and includes distributions with one or two parameters (i.e., mean and dispersion).

ting GLMs to data. In Sections 2.5 and 2.6, the statistical issues of assessing model goodness-of-fit to data and statistical inference of model parameters (i.e., hypothesis testing, confidence intervals), respectively, are presented and illustrated. In the last section, Section ??, technical details are covered. For more detailed descriptions of GLMs than given here see McCullagh & Nelder (1989), Fahrmeir & Tutz (2001), Dobson (1990), and Lindsey (1997), and specifically for categorical data see Agresti (2002, 2007).

## 2.2 The Three Components of a GLM

Model building using GLMs starts with initial decisions for the distribution of the outcome variable, the predictor or explanatory variables to include in the systematic component, and how to connect the mean of the response to the systematic component. Each of these three decisions is described in more detail in the following three sections.

### 2.2.1 *The Random Component*

A reasonable distribution for the response variable must be chosen. For example, in work by Espelage, Holt & Henkel (2003) on the effects of aggression during early adolescence, one way to measure the extent to which a child is aggressive (i.e., a bully) is a student's score on the self-report Illinois Bully Scale (Espelage & Holt 2001). The bully scores are typically thought of as a "continuous" or metric measure, because scores equal the mean of nine items each of which is scored from 0 to 4 (i.e., there are 37 possible scores). The distribution selected for the Illinois Bully Scale should be one that is appropriate for a continuous variable. An alternative way to measure "bullyness" is the number of students who say that a child is a bully (i.e., peer nominations). The peer nomination measure is a count and a distribution for a discrete integer variable should be selected.

The distribution selected is not the "true" distribution in the population, but is an approximation that should be a good representation of the probably distribution of the response variable. A good representation of the population distribution of a response variable should not only take into account the nature of the response variable (e.g., continuous, discrete) and the shape of the distribution, but it should also provide a good model for the relationship between the mean and variance. For example, a normal distribution might seem like a sensible distribution for the bully-scale; however, the bully scale is bounded with minimum equal to 0 and maximum equal to 4. The possible values of mean and variance depend on the bounds, as well as the shape of the distribution. For example, the mean of the bully scale can equal any value from 0 to 4. If values of the bully scale are uniformly distributed between 0 and 4, the mean would equal 2 and the variance of (a discrete random)

variable would equal 1.41. The value of the variance of a bounded scale depends on the shape of the distribution and the end points of the scale. In Chapter ??, the beta is a distribution for continuous variables with lower and upper bounds is presented along with beta-regression models.

Within the GLM framework, the distribution for a response variable can be any member of the natural exponential dispersion family. Members of the natural exponential family for continuous response variables include the normal, gamma, and inverse Gaussian distribution, and for discrete outcome variables the Poisson, Bernoulli, and binomial distributions. Other useful distribution some of which are in the exponential family and others that are not will be introduced later in the text. The distributions introduced in this chapter provide a representation of the distribution for many common measures found in psychological and social studies research.

A natural exponential dispersion distribution has two parameters, a *natural parameter*  $\theta$  and a *dispersion parameter*  $\phi$ . The parameter  $\theta$  conveys information about the location of the distribution (i.e., mean). When the distribution is expressed in its most basic or canonical form, the natural parameter  $\theta$  is a function of the mean  $\mu$  of the distribution. This function is known as the *canonical link*. Link functions are discussed in more detail in Section 2.2.3.

Variances of particular distributions in the exponential family equal a function of  $\mu$  and  $\phi$ . In GLMs, non-constant variance or heteroscedasticity is expected. The only exception is the normal distribution where the mean and variance are independent of each other and the variance equals the dispersion parameter (i.e.,  $\sigma^2 = \phi$ ). For distributions that have  $\phi = 1$ , the variance is solely a function of the mean (e.g., Poisson and Bernoulli distributions). In GLMs, the  $\phi$  parameter is often regarded as a nuisance parameter and attention is focused mostly on the mean. This will not be the case later in this book.

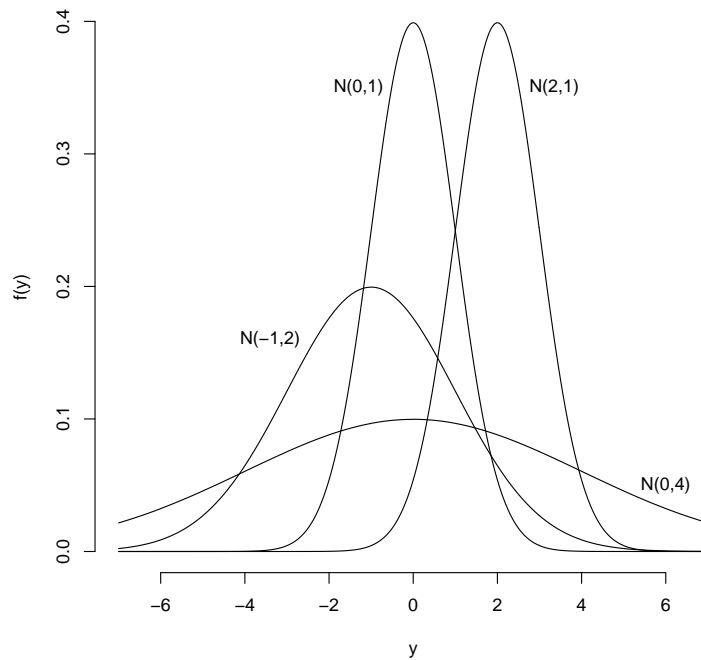
Below we review the basic characteristics of the normal, gamma, and inverse Gaussian distributions for continuous variables, and the Bernoulli, binomial and Poisson distributions for discrete variables. More technical details regarding the natural exponential distribution are given in Appendix ??. Other distributions are introduced later in the text as they are needed.

### 2.2.1.1 Normal Distribution

The most well known and familiar distribution for continuous random variables is the normal distribution. A normal distribution is characterized by its mean  $\mu$  and variance  $\sigma^2$ . The probability density function for the normal distribution is

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \quad \text{for } -\infty < y < \infty. \quad (2.1)$$

A particular normal distributions is represented by  $N(\mu, \sigma^2)$ . Typically, the parameter of most interest is the mean  $\mu$ . In GLM terminology,  $\theta = \mu$  is the natural parameter of the normal distribution and  $\phi = \sigma^2$  is the dispersion parameter.



**Fig. 2.1** Four examples of normal distributions  $N(\mu, \sigma^2)$  with different combinations of means and variances.

Examples of four normal distributions are given in Figure 2.1. Normal distributions are uni-modal and symmetric around the mean  $\mu$ . Note that two distributions with different means but the same variance (e.g.,  $N(0, 1)$  and  $N(2, 1)$ ) have the same shape and only differ in terms of their location. Alternatively, two distributions with the same mean but different variances have the same location but differ in terms of dispersion or spread of values around the mean (e.g.,  $N(0, 1)$  and  $N(0, 4)$ ). The mean or variance can be altered without effecting the other; that is, the mean and variance of a normal distributions are independent of each other.

Although measured variables are never truly continuous, a normal distribution is often a good representation or approximation of the distribution for many response variables, in part due to the Central Limit Theorem. Theoretical variables such as the random effects introduced in the next chapter are most commonly assumed to be normally distributed. The normal distribution also plays an important role as the sampling the distribution of parameter estimates (e.g., regression coefficients) and many test statistics.

### 2.2.1.2 Gamma Distribution

When the distribution of a response variable is not symmetric and values of the response variable are positive (e.g., reaction times), a normal distribution would be a poor representation of the distribution. An alternative distribution for skewed, non-negative responses is the gamma distribution.

A gamma distribution is often presented in terms of two parameters, a shape parameter and a scale parameter. Since our emphasis is on regression models, we will present the gamma distribution parameterized in terms of its mean  $\mu$  and dispersion parameter  $\phi$ . A particular gamma distribution will be represented as  $\text{Gamma}(\mu, \phi)$ . The probability density function for a gamma distribution is

$$f(y; \mu, \phi) = \frac{1}{\Gamma(\phi^{-1})} \left( \frac{1}{\mu\phi} \right)^{1/\phi} y^{1/\phi-1} \exp(-y/(\mu\phi)) \quad \text{for } y > 0, \quad (2.2)$$

where  $\Gamma$  is a *gamma function*. A gamma function can be thought of as a factorial function (i.e.,  $y! = y(y-1)(y-2)\dots 1$ ), except that it is for real numbers rather than integers.

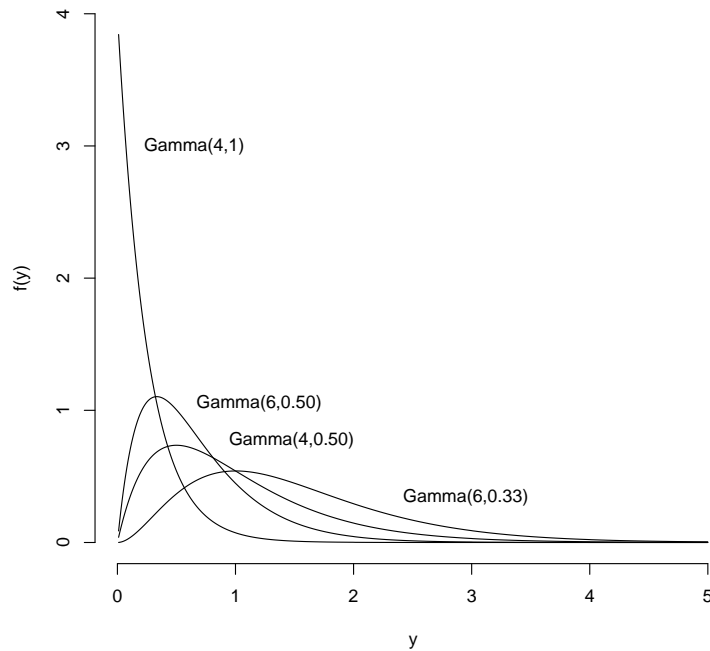
The natural parameter of the gamma distribution is  $\theta = 1/\mu$ . The parameters  $\mu$  and  $\phi$  can be any positive real number. To see the effect that  $\mu$  and  $\phi$  have on the shape of the distribution, four examples of Gamma distributions are given in Figure 2.2. The gamma distributions are positively skewed. For a given  $\mu$ , as  $\phi$  gets smaller, the distribution becomes less skewed (e.g.,  $\text{Gamma}(4, 1)$  and  $\text{Gamma}(4, 0.50)$ ). For a given  $\phi$ , as  $\mu$  gets larger, the distribution becomes less skewed (e.g.,  $\text{Gamma}(4, 0.50)$  and  $\text{Gamma}(6, 0.50)$ ).

Unlike the normal distribution where the mean and variance are independent of each other, for a gamma distribution the variance is a function of the mean and the dispersion parameter. The variance is a quadratic function of the mean; namely,

$$\sigma^2 = \mu^2 \phi.$$

To further illustrate this relationship, in Figure 2.3 the variance is plotted as a function of the mean for gamma distributions where  $\phi$  equals 2, 1, 0.50, 0.33 and 0.25. When responses are skewed, using a gamma distribution in a regression context not only implies heteroscedasticity, but it implies a specific relationship between the mean and variance should exist.

Special cases of the gamma distribution correspond to other well known skewed distributions for continuous random variables. When  $\phi = 1$  (e.g.,  $\text{Gamma}(4, 1)$  in Figure 2.2), the distribution is the exponential distribution with a rate parameter equal to  $1/\mu$ . The exponential distribution, a special case of the natural exponential family of distributions, is often used for rates of decay or decline. Another special case of the gamma distribution is the chi-squared distribution. A gamma distribution with  $\mu = \nu$  and  $\phi = 2/\nu$  is a chi-squared distribution where  $\nu$  equals the degrees of freedom of the chi-square distribution. For example, in Figure 2.2,  $\text{Gamma}(4, 0.50)$  is a chi-square distribution with  $\nu = 4$  and  $\text{Gamma}(6, 0.33)$  is chi-square with  $\nu =$



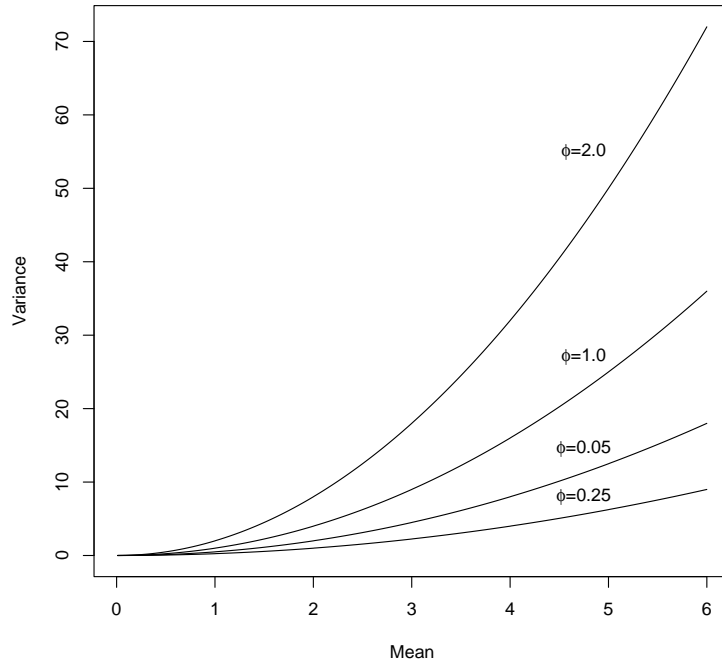
**Fig. 2.2** Examples of four Gamma distributions  $\text{Gamma}(\mu, \phi)$  with different combinations of mean and dispersion parameters.

6. Like the normal distribution, the chi-square distribution is also important as a sampling the distribution of many test statistics.

### 2.2.1.3 Inverse Gaussian Distribution

The inverse Gaussian distribution is probably the least familiar distribution to social scientists. Similar to the gamma distribution, the inverse Gaussian is a skewed distribution for non-negative continuous random variables. Although the normal (Gaussian) and inverse Gaussian distributions share some of the same properties, the name “inverse Gaussian” is a bit miss-leading in that this distribution is not derived from a normal distribution<sup>3</sup>

<sup>3</sup> The distribution was first derived to describe Brownian motion with positive drift (Chhikara & Folks 1988, Seshadri 1998). Brownian motion is basically the movement of particles over time where there is a tendency for particles to move more in one direction than another. The term “inverse” comes from the fact that the cumulate-generating function<sup>4</sup> for the time to cover a unit



**Fig. 2.3** The relationship between the mean and variance of gamma distributions where  $\phi$  ranges from 2 to 0.25.

The inverse Gaussian distribution has a number of different parameterizations (Chhikara & Folks 1988, Seshadri 1998). As before, we parameterize the distribution in terms of its mean  $\mu$  and dispersion parameter  $\phi$ . Both  $\mu$  and  $\phi$  are positive real numbers. The probability density for the inverse Gaussian is

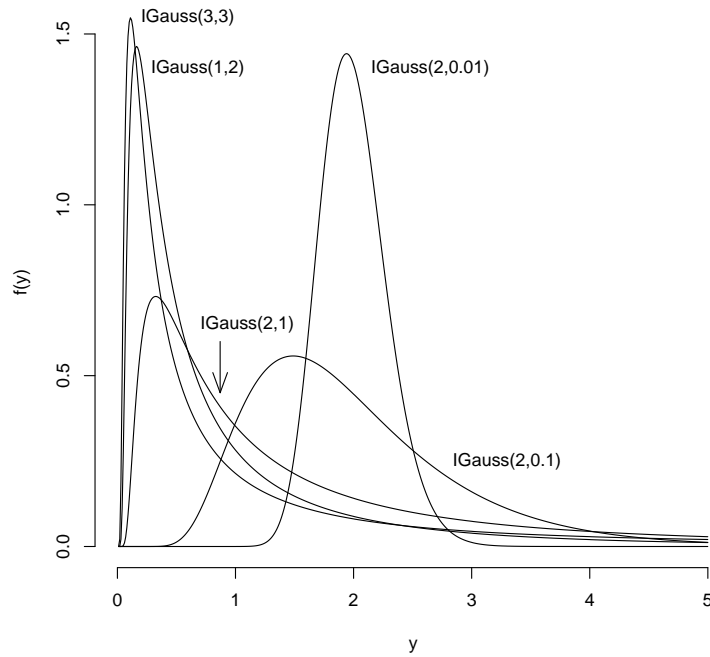
$$f(y) = \frac{1}{\sqrt{2\pi y^3 \phi}} \exp \left[ -\frac{(y-\mu)^2}{2\mu^2 \phi y} \right] \quad \text{for } y > 0. \quad (2.3)$$

We will represent a particular inverse Gaussian distribution as  $\text{IGauss}(\mu, \phi)$ .

Examples of inverse Gaussian distributions are given for different values of  $\mu$  and  $\phi$  in Figure 2.4. The dispersion parameter essentially controls the shape of the distribution. For example, compare the  $\text{IGauss}(2, 1.0)$ ,  $\text{IGauss}(2, 0.1)$  and  $\text{IGauss}(2, 0.001)$  that are given in Figure 2.4. As the dispersion parameter decreases, the inverse Gaussian distribution becomes more symmetric.

