Carolyn J. Anderson
Jay Verkuilen
Timothy R. Johnson

Applied Generalized Linear Mixed Models: Continuous and Discrete Data

For the Social and Behavioral Sciences

November 6, 2012

Springer

Contents

Part III Advanced and Special Topics

Notation

General conventions are

- Random variables are underlined, realizations are not.
- Small italic letters are scalars, small bold letters are vectors, capital bold letters are matrices.
- Greek letters indicate parameters.
- Hats are used to indicate estimates of parameters.
- ln is the natural logarithm

---

i	Index of level 1 observations (individual responses)
j	Index of level 2 observations (groups, clusters, etc.)
k	Index categories of discrete variables
q	Index of fixed effects regressors
r	Index of random effects regressors
n_j	Sample size within cluster j
N	Total sample size $\sum_j n_j$
K	Number of levels of categorical variable
Q	Number of fixed effects regressors
R	Number of random effects regressors
x	Constant scalar
$\mathbf{x}$	Constant vector
$\mathbf{X}$	Level 1 design matrix
z	Constant scalar
$\mathbf{z}$	Constant vector

$\mathbf{Z}$	Level 2 design matrix
$\mathbf{I}$	Identity matrix
$\mathbf{1}$	Vector of ones
y	Realization of outcome variable
$\underline{y}$	Random outcome variable
$\underline{\mathbf{y}}$	Random vector of outcome variables
$\underline{\mathbf{Y}}$	Random matrix of outcome variables
μ	Mean
σ^2	Variance
ρ	Correlation
$\text{var}(\underline{y})$	Variance of random variable $\underline{y}$
$\text{cov}(\underline{y})$	Variance of random variable $\underline{Y}$
θ	Generic parameter
$\hat{\theta}$	Estimator of scalar parameter θ
δ	Difference or indicator function
$\text{SE}(\hat{\underline{\theta}})$	Standard error of estimator $\hat{\theta}$
$\text{E}(\underline{x})$	Expectation of random variable x
$\text{E}(\underline{y}\|\underline{x})$	Conditional expectation
p	Observed proportion
f	Observed frequency
π_k	Probability
$\hat{\pi}_k$	Fitted probability
Pr	Probability
$\boldsymbol{\beta}$	Level 1 coefficient vector
$\underline{\boldsymbol{\gamma}}$	Level 2 coefficient vector
$\underline{\boldsymbol{\Sigma}}_j$	Within-group covariance matrix
ψ	Level 2 variance or covariance
$\boldsymbol{\Psi}$	Level 2 covariance matrix
$g(\cdot)$	Link function
η	Linear predictor
$f(y,\theta)$	Probability density/mass function
$b()$	Cumulant function of exponential family
$c()$	Normalization function of exponential family
ϕ	Dispersion parameter
$L(\theta;y)$	Likelihood function
$-2\ln L$	Negative twice the natural logarithm of the likelihood
G^2	Likelihood ratio chi square

X^2	Pearson chi square statistic
p-value	P-value
Bernoulli(π)	Bernoulli distribution
Beta(μ, ϕ)	Beta distribution
χ^2_ν	Chi-square distribution with ν degrees of freedom
Binomial(π, n)	Binomial distribution
F_{ν_1,ν_2}	F-distribution with degrees of freedom ν_1, ν_2
Gamma(μ, ϕ)	Gamma distribution
$N(\mu, \sigma^2)$	Normal distribution
Poisson(μ)	Poisson distribution
T_ν	Student's t-distribution with degrees of freedom ν

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[1]. The realization 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 this chapter, we go one step beyond the general linear model.

Under the general linear model, 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 are used to attempt 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 regres-

[1] Fisher (1928) was one of the first (if not the first) to realize 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).

sion, 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 the method in the program *Generalized Linear Interactive Modelling* or GLIM (Aitkin et al. 1989) in 1974 opened up the ability of researchers to design models to fit their data and to fit a wide variety of models, including those not previously proposed in the literature. Although the GLIM program is no longer supported, many common 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 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 specifying 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 *natural exponential family*, for short[2]. 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 a word 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

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

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, as well as to study differences between clusters.

In Section 2.2, the three components of a GLM are discussed in detail. In Sections 2.3, examples of continuous and discrete response 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 fitting 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. Technical details are covered in Appendix A. For more detailed descriptions of GLMs than given here see McCullagh & Nelder (1989), Fahrmeir & Tutz (2001), Dobson & Barnett (2008), 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 treated 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 values 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 **??**, beta-regression is presented for bounded scales where the response variable is a continuous.

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 distributions, and for discrete outcome variables the Poisson, Bernoulli, and binomial distributions. Other useful distributions some of which are in the exponential family and others that are not will be introduced later in the text as they are needed. The distributions introduced in this chapter are often reasonable representations of the distributions for many common measures found in psychological and behavioral science research.

A natural exponential dispersion distribution has two parameters, a *natural parameter* θ and a *dispersion parameter* ϕ. The parameter θ 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 θ is a function of the mean μ 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 μ and ϕ. 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 ϕ 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 A.

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 μ and variance σ^2. The probability density function for the normal distribution is

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

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

Examples of four normal distributions are given in Figure 2.1. Normal distributions are uni-modal and symmetric around the mean μ. 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 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 continuous 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 continuous responses is the gamma distribution.

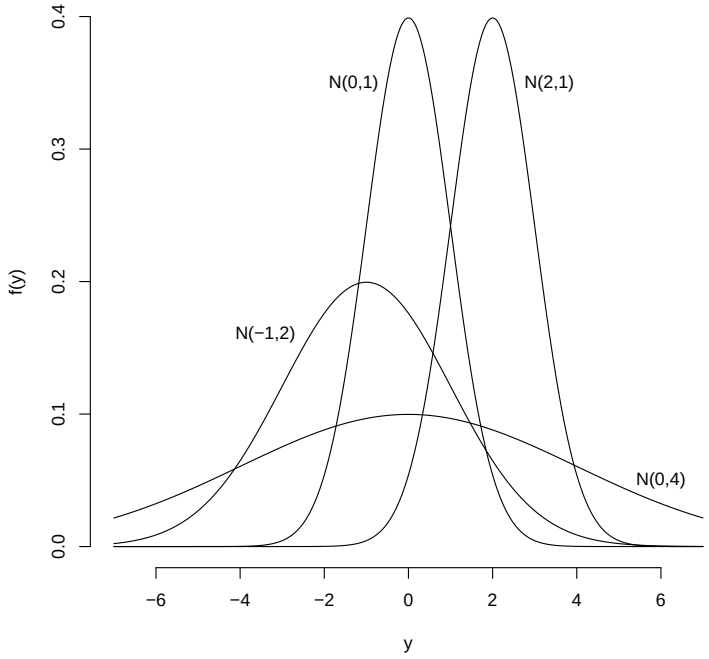

Fig. 2.1 Four examples of normal distributions $N(\mu, \sigma^2)$ with different combinations of means and variances.

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 μ and dispersion parameter ϕ. A particular gamma distribution will be represented as Gamma(μ, ϕ). The probability density function for a gamma distribution is

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

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

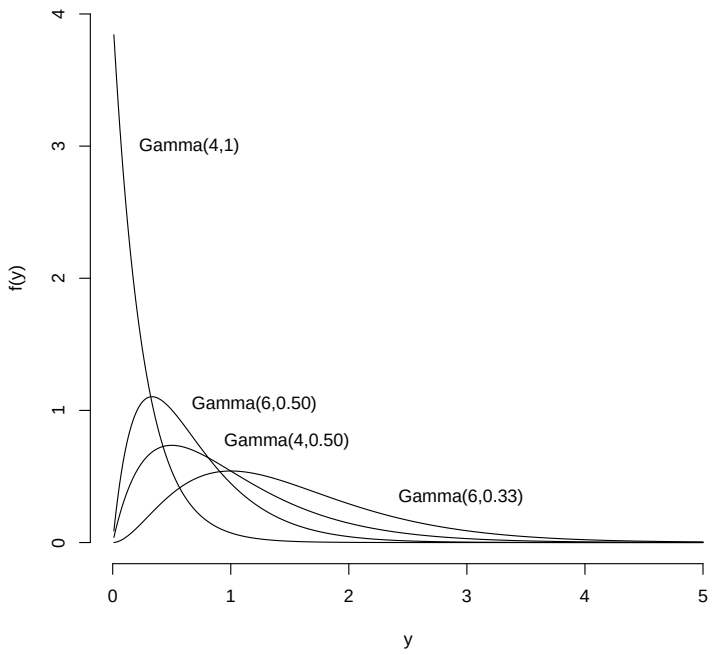

Fig. 2.2 Examples of four Gamma distributions Gamma(μ,ϕ) with different combinations of mean and dispersion parameters.

The natural parameter of the gamma distribution is $\theta = 1/\mu$. The parameters μ and ϕ can be any positive real number. To see the effect that μ and ϕ 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 μ, as ϕ gets smaller, the distribution becomes less skewed (e.g., Gamma$(4,1)$ and Gamma$(4,0.50)$). For a given ϕ, as μ gets larger, the distribution becomes less skewed (e.g., Gamma$(4,0.50)$ and 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,

$$\text{var}(\underline{y}) = \sigma^2 = \mu^2\phi.$$

To further illustrate this relationship, in Figure 2.3 the variance is plotted as a function of the mean for gamma distributions where ϕ 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.

Special cases of the gamma distribution correspond to other well known skewed distributions for continuous random variables. When $\phi = 1$ (e.g., 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 ν equals the degrees of freedom of the chi-square distribution. For example, in Figure 2.2, Gamma$(4,0.50)$ is a chi-square distribution with $\nu = 4$ and Gamma$(6,0.33)$ is chi-square with $\nu = 6$. 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 [3]

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 μ and dispersion parameter ϕ. Both μ and ϕ are positive real numbers. The probability density for the inverse Gaussian is

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

We will represent a particular inverse Gaussian distribution as IGauss(μ, ϕ).

[3] The inverse Gaussian 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 for the time to cover a unit of distance is inversely related to the function of the distance covered in a unit of time (Chhikara & Folks 1988).

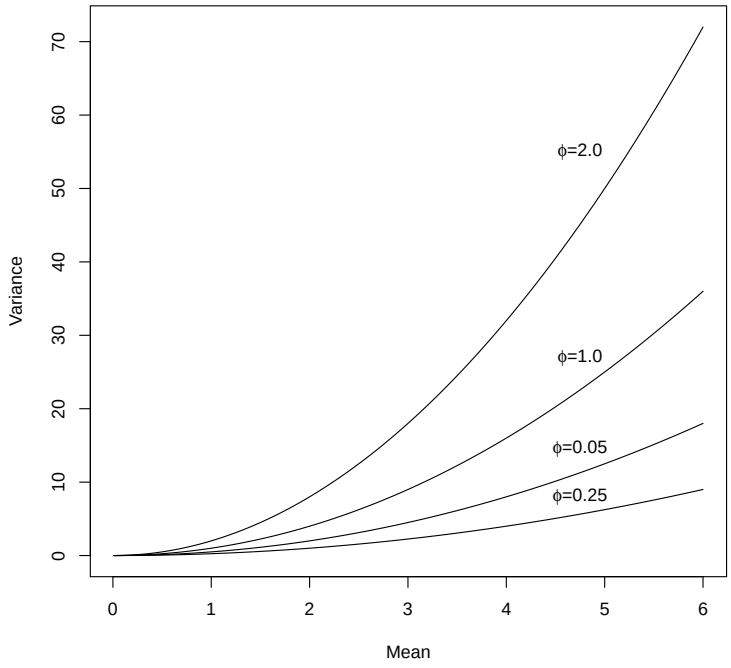

Fig. 2.3 The relationship between the mean and variance of gamma distributions where ϕ ranges from 2 to 0.25.

Examples of inverse Gaussian distributions are given for different values of μ and ϕ in Figure 2.4. The dispersion parameter essentially controls the shape of the distribution. For example, compare the IGauss(2,1.0), IGauss(2,0.1) and IGauss(2,0.01) that are given in Figure 2.4. As the dispersion parameter decreases, the inverse Gaussian distribution becomes more symmetric.

Some authors use the symbol σ^2 rather than ϕ to represent the dispersion parameter. We use ϕ here because σ^2 is typically used to represent variance; however, the variance of the inverse Gaussian distribution does not equal σ^2 nor ϕ. The variance of the distribution equals

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

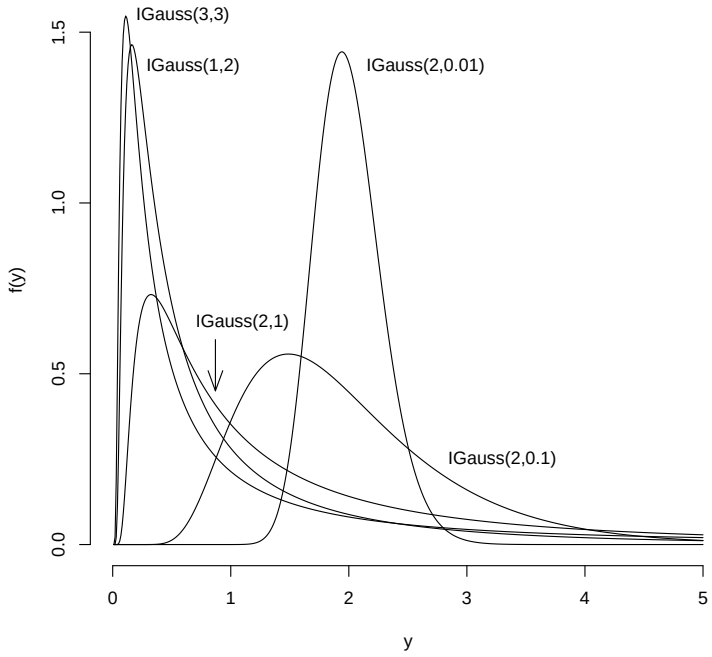

Fig. 2.4 Examples of inverse Gaussian distributions IGauss(μ, ϕ) with different mean and dispersion parameters.

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

2.2.1.4 Bernoulli Distribution

cja: We need to match notation between what's here and what's in binary chapter

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

$$\underline{y}^* = \begin{cases} 1 \text{ if an observation is in category one} \\ 0 \text{ if an observation is in category two} \end{cases}. \tag{2.4}$$

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

$$P(\underline{y}^* = y; \pi) = P(y; \pi) = \pi^y (1-\pi)^{1-y} \qquad \text{for } y = 0, 1. \tag{2.5}$$

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

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

In Figure 2.5, the curve showing this relationship is labeled $n = 1$. The variance reaches a maximum when $\pi = 0.5$, the point of maximum uncertainty.

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 π and a specific case of the binomial distribution will be represented as Binomial(π, n). When using the binomial distribution, interest is focused on estimating and modeling the probability π. The number of observations n 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

$$P(\underline{y} = y; \pi, n) = P(y; \pi, n) = \binom{n}{y} \pi^y (1-\pi)^{(n-y)} \qquad \text{for } y = 0, 1, \ldots, n. \tag{2.6}$$

The binomial coefficient

$$\binom{n}{y} = \frac{n!}{y!(n-1)!}$$

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

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

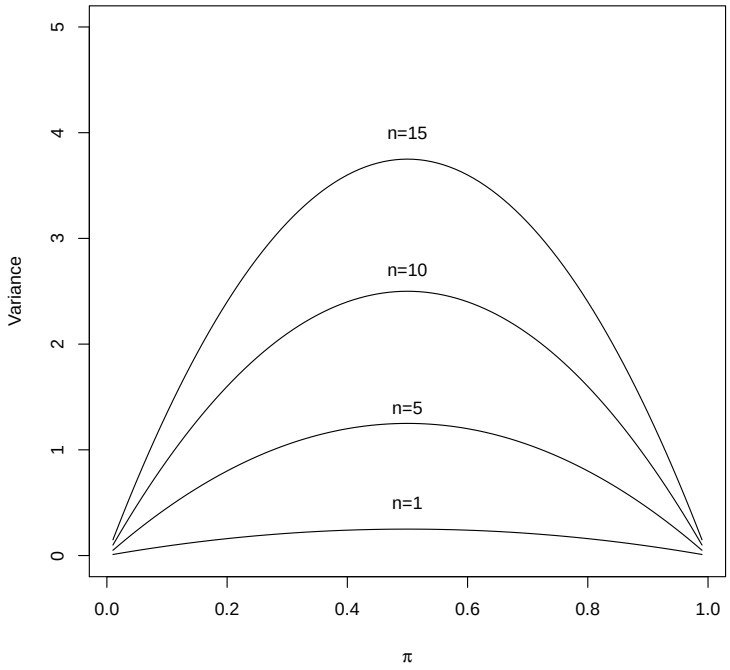

Fig. 2.5 Examples of the variance function for the binomial distribution with $n = 1, 5, 10, 15$.

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 Chapters 11 and **??**, 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. (2003), one way to measure the extent to which a child is a bully

is by peer nominations. In their 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 students are strictly speaking bounded by the number of students in the school. 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[4] μ 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., $y = 0,1,2,\ldots$). The probability that a Poisson random variable equals y is

$$P(\underline{y} = y;\mu) = P(y;\mu) = \frac{e^{-\mu}\mu^y}{y!} \qquad \text{for } y = 0,1,\ldots. \tag{2.7}$$

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 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

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

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 $x_1,\ldots,x_Q$ equal potential predictor variables. No restrictions are placed on the explanatory variables. They can be numeric or discrete. For discrete variables,

[4] Some authors use the symbol λ to represent the parameter of the Poisson distribution. Since the mean of the Poisson equals λ, we use μ as the parameter of the distribution.

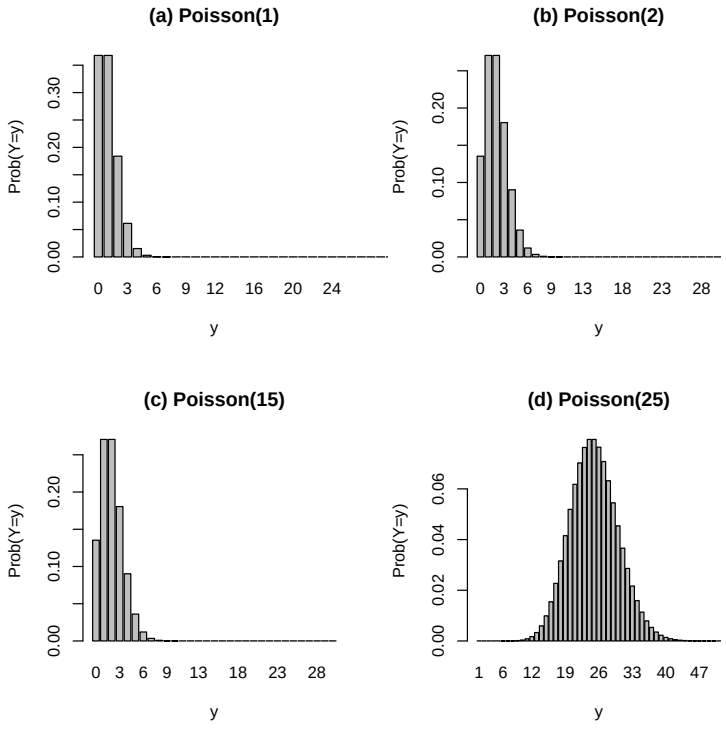

Fig. 2.6 Examples of Poisson distributions with different means.

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

$$\begin{aligned} \eta &= \beta_0 + \beta_1 x_1 + \ldots + \beta_Q x_Q \\ &= \boldsymbol{\beta}' \boldsymbol{x}, \end{aligned} \tag{2.8}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_Q)'$ is a vector of regression coefficients and $\boldsymbol{x} = (1, x_1, \ldots, x_Q)'$ is a vector of values on the predictor variables. Although η is a linear function of the xs, it may be nonlinear in shape. For example, η 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, Javdani et al. (2011) 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 can be modeled using a dummy code for whether a council existed in a particular judical circuit at each time point (i.e., at time point t, $x_{1t} = 1$ if council, $x_{1t} = 0$ if no council) and an interaction between the dummy code x_1 and time.

In normal linear regression models, most of the attention is given to η and finding the predictors or explanatory variables that best predict the mean of the response variable. This is also important in generalized linear models. Problems such as multicolinearity 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 $g(\ )$ connects the mean of the response variable to the linear predictor; that is,

$$g(\mu) = \eta. \tag{2.9}$$

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

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

The most natural and meaningful way to interpret model parameters is typically in terms of the scale of the data; namely, $\mu = g^{-1}(\eta) = g^{-1}(\beta_0 + \beta_1 x_1 + \ldots + \beta_Q x_Q)$. 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_is), then a distribution must be selected that describes the population distribution of transformed data. A transformation of the mean generally does not equal the mean of transformed values[5]; that is, $g(\mathrm{E}(\underline{y})) \neq \mathrm{E}(g(\underline{y}))$. As an example, suppose that we have a distribution with values (and probabilities)

[5] An exception is when $g(\mathrm{E}(\underline{y})) = \mathrm{E}(\underline{y}) = \mu = \eta$.

Table 2.1 Common link functions for different response variables. Note that Φ is the cumulative normal distribution.

Type of response variable		Link	$g(\mu)$	$g^{-1}(\eta)$
real	$\|y\| < \infty$	Identity	μ	η
real	$\|y\| < \infty$	Reciprocal	$1/\mu$ if $y \neq 0$ 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	$\ln(-\ln(1-\mu))$	$1-\exp(-\exp(\eta))$

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 $\ln(\mathrm{E}(\underline{y})) = \ln(4.1) = 1.411$; whereas, the mean of the logarithm equals $\mathrm{E}(\ln(\underline{y})) = 1.1\overline{74}$.