---

of distance is inversely related to the function of the distance covered in a unit of time (Chhikara & Folks 1988).



**Fig. 2.4** Examples of inverse Gaussian distributions  $\text{IGauss}(\mu, \phi)$  with different mean and dispersion parameters.

Some authors use the symbol  $\sigma^2$  rather than  $\phi$  to represent the dispersion parameter. We use  $\phi$  here because  $\sigma^2$  is typically used to represent variance; however, the variance of the inverse Gaussian distribution is neither  $\sigma^2$  nor  $\phi$ . The variance of the distribution equals

$$\text{var}(y) = \mu^3 \phi. \quad (2.4)$$

This is similar to the variance function of the Gamma distribution except that the variance for the inverse Gaussian increases more sharply as the mean increases.

#### 2.2.1.4 Bernoulli Distribution

Many response variables are clearly discrete, such as correct or incorrect, agree or disagree, true or false, sick or well, and fights or does not fight. The Bernoulli and binomial distributions apply to cases where the response variable can take one of

two possible values (i.e. a dichotomous response). Since the binomial distribution depends on the Bernoulli distribution, we start with the Bernoulli.

Let  $\underline{y}^*$  equal a Bernoulli random variable where

$$\underline{y}^* = \begin{cases} 1 & \text{if an observation is in category one} \\ 0 & \text{if an observation is in category two} \end{cases} \quad (2.5)$$

The parameter of the Bernoulli distribution is the probability  $\pi$  that an observation is in category one. The probability function for  $\underline{y}^*$  is

$$P(\underline{y}^* = y; \pi) = P(y; \pi) = \pi^y(1 - \pi)^{1-y} \quad \text{for } y = 0, 1. \quad (2.6)$$

The mean is  $\pi$ , the dispersion parameter for the Bernoulli is  $\phi = 1$ , and the variance is solely a function of the mean; specifically,

$$\text{var}(\underline{y}^*) = \pi(1 - \pi) = \mu(1 - \mu). \quad (2.7)$$

In Figure 2.5, the curve showing this relationship is labelled  $n = 1$ . The variance reaches a maximum when  $\pi = 0.5$ , the point of maximum uncertainty. This general shape of the variance function is often found for distributions for bounded scales.

### 2.2.1.5 Binomial Distribution

Sums of  $n$  independent observations from a Bernoulli distribution have a binomial distribution; that is,

$$\underline{y} = \sum_{i=1}^n \underline{y}_i^*$$

is a binomial random variable. The parameter of the binomial distribution is the probability  $\pi$  and a specific case of the binomial distribution will be represented as Binomial( $\pi, n$ ). When using the binomial distribution, interest is focused on estimating and modeling the probability  $\pi$ . The number of observations or “trials” is a known quantity. Binomial random variables can equal integer values from 0 to  $n$ . The probability that a binomial random variable equals  $y$  is

$$P(\underline{y} = y; \pi, n) = P(y; \pi, n) = \binom{n}{y} \pi^y (1 - \pi)^{(n-y)} \quad \text{for } y = 0, 1, \dots, n. \quad (2.8)$$

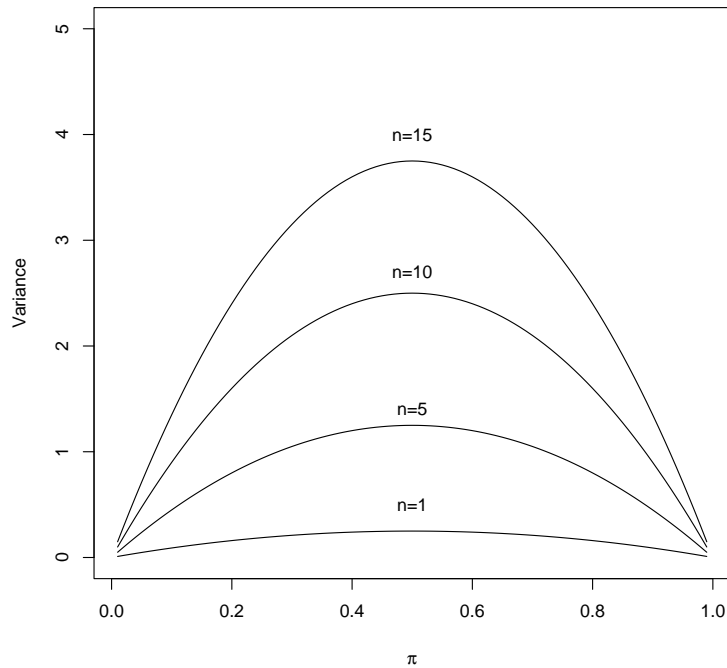
The binomial coefficient

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}$$

equals the number of ways to obtain the value of  $y$  from  $n$  trials. For  $n = 1$ , the Bernoulli distribution is the same as the binomial.

The mean and variance of a Binomial random variable equal

$$E(\underline{y}) = \mu = n\pi \quad \text{and} \quad \text{var}(\underline{y}) = n\pi(1 - \pi). \quad (2.9)$$



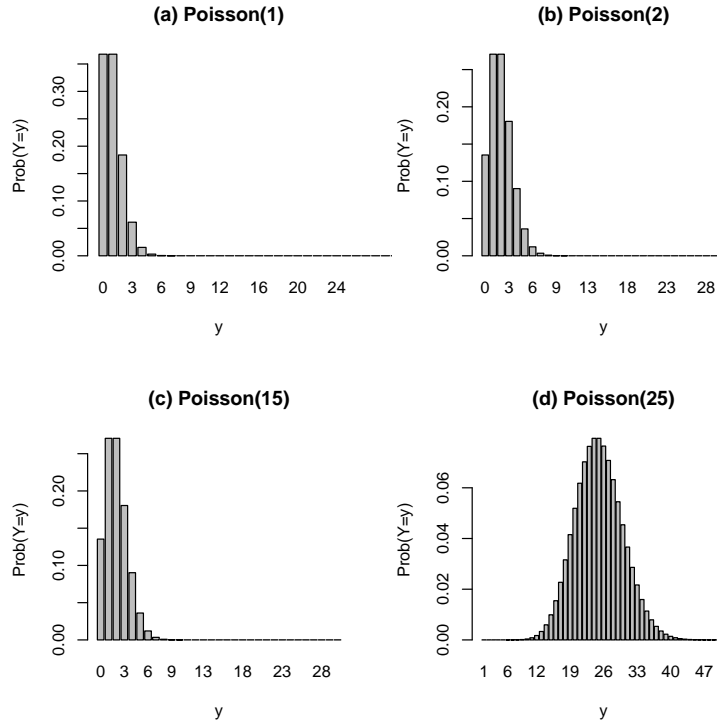
**Fig. 2.5** Examples of the variance function for the binomial distribution with  $n = 1, 5, 10, 15$ .

For the binomial distribution the dispersion parameter is  $\phi = 1/n$ . The variance function for the binomial distribution for different values of  $n$  are plotted in Figure 2.5. Regardless of  $n$ , the largest variance (i.e., point of maximum uncertainty) occurs when  $\pi = .5$ .

Not all discrete response variables have only two possible categories. In Chapter ??, the binomial distribution will be extended to the multinomial distribution for situations where there are two or more categories.

### 2.2.1.6 Poisson Distribution

Discrete variables can also be unbounded counts; that is, non-negative integers that do not necessarily have a maximum value. For example, in the research by Espelage et al. (2008), one way to measure the extent to which a child is a bully is by peer nominations. In this study, students in the school could nominate anyone in the school as a bully so that the number of bully nominations received by any one student are strictly speaking bounded by the number of students in the school.



**Fig. 2.6** Examples of Poisson distributions with different means.

However, since no student received bully nominations close to the maximum number of students in the school, we consider bully nominations as an unbounded count. In such situations where the response variable is a count, the Poisson distribution is often a good approximation of the distribution.

The parameter of the Poisson distribution is the mean<sup>5</sup>  $\mu$  and the dispersion parameter is  $\phi = 1$ . Let  $\underline{y}$  be a Poisson random variable where possible values of  $\underline{y}$  equal non-negative integers (i.e.,  $y = 0, 1, 2, \dots$ ). The probability that a Poisson random variable equals  $y$

$$P(\underline{y} = y; \mu) = P(y; \mu) = \frac{e^{-\mu} \mu^y}{y!} \quad \text{for } y = 0, 1, \dots \quad (2.10)$$

Figure 2.6 gives four examples of Poisson distributions with means of 1, 2, 15 and 25. The smaller the mean, the more positively skewed the distribution. In Figure 2.6 (d) where  $\mu = 25$ , the distribution is uni-modal and looks fairly symmetric. If we

<sup>5</sup> Some authors use the symbol  $\lambda$  to represent the parameter of the Poisson distribution. Since the mean of the Poisson equals  $\lambda$ , we use  $\mu$  as the parameter of the distribution.

had a response variable (integer values) with the distribution illustrated in Figure 2.6 (d), we might be tempted to use a normal distribution for the response variable. However, unlike the normal distribution, for a Poisson distribution the mean equals the variance

$$\mu = \sigma^2. \quad (2.11)$$

Using a normal distribution for a count would violate the assumption of equal variances. Heteroscedasticity is expected for counts.

### 2.2.2 The Systematic Component

The random component of a GLM accounts for unsystematic random variation in observations. The systematic component of a model is the fixed structural part of the model that will be used to explain systematic variability between means. The systematic component or linear predictor of a GLM is a linear function of explanatory or predictor variables. The linear predictor is the same as the right-side of a normal linear regression model.

Let  $x_1, \dots, x_Q$  equal potential predictor variables. No restrictions are placed on the explanatory variables. They can be numeric or discrete. For discrete variables, the  $x$ 's can be dummy codes, effect codes, or any coding deemed useful or appropriate to represent the categories of a variable. The linear predictor is

$$\begin{aligned} \eta &= \beta_0 + \beta_1 x_1 + \dots + \beta_Q x_Q \\ &= \boldsymbol{\beta}' \mathbf{x}, \end{aligned} \quad (2.12)$$

where  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_Q)'$  is a vector of regression coefficients and  $\mathbf{x} = (1, x_1, \dots, x_Q)'$  is a vector of values on the predictor variables. Although  $\eta$  is a linear function of the  $x$ s, it may be nonlinear in shape. For example,  $\eta$  could be a quadratic, cubic or higher-order polynomial. Spline functions are linear functions, but they generally are not linear in shape. Transformation of the predictors are also possible (e.g.,  $\ln(x)$ ,  $\exp(x)$ , etc.), as well as interactions (e.g.,  $x_1 x_2$ ).

When a predictor variable is discrete, the regression curve will be disjoint. For example, Allen, Todd and Anderson (in preparation) assessed whether outcomes of cases of domestic violence in the state of Illinois changed after the formation of councils that provided a coordinated response to domestic violence. In one study, they modeled the change over time in the rate of extensions of orders of protection. Before council formation, there was no change; however, after formation, there was a jump in the number of extensions and subsequently a slow increase from that point on. This disjoint function was modeled using a dummy code for whether a council existed in a particular judicial circuit (i.e.,  $x_1 = 1$  if council,  $x_1 = 0$  if no council) and an interaction between the dummy code  $x_1$  and time (time was measured as chronological year).

In normal linear regression models, most of the attention is given to  $\eta$  and finding the predictors or explanatory variables that best predict the mean of the response

variable. This is also important in generalized linear models and problems such as multicollinearity found in normal linear regressions are also problems in generalized linear models. Hypothesis testing and statistical inference for the regression coefficients is discussed after we cover the last component of a GLM, the link function.

### 2.2.3 The Link Function

The link function allows for a non-linear relationship between the mean of the response variable and the linear predictor. The link function  $g(\cdot)$  connects the mean of the response variable to the linear predictor; that is,

$$g(\mu) = \eta. \quad (2.13)$$

The link function should be monotonic (and differentiable). The mean in turn equals the inverse transformation of  $g(\cdot)$ ,

$$\mu = g^{-1}(\eta). \quad (2.14)$$

The most natural and meaningful way to interpret model parameters is typically in terms of the scale of the data, in which case we consider  $\mu = g^{-1}(\eta) = g^{-1}(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ . This is illustrated in the examples of GLMs in Sections 2.3.3 and 2.3.4.

It is important to note that the link relates the *mean* of the response to the linear predictor and this is different from transforming the response variable. If the data are transformed (i.e.,  $y_i$ s), then a distribution must be selected that describes the population distribution of transformed data. Except when  $g(E(\underline{y})) = E(\underline{y})$ , a transformation of the mean generally does not equal the mean of transformed values; that is,  $g(E(\underline{y})) \neq E(g(\underline{y}))$ . As an example, suppose that we have a distribution with values (and probabilities) of 1 (0.1), 2 (0.4), 3 (0.1), 4, (0.2), 7 (0.2), and 10 (0.1). The logarithm of the mean of this distribution is  $\ln(E(\underline{y})) = \ln(4.1) = 1.411$ ; whereas, the mean of the logarithm equals  $E(\ln(\underline{y})) = 1.174$ .

The value of the linear predictor  $\eta$  could potentially equal any real number, but the expected values of the response variable may be bounded (e.g., counts are non-negative; proportions are between 0 and 1). An important consideration in choosing a link function is whether the selected link will yield predicted values of the response that are permissible. For example, with non-negative data such as count data or reaction times, a common link is the natural logarithm.

A summary of common link functions that will yield allowable values for particular types of response variables and the corresponding inverses of the links are given in Table 2.1. If there are no restrictions on the response variable (i.e., they are real numbers that could be positive or negative), then an *identity link* might be chosen where the mean is identical to the linear predictor; that is,

$$\mu = \eta.$$

**Table 2.1** Common link functions for different response variables. Note that  $\Phi$  is the cumulative normal distribution. [We should combine this information with the information in Table 2.2.](#)

| Type of response variable | Link                                    | $g(\mu)$                                | $g^{-1}(\eta)$              |
|---------------------------|---|---|-----------------------------|
| real                      | $ y  < \infty$ Identity                 | $\mu$                                   | $\eta$                      |
| real                      | $ y  < \infty$ Reciprocal               | $1/\mu$ if $y \neq 0$<br>$0$ if $y = 0$ | $1/\eta$                    |
| non-negative              | $y \geq 0$ Log                          | $\ln(\mu)$                              | $\mu = \exp(\eta)$          |
| bounded                   | $0 \leq y \leq 1$ Logit                 | $\ln(\mu/(1-\mu))$                      | $\exp(\eta)/(1+\exp(\eta))$ |
| bounded                   | $0 \leq y \leq 1$ Probit                | $\Phi^{-1}(\mu)$                        | $\Phi(\eta)$                |
| bounded                   | $0 \leq y \leq 1$ Log-Log               | $\ln(-\ln(\mu))$                        | $\exp(-\exp(\eta))$         |
| bounded                   | $0 \leq y \leq 1$ Complementary Log-Log | $\ln(-\ln(1-\mu))$                      | $1 - \exp(-\exp(\eta))$     |

Alternatively, the inverse or *reciprocal link*,

$$1/\mu = \eta,$$

is a possibility.

For response variables that are bounded between 0 and 1 (e.g., proportions or bounded response scales), the expected values are also bounded between 0 and 1. In such cases, a common strategy is to use a cumulative distribution function of continuous random variables as link function. A cumulative response function equals the probability that a random variable is less than a particular value,  $P(\underline{y} \leq y)$  where  $\underline{y}$  is continuous. The value of  $P(\underline{y} \leq y)$  equals real numbers from 0 to 1 but possible values for  $\eta$  may span the real numbers. Common distributions used for this purpose are the logistic, normal and extreme value or Gumbal distributions. The cumulative distributions for these are plotted in Figure 2.7.

Since the normal and logistic distributions are symmetric around the mean, the corresponding links are symmetric around .5. The rate at which the curves above  $P(\underline{y} \leq y) = .5$  increase toward 1 is the same as the rate of decrease toward 0 when the probability is below .5. The link corresponding to the cumulative distribution function for the logistic distribution is *logit* link and equals the natural logarithm of the ratio; that is,

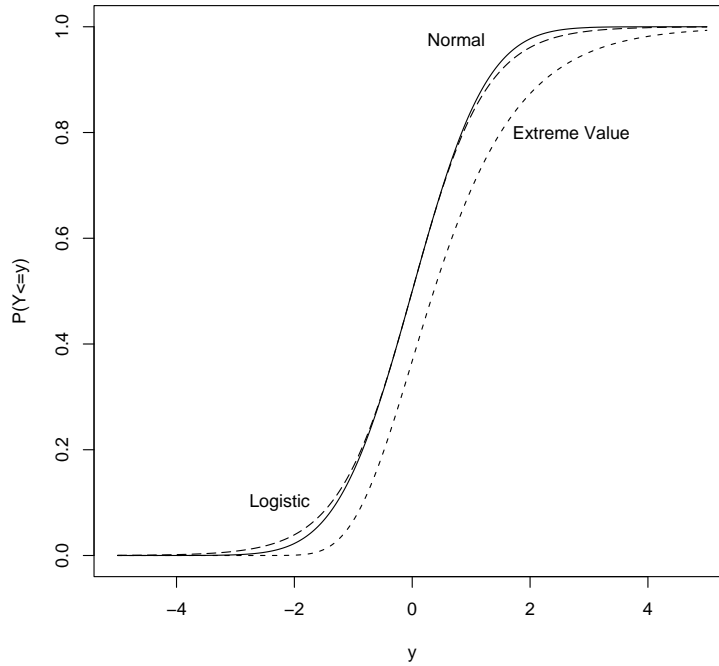
$$\text{logit}(\mu) = \ln(\mu/(1-\mu)). \quad (2.15)$$

When  $y$  is a proportion (i.e., probabilities are being modeled), the logit is the logarithm of odds. Alternatively, for a response variable  $y$  where  $0 \leq y \leq 1$ , one could use a *probit* link:

$$\text{probit}(\mu) = \Phi^{-1}(\mu), \quad (2.16)$$

where  $\Phi$  is the cumulative distribution function of the standard normal distribution. Note that the normal and logistic curves in Figure 2.7 are very similar. When modeling data, the choice between normal and logistic is minor in terms of model fit to data.

In the case of particular psychometric models (e.g., Thurstone's model model for paired comparison and the Bradley-Terry-Luce choice model), the choice of the link function is implied by the assumptions of the model. These models are discussed in more detail in Chapter ??.



**Fig. 2.7** Cumulative distribution functions for the standard normal, logistic (scale=.0625), and extreme value distributions.

The extreme value or Gumbel distribution is positively skewed such that  $P(\underline{y} \leq y)$  approaches 0 relatively quickly for smaller values of  $y$  but increases more slowly toward 1 for large values of  $y$ . The corresponding link is the *log-log link* and equals

$$\ln(-\ln(\mu)) = \eta.$$

If  $P(\underline{y} \leq y)$  approaches 0 more slowly and approaches 1 sharply, then a *complementary log-log link* could be employed:

$$\ln(-\ln(1 - \mu)) = \eta,$$

where  $(1 - \mu)$  is the complement of  $\mu$ .

When a distribution for a response variable is from the natural exponential family, there are special link functions known as *canonical link functions*. These links have desirable statistical properties that often make them preferable. In particular, with a canonical link the natural parameter equals the linear predictor (i.e.,  $\theta = \eta$ ) and sufficient statistics exist for the parameters. Table 2.2 gives the canonical link

function for members of the natural exponential distribution. Canonical links are often a good initial choice for a link function; however, in some cases, the canonical link make not be the best for the response variable. For example, in the study by Stine-Morrow et al. (2008) the response variable is reaction time. Reaction times are non-negative and skewed. The gamma distribution would be a good choice as a possible distribution, but the canonical link for the gamma, the inverse (i.e.,  $1/\mu$ ), yields negative predictions of reaction times when  $\eta < 0$ . With reaction times, an alternative link is the natural logarithm.

**Table 2.2** Distributions in the natural exponential family covered in this chapter or in later chapters. . . . [More can be added later if we want/need to.](#)

| Distribution     | Notation                   | Type of number | Range of $y$           | Canonical link | Dispersion parameter $\phi$ | Variance function | Probability $f(y; \mu, \phi)$ |
|------------------|----------------------------|----------------|------------------------|----------------|-----------------------------|-------------------|-------------------------------|
| Normal           | $N(\mu, \sigma^2)$         | real           | $-\infty < y < \infty$ | Identity       | $\sigma^2$                  | $\sigma^2$        | (2.1)                         |
| Gamma            | $\text{Gamma}(\mu, \phi)$  | real           | $0 < y$                | Inverse        | $\phi$                      | $\mu^2 \phi$      | (2.2)                         |
| Inverse Gaussian | $\text{IGauss}(\mu, \phi)$ | real           | $0 < y$                | $1/\mu^2$      | $\phi$                      | $\mu^3$           | (2.3)                         |
| Bernoulli        | $\text{Bernoulli}(\pi)$    | binary         | 0, 1                   | Logit          | 1                           | $\mu(1 - \mu)$    | (2.6)                         |
| Binomial         | $\text{Binomial}(\pi, n)$  | integer        | 0, 1, . . . $n$        | Logit          | $1/n$                       | $n\mu(1 - \mu)$   | (2.8)                         |
| Poisson          | $\text{Poisson}(\mu)$      | integer        | 0, 1, . . .            | Log            | 1                           | $\mu$             | (2.10)                        |

The ultimate decision on what link should be chosen depends on the nature of the response variable, theoretical considerations, and how well a model fits the data.

## 2.3 Examples of GLMs

In this section, we illustrate the formation of GLMs for a normal response variable, a positively skewed continuous variable, a binary response, and a count response. These examples are also used to illustrate assessing model goodness-of-fit to data and statistical inferential procedures common to GLMs. The modeling of data in this section is only a starting point. Each of the data sets has a clustered structure (e.g., responses nested within subjects, students nested within peer groups or classrooms). The clustered nature of the data is completely ignored and the conclusions presented here should not be taken seriously. In later chapters, we re-analyze each of these data sets using random effects to deal with the clustering and we reach different conclusions compared to those presented in this chapter.

### 2.3.1 A Normal Continuous Variable

The data for this example come from  $N = 358$  children in a study by Rodkin, Wilson & Ahn (2007) on social integration in classrooms. The response variable is a measure of a child's level of segregation with respect to mutual friendships within their classroom. Assuming normality of this measure is reasonable because the values of the response variable are (theoretically) continuous real numbers and the distribution within classrooms is likely to be uni-modal and roughly symmetric.

Potential predictor variables are gender, the child's ethnicity and the racial distribution in the classroom. The predictor variable gender is dummy coded (i.e.,  $\text{male} = 1$  for boys and 0 for girls), and ethnicity is effect coded with  $-1$  for European American and 1 for African American students. For the racial distribution variable, classrooms were categorized as having either a majority of students who were white, a majority who were black, or no clear majority (i.e., multicultural). Two coded orthogonal variables were used to represent the classroom racial distribution in the regression model. One code is for classroom majority where  $\text{CMaj} = 1$  if the majority of the students in the classroom are black,  $\text{CMaj} = -1$  if the majority are white, and  $\text{CMaj} = 0$  if there is no majority. The other code for classroom racial distribution is whether the classroom is multicultural where  $\text{MultC} = 1$  for a multicultural classroom and  $\text{MultC} = -0.5$  for either of the other classrooms.

The normal linear regression model for these data would typically be written as

$$\begin{aligned} \underline{\text{segregation}}_i &= \beta_0 + \beta_1(\text{male}_i) + \beta_2(\text{ethnicity}_i) + \beta_3(\text{CMaj}_i) \\ &\quad + \beta_4(\text{MultC}_i) + \beta_5(\text{ethnicity}_i)(\text{CMaj}_i) \\ &\quad + \beta_6(\text{ethnicity}_i)(\text{MultC}_i) + \underline{\epsilon}_i, \end{aligned}$$

where  $\underline{\epsilon}_i \sim N(0, \sigma^2)$  and  $\mathbf{x}_i = (1, (\text{male}_i), (\text{ethnicity}_i), \dots, (\text{ethnicity}_i)(\text{MultC}_i))'$ . The equivalent model written as a GLM is

$$\begin{array}{ll} \text{Random:} & \underline{\text{segregation}}_i | \mathbf{x}_i \sim N(\mu_i, \sigma^2) \\ \text{Link:} & \mu_i = \eta_i \\ \text{Linear predictor:} & \eta_i = \boldsymbol{\beta}' \mathbf{x}_i \end{array}$$

where  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_6)'$ . The GLM form emphasizes the fact we are modeling the mean conditional on predictor variables. It further emphasizes that three decisions are made. If the model does not fit the data well, then the normal distribution may be a poor representation of the distribution of the response, the identity may not be the best link function, the linear predictor may not include all relevant variables (or transformations of them), or some combination of these three.

The estimated parameters are reported in Table 2.3. Notice that the parameters for child's ethnicity, the interaction between ethnicity and classroom majority ( $\text{CMaj}$ ), and the interaction between ethnicity and multicultural ( $\text{MultC}$ ) are all significant. These results are not trustworthy because observations within classrooms dependent

**Table 2.3** Estimated parameters from a normal linear multiple regression model fit to the social segregation data from Rodkin et al. (2007).

| Parameter       | df | Estimate | Standard Error | t          |       | 95% Confidence intervals |       |
|-----------------|----|----------|----------------|------------|-------|--------------------------|-------|
|                 |    |          |                | (df = 295) | p     | Lower                    | Upper |
| Intercept       | 1  | 0.26     | 0.05           | 5.62       | < .01 | 0.17                     | 0.36  |
| male            | 1  | -0.06    | 0.07           | -0.81      | .42   | -0.19                    | 0.08  |
| ethnicity       | 1  | 0.22     | 0.04           | 6.11       | < .01 | 0.15                     | 0.30  |
| CMaj            | 1  | -0.05    | 0.04           | -1.39      | .16   | -0.13                    | 0.02  |
| ethnicity*CMaj  | 1  | -0.11    | 0.04           | -2.84      | < .01 | -0.18                    | -0.03 |
| MultC           | 1  | -0.07    | 0.06           | -1.25      | .21   | -0.19                    | 0.04  |
| ethnicity*MultC | 1  | 0.13     | 0.06           | 2.24       | .03   | 0.02                     | 0.25  |

and thus violate of the independence assumption required for statistical inference<sup>6</sup>. The observations within classrooms are most likely positively correlated; therefore, the standard error estimates are too small leading to the test statistics for parameters whose absolute values are too large. In other words, Type I error rates are inflated. We return to this example in Chapter ??.

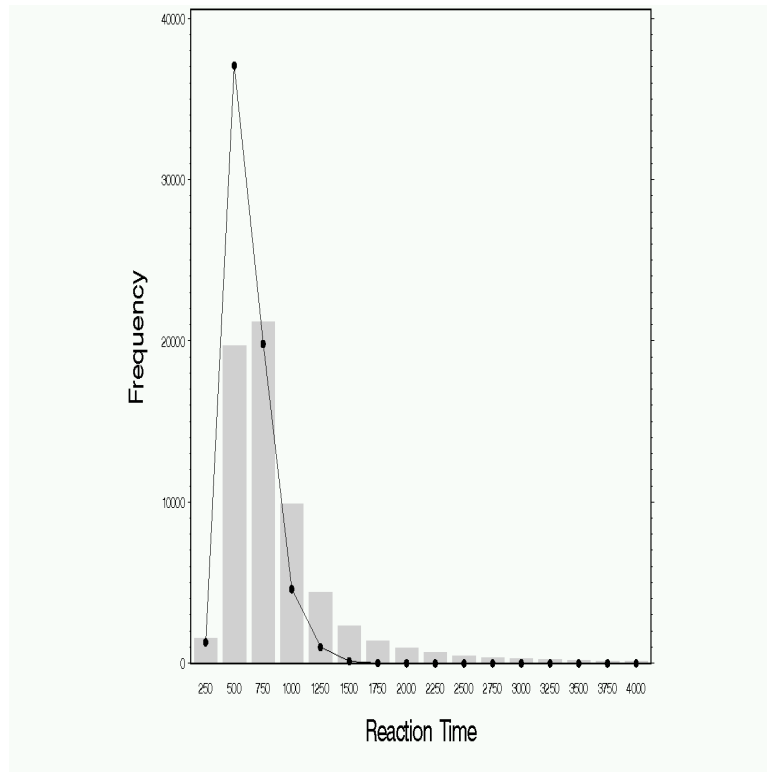
### 2.3.2 A Skewed Continuous Response Variable

The data for this example consists of a sub-set of data from  $N = 149$  elderly subjects in a study on cognition and aging from Stine-Morrow and colleagues. The procedures and data are similar to those reported in Stine-Morrow et al. (2008). Elderly individuals were presented with words on a computer monitor. The words were presented one at a time and a sequence of words made up a sentence. Each subject read multiple sentences and sentences could wrap over lines on the screen. A word would be presented and the subject would hit the space bar when they were ready of the next word. Of interest in this study is reading time measured in ml seconds between word presentation and the hitting of space bar. The reaction times are continuous and positively skewed as can be seen from the histogram of reaction times in Figure 2.8.

Given the nature of the response variable, two plausible distributions for these data are the gamma and inverse Gaussian distributions. Both of these distributions are positively skewed for non-negative continuous responses. Which distribution is better for the data may depend more on the relationship between the mean and variance and best determined empirically.

Predictor variables include textual variables and attributes of the subjects. The textual variables are the number of syllables in the word (`syll`), logarithm of the

<sup>6</sup> In normal linear regression of clustered data (where the dependencies within cluster are ignored), the estimated  $\beta$ s are actually reasonably good (consistent) estimates of the effects (–reference–). This is not true for all GLMs.



**Fig. 2.8** Histogram of the observed distribution of time taken by elderly participants to read a word and fitted values from a gamma regression the canonical link, the inverse.

word frequency (`logFreq`), inter-sentence boundary (`intSB`), and whether a new line is started (`newLine`). Subject attributes of interest are age, score on the North American Adult reading test (`NAART`), and measures of cognitive executive functioning. The latter includes overall mean response accuracy (`meanAcc`), response time for trials using the same task (`SwRTsame`), and task switching response time cost (`SWRTcost`). The structural part of the model will be a linear function of these textual and subject variables.

The last component of the model is a link function. The canonical link for the gamma is the inverse that in the context of reaction times is interpretable as the speed to read a word. The canonical link for the inverse Gaussian is  $1/\mu^2$ . In addition to the default link functions, the log link will also be considered because response times must be positive.

The family of GLMs that we fit to the reaction time data is

$$\underline{y}_i \sim f(y; \mu_i, \phi) \quad (2.17)$$

$$g(\mu_i) = \eta_i \quad (2.18)$$

$$\eta_i = \boldsymbol{\beta}' \mathbf{x}_i \quad (2.19)$$

where  $\mathbf{x}_i$  is a vector of values of the predictor variables,  $\boldsymbol{\beta}$  is a vector of regression coefficients,  $f(y; \mu_i, \phi)$  is either the gamma or the inverse Gaussian distribution. The links considered are  $1/\mu$ ,  $1/\mu^2$ , and  $\ln(\mu)$ . For example, a gamma regression using the inverse link is

$$\begin{aligned} (1/\mu_i) = \eta_i = & \beta_0 + \beta_1 (\text{syll}_i) + \beta_2 (\text{logFreq}_i) + \beta_3 (\text{intSB}_i) + \beta_4 (\text{newLine}_i) \\ & + \beta_5 (\text{age}_i) + \beta_5 (\text{NAART}_i) + \beta_6 (\text{meanAcc}_i) + \beta_6 (\text{SwRTsame}_i) \\ & + \beta_7 (\text{SwRTcost}_i). \end{aligned}$$

Additionally, reaction time  $\text{rt}_i \sim \text{Gamma}(\mu_i, \phi)$ , and the mean reaction time for response  $i$  is  $\mu_i = E(\text{rt}_i | (\text{syll}_i), (\text{logFreq}_i), \dots, (\text{SwRTcost}_i)) = 1/\eta_i$ .

Before fitting models to data, we deleted outliers with reaction times greater than 4,000 ml seconds, approximately 1% of the total  $N = 64,368$  responses. Regardless of the distribution, models with link  $1/\mu^2$  failed to converge and this link was deemed a poor choice. The predicted reaction times from the other four models are nearly identical. The correlations between the predictions with the same link function but different distributions equal .999 for the  $\ln(\mu)$  and  $1/\mu$  links. The correlations between predictions with different links with the same distribution equal to .989 for the gamma distribution and .986 for the inverse Gaussian.

Since the predicted values from the four models are so similar, only the fitted values from the gamma regression with the inverse link are plotted in Figure 2.8. The fitted values are represented by the dots connected by a solid line. The models over-predict the number of reactions times near 500 ml seconds by over 15,000, and under-predicts the reaction times greater than 500 ml seconds. In this data set, there are  $N = 149$  subjects each of who contribute reaction times for each of the 432 different words in the experiment. A model that takes into account random individual differences could improve the fit of the model to the data, as well as account for dependencies due to the nesting of observations within individuals.

All of the effects in the models are significant (except one in the inverse Gaussian model with the log link). Since we did not take into account the dependency in the data, the estimated parameters and standard errors are not reported here. We revisit this example in Chapter ??.