The value of the linear predictor η 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 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,

$$\mu = \eta.$$

Alternatively, the inverse or *reciprocal link*,

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

is a possibility. Such a link might be used with reaction times in which case η equals speed.

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 a continuous random variable as a 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 $\overline{0}$ to 1, but possible values for η may span the real numbers. Common distributions used for this purpose are the logistic, normal and the extreme value (Gumbal)

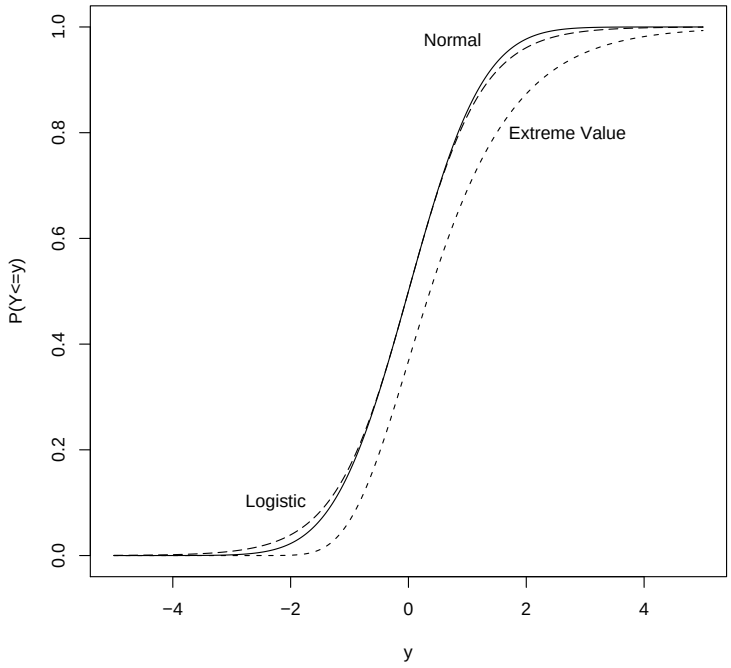

Fig. 2.7 Cumulative distribution functions for the standard normal, logistic (scale=.0625), and extreme value distributions.

distributions. The cumulative distributions for these three distributions 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 $P(y \leq y) = .5$ 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 the *logit* link and equals the natural logarithm of the ratio of μ divided by $1-\mu$; that is,

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

When y is a proportion (i.e., probabilities are being modeled), the logit is the logarithm of odds. A common alternative to the logit link for response variables

where $0 \leq y \leq 1$ is the *probit* link:

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

where Φ 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 typically minor in terms of model fit to data.

cja: Do we talk about all of these psychometric models?

In the case of particular psychometric models such as Thurstone's model for paired comparison, the Bradley-Terry-Luce choice model, the Rasch model from item response theory and random utility models, the choice of the link function (and systematic component) is implied by the assumptions of the psychometric model. These models are discussed in more detail in Chapters **??** and 11.

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 μ.

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 reasonable choice as a possible distribution; however, the canonical link for the gamma, the inverse (i.e., $1/\mu$), would yield negative predictions of reaction times when $\eta < 0$. With reaction times, an alternative link is the natural logarithm.

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

Table 2.2 Distributions in the natural exponential family covered in this chapter or in later chapters.

Distribution	Notation	Type of response	Range of y	Canonical link	Dispersion parameter ϕ	Variance function	Probability $f(y;\mu,\phi)$
Normal	$N(\mu,\sigma^2)$	real	$-\infty < y < \infty$	Identity	σ^2	σ^2	(2.1)
Gamma	Gamma(μ,ϕ)	real	$0 < y$	Inverse	ϕ	$\mu^2\phi$	(2.2)
Inverse Gaussian	IGauss(μ,ϕ)	real	$0 < y$	$1/\mu^2$	ϕ	μ^3	(2.3)
Bernoulli	Bernoulli(π)	binary	$0,1$	Logit	1	$\mu(1-\mu)$	(2.5)
Binomial	Binomial(π,n)	integer	$0,1,\ldots n$	Logit	$1/n$	$n\mu(1-\mu)$	(2.6)
Poisson	Poisson(μ)	integer	$0,1,\ldots$	Log	1	μ	(2.7)

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 reanalyze 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 $N = 302$ 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. A reasonable distribution for the measure of segregation is the normal distribution because the values of the response variable are (theoret-

ically) continuous real numbers and the distribution within classrooms is likely to be uni-model 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., $\texttt{male} = 1$ 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 was $\texttt{CMmaj} = 1$ if the majority of the students in the classroom are black, $\texttt{CMaj} = -1$ if the majority are white, and $\texttt{CMaj} = 0$ if there is no majority. The other code for classroom racial distribution is whether the classroom is multicultural where $\texttt{MultC} = 1$ for a multicultural classroom and $\texttt{MultC} = -0.5$ for either of the other classrooms.

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

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

where $\underline{\varepsilon}_i \sim N(0,\sigma^2)$. The equivalent model written as a GLM is

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

where $\boldsymbol{\beta}' = (\beta_0,\beta_1,\ldots,\beta_6)$ and $\boldsymbol{x}_i' = (1,\ \texttt{male}_i, \texttt{ethnicity}_i, \texttt{CMaj}_i, \texttt{MultC}_i, (\texttt{ethnicity}_i)(\texttt{CMaj}_i),\ (\texttt{ethnicity}_i)(\texttt{MultC}_i))$. Expressing the model as a GLM emphasizes the fact we are modeling the mean conditional on predictor variables[6]. This expression further emphasizes that three separate decisions were 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 the best link function, the linear predictor may not include all relevant variables (or transformations of them), or some combination of these three model components.

The estimated parameter are reported in Table 2.3. Notice that the parameters for child's ethnicity, the interaction between ethnicity and classroom majority

[6] For simplicity, we will drop $\boldsymbol{x}_i$ from the expression for the random component in subsequent models, except when we need to explicitly make a point regarding this conditioning (e.g., Section **??**.

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

						95% Confidence intervals	
Parameter	df	Estimate	Standard Error	t ($df = 295$)	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	2.23	.03	0.02	0.25

(`CMaj`), and the interaction between ethnicity and multicultural (`MultC`) are all significant. These results are not trustworthy because observations within classrooms are dependent and thus violate of the independence assumption required for statistical inference. The observations within classrooms are most likely positively correlated; therefore, the standard error estimates are too small leading to 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 7.

2.3.2 A Skewed Continuous Response Variable

The data for this example consists of a sub-set of data for $N = 149$ 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 for the next word. Of interest in this study is reading time measured in ml seconds between the presentation of a word and the hitting of the 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[7]. Both of these distributions are for non-negative continuous responses that are positively skewed . Which

[7] Another plausible distribution is the log-normal that is discussed in a later Chapter.

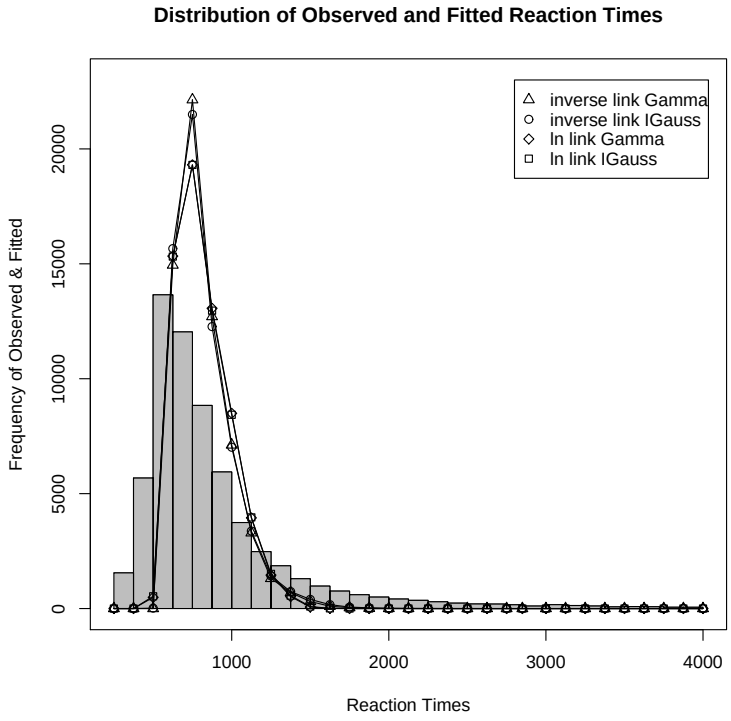

Fig. 2.8 Histogram of the observed distribution of time taken by elderly participants to read a word and the fitted values from models with the gamma distribution with inverse link (triangles), the inverse Gaussian with the inverse link (circles), the gamma with a log link (diamonds), and the inverse Gaussian with a log link (squares).

distribution is better for the data may depend more on the relationship between the mean and variance and may be 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 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 distribution is the inverse and 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 canonical link functions, the log link will also be considered because response times must be positive and the log link yields positive values.

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

$$\underline{rt}_i \sim f(\mu_i, \phi) \tag{2.10}$$

$$g(\mu_i) = \eta_i \tag{2.11}$$

$$\eta_i = \boldsymbol{\beta}' \boldsymbol{x}_i, \tag{2.12}$$

where $\underline{rt}_i$ is random responses (i.e., reaction time), $\boldsymbol{x}_i$ is a vector of values of the predictor variables, $f(\mu_i, \phi)$ is either the gamma or the inverse Gaussian distribution, and $\boldsymbol{\beta}$ is a vector of regression coefficients. Since each subject responded once to each word, the index i refers each subject by word combination. The links considered are $1/\mu$, $1/\mu^2$, and $\ln(\mu)$. For example, a gamma regression using the inverse link is

$$\underline{\texttt{rt}}_i \sim \text{Gamma}(\mu_i, \phi) \tag{2.13}$$

$$\begin{aligned}(1/\mu_i) &= \eta_i \\ &= \beta_0 + \beta_1\,(\texttt{syll}_i) + \beta_2\,(\texttt{logFreq}_i) + \beta_3\,(\texttt{intSB}_i) + \beta_4\,(\texttt{newLine}_i) \\ &\quad + \beta_5\,(\texttt{age}_i) + \beta_5\,(\texttt{NAART}_i) + \beta_6\,(\texttt{meanAcc}_i) + \beta_6(\texttt{SwRTsame}_i) \\ &\quad + \beta_7\,(\texttt{SWRTcost}_i),\end{aligned} \tag{2.14}$$

With the inverse link function, the mean reaction time for response i is $\mu_i = E(\underline{\texttt{rt}}_i|(\texttt{syll}_i), (\texttt{logFreq}_i), \ldots, (\texttt{SWRTcost}_i)) = 1/\eta_i$.

Before fitting models to data, we deleted outliers with reaction times greater than $4{,}000$ ml seconds, approximately 1% of the total $N = 63{,}357$ responses. Regardless of the distribution, when the link is $1/\mu^2$, the computing algorithm used to estimate the model parameters failed to converge. 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 both the $\ln(\mu)$ and $1/\mu$ links. The correlations between predictions with different links with the same distribution equal .989 for the gamma distribution and .986 for the inverse Gaussian.

The fitted values from the different models are plotted in Figure 2.8. The fitted values for the gamma with the inverse link (triangles), those from the inverse Gaussian with the log link (circles), those from the gamma with the log link (diamonds), and those from the inverse Gaussian with an inverse link (squares) are

plotted against reactions times. The models with the same link function are very similar to each other; however, all models fail to capture the nature of the observed distribution of reaction times.

In this data set, there are $N = 149$ subjects who each contribute 432 reaction times (i.e., one for each of the 432 words in the experiment). Subjects can be considered a random sample from the population of elderly adults, and words can also be considered a random sample of words from the English language. Trials of the experiment are nested both within subjects and words, both of which may contribute random variation to reaction times. This design is an example of *crossed random effects*. Potential random variation due to subjects and words should be simultaneously taken into consideration. A model that takes into account these two differences sources of variation could improve the fit of the model to the data, model differences between subjects, model

differences between words, and yield better standard errors for effect. We revisit this example in Chapter **??**.

cja: We could either put the follow-up to this data set in the chapter on bounded data (bounded from below & skewed) or we could have a chapter on skewed data, or reaction times or cross-random effects. What makes the most sense?

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 analyzed consist of measures on $N = 526$ fourth and fifth grade students. As an index of social status, children were asked who they think of as "cool". For this example, the response variable $\underline{\texttt{ideal}}_i$ equals the number of kids classified as a model or ideal student (i.e., popular or pro-social) among those who were nominated as being "cool". The response is dichotomous (i.e., whether the 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:

cja: I know that "cool." is grammatically correct but I like "cool". better. I just asked Carol Nickerson and she said that the the former is American and illogical and that the latter is British.

$$\texttt{Popularity}_i = \begin{cases} 1 \text{ high} \\ 0 \text{ low} \end{cases}, \qquad \texttt{Gender}_i = \begin{cases} 1 \text{ boy} \\ 0 \text{ girl} \end{cases}, \quad \text{and} \quad \texttt{Race}_i = \begin{cases} 1 \text{ black} \\ 0 \text{ white} \end{cases}.$$

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 π_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

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

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, $\underline{\texttt{ideal}}_i$ is binomially distributed. The linear probability model is

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

the logit model is

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

and the probit model is

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

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

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

Index	Nominating child's			Number of ideal students	Number of non-ideal students	Number of cases	Proportion of ideal students	Predicted Probabilities $\hat{\pi}_i$			Std. residuals from Logit model		95% confidence bands for logit π_i	
i	popularity	gender	race	who are "cool"	who are "cool"	n_i	p_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. This model can yield negative fitted values for the probabilities.

The estimated probabilities for the logit and probit models are nearly identical to each other and are very similar to those from the linear probability model. To show how similar the predictions are to each other and how well the models represent 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 are all very close to 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 Chapter10.

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 $\texttt{Popularity}_i$ and x_{2i} represent $\texttt{Gender}_i$ and x_{3i} represent $\texttt{Race}_i$. To emphasize that π_i is a function of the x_is, 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

$$\ln\left(\frac{\pi_i(x_{1i}, x_{2i}, x_{3i})}{1-\pi_i(x_{1i}, x_{2i}, x_{3i})}\right) = \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}. \tag{2.15}$$

The βs are most naturally interpreted in terms of odds ratios. Taking the inverse of the logarithm of (2.15) (i.e, the exponential) yields the odds that a cool-kid is an ideal student,

$$\frac{\pi(x_{1i}, x_{2i}, x_{3i})}{1-\pi_i(x_{1i}, x_{2i}, x_{3i})} = \exp[\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}]. \tag{2.16}$$

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

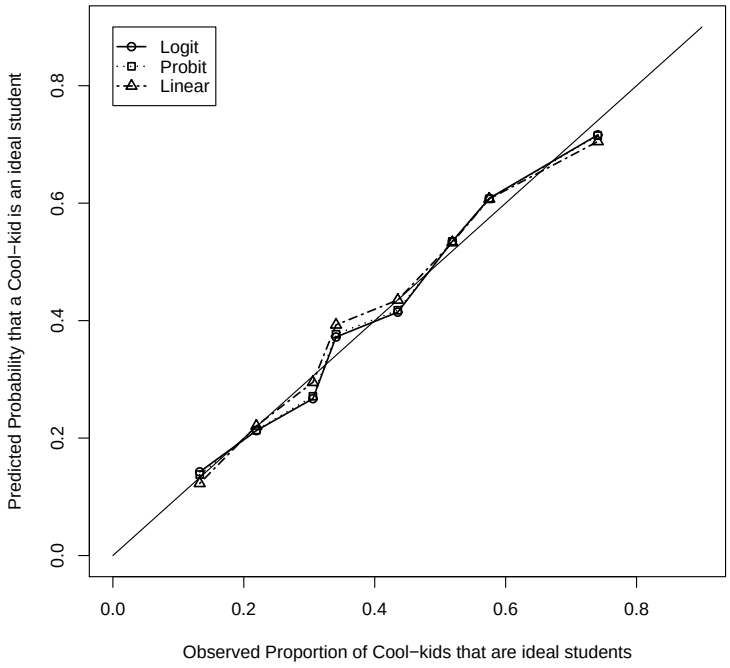

Fig. 2.9 Predicted probabilities from the logit, probit and linear probability model fit to the cool-kid data plotted against the observed proportions.

$$\begin{aligned}\frac{\pi((x_{1i}+1),x_{2i},x_{3i})}{1-\pi((x_{1i}+1),x_{2i},x_{3i})} &= \exp[\beta_0+\beta_1(x_{1i}+1)+\beta_2x_{2i}+\beta_3x_{3i}] \\ &= \exp[\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\beta_3x_{3i}]\exp(\beta_1). \quad (2.17)\end{aligned}$$

The relationship between the odds in (2.17) and in (2.16) is

$$\exp(\beta_1)\frac{\pi(x_{1i},x_{2i},x_{3i})}{1-\pi_i(x_{1i},x_{2i},x_{3i})} = \frac{\pi((x_{1i}+1),x_{2i},x_{3i})}{1-\pi((x_{1i}+1),x_{2i},x_{3i})}. \quad (2.18)$$

The value $\exp(\beta_1)$ equals how many times the odds are expected to change for a 1 unit increase in x_1. In other words, $\exp(\beta_1)$ is an odd ratios for a 1 unit change in x_1. The interpretation of β_1 does not depend of the specific value of x_{1i}, or either x_{2i} or x_{3i}.

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).

	Probit Model				Logit Model			
Effect	estimate	s.e.	Wald	p	estimate	s.e.	Wald	p
Intercept	0.0875	0.0860	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	—	—
df	4				4			
Deviance (p-value)	1.4944 (.83)				1.5955 (.81)			
X^2 (p-value)	1.4933 (.83)				1.5982 (.81)			
ln(likelihood)	−450.0575				−450.1081			

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 the odds for a child with low popularity. The odds that a boy nominates an ideal student are $\exp(-.4859) = 0.62$ times the odds for a girl, and the odds for a white student is $\exp(-1.4492) = 0.23$ 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 is $\exp(.4859) = 1/0.62 = 1.63$ 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."

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 Chapter **??** using more appropriate methods; however, we continue to use the "cook" kid–data in this chapter to illustrate GLM methodology.

cja: We can use these data to illustrate a 3-level model—works very nicely. This could be in Chapter **??** or a separate chapter on 3- and higher-level models.

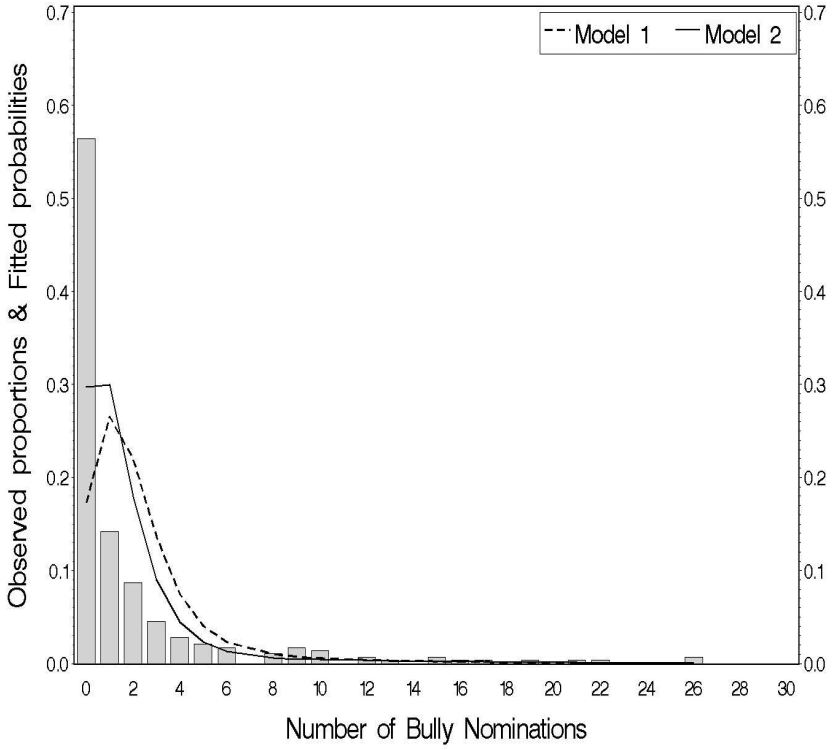

Fig. 2.10 Observed distribution and fitted values from two Poisson regression regression models. Model 1 only includes the bully scale as a predictor variable and Model 2 includes bully scale, gender, age and empathy.

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 (i.e., ln). The bully scale is the predictor variable in the systematic component. Our initial GLM for these data is

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

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

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

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

$$\widehat{\texttt{bullynom}}_i = \exp[-0.6557 + 0.8124 \, \texttt{bullyscale}_i].$$

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

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

and the model when a predictor is one unit larger is

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

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

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[8]. The model underpredicts 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.

[8] 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 2.6 Estimated parameters and fit statistics for simple and more complex Poisson regression models fit the peer nomination data (Espelage et al., 2004).

	Model 1					Model 2				
Parameter	Est.	SE	exp(β)	Wald	p	Est.	SE	exp(β)	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	$< .01$
df	287					283				
Deviance	1771.65					1701.92				
Pearson X^2	2968.56					2875.18				
ln(Likelihood)	153.30					188.17				

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

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).

Overdispersion can be caused by lacking all the 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 more predictor variables.

The variables gender, age, empathy (i.e., the perspective taking sub-scale of the Davis (1996) interpersonal reactivity index), and scores on a fighting self report measure 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 the dependency of observations within clusters. 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 10 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 more technical coverage is given in Appendices A and B.

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 distribution function for the Poisson distribution in equation (2.7) and the simple case of a single observation y. The likelihood function for the Poisson is

$$L(\mu;y) = \frac{e^{-\mu}\mu^{y}}{y!}. \tag{2.19}$$

Equations (2.7) and (2.19) are the same except the role of μ and y have been switched such that y is fixed and μ can vary in (2.19). For example, the likelihood given by (2.19) is plotted for $y = 2$ in Figure 2.11 (a). Notice that the maximum value of $L(\mu|2)$ occurs are $\mu = 2$.

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

$$L(\mu|y_1,\ldots,y_N) = \prod_{i=1}^{N} \frac{e^{-\mu}\mu^{y_i}}{y_i!}. \tag{2.20}$$

This is basically an application of the multiplicative rule in probability where the probability of independent events equals the product of the probabilities for each of the events. An example of this likelihood function is plotted in Figure 2.11 (b) for a sample of three observations $y_1 = 2$, $y_2 = 4$ and $y_2 = 4$ from a Poisson

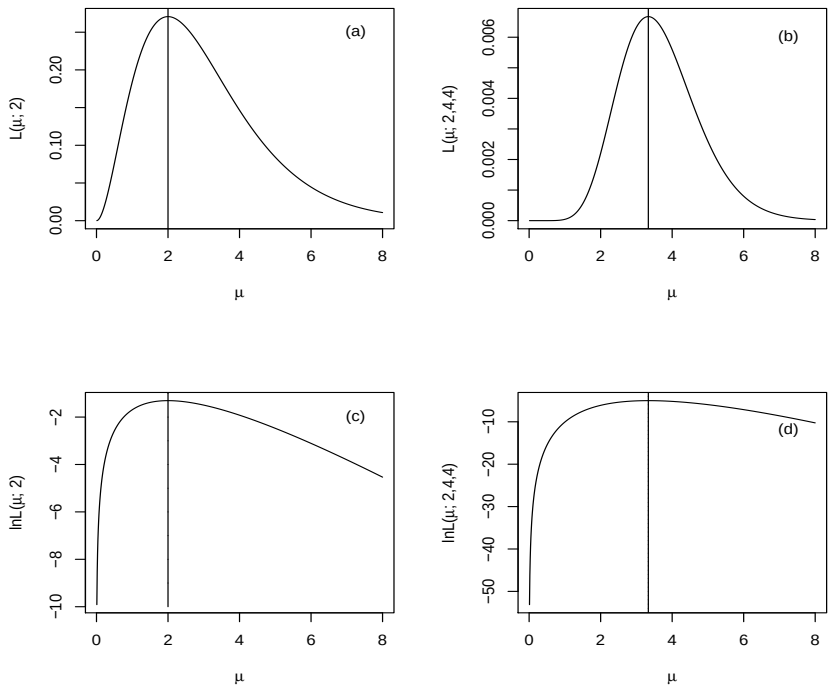

Fig. 2.11 The likelihood (top) and ln likelihood (bottom) for the Poisson distribution when $y = 2$ (left) and $y = 1, 4, 4$ (right) plotted as a function of possible values for the mean μ.

distribution. 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 μ conditional on predictor or explanatory variables. This model is $\mu_i = g^{-1}(\boldsymbol{\beta}'\boldsymbol{x}_i)$. The process and concepts are the same except that we replace μ_i by its model $g^{-1}(\boldsymbol{\beta}'\boldsymbol{x}_i)$ 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 μ_i equals $\hat{\mu}_i = g^{-1}(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_i)$.

Typically, estimation procedures use the natural logarithm (i.e., 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 Figure 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 (e.g., normal linear regression, ANOVA, ANCOVA), MLE of parameters requires an iterative algorithm. Common algorithms for finding maximum likelihood estimates are 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 updates the estimates on each iteration by solving a simple approximate problem. Parameters are repeatedly refined or up-dated 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[9] corresponds to the slope of the ln likelihood. There is one element in the score vector for each parameter to be estimated. When the maximum of the likelihood is achieved, the elements of the score vector (slopes) all equal zero. Consider the simple example in Figure 2.11 (c). When $\mu = 2$, the slope is flat (i.e., equal to 0). The Hessian matrix[10] 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 μ is closer to the MLE at $\mu = 2$.

The difference between Fisher scoring and Newton-Raphson algorithm 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 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

[9] The score vector or gradient is the vector of first partial derivatives of the ln likelihood function.

[10] The Hessian matrix is a matrix of second partial derivatives.

of a singular Hessian can usually be detected by the presence of outrageously large estimated standard errors. For example, in the cool–kid example, the popularity of a nominator is actually measured on a continuous scale that was dichotomized solely for the purpose of illustration of modeling with discrete predictor variables (i.e., a cross-classification or tabular data). Popularity is best entered as a numerical predictor; however, if popularity that has 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 parameters are larger than $100,000$; whereas, all other standard errors are less than 1.65.

A singular Hessian matrix 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[11]. 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.

[11] 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 *Dev*.

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

$$Dev = -2[\ln(L(M_1)) - \ln(L(M_y))], \tag{2.21}$$

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 $Dev > 0$.

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

$$X^2 = \sum_i \frac{(\mu_i - \hat{\mu}_i)^2}{\sqrt{\text{var}(\hat{\mu}_i)}}. \tag{2.22}$$

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 *Dev* 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 *Dev* are chi-square (McCullagh & Nelder 1989). For other distributions, the sampling distributions of X^2 and *Dev* are approximately chi-square for "large" samples. In these cases, model goodness-of-fit can be assessed statistically.

For the large sample (i.e., 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 whether they are 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 *Dev* 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 popularity 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 a very large sample would be required 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 *Dev* are approximately chi-square, the degrees of freedom equal the number of observations minus the number of unique parameters; that is,

$$df = \text{number of observations} - \text{number of unique parameters.}$$

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 *Dev* 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 global statistical tests could be 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 *Dev* 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 *Dev* depend on sample size. This is related to the issue of practice versus statistical significant. Model selection should not depend on a single statistic, without regard to the problem as a whole and the implications of the results.

2.5.2 Comparing Models

A researcher may be faced with selecting a "best" model from among a set of plausible or 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 cool-kid data? For the reaction time data, should we use a gamma or inverse Gaussian distribution and which link

function should be used? For the peer nominations of bullies, 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 11. In research on bullying in school, there is not a fully accepted way of measuring bullying; however, peer nominations of bullies is more acceptable than self reports. For the bully data set from Espelage et al. (2003), if it is claimed that the self report measure of bullying can be used in lieu of the peer nominations, then to support this claim the peer nomination should be used as a predictor variable of the self report measure.

cja: Ask Dorothy for reference

cja: or should this be the other way around?

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 (i.e., may capitalize on chance). Although not desirable as a final model, simple and complex models provide baselines against which to compare potential final 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 additional four predictors were implicitly set to 0. An example of non-nested models that have the same linear predictor but have different link functions and distributions for the response variable are reported in Table 2.8

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 different linear predictors because they require one model to be a special case of another. In some cases, they can be used to compare models with different distributions. This is possible when a distribution is a special case of a more general one, such

as chi-square is a special case of a gamma, the Poisson is a special case of the negative binomial distribution, or the beta-binomial distribution is a special case of the binomial distribution. In Chapter(s) 10 and **??** examples will be given for the latter situations.

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 nested model that has restrictions placed on its parameters and M_1 represent the full model. The likelihood ratio statistic equals

$$LR = -2[\ln(L(M_0)) - \ln(L(M_1))], \tag{2.23}$$

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

$$\begin{aligned} LR &= \underbrace{-2[\ln(L(M_0)) - \ln(L(M_y))]}_{Dev(M_0)} - \underbrace{(-2[\ln(L(M_1)) - \ln(L(M_y))])}_{Dev(M_1)} \\ &= -2[\ln(L(M_0)) - \ln(L(M_1)]. \end{aligned}$$

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 different 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 Baysian information criteria (*BIC*). These measures consider the distance between a "true" model and a model fit to the data.

The *AIC* equals

$$AIC = -2\ln(L(M_1)) + 2Q, \tag{2.24}$$

and the *BIC* equals

$$BIC = -2\ln(L(M_1)) + Q\ln(N), \tag{2.25}$$

where Q equals the number of parameters of a model and N the sample size. Smaller values of AIC and BIC indicate better models. Heuristically these measures can be thought of as penalizing a model based on it's complexity (i.e., balancing goodness-of-model fit to data and model 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[12]. This becomes relevant when some cases are excluded from a model due to missing values on some of the 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 a example, consider the Poisson distribution. The full logarithm of the likelihood is

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

[12] 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 (Espelage et al., 2004) 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`, `gender`, `empathy`, `age`, `fight`	6	−1040.49	2092.98	2114.98
Poisson	identity	`bullyscale`	2	−1044.53	2093.06	2100.39
Normal	log	`bullyscale`	3	−931.08	1868.16	1879.16
Normal	identity	`bullyscale`	3	−928.27	1862.54	1873.54

When finding the μ that maximizes the likelihood, the constant $\sum_{i=1}^{N} \ln(y_i!)$ can be ignored and only the first two terms (i.e., the kernel, $\sum_{i=1}^{N} y_i \ln(\mu) - N\mu$) are needed to obtain the maximum likelihood estimate of μ.

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 three 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. Little (in any) 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 were 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. The variance of these residuals tend to be 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 2.5th and 97.5th percentiles of the $N(0,1)$).

Other measures that focus more on the influence of single 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 βs), 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 the formation of confidence intervals for parameters 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; whereas, likelihood ratio tests require estimating two models. Whether a Wald or F test is used depends on whether a dispersion parameter ϕ 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 computed and compared to the F-distribution. The likelihood ratio test applies to models whether ϕ 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 at the maximum of the likelihood.

Wald Statistics

A property of MLE is that for large samples the sampling distribution of parameter estimates is asymptotically multivariate normal (MVN); that is,

$$\underline{\hat{\boldsymbol{\beta}}} \sim \text{MVN}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\hat{\beta}}). \tag{2.26}$$

The matrix $\boldsymbol{\Sigma}_{\hat{\beta}}$ is generally not a diagonal matrix; the estimates βs are typically correlated. Given the sampling distribution of $\hat{\boldsymbol{\beta}}$ in (2.26), confidence intervals for individual parameters or for linear combinations of them, and hypothesis tests can be conducted.

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

$$z = \frac{\hat{\beta}_q - \beta_q^*}{\text{ASE}_q}, \tag{2.27}$$

where ASE_q is the asymptotic standard error[13] of $\hat{\beta}_q$. The ASE is an estimate of $\sigma^2_{\hat{\beta}_q}$. If the null hypothesis is true, then $z \sim N(0,1)$. For example, in the logistic regression model of the cool-kid data, 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, p-value $< .01$.

The Wald test statistic[14] of $H_o : \beta_q = 0$ versus $H_a : \beta_q \neq 0$ equals the square of the z statistic in (2.27).

$$\text{Wald} = z^2 = \left(\frac{\hat{\beta}_q - \beta_q^*}{\text{ASE}_q} \right)^2. \tag{2.28}$$

When the null hypothesis is true, the Wald statistic in (2.28) 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 the probit and logit models are provided in Table 2.5. According to the Wald statistics for the logit model in the cook-kid example, the effect of gender,

[13] The ASEs are obtained in the estimation procedure (i.e., square root of the qth diagonal element of the inverse of the Hessian matrix) and are generally in the output from a program that fits GLMs to data

[14] A one-tailed test can be performed using the z-test statistic but only a two-tailed test can be performed when using the Wald statistic.

popularity and race are all significant (e.g., for gender, 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 (2.28) is actually a special case of a more general Wald statistic. The more general form can be used to simultaneously test multiple hypotheses such as whether the parameters in a set all equal zero, the equality of some parameters, contrasts between parameters, or any linear combination of the parameters. The multivariate Wald statistic is particularly useful for testing whether a categorical predictor with K levels is significant rather then performing separate tests for each of the the individual βs 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 β.

The hypothesis for Q^* simultaneous tests is

$$H_o : \boldsymbol{C}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \mathbf{0}, \tag{2.29}$$

where $\boldsymbol{C}$ is a $(Q^* \times Q)$ matrix of constants, $\boldsymbol{\beta}$ is a $(Q \times 1)$ vector of population parameters, and $\boldsymbol{\beta}^*$ is a $(Q \times 1)$ vector of hypothesized values of the parameters. If the null hypothesis in (2.29) is true, then

$$\text{Wald} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)' \boldsymbol{C}' (\boldsymbol{C} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} \boldsymbol{C}')^{-1} \boldsymbol{C} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \sim \chi^2_q. \tag{2.30}$$

Note that the matrix $\boldsymbol{C} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} \boldsymbol{C}'$ is the estimated covariance matrix for $\boldsymbol{C}(\underline{\hat{\boldsymbol{\beta}}} - \boldsymbol{\beta}^*)$.

If $\mathbf{C}$ is $(1 \times Q)$ vector with all 0s except for a 1 in the qth position, the test statistic in (2.30) reduces to (2.28) for testing the hypothesis $H_o : \beta_q = \beta_q^*$. Most computer programs have options to compute these statistics to test $H_o : \boldsymbol{C}\boldsymbol{\beta} = \mathbf{0}$; however, it is good to know what a program is doing and to be able to test hypotheses other than the default that is usually that a parameter equals $\mathbf{0}$ (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 logistic regression 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

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

Note that the matrix $\mathbf{C}$ has as many columns as (non-zero) parameters in the model. The number of rows of $\boldsymbol{C}$ 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 cool-kid example, the joint null hypothesis is

$$H_o : \boldsymbol{C\beta} = \begin{pmatrix} 0\ 1\ 0\ 0 \\ 0\ 0\ 1\ 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{2.31}$$

The alternative hypothesis is $H_a : \boldsymbol{C\beta} \neq \mathbf{0}$. o compute the test statistic in (2.30) requires an estimate of the covariance matrix of the parameter estimates. For the cool-kid example, this was obtained from the output when fitting the model to data and equals

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

Using $\boldsymbol{C}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}_{\hat{\beta}}$ in (2.30), the Wald statistic for the cool-kid null hypothesis (2.31) 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 ϕ is estimated, there is extra variability due to the estimation of ϕ that needs to be taken into account. For a single parameter and testing H_o : $\beta_q = \beta_q^*$, the test statistic is still (2.27); 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 (2.28)) 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; that is, $H_o : \beta_6 = 0$. 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$= .03$.

Linear combinations of parameters can simultaneously be tested using an F-test. To test the hypothesis that $H_o : \boldsymbol{C}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \mathbf{0}$, the test statistic equals the Wald statistic in (2.30) divided by the degrees of freedom for the test (i.e., by the number of rows in $\boldsymbol{C}$); that is,

$$F = \frac{\text{Wald}}{Q^*} = \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}^*)' \boldsymbol{C}' (\boldsymbol{C} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} \boldsymbol{C}')^{-1} \boldsymbol{C} (\boldsymbol{\beta} - \boldsymbol{\beta}^*)}{Q^*}. \tag{2.32}$$

As an example, consider the normal linear regression model of the social segregation in the classroom example and whether there is no interaction between

ethnicity and racial distribution; that is, $H_o : \beta_5 = \beta_6 = 0$. To perform this test, the matrix of linear combinations is defined as

$$\boldsymbol{C} = \begin{pmatrix} 0\,0\,0\,0\,0\,1\,0 \\ 0\,0\,0\,0\,0\,0\,1 \end{pmatrix}.$$

The matrix $\boldsymbol{C}$ has seven columns because there is a total of seven parameters (i.e., $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)'$). Since we are testing two parameters, the matrix $\boldsymbol{C}$ has two rows. An estimate of $\boldsymbol{\Sigma}_{\hat{\beta}}$ 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:

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.00221 & -0.00202 & -0.00021 & -0.00010 & 0.00075 & -0.00064 & -0.00021 \\ -0.00202 & 0.00466 & 0.00006 & 0.00015 & 0.00004 & 0.00063 & 0.00003 \\ -0.00021 & 0.00006 & 0.00134 & -0.00020 & -0.0004 & 0.00077 & -0.00003 \\ -0.00010 & 0.00015 & -0.00020 & 0.00143 & 0.00003 & 0.00022 & 0.00000 \\ 0.00075 & 0.00004 & -0.00037 & 0.00003 & 0.00344 & -0.0007 & 0.00020 \\ -0.00064 & 0.00063 & 0.00077 & 0.00022 & -0.0007 & 0.00352 & 0.00004 \\ -0.00021 & 0.00003 & -0.00003 & 0.00000 & 0.00019 & 0.00004 & 0.00143 \end{pmatrix}.$$

Using $\boldsymbol{C}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}_{\hat{\beta}}$ in (2.32), 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 Wald and F statistics only 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 model parameters. 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., $H_o : \beta = 0$), the latter are particularly useful for categorical predictor variables.

Suppose that we wish to test the hypothesis that $Q^*(< Q)$ regression coefficients all 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, and the maximum of the likelihood for a nested model that excludes the Q^* effects (i.e,. the βs corresponding to the Q^* effects are set equal to 0). The LR statistic equals

$$\text{LR} = -2(\ln(L(M_0)) - \ln(L(M_1))), \tag{2.33}$$

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 values are similar and the *LR* statistic is close to 0. If the null if false, then the nested model will have a smaller 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 ν equal to Q^*.

Revisiting the cool-kid example, we re-test hypothesis for gender and popularity for the logit model; namely, $H_o : \beta_2 = \beta_3 = 0$. 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 χ^2_2 has a p-value$< .01$. Note that the Wald test statistic for the same hypothesis (i.e., 29.86) was slightly smaller than the LR because the LR is more powerful than the Wald.

2.6.2 Confidence Intervals

Confidence intervals can be computed for parameters and linear functions of parameters. Interval estimates of parameters provide information regarding the precision of estimates by giving a range of plausible values for parameters and estimated means that are function of parameters. Confidence intervals for parameters of GLMs are presented followed by confidence bands for estimated means are discussed. The method presented for confidence intervals for means can be adapted to provide confidence intervals for linear functions of parameters.

Confidence Intervals for Parameter Estimates

The method for forming confidence intervals relies on the fact that maximum likelihood parameter estimates follow a normal distribution (i.e., $\underline{\hat{\beta}}_q \sim N(\beta_q, \text{var}(\hat{\beta}_q))$.

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

$$\hat{\beta}_q \pm z_{\alpha/2}\text{ASE}_q, \tag{2.34}$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$th percentile of the standard normal distribution, $N(0,1)$. For the logit model in the cool-kid example, a 95% confidence interval for the effect of `popularity`, β_3 is

$$0.7856 \pm 1.960(0.1667) \longrightarrow (0.49, 1.11). \tag{2.35}$$

When a link function other than the identity is used, a transformation of the end points of (2.34) is often more useful. For the logistic regression (i.e., logit model), 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 β_3, yields an interval for the odds ratio. In our cook-kid 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 ϕ is estimated such as the normal, inverse Gaussian or gamma distribution, a $(1-\alpha)100\%$ confidence interval for β_q can be formed as in (2.34) 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., $\nu =$ sample size $-$ number of parameters) should be used. Specifically,

$$\hat{\beta}_q \pm t_{(\nu,.975)} \text{ASE}_q. \tag{2.36}$$

For example, in the normal linear regression of the social segregation data in Section 2.3.1, a 95% confidence interval for the interaction parameter between ethnicity and a multicultural classroom, β_6, is

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

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 $\text{E}(\underline{\hat{y}}_i)$). The same can be done for any GLM.

Putting confidence bands on $\text{E}(\hat{y}_i)$ use two facts: (a) estimated regression coefficients follow a multivariate normal distribution as stated in (2.26), and (b) linear combinations of normally distributed random variables are themselves normally distributed random variables. The implication of these two facts is that

$$\underline{\hat{\eta}}_i = \boldsymbol{x}_i' \underline{\hat{\boldsymbol{\beta}}} \sim N(\eta_i, \sigma^2_{\hat{\eta}_i}). \tag{2.37}$$

where $\boldsymbol{x}_i' = (1, x_{1i}, \ldots, x_{Qi})$ is the ith row from the design matrix, and $\underline{\hat{\boldsymbol{\beta}}}' = \left(\underline{\hat{\beta}}_0, \underline{\hat{\beta}}_1, \ldots, \underline{\hat{\beta}}_Q\right)$. To make use of (2.37), requires an estimate of $\sigma^2_{\eta_i}$.