### 2.3.3 A Dichotomous Response Variable

The data for this example come from a study of by Rodkin et al. (2006) on the social status of children among their peers. The data consist of measures on  $N = 526$  fourth and fifth grade students. As an index of social status, children were asked they think is “cool.” For this example, the response variable `ideal`, equals the number of kids classified as a model or ideal student (i.e., popular or prosocial) among those who were nominated as being “cool.” Since the response is dichotomous (cool-kid

is an ideal student or not), the binomial distribution is the natural choice as the distribution of the response and we model the probability that an ideal student is nominated as being cool. The predictor or explanatory variables are the nominating child's popularity, gender and race. Each of the predictors are dummy coded as follows:

$$\text{popularity}_i = \begin{cases} 1 & \text{high} \\ 0 & \text{low} \end{cases}, \quad \text{gender}_i = \begin{cases} 1 & \text{boy} \\ 0 & \text{girl} \end{cases}, \quad \text{and} \quad \text{race}_i = \begin{cases} 1 & \text{black} \\ 0 & \text{white} \end{cases}.$$

The index  $i$  is used to represent a particular case or combination of the predictor variables (e.g., white girl who is popular is one combination),  $n_i$  equals the number of students nominated as "cool" by peers who had combination  $i$  on the predictor variables, and  $\pi_i$  equals the probability that a student nominates an ideal student as being cool. The cool-kid data are given in Table 2.4.

Our basic GLM for the cool-kid data is

$$\begin{aligned} \text{ideal}_i &\sim \text{Binomial}(\pi_i, n_i) \\ g(\pi_i) &= \eta_i \\ \eta_i &= \beta_0 + \beta_1 \text{Popularity}_i + \beta_2 \text{Gender}_i + \beta_3 \text{Race}_i, \end{aligned}$$

Three different link functions are illustrated: the identity (i.e.,  $\pi_i = \eta_i$ ), the logit (i.e.,  $\ln(\pi_i/(1 - \pi_i)) = \eta_i$ ), and the probit (i.e.,  $\Phi^{-1}(\pi_i) = \eta_i$ ). Putting the three components together leads to three different models. In all models,  $\text{incorrect}_i$  is binomially distributed. The linear probability model is

$$\pi_i = \beta_0 + \beta_1 \text{Popularity}_i + \beta_2 \text{Gender}_i + \beta_3 \text{Race}_i,$$

the logit model is

$$\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 \text{Popularity}_i + \beta_2 \text{Gender}_i + \beta_3 \text{Race}_i,$$

and the probit model is

$$\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 \text{Popularity}_i + \beta_2 \text{Gender}_i + \beta_3 \text{Race}_i.$$

The fitted values  $\hat{\pi}_i$  for each of these three models are reported in Table 2.4.

Given that  $\pi_i$  is the probability the student nominated by one child with explanatory variables equal to  $\text{Popularity}_i$ ,  $\text{Gender}_i$ , and  $\text{Race}_i$  is an ideal student, the number of ideal students nominated equals  $\text{ideal}_i = n_i \pi_i$ . However, the data given in Table 2.4 is collapsed over individuals with the same pattern on the predictor variables. These models can be fit to individual level data (i.e.,  $y = 0, 1$  and  $n_i = 1$ ) or to collapsed data (as in Table 2.4). Both ways lead to the same estimated probabilities and counts. More on this issue is discussed in Chapter ??.

**Table 2.4** Data of students nominated as “cool” who are model students and predictions from linear probability, probit and logit models.

| Index<br><i>i</i> | Nominating child's |        |       | Number of                        | Number of non-                   | Number                           | Proportion of Predicted Probabilities  |        |        | Std. residuals |         | 95% confidence |       |       |
|-------------------|--------------------|--------|-------|----------------------------------|----------------------------------|----------------------------------|--|--------|--------|----------------|---------|----------------|-------|-------|
|                   | popularity         | gender | race  | ideal students<br>who are “cool” | ideal students<br>who are “cool” | of cases<br><i>n<sub>i</sub></i> | ideal students<br><i>p<sub>i</sub></i> | Linear | Probit | Logit          | Pearson | Adjusted       | lower | upper |
| 1                 | low                | girl   | white | 70                               | 65                               | 135                              | .52                                    | .53    | .53    | .54            | -0.38   | -0.65          | .47   | .60   |
| 2                 | low                | girl   | black | 32                               | 114                              | 146                              | .22                                    | .22    | .21    | .21            | 0.19    | 0.30           | .17   | .27   |
| 3                 | low                | boy    | white | 47                               | 61                               | 108                              | .43                                    | .44    | .42    | .41            | 0.44    | 0.70           | .34   | .49   |
| 4                 | low                | boy    | black | 13                               | 85                               | 98                               | .13                                    | .12    | .14    | .14            | -0.28   | -0.36          | .10   | .19   |
| 5                 | high               | girl   | white | 80                               | 28                               | 108                              | .74                                    | .71    | .71    | .72            | 0.57    | 0.86           | .65   | .78   |
| 6                 | high               | girl   | black | 15                               | 29                               | 44                               | .34                                    | .39    | .38    | .37            | -0.43   | -0.53          | .29   | .46   |
| 7                 | high               | boy    | white | 46                               | 34                               | 80                               | .58                                    | .61    | .61    | .61            | -0.61   | -0.88          | .53   | .68   |
| 8                 | high               | boy    | black | 11                               | 25                               | 36                               | .31                                    | .29    | .27    | .27            | 0.52    | 0.62           | .20   | .35   |

The model with the identity link (i.e.,  $\pi_i = \eta_i$ ) is known as the *linear probability* model. This differs from normal linear regression in that the distribution for the response variable is the binomial distribution and not the normal distribution. The estimated probabilities of the linear probability model are given in Table 2.4. Although all of the estimated probabilities of this model were positive, this will not always be the case. Sometimes this model yields negative fitted value for the probabilities.

The estimated probabilities for the logit and probit models are nearly identical and are very similar to those from the linear probability model. To show how similar the predictions are, as well as given an ideal of how well the models are fitting the data, the estimated probabilities from the three models are plotted against the observed proportions in Figure 2.9. Note that perfect prediction corresponds to the solid line identity line. The predicted probabilities for the three models are basically on top of each other and all very close the observed values.

In Sections 2.5, more formal methods are presented for assessing whether the models provide a good representation of the data and for choosing among a set of plausible models. One advantage of the logit model is that the logit is the canonical link function for the binomial distribution and the interpretation of the models parameters is relatively straight forward. Furthermore, when the canonical link for the binomial distribution is used, the logistic regression model is special case of a Poisson regression. We exploit this relationship in Chapter??.

The estimated parameters and fit statistics for the linear, logit and probit models are reported in Table 2.5. A brief explanation is given here on how to interpret the parameters of a logit model and save more detailed discussion for Chapter ??. To make this discussion more general, let  $x_{1i}$  represent `Popularityi` and  $x_{2i}$  represent `Genderi` and  $x_{3i}$  represent `Racei`. To emphasize that  $\pi_i$  is a function of the  $x_i$ s, the probability is written as a function of them (e.g.,  $\pi_i(x_{1i}, x_{2i}, x_{3i})$ ). The logarithm of the odds equals

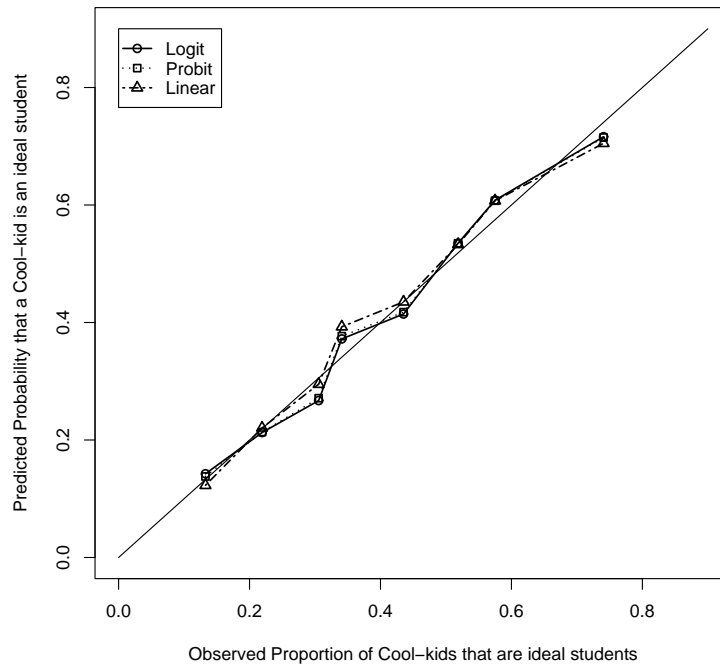
$$\ln \left( \frac{\pi_i(x_{1i}, x_{2i}, x_{3i})}{1 - \pi_i(x_{1i}, x_{2i}, x_{3i})} \right) = \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}. \quad (2.20)$$

The  $\beta$ s are most naturally interpreted in terms of odds ratios. Taking the inverse of the logarithm of (2.20) (i.e, the exponential) yields the odds that a cool-kid is an ideal students,

$$\frac{\pi(x_{1i}, x_{2i}, x_{3i})}{1 - \pi(x_{1i}, x_{2i}, x_{3i})} = \exp[\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}]. \quad (2.21)$$

If  $x_{1i}$  is one unit larger but  $x_{2i}$  and  $x_{3i}$  remain the same, the odds equals

$$\frac{\pi(x_{1i} + 1), x_{2i}, x_{3i})}{1 - \pi((x_{1i} + 1), x_{2i}, x_{3i})} = \exp[\beta_0 + \beta_1(x_{1i} + 1) + \beta_2 x_{2i} + \beta_3 x_{3i}]. \quad (2.22)$$



**Fig. 2.9** Predicted probabilities from the logit, probit and linear probability model fit to the cool-kid data plotted against the observed proportions.

The ratio of the two odds above is an odds ratio. In our case, dividing the odds in (2.22) by the odds in (2.21) equals  $\exp(\beta_1)$ . This interpretation does not depend of the specific value for gender or race.

For the cool-kid example, the estimated parameters for the logit (and probit) model are reported in Table 2.5. Using the estimated parameters from the logit model, the estimated odds that a highly popular child nominates an ideal student as cool are  $\exp(0.7856) = 2.19$  times larger than the odds for child with low popularity. The odds that a boy nominates an ideal student are  $\exp(-.4859) = 0.62$  times larger than the odds for a girl, and the odds for a white student are  $\exp(-1.4492) = 0.23$  times the odds for a black student. Since the value of the predictor in the numerator of the odds is somewhat arbitrary, we can switch the roles of gender and race and say that the odds that a girl nominates an ideal student are  $\exp(.4859) = 1/0.62 = 1.63$  times the odds for a boy, and the odds for a white student are  $\exp(1.4492) = 1/0.23 = 4.26$  times the odds for a black student. It is more likely that girls, whites and highly popular students will nominate a model or ideal student as being cool.

**Table 2.5** Model goodness-of-fit statistics and parameter estimates of the probit and logit models fit to the model cool kid data (Rodkin et al. 2006).

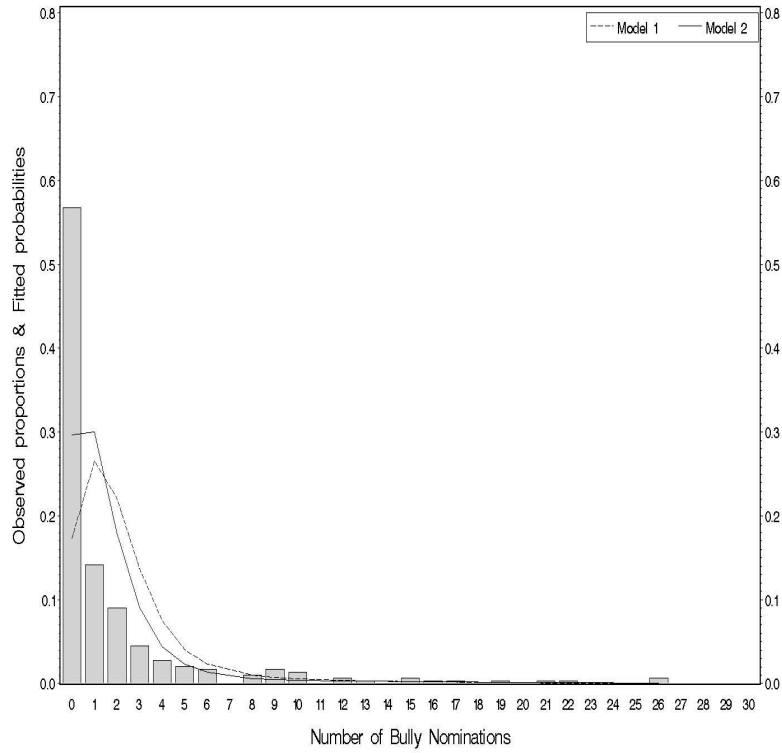
| Effect                     | Probit Model |        |       |          | Logit Model  |        |       |          |
|----------------------------|--------------|--------|-------|----------|--------------|--------|-------|----------|
|                            | estimate     | s.e.   | Wald  | <i>p</i> | estimate     | s.e.   | Wald  | <i>p</i> |
| Intercept                  | 0.0875       | 0.0875 | 1.04  | .31      | 0.1403       | 0.1397 | 1.01  | 0.32     |
| Popularity:                |              |        |       |          |              |        |       |          |
| High                       | 0.4804       | 0.1013 | 22.48 | < .01    | 0.7856       | 0.1667 | 22.22 | < .01    |
| Low                        | 0.0000       | 0.0000 | —     | —        | 0.0000       | 0.0000 | —     | —        |
| Gender:                    |              |        |       |          |              |        |       |          |
| Boy                        | −0.2955      | 0.0987 | 8.97  | < .01    | −0.4859      | 0.1640 | 8.77  | 0.01     |
| Girl                       | 0.0000       | 0.0000 | —     | —        | 0.0000       | 0.0000 | —     | —        |
| Race:                      |              |        |       |          |              |        |       |          |
| Black                      | −0.8817      | 0.1011 | 76.13 | < .01    | −1.4492      | 0.1701 | 72.55 | < .01    |
| White                      | 0.0000       | 0.0000 | —     | —        | 0.0000       | 0.0000 | —     | —        |
| <i>df</i>                  | 4            |        |       |          | 4            |        |       |          |
| Deviance( <i>p</i> -value) | 1.4944 (.83) |        |       |          | 1.5955 (.81) |        |       |          |
| $X^2$ ( <i>p</i> -value)   | 1.4933 (.83) |        |       |          | 1.5982 (.81) |        |       |          |
| ln(likelihood)             | −450.0575    |        |       |          | −450.1081    |        |       |          |

The students providing the nominations (i.e., the responses) in the cool-kid example are nested within peer groups and peer groups are further nested within classrooms. This nesting leads to responses from students that are highly positively correlated and  $\hat{\beta}$ s and estimated standard errors are biased. The estimated standard errors are too small, which in turn leads to test statistics for the parameters that are too large. In other words, the Type I error rates of statistical tests inflated. These data are re-analyzed in the next chapter using more appropriate methods; however, we continue to use the cool kid-data in this chapter to illustrate GLM methodology.

### 2.3.4 A Count Response Variable

The data from this example are from a study by Espelage et al. (2004) on the effects of aggression during early adolescence. The response variable, the extent to which a child is a bully, has been measured in two different ways. One method takes the average of responses to nine items from the Illinois Bully Scale (Espelage & Holt, 2001), and the other method uses the number of children who list another as being a bully. Bully nominations are viewed as a more objective measure than scores the Illinois Bully Scale (a self report measure). In this analysis, we will model bully nominations as the response variable with the bully scale scores as an explanatory or predictor variable.

The distribution of the peer nominations is given in Figure 2.10. Note that the distribution is very skewed and responses are non-negative integers. Since the response variable is a count, our initial choice of a distribution is the Poisson with its canonical link, the natural logarithm ln. The bully scale is the predictor variable



**Fig. 2.10** Observed distribution and fitted values from two Poisson regression models. Model 1 only includes the bully scale as a predictor variable and Model 2 includes bully scale, gender, age and empathy.

included in the systematic component. Our GLM model for these data is

$$\begin{aligned} \text{bullynom}_i &\sim \text{Poisson}(\mu_i) \\ \ln(\mu_i) &= \eta_i \\ \eta_i &= \beta_0 + \beta_1(\text{bullyscale})_i \end{aligned}$$

The parameter estimates and fit statistics are reported in Table 2.6. The fitted model equals

$$\ln(\widehat{\text{bullynom}}_i) = -0.6557 + 0.8124(\text{bullyscale})_i,$$

or using the inverse of  $\ln$  that gives the predicted counts,

$$\widehat{\text{bullynom}}_i = \exp[-0.6557 + 0.8124(\text{bullyscale})_i.]$$

Interpretation of the regression coefficients in a Poisson regression model is similar to that for normal linear regression in that we consider a one unit difference of

an explanatory variable and the corresponding difference between the predicted (fitted) expected values of the response variable (i.e., the estimated means); however, the effect of a predictor on the mean response is multiplicative rather than additive. Specifically, the Poisson regression model when a predictor has a value of  $x_i$  is

$$\mu_{i,x_i} = \exp[\beta_0 + \beta_1(x_i)] = e^{\beta_0} e^{\beta_1 x_i}, \quad (2.23)$$

and the model when a predictor is one unit larger is

$$\mu_{i,(x_i+1)} = \exp[\beta_0 + \beta_1(x_i + 1)] = e^{\beta_0} e^{\beta_1 x_i} e^{\beta_1}. \quad (2.24)$$

The expected mean count  $\mu_{i,(x_i+1)}$  is  $\exp(\beta_1)$  times the mean  $\mu_{i,x_i}$ ; that is, the predicted mean given  $(x_i + 1)$  is  $\exp(\beta_1)$  times the mean given  $x_i$ . In our example, for a one point larger on the bully scale, the mean number of nominations is  $\exp(0.8124) = 2.5$  times larger.