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

$$\hat{\sigma}^2_{\underline{\eta}_i} = \boldsymbol{x}_i' \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} \boldsymbol{x}_i. \tag{2.38}$$

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

$$\hat{\eta}_i \pm z_{\alpha/2} \hat{\sigma}_{\hat{\eta}_i}. \tag{2.39}$$

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

$$\hat{\eta}_i \pm t_{\nu,\alpha/2} \hat{\sigma}_{\hat{\eta}_i}, \tag{2.40}$$

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

Given the confidence interval for η, the confidence interval for $\mathrm{E}(\underline{\hat{y}}_i|\boldsymbol{x}_i) = \mu_i$ is found by applying the inverse of the link function to the end points of the confidence interval for η. For models where ϕ is known, the confidence interval for $\mathrm{E}(\hat{y}_i|\boldsymbol{x}_i)$ is

$$g^{-1}(\hat{\eta}_i - z_{\alpha/2}\hat{\sigma}_{\hat{\eta}_i}), \quad g^{-1}(\hat{\eta}_i + z_{\alpha/2}\hat{\sigma}_{\hat{\eta}_i}) \tag{2.41}$$

For the case when ϕ 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 π_i of an ideal student being nominated as cool. As an example, we find the confidence interval for a white boy with low popularity (i.e., $\pi_3 = \eta_3 = \boldsymbol{x}_3'\hat{\boldsymbol{\beta}}$) by first computing the linear predictor. In this case, $\mathbf{x} = (1,0,1,0)'$, $\hat{\beta} = (0.1403, 0.7856, -0.4859, -1.4492)'$, and

$$\hat{\eta}_3 = \mathbf{x}'\hat{\beta} = 0.1403 - 0.4859 = -0.3456.$$

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

$$\begin{aligned}\hat{\sigma}^2_{\hat{\eta}_3} &= (1,0,1,0) \begin{pmatrix} 0.01952 & -0.01078 & -0.01161 & -0.01194 \\ -0.01078 & 0.02778 & -0.00116 & 0.00223 \\ -0.01161 & -0.00116 & 0.02691 & 0.00246 \\ -0.01194 & 0.00223 & 0.00246 & 0.02895 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \\ &= 0.0232,\end{aligned}$$

and the standard error for $\hat{\eta}_3$ is $\sqrt{\hat{\sigma}^2_{\hat{\eta}_3}} = \sqrt{0.0232} = 0.1523$. The 95% confidence interval for η_3 is

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

and the 95% confidence band for π_3 is found by using the inverse transformation of the logit on the end points:

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

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 relationship 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:

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

where $f(y|\mu_i, \phi)$ is the distribution function for the response variable, $g(\ )$ is the link function, η_i is the linear predictor, β_q are regression coefficients, 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 **??** along with the type and range of response values, the canonical link functions and other information for each special case.

The material in this chapter on GLMs, estimation, assessment of model fit, and statistical inference are modified and expanded on in the subsequent chapters. 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 a 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 the 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 2.4, 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 2.4, 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. More skewed data: Reaction time data. Fit simpler models to the odyssey data by using fewer predictor variables. Indicate which distributions and links to try.

2.9. A dichotomous response variable—use linear, logit & probit link and compare.Could use bully data and dichotomize fight scale score into fight/no fight.

2.10. 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`$= 1$ of 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 probablity, logit and probit models to the data.
b Which do you think is the best. Why?
c Interpret the results of you favorite model.

2.11. 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 at cool.

a Fit a models 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.

2.12. We could use Jason Findley's data as an example for a log-linear model. The log-linear model is probably model for it; however, it could be viewed as a nested data and a multilevel model fit to it in the count chapter (as a continuation of this example).

a Fit a log-linear model to this data set.
b etc.

Chapter 3
Generalized Linear Mixed Effects Models

3.1 Introduction

A generalized linear mixed effects model (GLMM) is a GLM with fixed and random effects in the linear predictor. The term "mixed" in GLMM comes from the fact that both fixed effects and random effects are included in a model. The fixed effects are viewed as constant in the population; whereas, random effects are considered stochastic or variable[1]. The fixed effects convey systematic and structural differences in responses. The random effects convey stochastic differences between groups or clusters. The addition of random effects permits generalizations to the population from which clusters have been (randomly) sampled, accounts for differences between clusters, and accounts for within cluster dependency.

Modeling heterogeneity over clusters using only fixed effects often does not suffice. For example, the specific classrooms in the study by Rodkin et al. (2007) on social segregation are a sample of $N = 56$ third and fourth grade classrooms. Interest is not focused just on these 56 classes and the children within them, but on the larger population of similar classrooms and children. When only fixed effects are included in models, the results only generalize to the specific groups in the study. An additional problem of only including fixed effects is that parameters are needed for each cluster, and the addition of a clusters increases the number of parameters. Estimation of both parameters in common for all clusters and cluster-specific effects may lead to inconsistent estimates (Neyman & Scott 1948, Verbeke et al. 2001, Verbeke & Molenberghs 2000). Even if estimates are consistent, to achieve reasonable precision of the regression coefficients for fixed

[1] GLMMs can be thought of as a combination of frequentist and Bayesian notions (–reference–). From a frequentist perspective, parameters of models are viewed as fixed in the population and data are considered to be random. From a Bayesian perspective, parameters of models are viewed as random and data as fixed.

cluster-specific effects, the number of observations per cluster would need to be relatively large; however, in many studies cluster size is as small as 2 or 3. For example, in the studies by Espelage et al. (2003) and Rodkin et al. (2006) on bullying in schools, the clusters are peer groups and most groups are relatively small, and in studies by Kowal et al. (2002) and Kowal et al. (2004) on sibling relationships, families are the clusters and members of the cluster are two siblings within the family (i.e., the cluster size is two).

If the processes leading to responses of members within groups or clusters are essentially the same for each cluster, then fixed regression coefficients are the same for groups or clusters of observations. Although each group may only provide a small amount of information, combining the data over clusters to estimate these common effects leads to an increase in precision. GLMMs can handle unequal and small cluster sizes because data from all clusters are used to estimate fixed regression coefficients and their standard errors. This notion is referred to as *borrowing strength* (Kreft & de Leeuw 1998). Differences between the groups that are unaccounted for by the systematic or fixed effects are captured by random, unobserved variables. Instead of estimating specific values for these unobserved cluster-specific effects, the parameters of the distribution of the random effects are estimated (i.e., means, variances and covariances).

In terms of a multilevel random effects model, the Level 1 regression model can be any GLM discussed in the previous chapter; however, the regression coefficients in the linear predictor are potentially random. For example in the study by Espelage et al. (2003) on bullying where children are nested within peer groups and the number of nominations a child receives is the response variable, a GLM appropriate for count data should be selected (e.g., Poisson distribution with a log link), but the intercept and regression coefficients for predictors may vary randomly over peer groups (i.e., the clusters). The variability of parameters in the Level 1 regression model accounts for between cluster differences and dependency of observations within clusters.

To study how clusters differ (e.g., classrooms, elderly adults, peer groups) and how context effects members of clusters, the regression coefficients from the Level 1 models themselves are treated as response variables in linear regression models where the predictor variables are measured attributes of the clusters and random residual effects. The models for the random coefficients can capture systematic (fixed) and stochastic (random) differences between clusters. The models for the regression coefficients are the Level 2 models. The parameters of the Level 1 and 2 models are estimated simultaneously. As shown in this chapter, cluster-specific random terms in the Level 2 models account for within cluster dependency.

In Sections 3.2 and 3.3 a multilevel linear regression model for normally distributed response variables is developed starting with a paired-dependent t-test. In Section 3.4, a general notation is presented for GLMMs that can handle nonnormal repsonse variables, different link functions and more predictor variables is

presented. This general form is subsequently used in an example of a multilevel logistic regression model in Section 3.5. These examples are used to introduce additional notation, concepts and the basic properties of the models. To wrap up this introduction to GLMMs, the definition and distinctions between cluster-specific, population average models, and marginal models is explained. These distinctions have important implications in the modeling of data.

3.2 Normal Random Variables

Random effects ANOVA is an example of a GLMM and more specifically it is an example of a *hierarchical linear model*. A hierarchical linear model or HLM refers to the case of multilevel random effects models for normally distributed response variables. In an HLM, multiple linear regression models for normally distributed response variables are proposed for the Level 1 and Level 2 models. Inserting the Level 2 models for the regression coefficients back into the Level 1 model leads to a combined or *linear mixed model* (LMM) that includes both fixed and random effects. HLMs and the implied LMMs are introduced by example using a simple paired dependent t-test. Subsequently, this model is extended by adding of a second variable to model between cluster variability.

3.2.1 Paired Dependent t-test

The pair dependent t-tests focuses on whether two groups have the same mean response; however, the responses between the two groups are dependent. Dependency can result when two measures are recorded on the same individual (e.g., math and reading scores, pre-test and post-test measures). In the case of repeated measures, the individual is a "cluster"; that is, responses are nested within individuals. Dependency can also result when there is a single measure that is taken on both the members of a pair (e.g., husband and wife, two siblings in a family). In this case, the pair is a cluster (e.g., couple, family). Independence over pairs or clusters is assumed.

As an example, we use data from a study on anorexia taken from Hand et al. (1996) and perform a paired dependent t-test. The data consist of weights of $N = 72$ girls measured in pounds before and after treatment for anorexia. This is an example of a longitudinal study with only two time points. The data are plotted in Figure 3.1 with histograms for before and after treatment weights, a histogram of the change in weight, and a plot of weight by measurement occasion with data for each girl connected by a line. From the histograms, it appears that overall the

cja: Since Jay also uses this in binary chapter we should use the same reference(s).

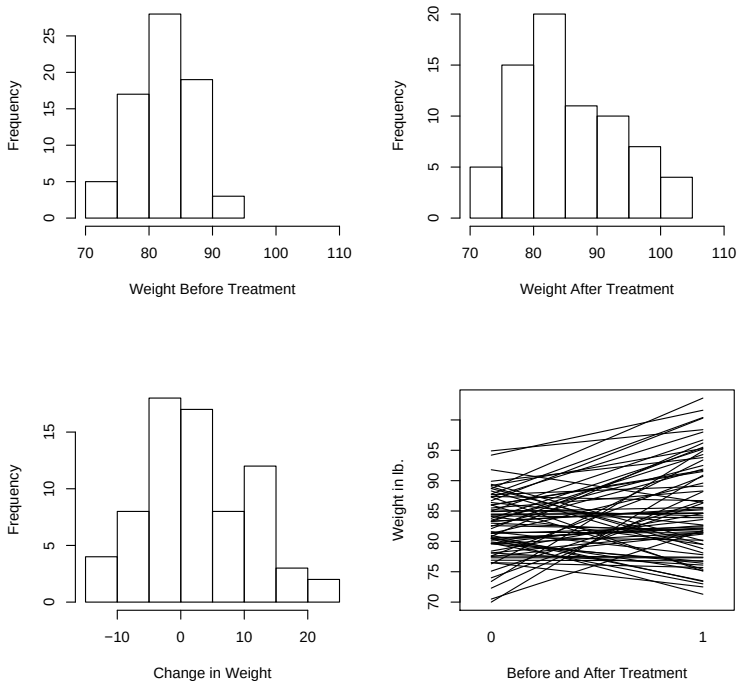

Fig. 3.1 Histograms and plot of weight in kilograms of girls before and after treatment for anorexia. Lines in plot in lower right corner connect the weights each girl.

girls are heavier after treatment. The variability in weights appears to be greater after the treatment than before. The line plot explicitly shows the change in weight for individual girls and shows that some girls' weight stayed the same, some decreased, and others increased. Also apparent from the line plot is the presence of considerable variability between girls both in terms of weight at the beginning of the study and how much their weight changed. The paired dependent t-test only assesses whether on average girls' weights changed over the course of the study.

The index j is used to indicate a cluster and the index i the observations within a cluster. In our anorexia example, y_{ij} equals the weight of girl j measured on occasion i (i.e., $i = 1$ for before treatment and $i = 2$ for after). The null hypothesis of the paired dependent t-test is $H_o : \mu_1 = \mu_2$. We assume that $\underset{\sim}{y}_{ij} \sim N(\mu_j, \sigma_j^2)$ and girls' weights are independent over girls. The two weight measurements for each girl are expected to be dependent across time. The test can be computed

Table 3.1 Sample statistics and paired dependent t-test for anorexia data set (Hand et al. 1996).

Before	After	Difference scores
$\bar{y_1} = 82.41$	$\bar{y_2} = 85.17$	$\bar{d} = 2.76$
$s_1^2 = 26.86$	$s_2^2 = 64.56$	$s_d^2 = 63.74$
	$\text{cov}(y_{1j}, y_{2j}) = 13.84$	
$se_{\bar{y_2}-\bar{y_1}} = \sqrt{(26.86+64.56-2(13.84))/72} = 0.94$		$se_{\bar{d}} = \sqrt{63.74/72} = 0.94$
$t = (85.17-82.41)/0.94 = 2.94$		t=2.76/0.94=2.94

using either sample statistics for the before and after measurements or by constructing difference scores for each girl (i.e., $d_j = y_{2j} - y_{1j}$). Table 3.1 contains sample statistics for before treatment, after treatment and difference scores, and it also contains computation of the t statistic. Comparing the test statistic, $t = 2.94$, to Students t-distribution with $\nu = 72 - 1 = 71$, $p < .01$. The data support the conclusion that on average girls gained weight after their treatment.

3.2.2 Paired t-test as a Random Intercept HLM

Before representing the paired t-test as a GLMM (or more specifically as an LMM), we start with a simple regression model that corresponds to a random effects ANOVA. This is the Level 1 model for the data. The response variable is $\underline{y}_{ij}$ that represents girl j's weight on measurement occasions i. The predictor variable in this case is measurement occasion or time that is dummy coded as $x_{1j} = 0$ for before and $x_{2j} = 1$ for after treatment. To help distinguish between parameters that are cluster specific (i.e., have a unique value for each cluster), we will use ω_{qj} to represent the regression coefficient for explanatory or predictor variable q for cluster j. In the current case, the girl-specific effect $\underline{\omega}_{0j}$ is a random intercept; that is, girls in the study differ in terms of their weight before treatment. The model for the alternative hypothesis of the t-test (i.e., $H_a : \mu_1 \neq \mu_2$) is

$$\text{Level 1 Model:} \qquad \underline{y}_{ij} = \underline{\omega}_{0j} + \beta_1 x_{ij} + \underline{\varepsilon}_{ij}, \tag{3.1}$$

where $\underline{\omega}_{0j}$ represents the random cluster or girl-specific effect for girl j, β_1 represents the fixed effect of treatment, and $\underline{\varepsilon}_{ij}$ represents the within girl (cluster) random component (i.e., the residual or error for girl j on occasion i). This model does

not include differential treatment effects. Before treatment (i.e., when $x_{ij} = 0$), the expected weight for girl j equals ω_{0j} and after treatment her expected weight equals $\omega_{0j} + \beta_1$. The parameter β_1 equals the expected change in weight.

The girls in the study are a random sample from a population of girls treated for anorexia. The cluster (girl) specific effect $\underline{\omega}_{0j}$ is an unobserved or latent random variable. A different sample of girls treated with anorexia would lead to different values of the ω_{0j}s. So long as samples are randomly drawn from the same population of girls, the sample mean and variance of the ω_{0j}s are estimates of the population mean and variance of $\underline{\omega}_{0j}$. The population parameters of the distribution of the $\underline{\omega}_{0j}$'s are fixed. The population means and variances are estimated and not the actual values of the ω_{0j}s for a specific sample of girls[2].

To describe the differences between girls' weights, a normal linear regression model for the intercept $\underline{\omega}_{0j}$ is formed and is the Level 2 model,

$$\text{Level 2 Model:} \qquad \underline{\omega}_{0j} = \beta_0 + \underline{\gamma}_{0j}, \tag{3.2}$$

where β_0 is the overall fixed effect for the intercept (i.e., the mean of the $\underline{\omega}_j$s), and $\underline{\gamma}_{0j}$ represents unaccounted random differences between girls in terms of their intercepts. The random term $\underline{\gamma}_{0j}$ can be viewed as a residual or error term that could reflect important variables not included in the model. Replacing $\underline{\omega}_{0j}$ in the Level 1 model (3.1) by its definition in the Level 2 model (3.2) yields the linear mixed model (LMM)

$$\text{Linear Mixed Model:} \qquad \underline{y}_{ij} = \underbrace{\beta_0 + \beta_1 x_j}_{\text{fixed}} + \underbrace{\underline{\gamma}_{0j} + \underline{\varepsilon}_{ij}}_{\text{random}}. \tag{3.3}$$

The model in (3.3) contains parameters that are fixed in the population (i.e., β_0 and β_1) and effects that are random (i.e., $\underline{\gamma}_{0j}$ and $\underline{\varepsilon}_{ij}$). Given the dummy coding for measurement occasion, the parameter β_1 equals the expected change in weight for the average girl.

The final assumptions needed to complete the model are those for the random or stochastic effects. These are $\underline{\gamma}_{0j} \sim N(0, \psi_{00})$ and independent over j, $\underline{\varepsilon}_{ij} \sim N(0, \sigma^2)$ and independent over i and j, and $\underline{\gamma}_{0j}$ and $\underline{\varepsilon}_{ij}$ are independent of each other. The independence between $\underline{\varepsilon}_{ij}$ and $\underline{\gamma}_j$ implies that their covariance equals zero (i.e., $\text{cov}(\underline{\varepsilon}_{ij}, \underline{\gamma}_j) = 0$). In sum, the assumptions for the random effects are

[2] The ω_{0j}'s can be estimated *after* the model has been fit to the data. We use such estimates later in this section.

Table 3.2 Estimated parameters and fit statistics of model (3.3) that corresponds to a t-test fit to the anorexia data set.

Effect	Parameter	Estimate (s.e.)
Intercept	β_0	82.41 (0.80)
Occasion (after)	β_1	2.76 (0.94)
var($\underline{\gamma}_j$)	ψ_{00}	13.84 (5.67)
var($\underline{\varepsilon}_{ij}$)	σ^2	31.87 (5.36)

$$\begin{pmatrix} \underline{\varepsilon}_{ij} \\ \underline{\gamma}_{0j} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \psi_{00} \end{pmatrix}\right) \; i.i.d., \tag{3.4}$$

where "*MVN*" stands for multivariate normal. The parameters of the model are the fixed effects (i.e., β_0 and β_1) and the variances of the random effects (i.e., ψ_{00} and σ^2).

The estimated parameters of the model[3] for the anorexia data are reported in Table 3.2. The equivalence between the paired dependent t-test and the model becomes apparent by comparing the sample statistics in Table 3.1 with the model parameters in Table 3.2. The equivalences are:

$$\begin{aligned} \bar{y_1} &= \hat{\beta}_0 = 82.41 \\ \bar{y_2} &= \hat{\beta}_0 + \hat{\beta}_1 = 82.41 + 2.76 = 85.17 \\ \text{cov}(\underline{y}_{1j}, \underline{y}_{2j}) &= \hat{\psi}_{00} = 13.84 \\ se_d &= \sqrt{2\hat{\sigma}^2/N} = \sqrt{2(31.87)/72} = 0.94. \end{aligned}$$

For the hypothesis $H_o : \beta_1 = 0$ the test statistics equals $\hat{\beta}_1/\text{s.e.} = 2.76/0.94 = 2.94$ and is exactly the same one as reported in Table 3.1. As shown below, these equivalencies stem from the assumptions of the model.

To derive the equivalencies between the LMM given in (3.3) and (3.4) and the paired dependent t-test, consider the difference between girl j's weight before and after the treatment based the LMM,

$$\begin{aligned} \underline{y}_{2j} - \underline{y}_{1j} &= (\beta_0 + \beta_1 + \underline{\gamma}_{0j} + \underline{\varepsilon}_{2j}) - (\beta_0 + \underline{\gamma}_{0j} + \underline{\varepsilon}_{1j}) \\ &= \beta_1 + \underline{\varepsilon}_{2j} - \underline{\varepsilon}_{1j}. \end{aligned} \tag{3.5}$$

[3] The model parameters were estimated using residual maximum likelihood estimation or "REML". This corresponds to the estimation method implicitly used in the t-test and typically used for ANOVA models. More on estimation for models is covered in Chapter 4.

Since the expected values of $\underline{\varepsilon}_{1j}$ and $\underline{\varepsilon}_{2j}$ both equal zero, the expected value of the difference is $\mathrm{E}(\underline{y}_{2j} - \underline{y}_{1ij}) = \beta_1$; that is, the mean weight difference between time points equals $\mu_2 - \mu_1 = \beta_1$. Testing $H_0 : \mu_1 - \mu_2 = 0$ is the same as testing $H_0 : \beta_1 = 0$.

To obtain the standard error of the expected difference (i.e., $\mathrm{E}(\underline{y}_{2j} - \underline{y}_{1j})$), note that based on (3.5) the variance of the difference $(\underline{y}_{2j} - \underline{y}_{1j})$ equals the variance of $(\underline{\varepsilon}_{2j} - \underline{\varepsilon}_{1j})$. Since $\underline{\varepsilon}_{2j}$ and $\underline{\varepsilon}_{1j}$ are independent, the variance of their difference equals the sum of their variances,

$$\mathrm{var}(\underline{y}_{2j} - \underline{y}_{1j}) = \mathrm{var}(\underline{\varepsilon}_{2j} - \underline{\varepsilon}_{1j}) = 2\sigma^2.$$

The variance of the mean difference is the variance of the difference divided by the sample size, (i.e., $\mathrm{var}(\mathrm{E}(\underline{y}_{1j} - \underline{y}_{2j})) = 2\sigma^2/N$), and the standard error of the mean is the square root, $\sqrt{2\sigma^2/N}$. Putting this all together gives the t-test of $H_0 : \mu_2 - \mu_1 = \beta_1 = 0$,

$$t = \frac{\hat{\beta}_1}{\sqrt{2\hat{\sigma}^2/N}}.$$

Examining the last equivalency (i.e., $\mathrm{cov}(\underline{y}_{ij}, \underline{y}_{i'j}) = \psi$) shows how random effects can account for dependency within clusters. The dependency within clusters is due to the fact each observation within a cluster has the same value for the random effect (i.e., γ_j). Using model (3.3), the covariance between observations i and i' in cluster j equals

$$\begin{aligned}
\mathrm{cov}(\underline{y}_{ij}, \underline{y}_{i'j}) &\equiv \mathrm{E}\left[(\underline{y}_{ij} - \mathrm{E}(\underline{y}_{ij}))(\underline{y}_{i'j} - \mathrm{E}(\underline{y}_{i'j}))\right] \\
&= \mathrm{E}\Big[(\beta_0 + \beta_1 x_{ij} + \underline{\gamma}_{0j} + \underline{\varepsilon}_{ij}) - (\beta_0 + \beta_1 x_{ij})) \\
&\quad \times ((\beta_0 + \beta_1 x_{i'j} + \underline{\gamma}_{0j} + \underline{\varepsilon}_{i'j}) - (\beta_0 + \beta_1 x_{i'j}))\Big] \\
&= \mathrm{E}\left[(\underline{\gamma}_{0j} + \underline{\varepsilon}_{ij})(\underline{\gamma}_{0j} + \underline{\varepsilon}_{i'j})\right] \\
&= \underbrace{\mathrm{E}(\underline{\gamma}_{0j}^2)}_{\psi_{00}} + \underbrace{\mathrm{E}(\underline{\varepsilon}_{ij}\underline{\varepsilon}_{i'j})}_{0} + \underbrace{\mathrm{E}(\underline{\gamma}_{0j}\underline{\varepsilon}_{ij})}_{0} + \underbrace{\mathrm{E}(\underline{\gamma}_{0j}\underline{\varepsilon}_{i'j})}_{0} \\
&= \psi_{00}. \qquad (3.6)
\end{aligned}$$

The first line is the definition of the covariance. In the second line, the elements of the covariance are replaced by their LMM model values from (3.3). The third line is a result of algebraic simplification and the fourth line uses the fact that the expectation of a sum equals the sum of expectations. The cross-product terms $\mathrm{E}(\underline{\varepsilon}_{ij}\underline{\varepsilon}_{i'j})$, $\mathrm{E}(\underline{\gamma}_{0j}\underline{\varepsilon}_{ij})$, and $\mathrm{E}(\underline{\gamma}_{0j}\underline{\varepsilon}_{i'j})$ all equal zero because the random effects are

independent of each other (i.e., the covariances are 0). In sum, the variance of the random intercept, $\text{var}(\underline{\gamma}_{0j}) = \psi_{00}$, is also equal to the covariance between observations within clusters.

The t and F tests for regression coefficients of GLMs that were covered in Chapter 2 apply to fixed effects in GLMMs; however, there are some important differences and considerations. Statistical inference for fixed effects and random effects of GLMMs are discussed in Chapter **??**.

3.3 Linear Regression Models for Random Coefficients

Modeling or accounting for systematic differences between clusters can be achieved by adding predictor or explanatory variables to the Level 2 model. In this section, the effect of adding explanatory variables to the Level 2 linear regression model for the intercept and the slope from the Level 1 model are studied. In the following sections, Level 2 model for the intercept is made more complex by adding predictor variables. Subsequently the slope is allowed to vary over clusters and its variability modeled. These additional complexities are illustrated by continuing the anorexia example.

In the anorexia data set, the girls (clusters) received one of three different treatments: cognitive/behavioral therapy, family therapy, or the standard (control) treatment. The mean weights for each treatment are plotted again time in Figure 3.2. There is a difference in mean weight between types of treatment before treatment and there appears to be an interaction between time and treatment type. Type of treatment may help to explain some of the variability between the girls both in terms of differences in their intercepts and the effectiveness of the treatment (i.e., the slopes of the Level 1 regression lines).

3.3.1 Modeling the Intercept

To include treatment type in the analysis, dummy codes for type of treatment are defined as follows[4]:

$$z_{1j} = \begin{cases} 1 \text{ if cognitive/behavioral} \\ 0 \text{ otherwise} \end{cases} \qquad z_{2j} = \begin{cases} 1 \text{ if family therapy} \\ 0 \text{ otherwise} \end{cases}.$$

[4] Dummy codes are not the only choice for coding of categorical predictors. Effect codes or other ways of coding variables could also be used. The choice should be based on easy of interpretation.

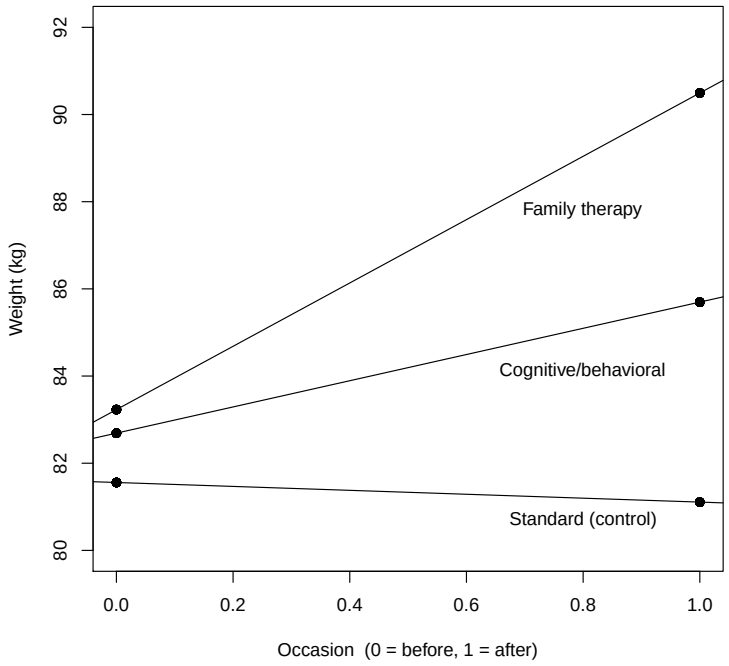

Fig. 3.2 Mean weight (dots) plotted against measurement occasions with different symbols representing type of treatment that a girl received.

The Roman letter z will be used for values of Level 2 predictor variables that describe characteristics or attributes of clusters. Since models for the slope will be introduced, ω_{qj}s will be used as the regression coefficient for Level 1 effects. The revised random intercept model is

$$\begin{aligned} \text{Level 1: } \underline{y}_{ij} &= \underline{\omega}_{0j} + \omega_{1j} x_{ij} + \underline{\varepsilon}_{ij} \\ \text{Level 2: } \underline{\omega}_{0j} &= \beta_{00} + \beta_{01} z_{1j} + \beta_{02} z_{2j} + \underline{\gamma}_{0j} \\ \omega_{1j} &= \beta_{10}. \end{aligned}$$

In the Level 2 models, the first sub-script on the βs indicates the effect in the Level 1 model, and the second sub-script indicates the variable in the Level 2 regression. For example, the sub-script "1" in β_{10} indicates this is the regression coefficient of

x_{ij}, and the "0" indicates the intercept of the Level 2 model. As a second example, the sub-script "0" in β_{02} indicates this belongs to the model for the intercept, and the "2" indicates that it is the coefficient for the second predictor in the model for the intercept.

The combined model or the LMM obtained by replacing $\underline{\omega}_{0j}$ and ω_{1j} in the Level 1 models by their Level 2 models is

$$\underline{y}_{ij} = \underbrace{\beta_{00} + \beta_{01}z_{1j} + \beta_{02}z_{2j} + \beta_{10}x_{ij}}_{\text{Structural}} + \underbrace{\underline{\gamma}_{0j} + \underline{\varepsilon}_{ij}}_{\text{Stochastic}} . \tag{3.7}$$

The assumption for the stochastic or random effects are

$$\begin{pmatrix} \underline{\varepsilon}_{ij} \\ \underline{\gamma}_{0j} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \psi_{00} \end{pmatrix} \right) \; i.i.d., \tag{3.8}$$

In the LMM, the structural or fixed effects part of the model accounts for systematic variation in the overall weight of girls (i.e., the intercept) due to treatment type, and the expected change in weight. The stochastic part of the model conveys information about random variation in the response variable that is not accounted for by the structural part of the model.

The LMM (and HLM) contain unobserved or latent variables (i.e., $\underline{\gamma}_{0j}$ and $\underline{\varepsilon}_{ij}$). It is important to note that the HLM and the implied LMM have implications for the observed data, in particular the distribution of $\underline{y}_{ij}$. Based on properties of the (multivariate) normal distribution, $\underline{y}_{ij}$ is normally distributed because it is a linear combination of normally distributed variables in (3.7). Since the expected values of $\underline{\gamma}_{0j}$ and $\underline{\varepsilon}_{ij}$ equal 0, the mean of $\underline{y}_{ij}$ is the fixed or structural part of the model (i.e., $\beta_{00} + \beta_{01}z_{1j} + \beta_{02}z_{2j} + \beta_{10}x_{ij}$). Lastly, since $\underline{\gamma}_{0j}$ and $\underline{\varepsilon}_{ij}$ are independent of each other, the variance of $\underline{y}_{ij}$ equals $\psi_{00} + \sigma^2$. In sum, the model for the data is

$$\underline{y}_{ij} \sim N\left((\beta_{00} + \beta_{01}z_{1j} + \beta_{02}z_{2j} + \beta_{10}x_{ij}), (\psi_{00} + \sigma^2)\right) \; i.i.d. \tag{3.9}$$

The (7.7) emphasizes that the βs, ψ_{00} and σ^2 are the parameters of the model that are to be estimated. The LMM has been collapsed or marginalized over the latent/unobserved effects and (7.7) is a *marginal model*.

The parameter estimates and fit statistics for the model defined by (3.7) and (3.8) are reported in Table 3.3 under the column labeled "Model 2". Parameter estimates and fit statistics for the first model that was defined by (3.3) and (3.4) are also reported in the table under the columns labeled "Model 1". Model 1 was originally estimated using a version of maximum likelihood estimation (i.e., REML) that gives the exact same results as a standard t-test. Unfortunately, the maximum

likelihood from this method and thus the AIC and BIC statistics are not comparable across models. Since we would like to compare the results over different models, Model 1 was re-estimated using full maximum likelihood estimation and it is these estimated parameters and fit statistics that are reported in Table 3.3.

Based on the information criteria, Model 2 is an improvement over Model 1 (i.e., $AIC = 950.2$ versus 958.1 and $BIC = 963.9$ versus 967.2). The additional parameters have accounted for some of the variance of $\underline{y}_{ij}$. In particular, the estimated variance ψ_{00} from Model 2 is smaller than that from Model 1 (i.e., 9.18 versus 13.65), but σ^2 remained essentially the same. In other words, some of the variance due to differences between girls is now systematic rather than random as in Model 1. Adding important Level 2 predictors for the intercept in a random intercept model will lead to an estimate of ψ_{00} that is smaller than the estimate in the model without the predictors.

The values of the estimated fixed effects for treatment are in the same order as the mean values of weight for each type of treatment as seen in Figure 3.2. From top to bottom, the largest is $\hat{\beta}_{02} = 5.53$ for family therapy, followed by $\hat{\beta}_{01} = 2.86$ for cognitive, and 0 for the standard or control. Model 2 implies three parallel regression lines, one for each treatment. The estimated regression lines from Model 2 are plotted in the top right side of Figure 3.3 where the top line corresponds girls who received family therapy, the middle line to cognitive therapy and the lowest line to control or standard treatment. These regression lines correspond to the estimated expected values (i.e., means) given by the marginal model (7.7). The observed means for each treatment are plotted in the top left of Figure 3.3. Comparing the data to Model 2, Model 2 is not a good representation of the data.

To gain greater insight into the marginal model and the underlying LMM (and HLM), we examine these for a specific girl, girl #7, who happened to receive cognitive therapy. Estimates of γ_{0j} and ε_{ij} were computed using the estimated βs, ψ_{00} and σ^2. In Figure 3.3, the data for girl #7 (dots), the marginal means (line labeled $E(y_{ij}|x_{i7}, z_{17}, z_{27})$, and the girl-specific means (line labeled $\mathrm{E}(y_{iy}|x_{i7}, z_{17}, z_{27}, \hat{\gamma}_{0y})$) are plotted. The marginal means $\mathrm{E}(y_{i7}|x_{i7}, z_{17}, z_{27})$ are the same for all girls who received cognitive therapy (i.e., $z_{1j} = 1$ and $z_{2j} = 0$). The underlying LMM model assumes the existence of unobserved or latent variables γ_{0j} and ε_{ij}. The γ_{0j}s adjust the marginal regression up or down, and in the case of girl #7, the cluster-specific regression line is higher and closer to her data[5]. The remainder of the difference is due to ε_{ij}.

According to the underlying LMM (and HLM), there is a cluster-specific regression line for each girl. The estimated girl-specific regressions based on Model 2 are plotted in top right of Figure 3.5. The cluster-specific regressions consist of

[5] The cluster-specific regression line may seem still too low in that $\hat{\gamma}_{07}$ could be larger and yield a better representation of girl #7's data. However, the estimated $\hat{\gamma}_{07}$ is a weighted average of her data and the estimated marginal means given by $\mathrm{E}(y_{i7}|x_{i7}, z_{17}, z_{27})$. This is a topic discussed in Chapter 4???.

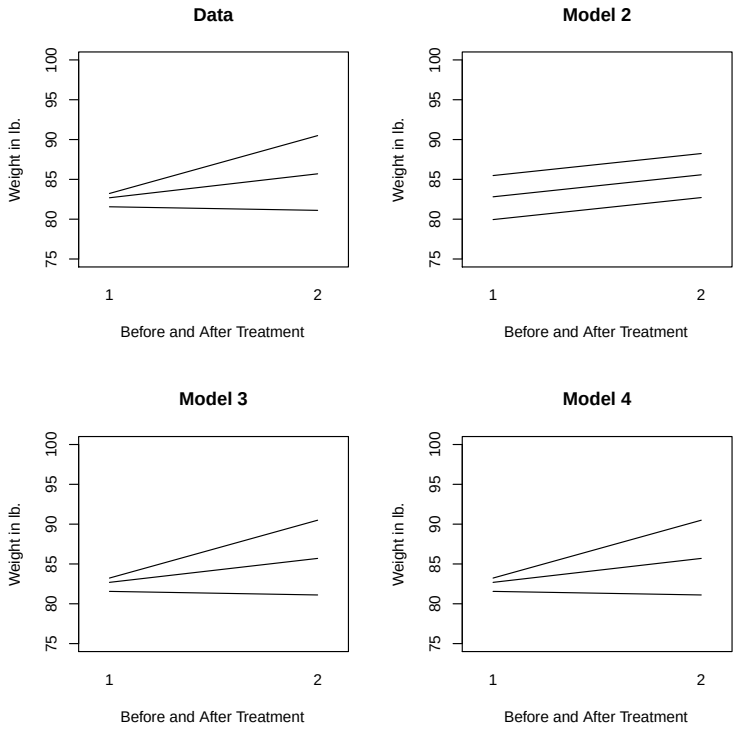

Fig. 3.3 Weight of girls plotted as a function of time with separate lines for type of treatment. The upper left plot is observed data and the other three are predictions of $E(y_{ij})$ from various models fit to the data.

a set of parallel lines and do not resemble the plot of the data. Model 2 can be improved by adding a Level 2 model for the slope such that the effectiveness of the treatment can vary over girls.

3.3.2 Modeling The Slope

In the plot of the data in Figure 3.5 (and Figure 3.1), most girls appear to have different slopes. In other words, the effectiveness of treatment differs over girls. Furthermore, Figure 3.2 suggests that treatment type may be an important predictor of the Level 1 model slope of x_{ij}. A Level 2 model will be specified for the

Table 3.3 Parameter estimates and fit statistics for random intercept and slope models fit to anorexia data using full maximum likelihood estimation.

	Model 1		Model 2		Model3		Model 4	
Effect	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)
Fixed Effects								
Intercept	82.41	(0.79)	79.95	(1.08)	81.56	(1.22)	81.56	(1.00)
Occasion:								
After	2.76	(0.93)	2.76	(0.93)	−0.45	(1.45)	−0.45	(1.45)
Before	0.00	.	0.00	.	0.00	.		
Type of Treatment:								
Cognitive			2.86	(1.35)	1.13	(1.68)	1.13	(1.38)
Family			5.53	(1.56)	1.67	(1.93)	1.67	(1.59)
Control			0.00	.	0.00	.	0.00	.
Occasion × Type:								
After × Cognitive					3.46	(1.99)	3.46	(1.99)
After × Family					7.71	(2.30)	7.71	(2.30)
After × Control					0.00	.	0.00	.
Before× Cognitive					0.00	.	0.00	.
Before× Family					0.00	.	0.00	.
Before× Control					0.00	.	0.00	.
Random Effects								
ψ_{00} (intercept)	13.65	(5.55)	9.18	(4.91)	11.31	(4.73)	11.31	(4.49)
ψ_{11} (slope)							24.87	(9.15)
σ^2	31.43	(5.24)	31.43	(5.24)	27.16	(4.53)	14.72	(4.76)
Fit Statistics								
Number of parameters	4		6		8		9	
$-2\ln(\text{Likeilihood})$	950.1		938.2		927.7		919.0	
AIC	958.1		950.2		943.7		937.0	
BIC	967.2		963.9		961.9		957.5	

slope of x_{ij} using treatment type. This will allow for the possibility of differential effectiveness of treatment based on the type of treatment that a girl received. It may be the case, the type of treatment will account for all of the between girl variation in terms of the slope of x_{ij} or will only account for part of the between girl variance. Two variations of this model are considered that reflect these two possibilities. In the first one, there is a fixed effect for the slope but it can have different values depending on treatment type. In the second one, both a fixed treatment type and a random effect for the slope is included the Level 2 model.

The next model, Model 3, has the same Level 1 model as Model 2, but now includes fixed effects for type of treatment. Model 3 is

$$\begin{aligned}
\text{Level 1:} \qquad & \underline{y}_{ij} = \underline{\omega}_{0j} + \omega_{1j} x_{ij} + \underline{\varepsilon}_{ij} \\
\text{Level 2:} \qquad & \underline{\omega}_{0j} = \beta_{00} + \beta_{01} z_{1j} + \beta_{02} z_{2j} + \underline{\gamma}_{0j} \\
& \omega_{1j} = \beta_{10} + \beta_{11} z_{1j} + \beta_{12} z_{2j}.
\end{aligned}$$

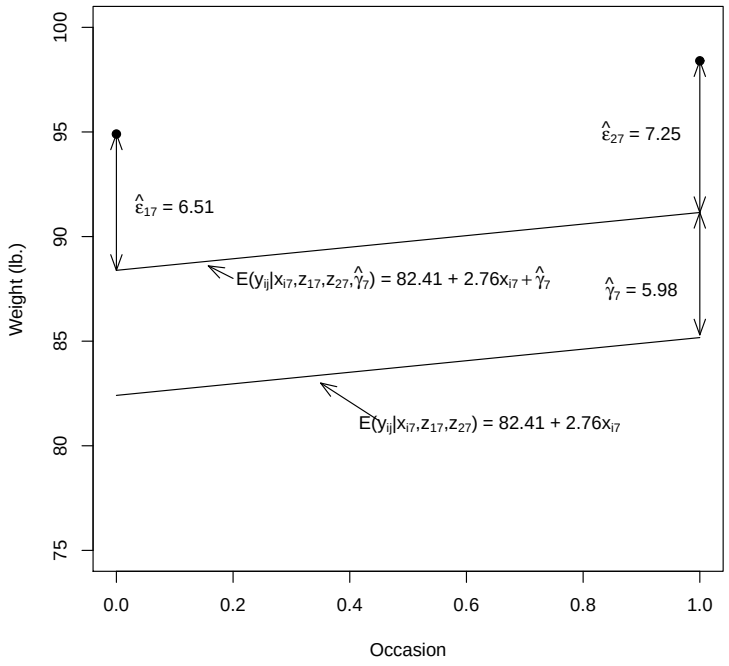

Fig. 3.4 Plot of data (dots), marginal regression line ($\mathrm{E}(y_{i7}|x_{i7}), z_{17}, z_{27}$), and cluster-specific regression line ($\mathrm{E}(y_{i7}|x_{17}, z_{17}, z_{27}, \hat{\gamma}_{07})$) for girl #7 who have cognitive therapy.

Since the model for ω_{1j} does not include a random term, the treatment type explains all the differences or variation between girls in regards to the effectiveness of their treatment. The explanatory variable for differences in the intercept need not be the same as those for the effectiveness of the treatment. For example, a good predictor of the intercept might be a girl's height and a good predictor of the slope might be the type of treatment that a girl receives.