**Table 2.6** Estimated parameters and fit statistics for simple and more complex Poisson regression models fit the peer nomination data.

| Parameter       | Model 1 |        |               |        |       | Model 2 |        |               |        |       |
|-----------------|---------|--------|---------------|--------|-------|---------|--------|---------------|--------|-------|
|                 | Est.    | SE     | $\exp(\beta)$ | Wald   | p     | Est.    | SE     | $\exp(\beta)$ | Wald   | p     |
| Intercept       | -0.6557 | 0.0888 |               | 54.52  | < .01 | -4.2457 | 0.6128 |               | 48.00  | < .01 |
| Bully scale     | 0.8124  | 0.0351 | 2.25          | 536.72 | < .01 | 0.7812  | 0.0543 | 2.18          | 206.74 | < .01 |
| Gender (female) |         |        |               |        |       | -0.3278 | 0.0942 | 0.72          | 12.12  | < .01 |
| Gender (male)   |         |        |               |        |       | 0.0000  | 0.0000 | .             | .      | .     |
| Empathy         |         |        |               |        |       | 0.1331  | 0.0515 | 1.14          | 6.68   | = .01 |
| Age             |         |        |               |        |       | 0.2574  | 0.0492 | 1.29          | 27.33  | < .01 |
| Fight scale     |         |        |               |        |       | 0.1533  | 0.0447 | 1.17          | 11.74  | < .01 |
| <i>df</i>       |         |        | 287           |        |       |         |        | 283           |        |       |
| Deviance        |         |        | 1771.65       |        |       |         |        | 1701.92       |        |       |
| Pearson $X^2$   |         |        | 2968.56       |        |       |         |        | 2875.18       |        |       |
| ln(Likelihood)  |         |        | 153.30        |        |       |         |        | 188.17        |        |       |

The simple Poisson regression model fails to give a good representation of the data as can be seen in Figure 2.10 where the dashed line shows the model predicted probabilities computed using the fitted mean counts<sup>7</sup>. The model under-predicts the number of observations with 5 or fewer nominations. The model may be failing to fit for a number of reasons. Recall that for the Poisson distribution,  $\mu = \sigma^2$ . Overdispersion occurs when  $\mu < \sigma^2$  and this is found in the bully data. Means and variances of the number of nominations for ranges of the bully scale are reported in Table 2.7. The means increase as bully scores increase (as expected), but the variances are much larger than the means. When data are over-dispersed, the standard errors for parameter estimates are too small, which in turn leads to test statistics for coefficients that are too large (i.e., higher Type I error rates).

<sup>7</sup> The predicted probabilities were computed using  $\hat{\pi}_i = \hat{\mu}_i/N$  where  $\hat{\mu}_i$  is the fitted mean count for observation  $i$  and  $N$  is the total sample size.

**Table 2.7** The means and variances of peer nominations for ranges of the bully scale scores.

| Bully score | Mean  | Variance |
|-------------|-------|----------|
| 0–0.9       | 1.42  | 12.70    |
| 1–1.9       | 4.86  | 84.49    |
| 2–2.9       | 7.20  | 179.03   |
| 3–4.0       | 12.20 | 301.70   |

Overdispersion can be caused by lacking necessary predictor variables, having correlated observations due to nesting or clustering of observations, or the wrong distribution for the data. All observations with the same value of the predictor variable are assumed to be independent values from the same Poisson distribution. This assumption is known as the *homogeneity* assumption. We may be able to overcome heterogeneity by adding predictor variables when the needed additional variables are available.

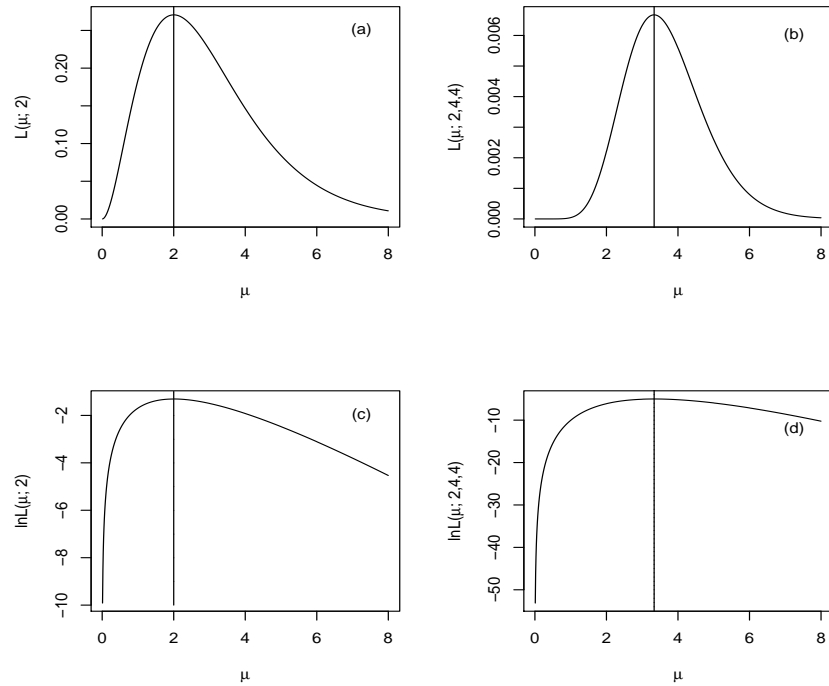
The variables gender, age, empathy (perspective taking sub-scale of the Davis (–reference–) measure of empathy), and scores on a fighting scale were all added to the model. The results of the second model are given in Table 2.6 under “Model 2”. Although Model 2 fits better than Model 1, it still fails to adequately represent the data. The solid line in Figure 2.10 shows that Model 2 still fails to capture the low end of the distribution.

Another potential problem with this analysis that could lead to overdispersion is dependence of observations. Students are nested or clustered within peer groups. This analysis has not taken this potential dependency into account. If there is dependency in the data, the standard errors will be too small and statistical tests will have higher Type I error rates. The statistical tests for the regression coefficients in this example are not to be trusted. When assessing and detecting problems with GLMs, we must consider our decisions regarding the distribution, systematic component and link function. We return to this example in Chapter ?? where we consider ways of including dependence (i.e., random effects), and exploring alternative distributions for the data.

## 2.4 Estimation

When using GLMs, having a basic understanding of how parameters are estimated can help detect problems and point to potential solutions. An overview of estimation is provided here and a more technical coverage is given in Section ??.

Maximum likelihood estimation (MLE) is typically used to estimate the parameters of GLMs. Maximum likelihood estimates are those that are most likely given the data. This is achieved by considering the probability density as a function of the parameters rather than as a function of data. Given data and a probability model (i.e., random component of a GLM), those parameters that yield a maximum value of the function are maximum likelihood estimates. For example, consider the distri-



**Fig. 2.11** The likelihood (top) and ln likelihood (bottom) for the Poisson distribution when  $y = 2$  (left) and  $y = 1, 4, 4$  (right) plotted as a function of possible values for the mean  $\mu$ .

tribution function for the Poisson distribution in equation (2.10) and the simple case of a single observation  $y$ . The likelihood function for the Poisson is

$$L(\mu; y) = \frac{e^{-\mu} \mu^y}{y!}. \quad (2.25)$$

Equations (2.10) and (2.25) are the same except the role of  $\mu$  and  $y$  have been switched such that  $y$  is fixed and  $\mu$  can vary in (2.25). For example, the likelihood given by (2.25) is plotted for  $y = 2$  in Figure 2.11 (a). Notice that the maximum value of  $L(\mu|2)$  occurs at  $\mu = 2$ .

Suppose that we have a sample of  $N$  independent observations  $y_1, \dots, y_N$  from Poisson( $\mu$ ). The likelihood function for the whole sample is the product of the individual likelihoods,

$$L(\mu|y_1, \dots, y_N) = \prod_{i=1}^N \frac{e^{-\mu} \mu^{y_i}}{y_i!}. \quad (2.26)$$

This is basically an application of the multiplicative rule in probability where the probability of independent events equals the produce of the probabilities for each of the events. An example of this likelihood function is plotted in Figure 2.11 (b). The maximum of the likelihood occurs at  $\mu = 3.33$ , the mean of 2, 4 and 4.

To estimate the parameters of a GLM, we specify a model for  $\mu$  conditional on predictor or explanatory variables. This model is  $\mu_i = g^{-1}(\boldsymbol{\beta}'\mathbf{x}_i)$ . The process and concepts are the same except that we replace  $\mu_i$  by its model  $g^{-1}(\boldsymbol{\beta}'\mathbf{x}_i)$  in the likelihood function. The likelihood function is then a function of the regression coefficients  $\boldsymbol{\beta}$ . Once the MLE of  $\boldsymbol{\beta}$  has been estimated (i.e.,  $\hat{\boldsymbol{\beta}}$ ), the MLE of  $\mu_i$  equals  $\hat{\mu}_i = g^{-1}(\hat{\boldsymbol{\beta}}'\mathbf{x}_i)$ .

Typically, estimation procedures use the ln of the likelihood because it is easier to work with. Examples of the  $\ln(L(\mu; y))$  for the Poisson distribution are given in Figures 2.11 (c) and (d). The maximum of the ln likelihood function occurs at the same value of the parameters as it does for the likelihood function.

Except for special cases of the general linear models (i.e., normal linear regression and ANOVA), MLE of parameters requires an iterative algorithm. Common algorithms for finding maximum likelihood estimates are the Newton-Raphson and Fisher scoring. These are iterative algorithms used to solve nonlinear equations. The algorithms start with an initial set of parameter estimates and up-dates the estimates on each iteration by solving a simple approximate problem. The up-dating is repeated until the algorithm converges and a maximum of the likelihood has been achieved. The parameter up-dating equations for Newton-Raphson and Fisher scoring equal current parameter estimates minus the product of the inverse of the *Hessian matrix* and the *score vector*.

The score vector<sup>8</sup> corresponds to the slope of the ln likelihood. The is one element in the score vector of each parameter to be estimated. When the maximum of the likelihood is achieved, the elements of the score vector (slopes) all equal zero. This is illustrated in our simple example in Figure 2.11 (c). When  $\mu = 2$ , the slope is flat, equal to 0. The Hessian matrix<sup>9</sup> conveys information about the rate of change of the likelihood. When a parameter estimate is far from the MLE, the rate of change will be larger. In our simple example, note that when  $\mu = 0$  in Figure 2.11 (c), the rate of change is larger than it is when  $\mu$  is closer to the MLE at  $\mu = 2$ .

The difference between Fisher scoring and Newton-Raphson is how the Hessian matrix is computed. The Newton-Raphson method computes the Hessian using data; whereas, Fisher scoring uses the expected value of the Hessian and equals the negative of Fisher's information matrix. Different algorithms for finding MLEs should all yield the same results.

Common problems to look for are lack of convergence, fitted values outside the permitted range (e.g., counts that are negative), and a singular or nearly singular Hessian matrix (i.e., there is no unique inverse of the Hessian). Problems are generally caused by the wrong model for the data. Estimation problems can generally be solved by modifying the model. For example, a linear probability model will yield

---

<sup>8</sup> The score vector or gradient is the vector of first partial derivatives of the ln likelihood function.

<sup>9</sup> The Hessian matrix is a matrix of second partial derivatives.

negative estimated probabilities whenever  $\eta_i < 0$  and probabilities greater than one whenever  $\eta_i > 1$ . In such cases, the estimation algorithm may fail to converge. A reasonable solution in such a case would be to use a different link function that would ensure that probabilities are within the permitted range of 0 to 1 (e.g., probit, logit).

If a model has too many predictor variables or the predictors are highly correlated, the Hessian may be singular (or nearly so). This indicates an unstable solution. Most computer programs will issue a warning or error message. The problem of a singular Hessian can be detected by the presence of outrageously large estimated standard errors. For example, in the cool-kid data, the popularity of a nominator is actually measured on a continuous scale that was dichotomized solely for the purpose of illustration of modeling and modeling concepts. Popularity is best entered as a numerical predictor; however, if popularity with 70 different values was entered into the model as a categorical predictor variable, the estimation fails. The Hessian is not singular (i.e., not “positive definite”). The estimated standard errors for many of the levels of popularity are 71,098.7 or 118,499.2; whereas, the standard errors for gender and race equal 0.35 and 0.19, respectively.

Multicollinearity can lead to the Hessian being nearly singular. This causes a problem for estimation because the estimation algorithms must take the inverse of the Hessian and there is no unique inverse for a singular matrix. In such cases, Fisher scoring will tend to perform better than Newton-Raphson since the expected value of the Hessian is used. Alternatively, a variation of Newton-Raphson, “ridge stabilized” Newton-Raphson, might also work<sup>10</sup>. Perhaps the best thing to do is to fix the source of the problem by modifying the model.

## 2.5 Assessing Model Goodness-of-Fit to Data

When making statistical inferences about populations, the data and the model are taken as given; however, uncertainty exists in the model specification itself (Burnham & Anderson 2002). Valid inference depends on using a model that is a good representation of data; therefore, choosing a model (or sub-set of models) should precede interpretation of parameter estimates.

Assessing model goodness-of-fit to data should never be based on a single statistic or statistical test. Evaluating model fit is best thought of as a process of gathering evidence for and against a model or a sub-set of plausible models. Three aspects that we consider here are examining global measures of goodness-of-fit to data, comparing competing models within a set of plausible models, and assessing local lack of fit.

General methods commonly used for assessing fit are described below. Other methods have been developed for particular types of models. The model specific methods are described in subsequent chapters in the context of particular models.

<sup>10</sup> Basically in a ridge stabilized regression, positive values are added to the diagonal of the Hessian to help keep it from being singular (–reference–).

### 2.5.1 Global Measures of Fit

Global measures of fit compare observed values of the response variable with fitted or predicted values. Two common measures are deviance ( $D$ ) and the generalized Pearson  $X^2$  statistic. Most computer programs for GLMs output values of both  $X^2$  and  $D$ .

Deviance is a global fit statistic that also compares observed and model fitted values; however, the exact function used depends on the likelihood function of the random component of the model. Deviance compares the maximum value of the likelihood function for a model, say  $M_1$ , and the maximum possible value of the likelihood function computing using the data. When the data are used in the likelihood function, the model  $M_y$  is saturated and has as many parameters as data points. The model  $M_y$  fits the data perfectly and gives the largest value possible for the likelihood. Deviance equals

$$D = -2[\ln(L(M_1)) - \ln(L(M_y))], \quad (2.27)$$

where  $L(M_1)$  and  $L(M_y)$  equal values of the likelihood function for models  $M_1$  and  $M_y$  (the data), respectively. If model  $M_1$  fits the data perfectly, the two values of the likelihood will be equal and  $Dev = 0$ . In practice where the model  $M_1$  is a summary of the information or structure in the data, the likelihood for  $M_1$  will be smaller than the likelihood using the observed data (i.e.,  $L(M_1) < L(M_y)$ ) and  $D > 0$ .

Another common global measure of fit is a generalized Pearson's  $X^2$  statistic,

$$X^2 = \sum_i \frac{(\mu_i - \hat{\mu}_i)^2}{\sqrt{\text{var}(\hat{\mu}_i)}}. \quad (2.28)$$

The greater the difference between observed and fitted values relative the the variance of the fitted values, the larger the value of  $X^2$ .

Both  $X^2$  and  $D$  can always be used as indices of fit. When data are normally distributed (i.e., the random component of the GLM is normal), then the sampling distribution of  $X^2$  and  $D$  are chi-square (McCullagh & Nelder 1989). For other distributions, the sampling distributions of  $X^2$  and  $D$  are approximately chi-square for "large" samples. In these cases, model goodness-of-fit can be assessed statistically.

For the large sample or asymptotic results to apply there must be a large number of individuals who have the same values on the variables in the model. How large is "large enough"? Consider the data as a cross-classification of variables including both discrete and/or essentially continuous variables. If there are 5 or more observations per cell (or for most cells), then the sampling distributions of  $X^2$  and  $D$  may be approximately chi-squared. This condition is easier to meet when all variables are discrete, but runs into problems when variables are (nearly) continuous. For example, in the cool-kid example, the cross-classification of type of kid (ideal or not) by popularity by gender by race has  $2 \times 2 \times 2 \times 2 = 16$  cells and the size of this table does not increase when additional subjects are added to the study. Adding more subjects to the study increases the number of observations per cell. If popu-

larity was treated a numeric or continuous variable, the size of the table would have been  $2 \times 70 \times 2 \times 2 = 560$  cells and would need a larger sample to have at least five observations per cell (there are only  $N = 526$  students in the study). Furthermore, adding an additional subject would likely increase the size of the table because the new observation may have a different value on the popularity measure.