The above HLM implies an interaction between measurement occasion (x_{ij}) and treatment type. This is readily seen in the LMM. Replacing the ω's in Level 1 by their Level 2 models yields the LMM

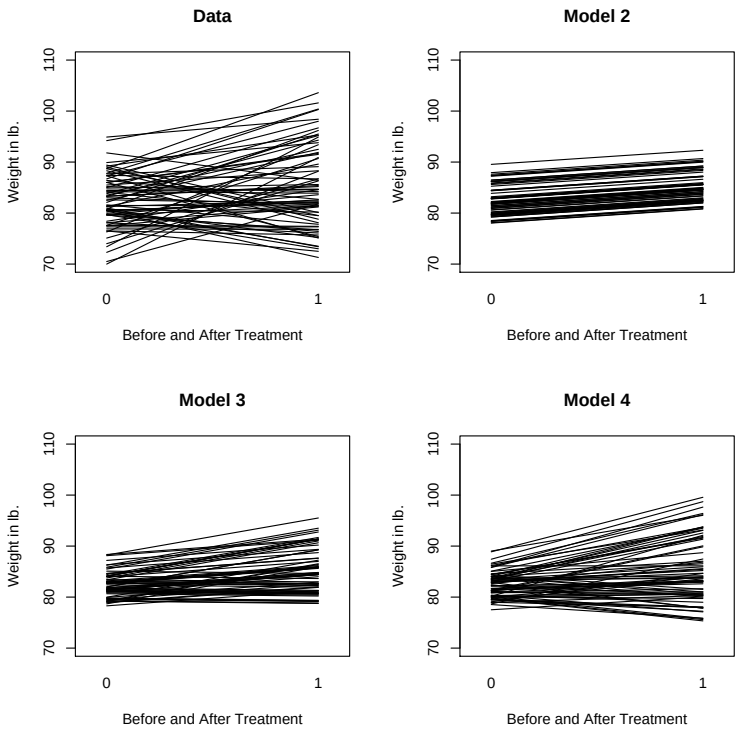

Fig. 3.5 Observed weight of girls (data) and the cluster–specific regressions for Models 2 – 4 plotted as a function of time with separate lines for each girl.

$$\begin{aligned} \underline{y}_{ij} &= \underbrace{(\beta_{00} + \beta_{01} z_{1j} + \beta_{02} z_{2j} + \underline{\gamma}_{0j})}_{\text{Model for intercept}} + \underbrace{(\beta_{10} + \beta_{11} z_{1j} + \beta_{12} z_{2j})}_{\text{Model for slope}} x_{ij} + \underline{\varepsilon}_{ij} \\ &= \beta_{00} + \beta_{01} z_{1j} + \beta_{02} z_{2j} + \beta_{10} x_{ij} + \underbrace{\beta_{11} z_{1j} x_{ij} + \beta_{12} z_{2j} x_{ij}}_{\text{Cross-level interactions}} + \underline{\gamma}_{0j} + \underline{\varepsilon}_{ij}, \end{aligned}$$

The LMM has a *cross-level interaction* that is an interaction between a Level 1 predictor variable and a Level 2 predictor. Cross-level interactions arise whenever a model for the slope of a Level 1 variable is a function of a Level 2 variable. In the anorexia example, a girl who underwent family theory may on average respond differently than a girl who had the standard treatment and this may differ from a girl who had cognitive therapy.

The estimated parameters and fit statistics for Model 3 are reported in Table 3.3. The estimated model is plotted in Figure 3.3 where the highest line is that for family therapy, followed by cognitive therapy and the lowest is the standard treatment. The cross-level interaction yields divergent lines for the estimated expected values. The lines in the plot for Model 3 have one of three different slopes. The greatest increase in weight is for family therapy (i.e., $\hat{\beta}_{10}+\hat{\beta}_{12}=-0.45+7.71=7.26$), followed by cognitive therapy $\hat{\beta}_{10}+\hat{\beta}_{11}=-0.45+3.46=3.01$. With the standard treatment, weight is actually predicted to decrease over time $\hat{\beta}_{10}=-0.45$).

Comparing the plot of data in the upper left graph in Figure 3.3 to Model 3, the two graphs are essentially identical. This occurs because there are six estimated βs and six means. The observed marginal means are fit perfectly by the marginal model derived from the HLM/LMM. In addition to capturing the marginal mean structure, Model 3 is a better model than either Model 1 or Model 2 in terms of the information criteria.

The variance for the intercept ψ_{00} is larger in Model 3 than in Model 2 (i.e., $\hat{\psi}_{00}=11.31$ versus 9.18), but the within cluster variance σ^2 that is smaller in Model 3 (i.e., $\hat{\sigma}^2=27.16$ versus 31.43). Adding the cross-level interactions changed the variation at both levels. Allowing slopes to vary often leads to changes the intercept variance, and in this case, the variance ψ_{00} increased. Even though the variance of the intercept increased, the total unexplained variation in girls' weights (i.e., $\widehat{var}(\underset{\sim}{y}_{ij}|x_{ij},z_{1j},z_{2j})=\hat{\psi}_{00}+\hat{\sigma}^2$) is smaller in Model 3 than either Models 1 or 2,

$$\begin{aligned}
&\text{Model 1:} && \widehat{var}(\underset{\sim}{y}_{ij}|x_{ij}) = 13.65+31.43=45.09\\
&\text{Model 2:} && \widehat{var}(\underset{\sim}{y}_{ij}|x_{ij},z_{1j},z_{2j}) = 9.18+31.43=40.61\\
&\text{Model 3:} && \widehat{var}(\underset{\sim}{y}_{ij}|x_{ij},z_{1j},z_{2j}) = 11.31+27.16=38.47
\end{aligned}$$

The subject-specific estimated regression lines are plotted in Figure 3.5. There are three different slopes (one for each type of treatment). The intercepts differ depending on treatment type plus a girl-specific random component; namely, the girl-specific intercepts equal

$$\begin{aligned}
&\text{Cognitive:} && \hat{\omega}_{0j}=81.56+1.13+\hat{\gamma}_{0j}\\
&\text{Family:} && \hat{\omega}_{0j}=81.56+1.67+\hat{\gamma}_{0j}\\
&\text{Standard:} && \hat{\omega}_{0j}=81.56+\hat{\gamma}_{0j}.
\end{aligned}$$

The girl-specific regressions from Model 3 are more similar to the data than those from Model 2; however, from looking at the data, it appears that there are more than just three slopes as predicted by Model 3. Model 3 can be modified by adding

a random term to the model for the regression coefficient for measurement occasion (i.e., the slope of x_{ij}).

The random intercept and slope model, Model 4, is

$$\begin{aligned} \text{Level 1: } \underline{y}_{ij} &= \omega_{0j} + \underline{\omega}_{1j} x_{ij} + \underline{\varepsilon}_{ij} \\ \text{Level 2: } \underline{\omega}_{0j} &= \beta_{00} + \beta_{01} z_{1j} + \beta_{02} z_{2j} + \underline{\gamma}_{0j} \\ \underline{\omega}_{1j} &= \beta_{10} + \beta_{11} z_{1j} + \beta_{12} z_{2j} + \underline{\gamma}_{1j}. \end{aligned}$$

The LMM for Model 4 is obtained by replacing $\underline{\omega}_{0j}$ and $\underline{\omega}_{1j}$ by their Level 2 models; that is,

$$\begin{aligned} \underline{y}_{ij} &= (\beta_{00} + \beta_{01} z_{1j} + \beta_{02} z_{2j} + \underline{\gamma}_{0j}) + (\beta_{10} + \beta_{11} z_{1j} + \beta_{12} z_{2j} + \underline{\gamma}_{1j}) x_{ij} + \underline{\varepsilon}_{ij} \\ &= \underbrace{\beta_{00} + \beta_{01} z_{1j} + \beta_{02} z_{2j} + \beta_{10} x_{ij} + \beta_{11} z_{1j} x_{ij} + \beta_{12} z_{2j} x_{ij}}_{\text{Structural}} + \underbrace{\underline{\gamma}_{0j} + \underline{\gamma}_{1j} x_{ij} + \underline{\varepsilon}_{ij}}_{\text{Stocahstic}}, \end{aligned} \tag{3.10}$$

where

$$\begin{pmatrix} \underline{\varepsilon}_{ij} \\ \underline{\gamma}_{0j} \\ \underline{\gamma}_{1j} \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 & \\ 0 & \psi_{00} & 0 \\ 0 & 0 & \psi_{11} \end{pmatrix} \right) \; i.i.d.$$

If the girls' weights before treatment ($\underline{\omega}_{0j}$) are unrelated to effect of treatment (i.e., $\underline{\omega}_{1j}$), then we expect the random intercepts to be independent of the random slopes. Furthermore, with only two observations per cluster, the covariance ψ_{01} must be (and was) set to 0 so that the model is identified[6].

Even though the marginal means from Model 4 are the same as those from Model 3 (i.e., see Figure 3.3), Model 4 has smaller values of AIC and BIC. The reason has to do with the modeling of the variance. In terms of the underlying model (i.e., HLM and implied LMM), the cluster specific regressions plotted in the lower right corner of Figure 3.5 are more similar to the data than Models 2 or 3. In the underlying model, the girls can now have their own unique effect. However γ_{0j} and γ_{1j} are not parameters of the marginal model but are hypothetical unobserved variables and the estimates of them are biased (i.e., the variance between $\hat{\gamma}_{0j}$'s is too small). The HLM and LMM imply the marginal model and this is what is estimated. The variances of γ_{0j} and γ_{1j} are estimated. To see the effect that the random slope from the underlying model (i.e., HLM and LM) has on the variance of $\underline{y}_{ij}$ in the marginal model, the (marginal) variance is found by starting with the LMM:

[6] The assumption that $\psi_{01} = 0$ can be relaxed by, for example, setting $\psi_{11} = 2\psi_{00}$ and then estimating ψ_{00} and ψ_{10}. Doing so leads to an even better fitting model.

$$\begin{aligned}
\text{var}(\underline{y}_{ij}) &= \text{E}[(\underline{y}_{ij} - \text{E}(\underline{y}_{ij}))^2] \\
&= \text{E}\left[\Big((\beta_{00} + \beta_{01}z_{1j} + \beta_{02}z_{2j} + \beta_{10}x_{ij} + \underline{\gamma}_{0j} + \underline{\gamma}_{1j}x_{ij} + \underline{\varepsilon}_{ij}) \right. \\
&\qquad \left. -(\beta_{00} + \beta_{01}z_{1j} + \beta_{02}z_{2j} + \beta_{10}x_{ij})\Big)^2\right] \\
&= \text{E}[(\underline{\gamma}_{0j} + \underline{\gamma}_{1j}x_{ij} + \underline{\varepsilon}_{ij})^2] \\
&= \psi_{00} + \psi_{11}x_{ij}^2 + \sigma^2.
\end{aligned}$$

The variance of $\underline{y}_{ij}$ is a function of x_{ij}.

Even though the response variable is assumed to be normally distributed, models with random slopes lead to heteroscedasticity — the variance depends on the value(s) of predictor variable(s). When both the intercept and slope are random, the variance of the response variable is a quadratic function of the x_{ij}s and the variances and covariances of the random effects. Since $\text{var}(\underline{y}_{ij})$ is a quadratic function and the variances ψ_{00} and σ^2 are always positive, the model will yield $\widehat{\text{var}}(\underline{y}_{ij})$ that is positive.

3.4 Generalized Linear Mixed Models

In the previous section, models were presented by specifying a Level 1 model and multiple Level 2 models. To extend the material in the previous section to more general cases (e.g., more predictors, non-normally distributed response variables, and alternative link functions), a more general expression of the models is required. Two ways of doing this are presented: one uses a multilevel framework where the Level 1 model is a GLM, and the other simply adds random effects to the linear predictor of a GLM. The multilevel random effects approach explicitly takes the hierarchical structure of the data into account; whereas, the random coefficients approach need not. The resulting parameter estimates, interpretations, test statistics, marginal models, and model goodness-of-fit to data are the same. Since both ways of expressing GLMMs are used in this text, GLMMs as multilevel random effects models and as random coefficients models are both presented below.

For both notations, a GLMM is conditional on the random effects; that is, it is a regression model for data within a cluster. The conditioning will be explicitly indicated in this section; that is, we will use "$\underline{y}_{ij}|\underline{\boldsymbol{\gamma}}_j$" to refer to the cluster specific regression model rather than "$\underline{y}_{ij}$".

3.4.1 Multilevel Normal Response Variable

Before presenting more general forms, the example of the previous section is expressed using multilevel GLMM notation specific to normal response variables. Using the most complex model for the anorexia data, Model 4, the Level 1 model is a GLM for responses within a cluster. The Level 1 model is

$$\begin{aligned} \underline{y}_{ij}|\underline{\boldsymbol{\gamma}}_j &\sim N\,(\underline{\mu}_{ij}, \sigma^2)\ i.i.d. \\ \underline{\mu}_{ij} &= \underline{\eta}_{ij}. \\ \underline{\eta}_{ij} &= \underline{\omega}_{0j} + \underline{\omega}_{1j} x_{ij}. \end{aligned}$$

This differs from a regular GLM in that the linear predictor and hence the mean are themselves random variables — random across clusters. The distribution of the response variable $\underline{y}_{ij}$ is conditional on the unobserved random cluster specific effects. Also note that within a cluster (i.e., given that $\underline{\boldsymbol{\gamma}}_j$), observations are independent. The linear predictor contains random coefficients, $\underline{\omega}_{0j}$ and $\underline{\omega}_{1j}$.

The Level 2 model focuses on modeling differences between clusters (girls) in terms of their overall level (i.e., $\underline{\omega}_{0j}$) and the effect of treatment or time (i.e., $\underline{\omega}_{1j}$), specifically,

$$\begin{aligned} \underline{\omega}_{0j} &= \beta_{00} + \beta_{01} z_{1j} + \beta_{02} z_{2j} + \underline{\gamma}_{0j} \\ \underline{\omega}_{1j} &= \beta_{10} + \beta_{11} z_{1j} + \beta_{12} z_{2j} + \underline{\gamma}_{1j}. \end{aligned}$$

where

$$\underline{\boldsymbol{\gamma}}_j = \begin{pmatrix} \underline{\gamma}_{0j} \\ \underline{\gamma}_{1j} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{00} & 0 \\ 0 & \psi_{11} \end{pmatrix} \right)\ i.i.d.$$

In Model 4 for the anorexia data, the covariance between $\underline{\gamma}_{0j}$ and $\underline{\gamma}_{1j}$ was set to 0; however, in most cases covariances between random effects are estimated.

3.4.2 Multilevel GLMM

The expression of the model for normal response variables is a special case of a more general form of a multilevel GLMM; that is,

$$\begin{aligned} \text{Level 1} \qquad & \underline{y}_{ij}|\underline{\boldsymbol{\gamma}}_j \sim f(\underline{\mu}_{ij}, \sigma^2)\ i.i.d. \\ & g(\underline{\mu}_{ij}) = \underline{\eta}_{ij} \\ & \underline{\eta}_{ij} = \underline{\omega}_{0j} + \underline{\omega}_{1j} x_{1ij} + \ldots \underline{\omega}_{Pij} x_{Pij}, \end{aligned} \tag{3.11}$$

where the distribution function $f(\,)$ is a member of the exponential family, σ^2 is the within cluster variance of $\underline{y}_{ij}$, and $g(\,)$ is a link function. The Level 1 model is a GLM where the regression coefficients in the linear predictor may be random.

The Level 2 model consists of normal linear regression models for the Level 1 (random) regression coefficients:

$$\begin{aligned} \text{Level 2} \qquad \underline{\omega}_{0j} &= \beta_{00} + \beta_{01} z_{1j} + \ldots + \beta_{0Q} z_{Qij} + \underline{\gamma}_{0j} \\ \underline{\omega}_{1j} &= \beta_{10} + \beta_{11} z_{1j} + \ldots + \beta_{1Q} z_{Qj} + \underline{\gamma}_{1j} \\ &\vdots \\ \underline{\omega}_{Pj} &= \beta_{P0} + \beta_{P1} z_{1j} + \ldots + \beta_{PQ} z_{Qj} + \underline{\gamma}_{Pj}, \end{aligned} \tag{3.12}$$

where $\underline{\boldsymbol{\gamma}}_j \sim MVN(\mathbf{0}, \boldsymbol{\Psi})$ *i.i.d.*, or

$$\begin{pmatrix} \underline{\gamma}_{0j} \\ \underline{\gamma}_{1j} \\ \vdots \\ \underline{\gamma}_{Pj} \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{00} & \psi_{10} & \ldots & \psi_{P0} \\ \psi_{10} & \psi_{22} & \ldots & \psi_{P1} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{P0} & \psi_{P1} & \ldots & \psi_{PP} \end{pmatrix} \right) \; i.i.d.$$

The Level 2 model is essentially a multivariate multiple regression, $\underline{\boldsymbol{\omega}}_j = \mathbf{B}\mathbf{z}_j + \underline{\boldsymbol{\gamma}}_j$. If a Level 1 effect is not random, for example x_{pij}, then the corresponding $\underline{\gamma}_{pj} = 0$ for all clusters and $\psi_{pp} = 0$, as well as all covariances involving Level 1 predictor p. The predictors at Level 2 may be different for different $\underline{\omega}_{pj}$s (i.e., some of the β_{pq} may equal or be set to 0).

Replacing the $\underline{\omega}_{pj}$s in the Level 1 linear predictor by Level 2 models in (3.12) yields the following linear predictor

$$\underline{\eta}_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\underline{\boldsymbol{\gamma}}_j, \tag{3.13}$$

where $\boldsymbol{\beta}$ contains all the fixed regression coefficients

$$\boldsymbol{\beta}' = (\beta_{00}, \beta_{01}, \ldots, \beta_{10}, \beta_{11}, \ldots, \beta_{PQ}),$$

and elements of the vector $\mathbf{x}_{ij}$ are *all* the predictor variables. These consist of all Level 1 predictors, Level 2 predictors, and products of Level 1 and Level 2 predictors (for cross-level interactions),

$$\mathbf{x}'_{ij} = (1, z_{1j}, \ldots, x_{1j}, x_{1ij} z_{1j}, \ldots, x_{Pij} z_{Qj}).$$

The vector $\boldsymbol{z}_{ij}$ contains the Level 1 variables that have random effects; that is, $\boldsymbol{z}'_{ij} = (1, x_{1ij}, \ldots, x_{Pij})$. If a variable in the Level 1 regression model does not have a random coefficient, then $\boldsymbol{z}_{ij}$ would not contain the corresponding predictor. For example, suppose that the effect for variable p is not random, then $\boldsymbol{z}_{ij}$ would exclude x_{pij}.

3.4.3 Random Coefficients Model

A *random coefficients model* is a GLM with a linear predictor that contains random coefficients, and perhaps also fixed ones. The GLMM is

$$\begin{aligned} \underline{y}_{ij} | \underline{\boldsymbol{\gamma}}_j &\sim f(\underline{\mu}_{ij}, \sigma^2) \\ g(\underline{\mu}_{ij}) &= \underline{\eta}_{ij} \qquad (3.14) \\ \underline{\eta}_{ij} &= \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\underline{\boldsymbol{\gamma}}_j \\ \underline{\gamma}_{ij} &\sim MVN(\boldsymbol{0}, \boldsymbol{\Psi}) \; i.i.d., \qquad (3.15) \end{aligned}$$

where $\boldsymbol{x}_{ij}$ consists of all predictor variables, $\boldsymbol{z}_{ij}$ contains all predictors that have random effects, $\boldsymbol{\beta}$ are fixed regression coefficients for the elements of $\boldsymbol{x}_{ij}$, and $\underline{\boldsymbol{\gamma}}_j$ are random regression coefficients for the elements of $\boldsymbol{z}_{ij}$.

The random coefficients model does not require specifying models at multiple levels; however, specifying models hierarchically as done in this chapter will lead to a random coefficients model. With multilevel or hierarchical data, $\boldsymbol{x}_{ij}$ contains all that predictor variables at all levels of the hierarchy, and $\boldsymbol{z}_{ij}$ contains those variables at lower levels of the hierarchy that have random effects. The multilevel and random coefficients expressions of GLMMs are used in the following example and subsequent chapters.

3.5 GLMM for a "Cool" Dichotomous Response Variable

In Chapter 2, a logistic regression model was fit to the "cool" kid data (Rodkin et al. 2006) without taking into account the nested or clustered structure of the data. Recall that in this example, kids nominated as "cool" by their fellow students who were classified as being an ideal or model student (from the teacher's perspective). Each student could nominate up to three others as being "cool". The response variable in this example was whether the identified "cool" student is an ideal student. The nominators are nested within classrooms ($N = 56$ classes).

These data are reanalyzed below by adding random effects to the linear predictor of a logistic regression model.

3.5.1 Fixed or Random Intercept

The index j refers to a classroom (i.e., a cluster) and i refers to a particular nominator in classroom j (i.e., a student or member of a cluster). Let π_{ij} be the probability that an ideal student is nominated as "cool" by student i in classroom j, and n_i^* equals the number kids that student i nominated as "cool." Since the response is a dichotomous classification (whether a "cool" kid is an ideal student), the binomial distribution with a logit link is the natural choice for these data. The Level 1 model is

$$\ln\left(\frac{\underline{\pi}_{ij}}{1-\underline{\pi}_{ij}}\right) = \underline{\omega}_{0j} + \omega_{1j}\,\texttt{Popluarity}_{ij} + \omega_{2j}\,\texttt{Gender}_{ij} + \omega_{3j}\,\texttt{Race}_{ij}. \tag{3.16}$$

Note that all of the predictor variables describe the nominator of the "cool" kid. We start with a relatively simple Level 2 model to describe differences between classrooms and only fit a random intercept with no predictor variables. The other coefficients are fixed and do not include a random term in the models for them. The Level 2 model is

$$\begin{aligned}\underline{\omega}_{0j} &= \beta_{00} + \underline{\gamma}_{0j}\\ \omega_{1j} &= \beta_{10}\\ \omega_{2j} &= \beta_{20}\\ \omega_{3j} &= \beta_{30}.\end{aligned}$$

According to the model, all the differences between classrooms are due to heterogeneity of the overall level of the linear predictor (i.e. $\underline{\gamma}_{0j}$). The random term $\underline{\gamma}_{0j}$ shifts the linear predictors up or down, and the set of linear predictors for $j = 1,\ldots,N$ consist of parallel lines.

The model above can also be expressed succinctly as a random coefficients models as follows:

$$
\begin{aligned}
\underline{\texttt{ideal}}_{ij}|\underline{\gamma}_{0j} &\sim \text{Binomial}(\underline{\pi}_{ij}, n_j^*) \\
\text{logit}(\underline{\pi}_{ij}) &= \underline{\eta}_{ij} \\
\underline{\eta}_{ij} &= \beta_{00} + \beta_{10}\,\texttt{Popularity}_{ij} + \beta_{20}\,\texttt{Gender}_{ij} + \beta_{30}\,\texttt{Race}_{ij} + \underline{\gamma}_{0j} \\
\underline{\gamma}_j &\sim N(0, \psi_{00})\ i.i.d.
\end{aligned}
$$

The estimated parameters, standard errors and fit statistics from the GLM from Chapter 2 and those from random intercept model given above are reported in Table 3.4 under the columns labeled Model 1 and Model 2, respectively. In this chapter we have not collapsed the data into a tabular form as presented in Chapter 2 because we need to know the classroom in which students are members to be able to add classroom characteristics as predictors of difference between the classrooms (clusters). For the GLM, whether data are or are not collapsed when analyzed does not effect the parameter estimates and standard errors, but it does effect the value of the computed likelihoods. The value of the maximum likelihoods will differ by a constant value when data are not collapsed; however, the difference between likelihoods will be the same (i.e., likelihood ratio tests are unaffected). A more detailed discussion of collapsed versus uncollapsed data is given in Chapter **??** that covers models for dichotomous variables in detail.

Comparing Model 1 that only contains fixed effects to Model 2 that has a random intercept model given, we see that all of the standard errors of the parameter estimates are noticeably larger in Model 2. For example, the standard error for $\hat{\beta}_{10}$ increases from 0.17 to 0.21. In the fixed effect model, β_{10} for `Popularity` is significant (i.e., Wald$= (0.79/0.17)^2 = 22.21$, $df = 1$, $p < .01$), but in the random intercept model it is not significant (i.e., Wald$= (0.11/0.21)^2 = 0.26$, $df = 1$, $p > .05$). Also note that the parameter estimates themselves are different. For binomial data, ignoring the clustered nature of the data leads to biased parameters estimates of fixed effects (Demidenko 2004)[7]. This example demonstrates the importance of taking into account the nested structure of the data.

Model 2 was further refined by dropping `Popularity` from the model. The results are reported under the columns labeled Model 3 in Table 3.4. Comparing Models 2 and 3, we find that the maximum of the likelihood hardly changes (i.e., $-2\ln(\text{Likelihood}) = 696.25$ versus 695.99), *AIC* is smaller when popularity is not included (i.e., 704.25 versus 705.99), and the estimated parameters for gender and race and their standard errors are nearly the same.

[7] This is not true of all members of the exponential family. For example, the GLMs for responses that are normal and for counts that are Poission distributed yield consistent, unbiased estimates of fixed effects when clustering is ignored.

Table 3.4 Estimated parameters, standard errors and fit statistics for various logistic regression models fit to the "cool" kid data (Rodkin et al. 2006).

	Model 1		Model 2		Model 3		Model 4	
Effect	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)
Fixed Effects:								
Intercept	0.14	(0.14)	0.31	(0.29)	0.36	(0.28)	0.27	(0.27)
Popularity:								
High	0.79	(0.17)	0.11	(0.21)				
Low	0.00	.	0.00	.				
Gender:								
Boy	−0.49	(0.16)	−0.65	(0.21)	−0.66	(0.21)	−0.64	(0.21)
Girl	0.00	.	0.00	.	0.00	.		
Race:								
Black	−1.45	(0.17)	−1.31	(0.32)	−1.31	(0.33)	−1.08	(0.33)
White	0.00	.	0.00	.				
Class aggression							−0.65	(0.23)
Random Effects:								
var(Intercept) $= \psi_{00}$			2.21	(0.69)	2.30	(0.70)	1.96	(0.61)
Model Information								
Number parameters	4		5		4		5	
$-2\ln(\text{Likelihood})$	782.22		695.99		696.25		688.65	
AIC	790.22		705.99		704.25		698.65	

3.5.2 Adding Level 2 Predictors

For the "cool" kid data, a potential predictor for the intercept is a measure of the amount of aggression in a classroom where higher scores correspond to higher levels of aggression. In the next model, classroom aggression is added to the Level 2 model for the intercept. Using the revised Level 1 model (without popularity) and adding $\texttt{ClassAgg}_j$ leads to the following multilevel random effects logistic regression model

$$\text{Level 1:} \qquad \ln\left(\frac{\pi_{ij}}{(1-\pi_{ij})}\right) = \underline{\omega}_{0j} + \omega_{1j}\,\texttt{Gender}_{ij} + \omega_{2j}\,\texttt{Race}_{ij} \tag{3.17}$$

$$\begin{aligned} \text{Level 2:} \qquad \underline{\omega}_{0j} &= \beta_{00} + \beta_{01}\,\texttt{ClassAgg}_j + \underline{\gamma}_{0j} \\ \omega_{1j} &= \beta_{10} \\ \omega_{2j} &= \beta_{20} \end{aligned} \tag{3.18}$$

Replacing the Level 1 regression coefficients in (3.17) by their values in the Level 2 models (3.18) yields the mixed linear predictor

$$\underline{\eta}_{ij} = \beta_{00} + \beta_{01}\,\texttt{ClassAgg}_j + \beta_{10}\,\texttt{Gender}_{ij} + \beta_{20}\,\texttt{Race}_{ij} + \underline{\gamma}_{0j}. \quad (3.19)$$

The results for this model, Model 4, are reported in Table 3.4. Note that the variance of the random intercept has decreased from $\hat{\psi}_{00} = 2.30$ to 1.96. Classroom aggression has helped to explain some of the differences between classes and accounted for approximately $(2.30 - 1.96)/2.30 \times 100\% = 15\%$ of the between class variance of the linear predictors. The estimated parameter for classroom aggression equals $\hat{\beta}_{01} = -0.65$, and indicates that holding all other variables constant (including random effects), within a classroom the odds that an ideal student is selected as cool is $\exp(-0.65) = 0.52$ times the odds when classroom aggression is one unit smaller. Alternatively, for a given classroom and all other effects held constant, the odds that an ideal student is nominated is $1/0.52 = 1.92$ times than the odds when classroom aggression is one unit larger. In other words, higher classroom aggression is associated with lower chances that an ideal student is selected as "cool."

3.6 Cluster-specific, Population Average, and Marginal Models

An important distinction needs to be made between cluster-specific, population average models, and marginal models. The distinction between these model families is important for estimation and interpretation of model parameters, and for decisions regarding what type of model to use in specific applications.

3.6.1 Model Types

The models discussed in this chapter (and most of this text) are known as *cluster-specific* or *subject-specific* models. In a cluster-specific model, the response variable is modeled conditional on (hypothetical) cluster-specific random effects. The underlying model is proposed for data within a cluster. For clarity in this section, the conditional nature of the model will be made explicit by using $\underline{y}_{ij}|\underline{\boldsymbol{\gamma}}_j$ to represent the response variable in the cluster-specific model. The basic processes operating within a cluster or context are assumed to be the same in different clusters, and the same functional form for the response holds for all clusters. Differences between groups are in terms of the random or unobserved cluster-specific effects (i.e., $\underline{\boldsymbol{\gamma}}_j$). When modeling the response, the variance explained by the model is broken down into that between groups and within groups and hence these models are sometimes referred to a *variance components*.

In a *population average model* responses $\underline{y}_{ij}$ are modeled without explicitly modeling the differences between clusters. The dependency induced by clustered data is dealt with by assuming a form or pattern for the within cluster covariance matrix and this form is assumed to be the same for all clusters. Given the way of dependency is dealt with dependency, the models are also referred to as *variance pattern* models. The population average models are essentially averaging data over heterogeneous clusters, and thus are also sometimes referred to as marginal models.

In this text, *marginal models* will be taken to refer to models that are not conditional on unobserved, random effects. This includes population average models, as well as models for data implied by random effects or cluster-specific GLMMs. For example, the model in (7.7) is a marginal model that is implied by a cluster-specific random intercept model. This marginal model indicates the distribution, mean and variance of the $\underline{y}_{ij}$. The variance in the marginal model is a function of variance due to between groups and within groups. In the case of the population average models, between group/cluster heterogeneity is not modeled but the nature of the relationship observations within a cluster is assumed. In the case of the GLMMs, integrating over $\underline{\boldsymbol{\gamma}}_j$ or collapsing over clusters is done during the estimation of the model parameters. When dealing with GLMMs, the marginal models implied by cluster-specific models and population average models are often not the same.

cja: Not true for all marginal models.

3.6.2 Interpretation of Parameters

Whether a model conditions on $\underline{\boldsymbol{\gamma}}_j$ or not has implications for interpreting the fixed effects parameters. Since the cluster-specific models are conditional, when interpreting the fixed effects, β_{pq}'s, they refer to effects within a cluster. For example, consider the logistic regression for the "cool" kids for classroom j (Model 3 in Table 3.4) where

$$\mathrm{E}(\underline{y}_{ij}|\gamma_{0j}) = \frac{\exp(\beta_{00} + \beta_{10}\mathrm{Gender}_{ij} + \beta_{20}\mathrm{Race} + \gamma_{0j})}{1 + \exp(\beta_{00} + \beta_{10}\mathrm{Gender}_{ij} + \beta_{20}\mathrm{Race} + \gamma_{0j})}.$$

Typically fixed effects parameters are interpreted by considering the effect for a unit change in the predictor variable holding *all* other predictors fixed. In cluster-specific regression models, this includes holding γ_{0j} constant; therefore, when interpreting the fixed effects parameters from the above model, a β_{pq} pertains to students within classroom j or students within classrooms with the same values of γ_{0j}.

In contrast to cluster-specific models, the fixed effects parameters in population average models pertain to the average Level 1 unit in the whole population. In the "cool" kid example, the β_p's refer to the average student and does not depend on their classroom (or classroom effects). For the logistic regression example, a population average logistic regression model analogous to Model 3 in Table 3.4 is

$$\mu_{ij}^{\text{POP}} = \text{E}(\underline{y}_{ij}) = \frac{\exp(\beta_0 + \beta_1 \text{Gender}_{ij} + \beta_2 \text{Race}_{ij})}{1 + \exp(\beta_0 + \beta_1 \text{Gender}_{ij} + \beta_2 \text{Race}_{ij})}. \tag{3.20}$$

The βs in this model are interpreted by considering the effect for a unit change in the predictor holding all other observed effects constant. This interpretation does not require that students come from the same classroom or have equal classroom effects.

The marginal mean given in (3.20 from the population average model differs from that that based on the cluster-specific model. Assuming that clusters are a random sample from a population, the marginal mean implied by a cluster-specific regression equals

$$\begin{aligned} \mu_{ij}^{\text{CS}} &= \int_{\gamma_{0j}} \frac{\exp(\beta_{00} + \beta_{10}\text{Gender}_{ij} + \beta_{20}\text{Race} + \underline{\gamma}_{0j})}{1 + \exp(\beta_{00} + \beta_{10}\text{Gender}_{ij} + \beta_{20}\text{Race} + \underline{\gamma}_{0j})} F(\underline{\gamma}_{0j}) \\ &= \int_{\gamma_{0j}} \text{E}(\underline{y}_{ij}|\underline{\gamma}_j) F(\underline{\gamma}_{0j}), \end{aligned} \tag{3.21}$$

where the integral is basically a weighted sum of the conditional means over all possible values of γ_{0j}, and the weights are values of $F(\gamma_{0j})$, the normal distribution function with mean 0 and variance ψ_{00}.

In general, the means μ_{ij}^{POP} and μ_{ij}^{CS} are not equal. Recall that in a GLM(M), $\text{E}(\underline{y}_{ij}) = \mu_{ij} = g^{-1}(\eta_{ij})$; therefore,

$$\mu_{ij}^{\text{POP}} \neq \mu_{ij}^{\text{CS}} \tag{3.22}$$

$$g^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\beta}) \neq \int \ldots \int g^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma}_j) F(\boldsymbol{\gamma}_j). \tag{3.23}$$

The non-equality between the means further implies that the estimated parameters of the population and cluster-specific models differ. For example, in the "cool" kid example, the parameter estimates from the cluster-specific random effects logistic regression model equal $\hat{\beta}_{00} = 0.31$, $\hat{\beta}_{10} = -0.66$ and $\hat{\beta}_{20} = -1.31$; whereas, those from the population average model[8] are $\hat{\beta}_0 = 0.34$, $\hat{\beta}_1 = -0.48$, and $\hat{\beta}_2 = -1.14$. Comparing the estimated effects, the population average model β's are smaller

[8] The population-average model was estimated using generalized estimating equations that takes into the account the dependency in the data by using a working correlation matrix estimated

(closer to zero) than those from the cluster-specific model. For logistic regression, the estimated β's from a population average models will be closer to zero than those from the random effects models (**?**).

In general, for GLMMs the population average models and cluster-specific models will have different interpretations of the fixed effects coefficients and different parameter estimates. The most notable exception is for normally distributed response variables using the identity link function. Using the identity link, and the linearity of the model, we find that the marginal means are equal,

$$\begin{aligned} \mathrm{E}(y_{ij}) = \overbrace{\int_{\boldsymbol{\gamma}_j} \mathrm{E}(\underline{y}_{ij}|\underline{\boldsymbol{\gamma}}_j) F(\underline{\boldsymbol{\gamma}}_j)}^{\text{Cluster-specific}} &= \overbrace{\eta_{ij}}^{\substack{\text{Population} \\ \text{average}}} \\ \mathrm{E}(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\underline{\boldsymbol{\gamma}}_j) &= \mathrm{E}(\boldsymbol{x}'_{ij}\boldsymbol{\beta}) \\ \boldsymbol{x}'_{ij}\boldsymbol{\beta} &= \boldsymbol{x}'_{ij}\boldsymbol{\beta}. \end{aligned}$$

Taking the anorexia data as an example, the mean for a girl j from the most complex model is

$$\mathrm{E}(\underline{y}_{ij}|\gamma_{0j}, \gamma_{1j}) = \beta_{00} + \beta_{10}x_{ij} + \beta_{01}z_j + \beta_{02}z_j + \gamma_{0j} + \gamma_{1j}x_{ij},$$

and the overall or marginal mean that is obtained by assuming that the girl effects are a random sample from $MVN(\mathbf{0}, \boldsymbol{\Psi})$ and collapsing (integrating) over $\underline{\gamma}_{0j}$ and $\underline{\gamma}_{1j}$ equals

$$\mathrm{E}(\underline{y}_{ij}) = \beta_{00} + \beta_{10}x_{ij} + \beta_{01}z_j + \beta_{02}z_j.$$

The latter model is the same as a population average model. The βs in the cluster specific model are the same as those in the population average model. The interpretation of, for example, β_{10} as being the effect of measurement occasion for girl j is also the effect of measurement occasion in the population of girls treated for anorexia.

3.6.3 Choosing a Model Family

When interest is focused on explaining differences between clusters or groups, as well as accounting for within cluster heterogeneity and the interplay between Level 1 and Level 2 units, the cluster-specific models are better suited for this pur-

from the data. The students within the classes were treated as "exchangeable". The estimated correlation was .28.

pose than the population average model. This is often the case in psychology and the behavioral sciences where understanding the effect of context on individuals is often of interest. For example, in the "cool" kid study, students' tendencies to choose ideal students as being "cool" differed over classrooms depending on the level of aggression in their class. Higher levels of aggression were associated with lower tendencies to choose an ideal student. In the anorexia example, the differences between girls' overall weight and change in weight depended on the type of treatment they received.

When the focus is on making generalizations to the population of Level 1 units but data are clustered, a population average model would be preferable. Such a model should still account for the dependency that is induced by the clustering of data. In the anorexia example, if we are only interested in average weight change and not how girls differ from each other, then the paired dependent t-test suffices. In the case of normally distributed responses, the choice is less critical than in other cases when is comes to values of estimated βs and their interpretation. With these other cases, the goals of study need to be clear and the choice between cluster-specific and population average models needs to be made to match the goals. For example, the population average approach clearly does not meet the researchers' goals in the "cool" kid study where the interplay between students and classroom environment is of primary interest.

3.6.4 Estimation

We observe y_{ij}, x_{ij}, z_j and the variance and covariance of the y_{ij}s. This is the data — it is marginal to unobserved effects. For cluster-specific GLMMs, the model makes implications for the distribution of $\underline{y}_{ij}$ as a function of the x_{ij}s and z_js and the variance and covariances of $\underline{y}_{ij}$ data. For the anorexia data, a GLMM for normally distributed data with an identity link, we were able to derive the distribution for y_{ij}, its mean and the variance and covariance between y_{ij}s. For other GLMMs, the same is true; however, expressing the mean requires integration and typically there are not closed forms for these integrals. These issues have implications for estimation. Estimation is much more difficult for the non-normal cases, but even fitting more complex models to data for normally distributed response variables can be challenging. A number of different estimation methods exist, as well as, different algorithms that implement these methods. Different estimations methods lead to different parameter estimates and estimates of standard errors. These topics are dealt with in more detail in Chapter 4 where not only are the estimation methods and algorithms discussed but also common estimation problems and potential solutions.

For population average models, a common estimation method is to use generalized estimating equations or GEEs. These deal with dependency within clustered data by estimating a within cluster correlation matrix based on the data and a pattern specified that is specified for it. A common choice for the correlation matrix is an "exchangeable" one where the Level 1 units are essentially the same so a single correlation is estimated that equals the correlation between any two units within a cluster. Another choice is "unstructured" where no constraints are placed on the correlation matrix other than it is a proper correlation matrix. This choice is only feasible for cases where cluster size is relative small; otherwise, the number of correlations that must be estimated becomes extremely large. A third common choice is "autoregressive" and is most appropriate for longitudinal data. In population average model for the "cool" kid data an exchangeable correlation makes was used and the common correlation between students within classes was estimated to be .28.

A disadvantage to using a population average model estimated by GEEs is that the parameter estimates are not maximum likelihood estimates — there is not likelihood that is maximized. This caused problems for model assessment and precludes the possibility of likelihood based inference. For more information on GEEs see **?** or specifically on GEEs for categorical data see **?** or Agresti (2007).

cja: Should be mention anything about marginal models estimated by MLE? Such as stuff by Bergsma and colleagues? I didn't want to get into too much detail here but just give a general sense of what's going on.

3.7 Summary

The basic concepts of models with random coefficients were introduced and illustrated in this chapter. From a multilevel perspective, the Level 1 models are standard GLMs except that the regression coefficients in the linear predictor may be random. The Level 2 models are linear regressions with random effect terms for the coefficients of the Level 1 linear predictor. These Level 2 regressions account for systematic and unsystematic differences between groups. The random cluster-specific effect terms account for dependency within clusters. Including effects in the Level 2 models for predictors in the Level 1 model lead to cross-level effects.

All GLMMs have the same basic form

$$\begin{aligned} \underline{y}_{ij}|\underline{\boldsymbol{\gamma}}_j &\sim f(\underline{\mu}_{ij}, \sigma^2) \; i.i.d. \\ g(\underline{\mu}_{ij}) &= \underline{\eta}_{ij} \\ \underline{\eta}_{ij} &= \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\underline{\boldsymbol{\gamma}}_j \\ \underline{\boldsymbol{\gamma}}_j &\sim MVN(\boldsymbol{0}, \boldsymbol{\Psi}) \; i.i.d., \end{aligned}$$

where $f(\,)$ a natural exponential distribution, $g(\,)$ is the link function, $\underline{\eta}_{ij}$ is a random linear predictor, $\boldsymbol{\beta}$ is a $(P \times 1)$ vector of fixed regression coefficients, $\boldsymbol{x}_{ij}$ is a vector of all predictor variables, $\boldsymbol{z}_{ij}$ is vector of all variables that have random coefficients, and $\underline{\boldsymbol{\gamma}}_j$ is a vector of random cluster-specific effects that follows a multivariate normal distribution.

The examples presented in this chapter illustrate an important lesson when dealing with random effects models (not just logistic regression). In particular, what is specified for the fixed part of the model can effect the results for the random part and what specified for the random part of model can effect the results of the fixed part of the the model. For example, in the "cool" kid example when a random intercept was added to the fixed effect model, regression coefficient for $\texttt{Popularity}_{ij}$ dropped in value and its standard error increased to the point that it was no longer significant. This leads to a quandary or "Catch-22"[9], because as was hinted at in this chapter and will be discussed in more detail in Chapter **??**, valid statistical inference for fixed effects require that the random part of the model is correct and statistic inference for the random part require that the fixed part of the model is correct.

Adding random effects to linear predictors of GLMs introduces a number of difficulties (e.g., estimation, model assessment) and differences relative to standard GLMS (e.g., interpretation of parameters, statistical inference for parameters). However, the benefits of using GLMMs far outweigh the additional problems encountered from introducing random effects into models. In the subsequent foundational chapters, procedures and tools for dealing with these added complexities in the use of GLMMs are presented.

Problems & Exercises

3.1. For this problem use the anorexia data analyzed in this chapter. If the null hypothesis of the paired dependent t-test is true (i.e., $H_0 : \mu_1 = \mu_2$), what model should fit the data (this model was not reported in this chapter)?

(a) Report this model as a multilevel model.
(b) Report the corresponding linear mixed model.
(c) Give the marginal model implied by (a) and (b).

[9] "Catch-22" is a phrase made popular by Joseph Heller's (1961) novel *Catch-22* and refers to the situation where a paradoxical rule or a set of conditions creates problems such that there is no way to resolve the problem.