When the sampling distribution of  $X^2$  and  $D$  are approximately chi-square, the degrees of freedom equal the number of observations minus the number of unique parameters; that is,

$$df = \text{number of observations} - \text{number of unique parameters.} \quad (2.29)$$

In our cool-kid example, since there are 8 possible logits and 4 estimated parameters, the model degrees of freedom equal  $df = 8 - 4 = 4$ . Since the smallest cell count is 11, the sampling distribution of Person's  $X^2$  and deviance are likely to be well approximated by a chi-square distribution. The deviance and Pearson's  $X^2$  for the probit model have  $p$ -values both equal .83, and those for the logit model both equal .81. These models seem to fit the data particularly well; however, nesting has been ignored. These statistical tests are misleading.

A further consideration when examining the fit of a model statistically is that when the sample is very large and the global fits statistics  $X^2$  and  $D$  have (approximate) chi-square sampling distributions, the lack of model fit to data may be significant even when the model is a good representation of the data. The values of  $X^2$  and  $D$  depend on sample size. This is related to the issue of practice versus statistical significant. Model selection should not depend on a single statistic or without regard to the problem as a whole.

### 2.5.2 Comparing Models

A researcher may be faced with selecting a “best” model from among a set of plausible and competing models. GLMs may differ in terms of the variables included in the linear predictor, the link function, or the random component. For example, should the probit or logit link be used for the psycholinguistics data? Should we use a gamma or inverse Gaussian distribution for the reaction time data and which link function should be used? For the peer nominations of bullies example, do we only need the bully scores as a predictor or should we include the additional predictor variables?

One aspect of the choice among models is based on substantive theory and the goal of an analysis. If one posits an underlying model that implies a probit model, then the probit should be selected. Psychological models that imply models for data are discussed in Chapters ?? and — maybe some other chapters —. For the bully dataset, if a researcher wishes to use a self report bully scale rather than the peer nominations, the models in a set should include the bully scores as predictor of the peer nominations

Other aspects of model selection take into consideration model goodness-of-fit to data and parsimony. There is a trade off between model goodness-of-fit to data and models that summarize the essential structure in the data. Models that are either too simple or too complex are not useful. A model that is too simple may not be a good representation of the information in the data and a model that is too complex does not provide enough of a summary of the information in the data to be useful. Although not desirable as a final model, the complex model provides a baseline against which to compare simpler models.

How models can be compared depends on whether the models are nested or non-nested. Nested models are special cases of more complex models. For example, Model 1 for the bully data that only includes the bully scale score is a special case of Model 2 that includes the bully scale score and four other predictor variables. In Model 1, the parameter estimates for the other four were implicitly set to 0. An example of non-nested models are the probit and logit models for the psycholinguistics data.

Likelihood ratio tests can be used to determine whether the fit of the model to data is statistically different between two models where one model is nested within the other. Information criteria, weigh both goodness-of-fit of the model to data and model complexity. Information criteria can be used to compare nested or non-nested models. In this section, likelihood ratio tests are discussed followed by information criteria.

### Likelihood Ratio Tests

Likelihood ratio tests are most often used to compare models with with different linear predictors because they require one model to be a special case of another. In a few cases, they can be used to compare models with different distributions, but this is more the exception than the rule. Later in Chapter(s) ?? and ?? examples will be given for this latter situation.

When one model is a special case of a more complex or “full” model, likelihood ratio tests can be used to assess whether the difference in model fit to data is statistically large. The likelihood ratio test is a conditional test in that given the full model fits the data, it tests whether the nested (simpler) model also fits the data. Let  $M_o$  represent the null or nest model that has restrictions placed on its parameters and  $M_1$  represent the full model. The likelihood ratio statistic equals

$$LR = -2[\ln(L(M_o)) - \ln(L(M_1))], \quad (2.30)$$

where  $L(M_o)$  and  $L(M_1)$  are maximum values of the likelihood function for the nested and full models.

To provide further insight into the  $LR$  test, the likelihood ratio test statistic can also be found by taking the differences between the two models’ deviances, because

$$LR = \underbrace{-2[\ln(L(M_o)) - \ln(L(M_y))]}_{D(M_o)} - \underbrace{(-2[\ln(L(M_1)) - \ln(L(M_y))])}_{D(M_1)}$$

$$= -2[\ln(L(M_0)) - \ln(L(M_1))].$$

Although the distributions of the global fit statistics may not be chi-square, the difference between them may be approximated by a chi-square distribution where the degrees of freedom equal the difference between the number of parameters in each model (i.e., the number of restrictions placed on the parameters of  $M_1$  to achieve  $M_0$ ).

As an example, consider the two models fit to the bully nomination data that are labeled as Model 1 (only bully scores as a predictor) and Model 2 (bully scores, gender, empathy, age, and score on a fight scale are all used as predictors). Model 1 is nested within Model 2 and  $LR = 1771.65 - 1701.92 = 69.73$  with  $\nu = 287 - 283 = 4$ . Comparing 69.73 to a chi-square distribution with  $\nu = 4$  gives  $p < .01$  that can be taken as evidence in favor of the more complex model, Model 2.

### Information Criteria

Information criteria can compare nested and non-nested models. The models can differ with respect to their linear predictors, link functions and distributions of the response variables. The two that are given here are Akaike's information criteria (*AIC*) and the Bayesian information criteria (*BIC*). These measures consider the distance between a "true" model and a model fit to the data. They try to balance goodness-of-model fit to data and model complexity.

The *AIC* equals

$$AIC = -2\ln(L(M_1)) + 2Q, \quad (2.31)$$

and the *BIC* equals

$$BIC = -2\ln(L(M_1)) + Q\ln(N), \quad (2.32)$$

where  $Q$  equals the number of parameters of a model and  $N$  the sample size. The smaller the value the better the model. Heuristically these measures can be thought of as penalizing a model based on their complexity; however, there is a theoretical basis for the penalties. The thorough discussion of these and other information criteria can be found in Burnham & Anderson (2002).

When models differ in terms of their linear predictors or link functions, computing *AIC* and *BIC* statistics is straightforward. For example the *AIC* and *BIC* statistics for the two models fit to the peer nomination data were computed using the statistics given in Table 2.6. For Model 1,  $AIC = -2(153.30) + 2(2) = -302.60$  and  $BIC = -2(153.30) + 2(2)\ln(289) = -295.27$ , and for Model 2,  $AIC = -364.34$  and  $BIC = -342.34$ . Comparing the two *AIC*s, the better model appears to be Model 2 and comparing the two *BIC*s, yields the same conclusion. This will not always be the case. Different information criteria can yield different conclusions.

Some caution is warranted when using *AIC* and *BIC* to compare models. The same data should be fit by models that are being compared<sup>11</sup>. This becomes relevant when some cases are excluded from a model due to missing values on some of the

---

<sup>11</sup> This is also true for *LR*

**Table 2.8** The full ln likelihood, AIC and BIC statistics for Model 1, Model 2 and two others fit to the peer nomination data where  $N = 289$ .

| Distribution | Link     | Predictors                                    | Number of parameters | Full ln(like) | AIC     | BIC     |
|--------------|----------|---|----------------------|---------------|---------|---------|
| Poisson      | ln       | bullyscale                                    | 2                    | -1075.36      | 2154.72 | 2162.05 |
| Poisson      | ln       | bullyscale,<br>gender, empathy,<br>age, fight | 6                    | -1040.49      | 2092.98 | 2114.98 |
| Poisson      | identity | bullyscale                                    | 2                    | -1044.53      | 2093.06 | 2100.39 |
| Normal       | identity | bullyscale                                    | 3                    | -931.08       | 1866.16 | 1873.49 |

variables. Attention should also be paid to ensure that the correct or *full* ln likelihood is used to compute *AIC* or *BIC* when comparing models with different distributions. For some distributions, the full logarithm of the likelihood has an additive constant that only depends on the data. Regardless of the link or what is included in the linear predictors, this additive constant is the same; therefore, some programs only use the *kernel* of the likelihood (i.e., the logarithm of the likelihood without the additive constant). As an example, consider the Poisson distribution. The full logarithm of the likelihood is

$$\begin{aligned} \ln(L(\mu; \mathbf{y})) &= \ln \left( \prod_{i=1}^N \frac{e^{-\mu} \mu^{y_i}}{y_i!} \right) \\ &= \sum_{i=1}^N y_i \ln(\mu) - N\mu - \underbrace{\sum_{i=1}^N \ln(y_i!)}_{\text{constant}}. \end{aligned}$$

When finding the  $\mu$  that maximizes the likelihood, the constant  $\sum_{i=1}^N \ln(y_i!)$  can be ignored and only the first two terms used (i.e.,  $\sum_{i=1}^N y_i \ln(\mu) - N\mu$ ) when estimating  $\mu$ .

In the bully nomination example, the ln likelihoods reported in Table 2.6 do not include the additive constants. In this example the additive constant equals 1, 228.66. The full likelihoods, AIC and BIC for Model 1, Model 2 and two additional models are contained in Table 2.8. In terms of *AIC* and *BIC*, it appears that the best model is the one with a normal distribution and identity link function. Very little weight should be placed on these results, because none of these models are acceptable. We found that neither Model 1 nor Model 2 fit the data, there is obvious overdispersion (making the normal distribution inappropriate), and that we have ignored the fact that the children in this study are nested within peer groups.

### 2.5.3 Local Measures of Fit

Part of determining whether a model is representative of the structure in the data includes examining local model miss-fit and looking for influential observations. Models may represent most of the data well, except for a sub-set of observations, and potential improvements to the model sometimes can be found by looking for systematic relationships in the residuals or identifying cases with particularly large residuals. Such observations may have too much influence in terms of goodness-of-model fit to data and/or on estimated parameters.

With respect to model fit to data, standardized residuals can be examined. Two common residuals are Pearson residuals and deviance residuals and these should be normally distributed. These residuals tend to be too small (i.e., variance too small) relative to the standard normal distribution. There are adjusted versions of both of these such that if the model fits the data well, the adjusted residuals should be distributed as  $N(0, 1)$ . In our cool kid example, the Pearson and adjusted Pearson residuals are reported in Table 2.4. Although the adjusted Pearson residuals are larger than the unadjusted, they are all small; that is, they are all between  $-1.96$  and  $1.96$  (the 2.75% and 97.25% percentiles of the  $N(0, 1)$ ).

Other measures that focus more on influential observations are based on the strategy of removing an observation, re-fitting the model, and computing a statistic. The statistics include global measures of fit (e.g.,  $X^2$ , deviance), regression coefficients (i.e., the  $\beta$ s), diagonal elements of the Hessian or Hat matrix, and others. Such statistics are computed for each observation. When the value of a computed statistic for a case deviates from the values computed for most of the other observations, the case may be an influential observation. Influential observations maybe outliers in the design space and/or values that are not fit well by the model.

## 2.6 Statistical Inference for Model Parameters

Statistical inference for model parameters primarily includes hypothesis testing and the formation of confidence intervals. We discuss Wald,  $F$ , and likelihood ratio tests for parameters, as well as formation of confidence intervals for parameter and predicted means. Confidence intervals give a sense of the precision of estimation.

### 2.6.1 Hypothesis Testing

Statistical inference of parameters can be performed using Wald tests,  $F$  tests and likelihood ratio tests. Wald and  $F$  tests require only fitting a single model and are useful as a first look at model parameters. Whether a Wald or  $F$  test is used depends on whether a  $\phi$  parameter is estimated. For example, in a Poisson regression the dispersion parameter is known (i.e.,  $\phi = 1$ ), so a Wald statistic would be used and it

would be compared to a chi-squared distribution. In normal linear regression where the dispersion parameter is estimated (i.e.,  $\phi = \sigma^2$ ), extra variability is introduced by having to estimate the variance. An  $F$  statistic should be compared to the  $F$ -distribution.

The likelihood ratio test applies to models whether  $\phi$  is known or estimated. Likelihood ratio tests are more powerful than Wald and  $F$ , because they use information from the likelihood at both the point of the null hypothesis and the maximum likelihood estimate. The likelihood ratio tests require estimating two models.

### Wald Statistics

A property of MLE is that the sampling distribution of parameter estimates is asymptotically (i.e., for large samples) approximately multivariate normally (MVN) distributed; that is,

$$\hat{\underline{\beta}} \sim \text{MVN}(\underline{\beta}, \underline{\Sigma}_{\hat{\beta}}). \quad (2.33)$$

The matrix  $\underline{\Sigma}_{\hat{\beta}}$  is generally not a diagonal matrix; the estimates  $\beta$ s are typically correlated. Given the sampling distribution of  $\hat{\underline{\beta}}$  in (2.33), hypothesis tests can be conducted and confidence intervals for individual parameters, sets of parameters, or linear combinations can all be computed.

Since  $\hat{\underline{\beta}}$  is MVN, then for the  $q$ th parameter,  $\hat{\beta}_q \sim N(\beta, \sigma_{\hat{\beta}_q}^2)$ . This fact can be used to test  $H_o : \beta_q = \beta_q^*$  by forming a  $z$ -statistic,

$$z = \frac{\hat{\beta}_q - \beta_q^*}{\text{ASE}_q}, \quad (2.34)$$

where  $\text{ASE}_q$  is the asymptotic standard error<sup>12</sup> of  $\hat{\beta}_q$ . The ASE is an estimate of  $\sigma_{\hat{\beta}_q}^2$ . If the null hypothesis is true, then  $z \approx N(0, 1)$ . For example, in the cool-kid example, the test statistic for the hypothesis that there is no effect of gender (i.e.,  $H_o : \beta_3 = 0$  versus  $H_a : \beta_3 \neq 0$ ) equals  $z = -.4859/0.1640 = -2.96$  and compared to a standard normal distribution has a  $p$ -value  $< .01$ .

The Wald statistic that tests an equivalent test the  $z$  statistic in (2.34) is

$$\text{Wald} = z^2 = \left( \frac{\hat{\beta}_q - \beta_q^*}{\text{ASE}_q} \right)^2. \quad (2.35)$$

When the null hypothesis is true, the Wald statistic in (2.35) has an approximate chi-square distribution with  $\nu = 1$  degree of freedom. Since the sampling distribution of a Wald statistic is chi-square, these statistics are sometimes referred to as “chi-square statistics.” The Wald statistics for each of the regression coefficients in both

<sup>12</sup> The ASEs are obtained in the estimation procedure (i.e., square root of the  $q$ th diagonal element of the inverse of the Hessian matrix) and are generally in the output from a program that fits GLMs to data

the probit and logit models are provided in Table ???. According to the Wald statistics in the cook-kid example, the effect of gender is significant (i.e.,  $\text{Wald} = (-2.96)^2 = 8.77$ ,  $\nu = 1$ ,  $p < .01$ ).

Although either the  $z$  or Wald can be used to test a single hypothesis, the Wald statistic in (2.35) is actually a special case of a more general Wald statistic. The more general form can be used to simultaneously test multiple hypotheses such whether as a set of parameters all equal zero, the equality between parameters, contrasts between them, and linear combinations of the parameters. The multivariate Wald statistics are also useful for testing whether a categorical predictor with  $K$  levels is significant rather than performing separate tests for each of the the individual  $\beta$ s that would require  $K - 1$  tests of the dummy or effect codes. Another use of the multivariate Wald statistic for categorical predictors is testing whether two (or more) levels have the same  $\beta$ .

The hypothesis for  $Q^*$  simultaneous tests is

$$H_o : \mathbf{C}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = \mathbf{0}, \quad (2.36)$$

where  $\mathbf{C}$  is a  $(Q^* \times Q)$  matrix of constants and  $\boldsymbol{\beta}$  is a  $Q \times 1$  vector of model parameters. If the null hypothesis in (2.36) is true<sup>13</sup>, then

$$\text{Wald} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)' \mathbf{C}' (\mathbf{C} \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} \mathbf{C}')^{-1} \mathbf{C} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \sim \chi_q^2. \quad (2.37)$$

If  $\mathbf{C}$  is  $(1 \times Q)$  vector with all 0s except for a 1 in the  $q$ th position, the test statistic in (2.37) reduces to (2.35) for testing the hypothesis  $H_o : \beta_q = \beta_q^*$ . Most computer programs have options to compute these statistics to test  $H_o : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ ; however, it is good to know what the program is doing and to be able to test hypotheses other than the default (i.e., specify a value for  $\boldsymbol{\beta}^*$  that was perhaps obtained from a previous study or implied by psychological theory).

In our cool-kid example, if we wanted to test whether the two variables popularity and gender were significant, the hypothesis would be  $H_o : \beta_1 = \beta_2 = 0$  and  $\mathbf{C}$  could be defined as

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Note that the matrix  $\mathbf{C}$  has as many columns as (non-zero) parameters in the model. The number of rows of  $\mathbf{C}$  equals the number of tests that in turn equals the degrees of freedom (i.e.  $\nu = Q^*$ ). Using this definition of  $\mathbf{C}$  for our psycholinguistics example, the joint null hypothesis is

$$H_o : \mathbf{C}\boldsymbol{\beta} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \quad (2.38)$$

<sup>13</sup> The matrix  $\mathbf{C}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}\mathbf{C}'$  is the covariance matrix for  $\mathbf{C}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ .

The alternative hypothesis is  $H_a : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{0}$ . To compute the test statistic in (2.37) requires an estimate of the covariance matrix of the parameter estimates. For the bully example, this was obtained from the output when fitting the model to data and equals

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \begin{pmatrix} 0.01952 & -0.01078 & -0.01161 & -0.01194 \\ -0.01078 & 0.02778 & -0.00116 & 0.00223 \\ -0.01161 & -0.00116 & 0.02691 & 0.00246 \\ -0.01194 & 0.00223 & 0.00246 & 0.02895 \end{pmatrix}.$$

Using  $\mathbf{C}$ ,  $\hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$  in (2.37), the Wald statistic for the psycholinguistics null hypothesis (2.38) equals 29.86 and compared to a chi-square distribution with  $\nu = 2$  (the number of rows in  $\mathbf{C}$ ) has a  $p$ -value  $< .01$ .

### F-Tests

For models where  $\phi$  is estimated, there is extra variability due to the estimation of  $\phi$  that needs to be taken into account. For a single parameter and testing  $H_o : \beta_1 = \beta_q^*$ , the test statistic is still (2.34); however, the sampling distribution of it is Student's  $t$ -distribution with  $\nu = N - Q$ . Alternatively, rather than using Student's  $t$ , we could square the test statistic (i.e., compute (2.35)) and compare the result to an  $F$ -distribution with  $\nu_1 = 1$  and  $\nu_2 = N - Q$ .

As an example, consider the social segregation in the classroom example where a normal linear regression model was fit to the data. Suppose that we wish to test whether the interaction between a multicultural classroom and ethnicity is significant,  $H_o : \beta_6 = 0$ . The test statistic equals  $0.1327/0.05936 = 2.24$  that compared to a  $t$ -distribution with  $\nu = 302 - 7 = 295$  has a  $p$ -value = .03.

Sets and linear combinations of parameters can be simultaneously tested using an  $F$ -test. To test the hypothesis that  $H_o : \mathbf{C}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \mathbf{0}$ , the test statistic equals the Wald statistic in (2.37) divided by the degrees of freedom for the test (i.e., by the number of rows in  $\mathbf{C}$ ); that is,

$$F = \frac{\text{Wald}}{Q^*} = \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}^*)' \mathbf{C}' (\mathbf{C} \hat{\boldsymbol{\Sigma}} \mathbf{C}')^{-1} \mathbf{C} (\boldsymbol{\beta} - \boldsymbol{\beta}^*)}{Q^*}. \quad (2.39)$$

As an example, consider the social segregation in the classroom example and we want to test the hypothesis that there is no interaction between ethnicity and racial distribution; that is,  $H_o : \beta_5 = \beta_6 = 0$ . To perform the test, first define the matrix of linear combinations for the test,

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Since we are testing two parameters, the matrix  $\mathbf{C}$  has two rows. Since there are a total of seven parameters (i.e.,  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)'$ ), the  $\mathbf{C}$  matrix has seven columns. An estimate of  $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$  is also required. From the output from fitting the normal linear regression model to the data, the following estimate of the covariance

matrix was obtained:

$$\hat{\Sigma} = \begin{pmatrix} 0.00221 & -0.00202 & -0.00021 & -0.00010 & 0.00075 & -0.00064 & -0.00021 \\ -0.00202 & 0.00466 & 0.00006 & 0.00015 & 0.00004 & 0.00063 & 0.00003 \\ -0.00021 & 0.00006 & 0.00134 & -0.00020 & -0.0004 & 0.00077 & -0.00003 \\ -0.00010 & 0.00015 & -0.00020 & 0.00143 & 0.00003 & 0.00022 & 4.8E-6 \\ 0.00075 & 0.00004 & -0.00037 & 0.00003 & 0.00344 & -0.0007 & 0.00020 \\ -0.00064 & 0.00063 & 0.00077 & 0.00022 & -0.0007 & 0.00352 & 0.00004 \\ -0.00021 & 0.00003 & -0.00003 & 4.8E-6 & 0.00019 & 0.00004 & 0.00143 \end{pmatrix}.$$

Using  $\mathbf{C}$ ,  $\hat{\beta}$  and  $\hat{\Sigma}_{\hat{\beta}}$  in (2.39), we obtain an  $F = 6.64$  that compared to an  $F_{2,295}$  distribution has a  $p$ -value  $< .01$ .

### Likelihood Ratio Tests

Likelihood ratio (LR) statistics can be used to test the same kinds of hypotheses as Wald and  $F$  statistics. The latter use information at the maximum of the likelihood; whereas, LR statistics are based on the value of the likelihood at the null hypothesis and at the maximum of the likelihood. As a result the LR statistics are more powerful.

A LR test involves placing restrictions on parameters of a model. The model without restrictions is the “full model” and the model with restrictions on parameters is the “nested” model. The nested model must be a special case of the full model. Restrictions include settings some regression coefficients equal to zero or placing equality restrictions on them. Although the former is the most common (i.e.,  $H_0 : \beta = 0$ ), the later are particularly useful for categorical predictor variables.

Suppose that we wish to test the hypothesis that  $Q^* (< Q)$  regression coefficients equal 0. To compute an LR statistic for this test requires the maximum of the likelihood function for the full model that includes the  $Q^*$  effects in which case the corresponding  $\beta$  are estimated, and the maximum of the likelihood for a nested model that excludes the  $Q^*$  effects which sets the  $\beta$  of interest equal to 0. The LR statistic equals

$$LR = -2(\ln(L(M_0)) - \ln(L(M_1))), \quad (2.40)$$

where  $L(M_0)$  is the maximum of the likelihood for the nested model and  $L(M_1)$  is the maximum of the likelihood of the full model. If the null hypothesis is true, then both likelihood are similar in value such that  $LR$  statistics close to 0. If the null is false, then the nested model will have a larger value of the likelihood and the  $LR$  statistic will be larger. When the null hypothesis is true, the sampling distribution of an  $LR$  statistic is chi-square with degrees of freedom  $v$  equal to  $Q^*$ .

Revisiting the cool-kid example, we re-test hypothesis for gender and popularity for the logit model; namely,  $H_0 : \beta_2 = \beta_3 = 0$ . The  $\ln$  likelihood for the full model is reported in Table 2.5 (i.e.,  $\ln(\text{likelihood}) = -450.1081$ ). For the null model, dropping the effects gender and popularity from the logit model yields  $\ln(\text{likelihood}) = -465.5907$ . The test statistic equals

$$\text{LR} = -2(-465.5907 - (-450.1081)) = 30.97$$

and compared to  $\chi_2^2$  has a  $p$ -value  $< .01$ . Note that the equivalent Wald test statistic for this hypothesis was 29.86. The Wald is smaller because the LR is more powerful.

### 2.6.2 Confidence Intervals

Interval estimates of parameters provide information regarding the precision of estimates by giving a range of plausible values for parameters and estimated means (function of parameters). Confidence intervals for parameters of GLMs are presented followed by confidence bands for estimated means are discussed.

#### Confidence Intervals for Parameter Estimates

Confidence intervals can be placed on parameters and linear functions of parameters. The method presented for confidence intervals for means can be adapted to provide confidence intervals for linear functions of parameters. In this section, the focus is on confidence intervals for parameters. The method for forming confidence intervals relies on the fact that maximum likelihood parameter estimates follow a normal distribution (i.e.,  $\hat{\beta}_q \sim N(\beta_q, \text{var}(\hat{\beta}_q))$ ).

For models where  $\phi$  is known such as the Poisson where  $\phi = 1$  or the binomial where  $\phi = 1/n$ , a  $(1 - \alpha)100\%$  confidence interval for  $\beta_q$  is

$$\hat{\beta}_q \pm z_{\alpha/2} \text{ASE}_q, \quad (2.41)$$

where  $z_{\alpha/2}$  is the  $(1 - \alpha/2)$ th percentile of the standard normal distribution,  $N(0, 1)$ . In the cool-kid example, a 95% confidence interval for the popularity,  $\beta_3$  is

$$0.7856 \pm 1.960(0.1667) \longrightarrow (0.49, 1.11). \quad (2.42)$$

When a link function other than the identity is used, a transformation of the end points of (2.42) is often more useful. In our logistic regression, a more useful confidence interval is one for the odds ratio. Since odds ratios equal  $\exp(\beta_3)$ , taking the exponential of the end points of a confidence interval for  $\beta$ , yields an interval for the odds ratio. In our psycholinguistics example, the 95% confidence interval for the odds ratio for the interaction is  $(\exp(0.49), \exp(1.11)) \longrightarrow (1.58, 3.04)$ .

For models where  $\phi$  is estimated such as the normal, inverse Gaussian or gamma distribution, a  $(1 - \alpha)100\%$  confidence interval for  $\beta_q$  can be formed as in (2.41) except that instead of using a value from the standard normal distribution, the  $(1 - \alpha/2)$ th percentile of the  $t$ -distribution with  $\nu = N - Q$  (i.e.,  $\nu = \text{Sample size} - \text{Number of parameters}$ ) should be used. Specifically,

$$\hat{\beta}_q \pm t_{(\nu, .975)} \text{ASE}_q. \quad (2.43)$$

For example, in the social segregation example in Section 2.3.1 where a normal linear regression was fit to data, a 95% confidence interval for the interaction parameter between ethnicity and a multicultural classroom,  $\beta_6$  is

$$0.1327 \pm 1.968(0.05936) \longrightarrow (0.02, 0.25), \quad (2.44)$$

where  $\hat{\beta}_6 = 0.1327$ ,  $t = 1.968$  is the 97.5th percentile of  $t$ -distribution with  $\nu = 302 - 7 = 295$ . and 0.05936 is the estimated standard error of  $\hat{\beta}_6$ .

### Confidence Bands for Predicted Means

In normal linear regression it is common to place confidence bands on regression lines (i.e., for  $E(\hat{y}_i)$ ). The same can be done for any GLM.

Putting confidence bands on  $E(\hat{y}_i)$  use two facts: (a) estimated regression coefficients follow a multivariate normal distribution as stated in (2.33), and (b) linear combinations of normally distributed random variables are themselves normally distributed random variables. The implication of these two facts is that

$$\hat{\eta}_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}} \sim N(\eta_i, \sigma_{\eta_i}^2). \quad (2.45)$$

where  $\mathbf{x}'_i = (1, x_{1i}, \dots, x_{Qi})$  is the  $i$ th row from the design matrix, and  $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_Q)$ . To make use of (2.45), an estimate of  $\sigma_{\eta_i}^2$  is needed.

Once a GLM is fit to data, an estimate of the covariance matrix of the  $\beta$ s,  $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$  is available. Using facts about linear combinations of random variables (in this case the  $\hat{\beta}$ s), the estimated variance of  $\hat{\eta}_i$   $\hat{\sigma}_{\eta_i}^2$  equals

$$\hat{\sigma}_{\eta_i}^2 = \mathbf{x}'_i \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} \mathbf{x}_i. \quad (2.46)$$

For models where  $\phi$  is known, a  $(1 - \alpha)100\%$  confidence interval for  $\eta_i$  is

$$\hat{\eta}_i \pm z_{\alpha/2} \hat{\sigma}_{\eta_i}. \quad (2.47)$$

When  $\phi$  is estimated, a  $(1 - \alpha) \times 100\%$  confidence interval for  $\eta_i$  is

$$\hat{\eta}_i \pm t_{\nu, \alpha/2} \hat{\sigma}_{\eta_i}, \quad (2.48)$$

where  $t$  is from Students  $t$ -distribution with  $\nu = N - Q$ .

Given the confidence interval for  $\eta$ , the confidence interval for  $E(\hat{y}_i | \mathbf{x}_i) = \mu_i$  is found by applying in inverse of the link function to the end point of the confidence interval for  $\eta$ . Specifically, for models where  $\phi$  is known, the confidence interval for  $E(\hat{y}_i | \mathbf{x}_i)$  is

$$g^{-1}(\hat{\eta}_i - z_{\alpha/2} \hat{\sigma}_{\eta_i}), \quad g^{-1}(\hat{\eta}_i + z_{\alpha/2} \hat{\sigma}_{\eta_i}) \quad (2.49)$$

For the case when  $\phi$  is estimated,  $z_{\alpha/2}$  is replace by  $t_{\nu, \alpha/2}$ .

As an example, consider the cool-kid example where a logit model was fit to the data. The model parameters are reported in Table 2.5, and the last two columns of Table 2.4 contain 95% confidence intervals for the probability  $\pi_i$  of an ideal student being nominated as cool. As an example, we find the confidence interval for a white boys with low popularity  $\pi_3$  by first computing the linear predictor. In this case,  $\mathbf{x} = (1, 0, 1, 0)'$ ,  $\hat{\boldsymbol{\beta}} = (0.1403, 0.7856, -0.4859, -1.4492)'$ , and

$$\hat{\eta}_3 = \mathbf{x}'\hat{\boldsymbol{\beta}} = 0.1403 - 0.4859 = -0.3456. \quad (2.50)$$

The estimated variance for  $\hat{\eta}_3$  equals

$$\begin{aligned} \hat{\sigma}_{\hat{\eta}_2}^2 &= (1, 0, 1, 0) \begin{pmatrix} 0.01952 & -0.01078 & -0.01161 & -0.01194 \\ -0.01078 & 0.02778 & -0.00116 & 0.00223 \\ -0.01161 & -0.00116 & 0.02691 & 0.00246 \\ -0.01194 & 0.00223 & 0.00246 & 0.02895 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \\ &= 0.0232, \end{aligned}$$

and the standard error for  $\hat{\eta}_2$  is  $\sqrt{\hat{\sigma}_{\hat{\eta}_2}^2} = \sqrt{0.0232} = 0.1523$ . The 95% confidence interval for  $\eta_2$  is

$$-0.3456 \pm 1.96(0.1523) \longrightarrow (-0.6442, -0.0470), \quad (2.51)$$

and the 95% confidence band for  $\pi_2$  is found by using the inverse transformation of the logit on the end points:

$$\left( \frac{\exp(-0.6442)}{1 + \exp(-0.6442)}, \frac{\exp(-0.0470)}{1 + \exp(-0.0470)} \right) \longrightarrow (0.34, 0.49). \quad (2.52)$$

In our cool-kid example, all 8 observed proportions fall within their 95% confidence bands (see Table 2.4). Since all of the proportions are well fit by the logit model and our global goodness-of-fit test statistics were not significant, it is tempting to conclude that the logit model is a good model for the cool-kid data; however, all of these statistical tests and confidence statements for the cool-kid data are not valid. The assumption of independent observations required for these tests and confidence statements has clearly been violated (i.e., students nested within peer groups within classrooms).

## 2.7 Summary

The GLM framework allows us to separate the decisions regarding how the response variable is distributed, what predictor variables should be included, and how the mean of the response is related to the linear function of the predictors. The decoupling of these decisions enables researchers to better capture the nature of the rela-

tionship between the response and predictor variables in an efficient manner. These three decisions are apparent by writing a GLMs in terms of the three components.

All generalized linear models are of the form:

$$\begin{aligned} y_i &\sim f(y|\mu_i, \phi) \\ g(\mu_i) &= \eta_i \\ \eta_i &= \sum_q \beta_q x_{iq} = \boldsymbol{\beta}' \mathbf{x}_i, \end{aligned}$$

where  $f(y|\mu_i, \phi)$  is the distribution function for the response variable,  $g(\cdot)$  is the link function,  $\eta_i$  is the linear predictor,  $\beta_q$  are regression coefficient, and  $x_{iq}$  are values of the predictor variables.

A summary of members of common distributions that are special cases of the natural exponential family are given in Table A.1 along with the type and range of response values, the canonical link functions and other information for each special case.

The examples used in this chapter illustrated the construction of GLMs for different types of data; however, the GLMs did not incorporate the nested structure of the data. The models lacked the ability to deal with dependent observations. This is remedied in the remainder of this text.

## Problems & Exercises

*These need to be fixed up, but they give a general idea of what data we have. These exercises can run through the book. More will be added.*

**2.1.** Give examples of response variables whose distribution might be best represented by the following distributions: (a) normal, (b) gamma, (c) inverse Gaussian, (d) beta, (e) binomial, and (f) Poisson.

**2.2.** Fit linear regression model to Allen data ignoring the fact that there eight level 2 units (STILL NEED TO GET NICOLE'S DATA)

**2.3.** Use the bully data set from Espelage et al. (2003) and fit a linear regression model to the data where empathy scores are the response variable and the bully scale score, the fight scale score, and gender are possible explanatory variables.

**2.4.** The study by Rodkin et al. (2007) of racial segregation in classrooms included three other measures of segregation based on sociometric data. The other measures were based on responses children made to questions about their peer groups, who they like, and who they like the least. Fit linear regression models to the measures

of segregation based on peer group affiliation using the same predictor variables as used in Section 2.3.1.

**2.5.** Use the racial segregation data and do problem ??, except use as the response the sociometric measure based on who children like the most.

**2.6.** Use the racial segregation data and do problem ??, except use as the response the sociometric measure based on children that a child dislikes.

**2.7.** A data set that is skewed... Perhaps some of Nicole's data on domestic violence from NIJ tech report. . . [once we have them written up for publication which is probably end of summer/early fall.](#)

**2.8.** A dichotomous response variable—use linear, logit & probit link and compare. [Could use bully data and dichotomize fight scale score into fight/no fight.](#)

**2.9.** The data from this problem comes from a study by Rodkin et al. (2006) with  $N = 526$  fourth to sixth graders who were nominated as being among the “coolest” kids in their class. The response variable is whether a tough kid is nominated as “cool.” The possible explanatory variables the child's race ( $race = 1$  if student is African American and 0 Caucasian), standard score for popularity of the nominator ( $pop$ ), child's peer group gender ( $gender = 1$  for boy group, 0 for girl group), and the location of the study ( $site = 1$  mid-west, 0 south).

- a Fit linear probability, logit and probit models to the data.
- b Which do you think is the best. Why?
- c Interpret the results of your favorite model.

**2.10.** Rather than using logit and probit models for the cool-kid data in Table 2.4, use Poisson regression to model the number of ideal kids nominated as cool.

- a Fit a model with main effects and two-way interactions.
- b Which model fit in part [a] is the same as the logit model given in the text? Show the relationship between the logit model and the Poisson regression model that are equivalent.
- b Fit a model with all main effect, two-way interactions and a three-way interaction. What do you notice? Explain.

## Appendix A

# The Natural Exponential Dispersion Family of Distributions

Different author's use slightly different notation for representing the natural exponential family. Our notation basically follows McCullagh and Nelder (1989) and we consider the two parameter version or the natural exponential dispersion family.

To introduce the exponential family, we will start with the familiar normal distribution, and put it into the basic or canonical form of the natural exponential distribution. The normal distribution function is

$$f(y; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left[ \frac{-(y - \mu)^2}{2\sigma^2} \right], \quad (\text{A.1})$$

where  $y$  is the response variable,  $\mu$  is the mean, and  $\sigma^2$  is the variance. Taking the  $\ln$  and  $\exp$  of the first term on the right-hand side of equation(A.1) and multiplying the squared term yields,

$$\begin{aligned} f(y; \mu, \sigma^2) &= \exp \left[ \ln \left( (2\pi\sigma^2)^{-1/2} \right) + \frac{2y\mu - \mu^2 - y^2}{2\sigma^2} \right] \\ &= \exp \left[ \frac{y\mu - \mu^2/2}{\sigma^2} + \left( \ln \left( (2\pi\sigma^2)^{-1/2} \right) - y^2/(2\sigma^2) \right) \right] \\ f(y; \theta, \phi) &= \exp \left[ \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right], \end{aligned} \quad (\text{A.2})$$

where  $b(\cdot)$  and  $c(\cdot)$  are functions. Sometimes, a weight parameter  $w$  is included such that rather than  $\phi$ , the dispersion is  $\phi/w$ .

For the normal distribution these functions are defined as

$$\begin{aligned} \theta &= \mu \\ b(\theta) &= \mu^2/2 \\ c(y, \phi) &= \ln \left( (2\pi\sigma^2)^{-1/2} \right) - y^2/(2\sigma^2). \end{aligned}$$

The parameter  $\theta$  is the *canonical or natural parameter* and it is a function of the mean. In the case of the normal distribution,  $\theta = \mu$ , and for the Poisson distribution  $\theta = \exp(\mu)$ . The function  $c(y, \phi)$  is a normalization term that ensures that probabilities sum to one. The function  $b(\theta)$  is known as the cumulant function. The mean and variance of the distribution can be obtained from the first and second derivatives of the likelihood with respect to  $\theta$ . This is shown below.

In some cases the dispersion parameter  $\phi$  may be weighted. As an example, consider the normal distribution where  $\phi = \sigma^2$ . If instead of single observations, we consider a sampling distribution of means of size  $N$  where observations are from the normal distribution with mean  $\mu$  and variance  $\sigma^2$ , then  $\phi = \sigma^2/n$ . Alternatively, consider the binomial distribution,  $\phi = 1/n$ , the number of trials.

As a second example of canonical form of the natural exponential family, consider the Poisson distribution. By taking the exponential and logarithm, we obtain

$$\begin{aligned} P(y; \mu) &= \frac{\mu^y e^{-\mu}}{y!} \\ &= \exp \left[ \ln \left( \frac{\mu^y e^{-\mu}}{y!} \right) \right] \\ &= \exp [y \ln(\mu) - \mu - \ln(y!)]. \end{aligned}$$

This last line has the same form as the canonical form of the exponential family given in (A.2), where

$$\begin{aligned} \theta &= \ln(\mu) \\ \phi &= 1 \\ b(\theta) &= \exp(\ln(\mu)) = \exp(\theta) \\ c(y, \phi) &= -\ln(y!). \end{aligned}$$

Given the canonical form of a member of the exponential family, the canonical link function is the function of  $\mu$  that yields the  $\theta$ . For example, with the Poisson distribution, the canonical link is the  $\ln$ , because  $\ln(\mu) = \theta$ . The canonical link of the normal distribution is the identity, where the mean is identical to the natural parameters (i.e.,  $\mu = \theta$ ). Furthermore, canonical links are those such that  $\theta = \eta$  in the GLM. In a GLM with a canonical link, there exist sufficient statistics for regression parameters (i.e., the  $\beta$ 's).

The specifications for the normal, Poisson and other common distributions that are members of the exponential dispersion family are given in Table A.1, includes ones that will be covered later in the text.

**Table A.1** Distributions in the natural exponential family covered in this chapter or in later chapters along with their canonical link functions.

| Distribution                   | Notation                   | Type of number | Range of $y$           | Canonical link | Dispersion parameter $\phi$ | Cumulant function $b(\theta)$ | Mean $E(y) = \mu = b'(\theta)$ | Variance function $b''(\theta)\phi$ | Probability density or mass $f(y; \mu, \phi)$ |
|--------------------------------|----------------------------|----------------|------------------------|----------------|-----------------------------|-------------------------------|--------------------------------|-------------------------------------|---|
| Normal                         | $N(\mu, \sigma^2)$         | real           | $-\infty < y < \infty$ | Identity       | $\sigma^2$                  | $\theta^2/2$                  | $\theta$                       | $\sigma^2$                          | (2.1)   |
| Gamma                          | $\text{Gamma}(\mu, \phi)$  | real           | $0 < y$                | Inverse        | $\phi$                      | $-\ln(-\theta)$               | $-1/\theta$                    | $\mu^2\phi$                         | (2.2)   |
| Inverse Gaussian               | $\text{IGauss}(\mu, \phi)$ | real           | $0 < y$                | $1/\mu^2$      | $\phi$                      | $-(-2\theta)^{1/2}$           | $(-2\theta)^{-1/2}$            | $\mu^3$                             | (2.3)   |
| Bernoulli                      | $\text{Bernoulli}(\pi)$    | binary         | 0, 1                   | Logit          | 1                           | $\ln(1 + e^\theta)$           | $e^\theta / (1 + e^\theta)$    | $\mu(1 - \mu)$                      | (2.6)   |
| Binomial                       | $\text{Binomial}(\pi, n)$  | integer        | 0, 1, ..., $n$         | Logit          | $1/n$                       | $\ln(1 + e^\theta)$           | $e^\theta / (1 + e^\theta)$    | $n\mu(1 - \mu)$                     | (2.8)   |
| Poisson                        | $\text{Poisson}(\mu)$      | integer        | 0, 1, ...              | Log            | 1                           | $e^\theta$                    | $e^\theta$                     | $\mu$                               | (2.10)  |
| Negative Binomial <sup>†</sup> | $\text{NegBin}(p, r)$      | integer        | 0, 1, ...              | Log            | $\phi$                      |                               |                                |                                     |   |

<sup>†</sup> The parametrization as a member of the exponential family is in terms of probability of  $y$  “failures” before the  $r$ th “success”. For the parametrization in terms of  $\mu$  and  $\phi$ ,  $\mu = (1 - p)/(r^{-1}p)$  and  $\phi = 1/r$  (Hilbe 2007).

### A.0.1 Likelihood, Score & Information

The likelihood of the natural exponential family equals (A.2) except that we consider  $y$  as given and  $\theta$  and  $\phi$  as unknown. The logarithm of the likelihood is easier to work, that is,

$$\begin{aligned} L(\theta, \phi|y) &= \ln(f(y; \theta, \phi)) \\ &= \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \end{aligned} \quad (\text{A.3})$$

To derive the mean and variance for the natural exponential family, two standard results from likelihood theory are used. The first is

$$E \left[ \frac{\partial L(\theta, \phi|y)}{\partial \theta} \right] = 0, \quad (\text{A.4})$$

where  $\partial L(\theta, \phi|y)/\partial \theta$  is known as the score or gradient and is often represented at  $\mathbf{\Delta}$ .

The second result is

$$E \left[ \frac{\partial^2 L(\theta, \phi|y)}{\partial \theta^2} \right] + E \left[ \left( \frac{\partial L(\theta, \phi|y)}{\partial \theta} \right)^2 \right] = 0. \quad (\text{A.5})$$

The first partial derivative of (A.3) with respect to  $\theta$  is

$$\frac{\partial L(\theta, \phi; y)}{\partial \theta} = \frac{y - b'(\theta)}{\phi}. \quad (\text{A.6})$$

To obtain the mean, (A.6) is used in (A.4) and the resulting equation solved for  $E(y)$ ; that is,

$$\begin{aligned} E \left[ \frac{y - b'(\theta)}{\phi} \right] &= 0 \\ E(y) &= \mu = b'(\theta). \end{aligned} \quad (\text{A.7})$$

The second partial derivative of (A.3) equals

$$\frac{\partial^2 L(\theta, \phi|y)}{\partial \theta^2} = \frac{b''(\theta)}{\phi}. \quad (\text{A.8})$$

The  $(Q \times Q)$  matrix with elements equal to second partial derivatives in (A.6) is the Hessian matrix and often represented as  $\mathbf{H}$ . Using (A.6) and (A.8) in (A.5) yields an equation that can be solved for the variance; that is,

$$E \left[ \frac{b''(\theta)}{\phi} \right] + E \left[ \left( \frac{y - b'(\theta)}{\phi} \right)^2 \right] = 0$$

$$-\frac{b''(\theta)}{a(\phi)} + \frac{\text{var}(\underline{y})}{\phi^2} = 0$$

$$\text{var}(\underline{y}) = b''(\theta)\phi. \quad (\text{A.9})$$

Equation (A.9) is known as the *variance function*.

An important property of the natural exponential family is that the variance depends on  $\theta$  and hence on the mean, as well as on  $\phi$ . Of all the member of the family of natural exponential distribution the variance function for the normal distribution is the only one that does not depend on its mean, because  $b''(\theta) = 1$  so that  $\text{var}(\underline{y}) = \sigma^2$ .

The values for  $\phi$ ,  $b'(\theta)$  and  $b''(\theta)$  are give in Table A.1 for various members of the natural exponential family.

### A.0.2 Estimation

The Newton-Raphson algorithm starts with some initial estimates of the parameters,  $\boldsymbol{\beta}^{[0]}$ , then iteratively up-dates the parameters until they do not change. The up-dating equation is

$$\boldsymbol{\beta}^{[t+1]} = \boldsymbol{\beta}^{[t]} - \mathbf{H}^{[t]-1} \boldsymbol{\Delta}^{[t]}, \quad (\text{A.10})$$

where  $\boldsymbol{\beta}^{[t+1]}$  is the vector of up-dated parameters,  $\boldsymbol{\beta}^{[t]}$  is the current estimate of parameters,  $\mathbf{H}^{[t]}$  is the Hessian matrix computed using the current estimates of the parameters, and  $\boldsymbol{\Delta}^{[t]}$  is the gradient computed using the current estimates of the parameters. The Hessian matrix provides an estimate of the covariance matrix for the parameters; namely,  $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = -\mathbf{H}^{-1}$ .



## Appendix A

### Index of Data Sets

| Data Set            | Type of Response | Section/Problem                                |
|---------------------|------------------|--|
| bully-nominations   | count            | Sect. 2.3.4, Prob. 2.3                         |
| cool-kids           | dichotomous      | Prob. 2.9                                      |
| Oydessy-of-the-Mind | continuous       | Sect. 2.3.2                                    |
| parents-n-kids      | continuous       |  |
| psycholinguistics   | dichotomous      | Sect. 2.3.3, Prob. 2.10                        |
| Racial-Segregation  | continuous       | Sect. 2.3.1 Prob. 2.4, Prob. 2.5,<br>Prob. 2.6 |



## References

- Agresti, A. (2002), *Categorical Data Analysis*, second edn, Wiley, NY.
- Agresti, A. (2007), *An Introduction to Categorical Data Analysis*, second edn, Wiley, NY.
- Burnham, K. P. & Anderson, D. R. (2002), *Model Selection and Multimodel Inference*, second edn, Springer, NY.
- Chhikara, R. S. & Folks, J. L. (1988), *The Inverse Gaussian Distribution*, Marcel Dekker, NY.
- Cohen, J. (1968), 'Multiple regression as a general data-analytic system', *Psychological Bulletin* **70**(6), 426–443.
- Demidenko, E. (2004), *Mixed Models: Theory and Applications*, Wiley, ?not in book?
- Dobson, A. J. (1990), *An Introduction to Generalized Linear Models*, Chapman and Hall, London.
- Espelage, D. & Holt, M. K. (2001), 'Bullying and victimization during early adolescence', *Journal of Emotional Abuse* **2**, 123–142.
- Espelage, D. L., Holt, M. K. & Henkel, R. R. (2003), 'Examiantion of peer-group contextual effects on agression during early adolescence', *Child Development* **74**, 205–220.
- Fahrmeir, L. & Tutz, G. (2001), *Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd Ed*, Springer.
- Fisher, R. A. (1928), 'The general sampling distribution of the multiple correlation coefficient', *Proceedings of the Royal Society of London. Series A*, **1**, 654–673.
- Fisher, R. A. (1934), 'Statistics in agricultural research', *Supplement to the Journal of the Royal Statistical Society* **1**, 51–54.  
**URL:** <http://www.jstor.org/stable/2983596>
- Hand, D. J., Daly, F., McConway, K., Lunn, D. & Ostrowki, E. (1996), *A Handbook of Small Data Sets*, Chapman & Hall, London.
- Hilbe, J. M. (2007), *Negative Binomial Regression*, Cambridge, NY.
- Kowal, A. K., Kramer, L., Krull, J. L. & Crick, N. R. (2002), 'Chidern's perceptions of fairness of parental perferential treatment and their socioemotional well-being', *Journal of Family Psychology* **16**, 297–306.
- Kowal, A. K., Krull, J. L. & Kramer, L. (2004), 'How the differential treatment of siblinds is linked with parent-child relationship quality', *Journal of Family Psychology* **18**, 658–665.
- Lindsey, J. K. (1997), *Applying Generalized Linear Models*, Springer, NY.
- McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models, 2nd Ed.*, Chapman and Hall, London.
- McCulloch, C. E. & Searle, S. R. (2001), *Generalized, Linear, and Mixed Models*, Wiley, New York.
- Nelder, J. & Wedderburn, R. W. M. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society, A* **135**, 370–384.
- Neyman, J. & Scott, E. L. (1948), 'Consistent estimates based on partially consistent observations', *Econometrica* **16**, 1–32.

- Rodkin, P. C., Farmer, T. W., Pearl, R. & Acker, R. V. (2006), 'They're cool: Social status and peer group support for aggressive boys and girls', **15**, 175–204.
- Rodkin, P. C., Wilson, T. & Ahn, H.-J. (2007), Social intergration between african american and european american children in majority black, marjority white, and multicultural elementary classrooms, in P. C. Rodkin & L. D. Harnish, eds, 'New Directions for Child and Adolescent Development', San Fransico, chapter 3, pp. 25–42.
- Scheffe, H. (1959), *The Analysis of Variance*, Wiley, ??
- Seshadri, V. (1998), *The Inverse Gaussian Distribution: Statistical Theory and Applications*, Springer, NY.
- Stine-Morrow, E. A. L., Miller, L. M. S., Gagne, D. D. & Hertzog, C. (2008), 'Self-regulated reading in adulthood', *Psychology and Aging* **23**, 131–153.
- Verbeke, G. & Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, Springer, NY.
- Verbeke, G., Spiessen, B. & LeSaffre, E. (2001), 'Conditional linear mixed models', *The American Statistician* **55**, 25–34.
- Wishart, J. (1934), 'Statistics in agricultural research', *Supplement to the Journal of the Royal Statistical Society* **1**(1), 26–61.  
**URL:** <http://www.jstor.org/stable/2983596>