(d) Fit this model to the data using REML estimation and interpret parameters estimates.
(e) How do the variance estimates from part (d) compare to those from Model 1 in Table 3.2? Explain.
(f) In terms of sample statistics, what do the model parameters equal? If needed, additional sample statistics can be computed from those in Table 3.1.

3.2. For this problem use the anorexia data analyzed in this chapter. Consider Model 4 described in the text. Fit a model with the same Level 2 model for the slope but no predictor variables in the model for the intercept.

(a) Report this model as a multilevel model.
(b) Report the corresponding linear mixed model.
(c) Give the marginal model implied by (a) and (b).
(d) Fit this model to the data using ML estimation.
(e) Compare the estimated model to Model 4 in terms of population average model, variance explained, and how well the model fits the data (i.e., AIC and BIC).

3.3. A model for the anorexia data was mentioned where $\psi_{11} = 2\psi 00$ could be fit and ψ_{10} could be estimated.

Based on the parameter estimate for Model 4, is it reasonable for $\psi_{11} = 2\psi_{00}$? Fit the model and compare it to Model 4.

3.4. Use Model 4 for that was fit to the anorexia data.

What is the variance of the girls' weight before treatment? What is the variance after treatment?
Compare these model based estimates to the variances computed using the data (those in Table 3.1).

3.5. Compute ICC for random intercept models fit to anorexia data (Model 1 and 2). Idea here is to start thinking about this.

3.6. For this problem fit the last model in Table 3.4 except do not fit a fixed effect for gender. Compare this model to the one in the table.

3.7. Re-do logistic regression and normal examples from Chapter **??** and compare results....this will be a number of exercise and detail instructions on what to do will be given.

Chapter 7
Linear Mixed Models for Normal Variables

Models, methods and procedures for normally distributed response variables are the most well known and developed in the GLMM family. The models go by different names in different fields, including *hierarchical linear models* (HLM) in education and social sciences, *variance component models* in biometrics, *linear mixed models* (LMM) in statistics, and *intercept and slopes as outcomes*. Some simple models for normal responses were introduce in Chapter 3, and this chapter covers more complex models and procedures that are particularly useful when analyzing normally distributed response variables. A general approach to building models for clustered data is discussed and illustrated. Many of the issues that arise for normal response variables apply to other types of data and random effects models.

The models for normal response variables can be represented in three alternative, but equivalent ways: as an HLM, an LMM, and as a GLMM. The HLM and LMM are primarily used in this chapter for exposition. The general model can be summarized as a GLMM,

$$\begin{aligned} \underline{y}_{ij}|\underline{\boldsymbol{\gamma}}_j &\sim N(\underline{\mu}_{ij}, \sigma^2)\ i.i.d. \\ \underline{\mu}_{ij} &= \underline{\eta}_{ij} \\ \underline{\eta}_{ij} &= \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\underline{\boldsymbol{\gamma}}_j \\ \underline{\boldsymbol{\gamma}}_{ij} &\sim MVN(\boldsymbol{0}, \boldsymbol{\Psi})\ i.i.d., \end{aligned} \tag{7.1}$$

where $\boldsymbol{x}_{ij}$ consists of all predictor variables, $\boldsymbol{z}_{ij}$ consists of a subset of the variables in $\boldsymbol{x}_{ij}$ that have random between cluster differences, and $\underline{\boldsymbol{\gamma}}_j$ is a vector of the random effects that account for unsystematic heterogeneity between clusters. The distributional assumptions for the random effects are $\underline{\boldsymbol{\gamma}}_j \sim MVN(\boldsymbol{0}, \boldsymbol{\Psi})\ i.i.d.$

Regardless of whether the model is written as a GLMM, HLM or LMM, due to the normality assumptions, a unique aspect of this model is that the implied marginal (unconditional) model based on (7.1) is also the population average models; that is,

$$\underline{y}_{ij} \sim N(\boldsymbol{x}'_{ij}\boldsymbol{\beta},\ (\boldsymbol{z}'_{ij}\boldsymbol{\Psi}\boldsymbol{z}_{ij} + \sigma^2))\ i.i.d.$$

In other words, when collapsing the cluster-specific model over the unobserved random coefficients (i.e., $\underline{\boldsymbol{\gamma}}_j$), the distribution of the response variable $\underline{y}_{ij}$ is normal where the mean equals a linear combination of the fixed effects (i.e., $\mathrm{E}(\underline{y}_{ij}) = \mu_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$), and the variance is a function of the variances and covariances of the random effects (i.e., $\mathrm{var}(\underline{y}_{ij}) = \boldsymbol{z}'_{ij}\boldsymbol{\Psi}\boldsymbol{z}_{ij} + \sigma^2$).

Data analysis should start by a examination the data, and in Section 7.1, various useful exploratory techniques are presented for LLMs. Some of these can be adapted to other GLMMS. Section 7.2 presents the basic steps in modeling clustered or hierarchically structured data and in the process adds complexity to models for normal responses. In Section 7.3, extensions of R^2 from normal linear regression are presented as one way to assess how well a model represents the data. The material in the final two sections, Section 7.4 and **??**, are discussed in the context of normally distributed response variables; however, they also apply to all other GLMMs. In Section 7.4, the issue of how and whether to center Level 1 predictor variables is discussed, and in Section 7.5, the two level model is extended three levels.

7.1 Exploratory Data Analysis

Whether a researcher has preconceived hypotheses or is conducting an exploratory analysis, a good starting point for any analysis is to "look" at the data. This should include examining the univariate distribution of each variable to possibly identify anomalous observations, and unusual or unexpected aspects of the variables. For example, in study by Espelage et al. (2003) where middle school children are nested within peer groups, gender might seem like a Level 1 variable; however, a closer look reveals that most peer groups are either all girl or all boys (i.e., gender is a Level 2 variable).

In Chapter 3, a number of implications for data based on HLMs/LMMs were found that could be useful in an exploratory data analysis and more generally to linear predictors of GLMMS (i.e., $\underline{\eta}_{ij}$). The exploratory data analysis (EDA) described here can help to detect possible fixed and random effects and may lead to a reasonable preliminary model or sub-set of models. The latter are especially useful in more exploratory analyses. Specific procedures, most of which are graphical,

depend on the nature and role of the predictor variables. The EDAs discussed and illustrated below are not the only possibilities, but they are ones that we have found to be particularly useful. An added benefit of looking at data is that a researcher can further assess whether the final model fit to the data yields fitted values that resemble the data and capture the main features of the data.

In Section 7.1.1, methods focusing on the Level 1 model of an HLM are presented, and in Section 7.1.2, methods to identify potential Level 2 are presented. The methos presented in both of these sections provide information about possible random effects. An EDA may suggest a small number of preliminary models and Section 7.1.3 covers how to examine the goodness-of-fit of these preliminary cluster specific models. Further methods for exploratory analysis of clustered data, in particular for predictors of random effects structures in the context of longitudinal data that can be used for other types of clustered data can be found in Verbeke & Molenberghs (2000) and **?**.

7.1.1 Graphing Level 1 Predictors

The first EDAs presented are ways of graphing the response variable versus potential Level 1 predictors. The goal is to try to identify effects that contribute to predicting the mean of the response variable conditional on observed covariates or predictors. In Section 7.1.1.1, graphs that are appropriate for numerical or continuous predictors are presented and in Section 7.1.1.2 those for discrete predictors are considered.

7.1.1.1 Numerical/Continuous Predictors

Recall that the LMM model for the response variable is conditional on cluster-specific effects. Plotting the values of the response variable versus potential Level 1 (within cluster) predictors for each cluster can reveal information about whether and how the predictors are related to the response, whether there are interactions between predictors, whether there is variation between clusters, show the nature of between cluster variation, whether there are anomalous clusters or observations within clusters, and possibly reveal other unexpected information about the data.

An example of this first type of graph uses data from a study by **?** on whether optimistic expectancies of first year law students can predict cell-mediated immunity (CMI). Student's CMI and optimism were measured at most five times during their first year of law school. Students are the clusters and the multiple measurements of CMI are nested within students. In Figure 7.1.1.1, CMI is plotted against optimism for a sample of 24 of out the 121 students in the study. The lines in the

Table 7.1 The distribution of the number of observations per student in the study of immunity and optimism (**?**).

n_j	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6	4.96	6	4.96
2	10	8.26	16	13.22
3	8	6.61	24	19.83
4	21	17.36	45	37.19
5	76	62.81	121	100.00

graphs are linear regressions. For those students (clusters) with multiple measures, the assumption of a linear relationship between CMI and optimism seems reasonable. If data are not linear, other curves such as cubic loess can be overlaid in the figures.

Noticeable in Figure **??** is that the number of observations per student n_j varies from 1 to 5. The distribution of the number of observations per student in the entire sample is reported in Table 7.1. Most students (i.e., 80.17%) have 4 or 5 observations and only 5% have only one observation[1]. If there are many observations per cluster, then it is possible to assess whether the distribution of data within clusters is approximately normal, as well as whether the variance is constant for each value of the predictor. Even with 4 or 5 observations per cluster, departures from normality may still be suggested. Examples where it is possible to detect departures from normality are included in the exercises at the end of this chapter and in Chapter 8.

Graphs such as Figure 7.1.1.1 also reveal between cluster information. The strength of the relationship and direction of the linear regressions appears to differ over clusters (i.e, students). To further examine this, in Figure 7.1.1.1, the cluster-specific linear regressions for all $N = 121$ students in the sample are plotted in the same figure[2]. As would be expected if the intercept is random, considerable variability in the overall level of CMI when optimism equals 0 is found. There is also considerable variability in the slope of the regressions and suggests that the effect of optimism differs over students. Another aspect of the data that will be important for modeling is the fact that different students appear to have on average different levels of optimism (i.e., the location on horizonal axis in Figures 7.1.1.1 and 7.1.1.1). For example, in Figure 7.1.1.1, student #68191 has on average low optimism scores, student #72195 has high scores, and student #62159 has scores in the middle. We return to this issue in Section 7.4 where centering of within cluster (Level 1) predictor variables is discussed.

[1] The data from all $N = 121$ students are used in the analysis and fitting models to data.

[2] xxxx refer to these are spaghetti plots. . . or is it plots that join points or overall smooth curves?

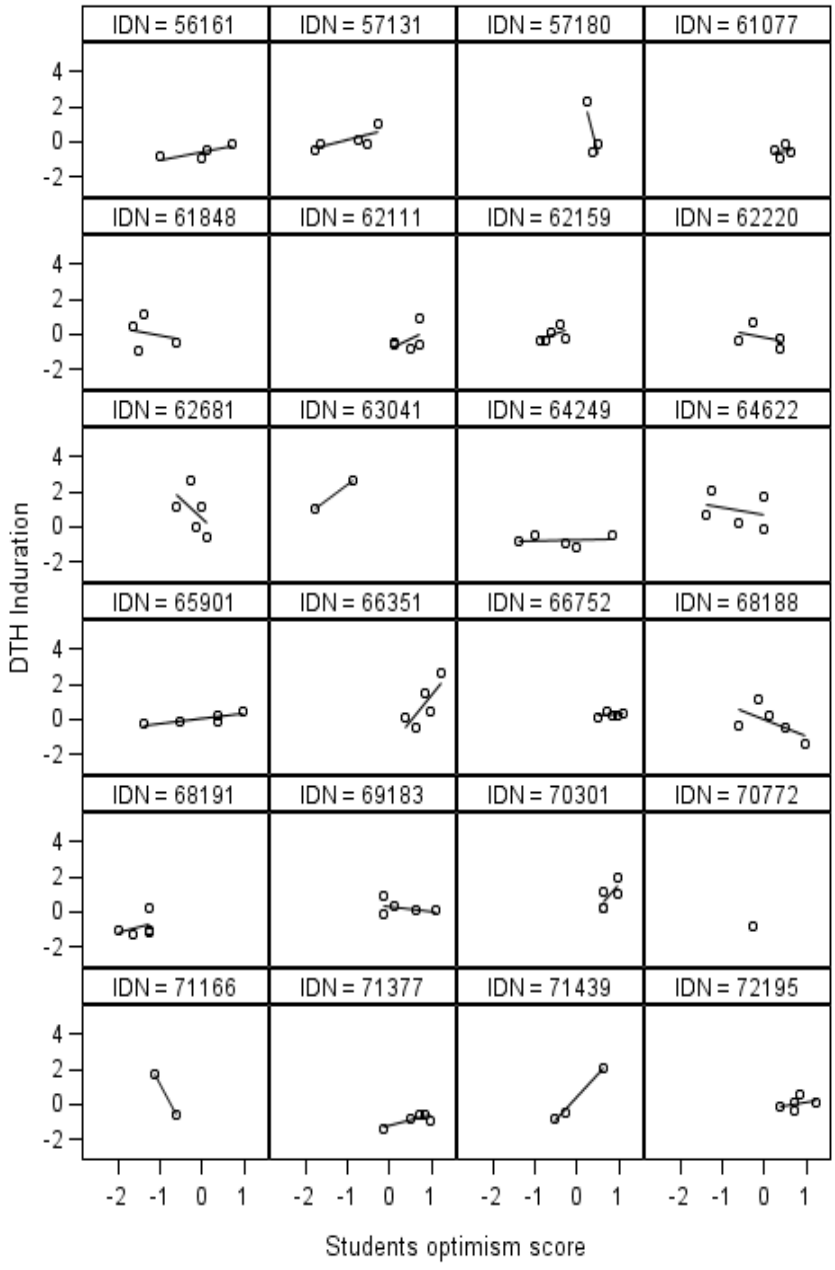

Fig. 7.1 Plot of a sample of student-specific data from **?** where a measure of immunity as measured by HTH Induration is plotted against students' levels of optimism with linear regressions drawn in each graph.

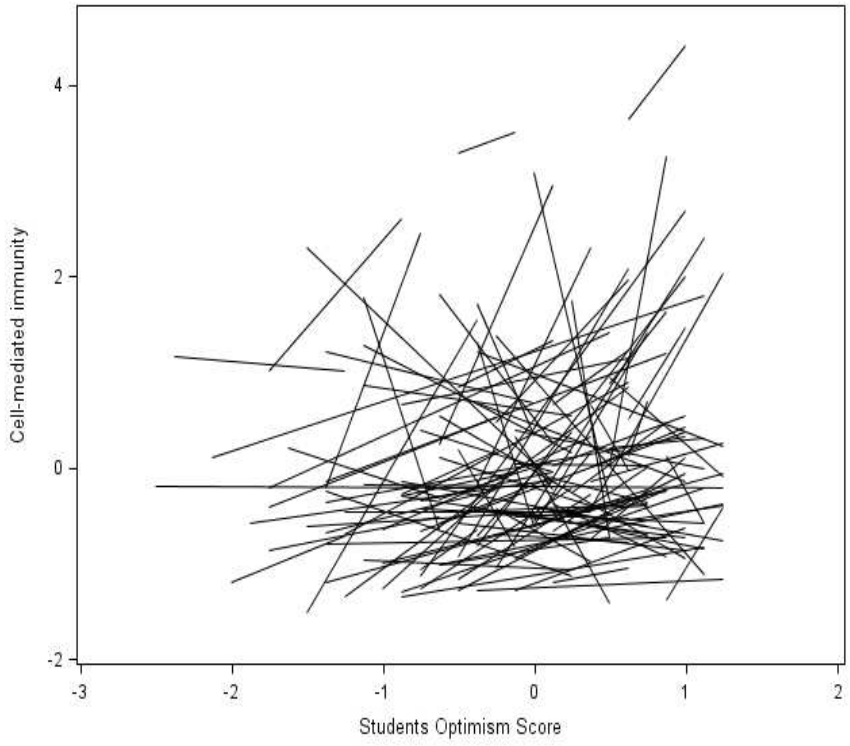

Fig. 7.2 Overlay of linear regressions fit to each student's data.

7.1.1.2 Discrete Predictors

For discrete Level 1 or within cluster predictors, plotting the response variable by levels of the discrete predictor tends not to be particularly informative because there will be multiple points for each cluster for each level of the discrete variable. Rather than plot the data as in the previous section, a more informative graph is to plot the mean response for each cluster for each level of the discrete predictor. Examining means is reasonable based on the fact that the fixed effects represent the means of the response variable (i.e., $\mathrm{E}(\underset{\sim}{y}_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta}$).

As an example, consider the classroom segregation data from Rodkin et al. (2007) that was introduced in Chapter 2. In this data set, the response variable is measure of a child's level of segregation with respect to mutual friendships within their classroom. Of interest is whether and how a child's ethnicity and

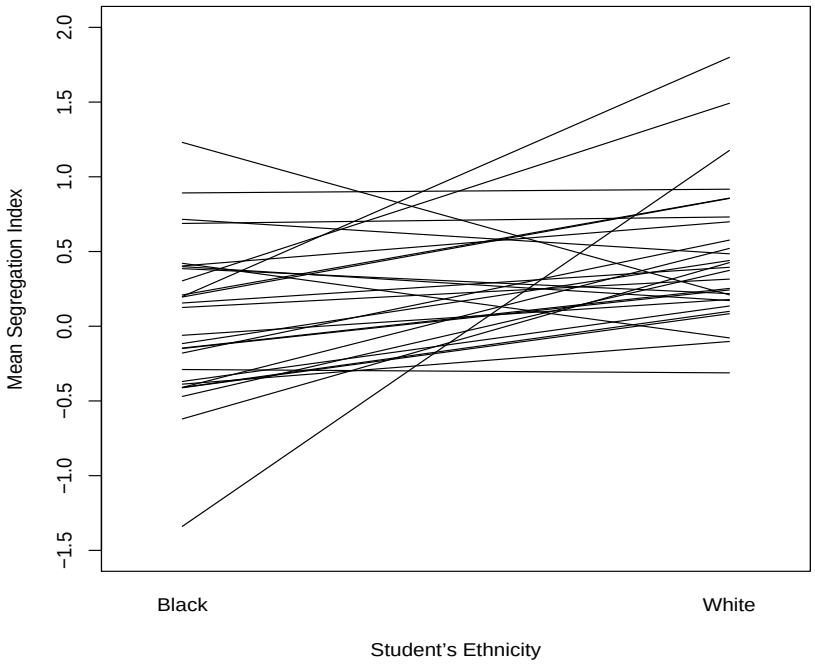

Fig. 7.3 Mean of the segregation index from Rodkin et al. (2007) plotted versus student's ethnicity where lines connect points for each classroom.

gender are related to segregation. Figure 7.1.1.2 is a plot of the class (cluster) mean segregation index for each ethnicity where a line connects the means for each classroom, one line per classroom. The vertical spread between the lines indicates between classroom variation in terms of intercepts and is consistent with a random intercept model. If ethnicity is a predictor of segregation, then the lines should have non-zero slopes. Some of the lines are flat, a few decrease, but most show positive slopes (i.e., higher mean segregation for white children). Overall only 5 out of 33 have slopes markedly different from the rest. Taken together, this suggests that the effect of ethnicity maybe a fixed effect but not random.

As a second example, Figure 7.1.1.2 is a plot of classroom mean segregation for boys and girls where the means have been connected by lines. The variability in the level of the classroom means differ between girls and boys; that is, the intercept variance may not be constant for gender. Most noticeable is the appearance

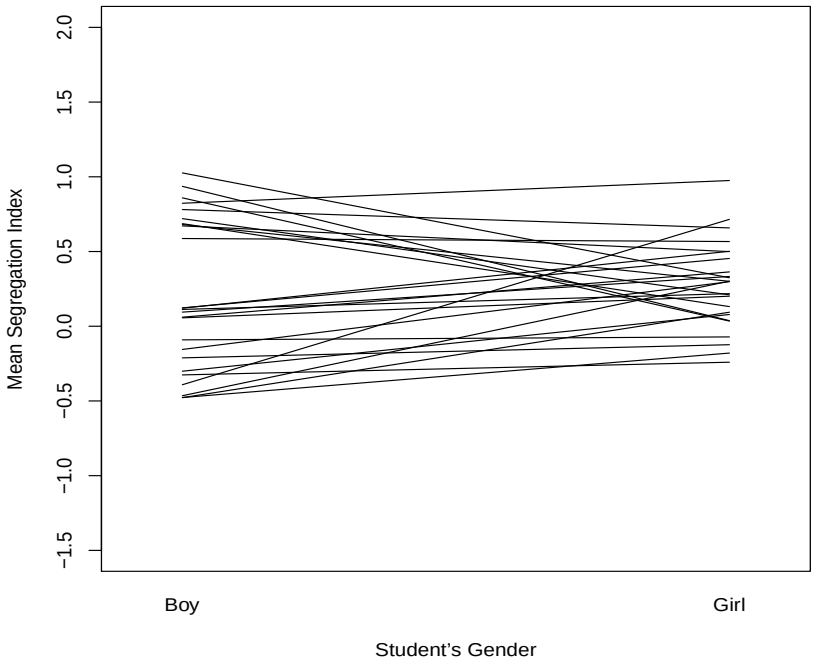

Fig. 7.4 Mean of the segregation index from Rodkin et al. (2007) plotted versus student's ethnicity (a) and versus student's gender (b) where lines connect points for each classroom.

of possibly two groups of boys—those with high segregation indices and those with lower values. It would be important and informative to ascertain what distinguishes these two types of boys and what accounts for differences in variability between boys and girls. It is possible that the critical variable(s) that distinguishes these classrooms may not have been collected; however, with this data set, it may be the case that there is a gender by ethnicity interaction. To examine this possibility, the mean segregation indices for boys and girls of each ethnicity are plotted in Figure 7.1.1.2 with approximate 95% confidence intervals on the means. The standard errors used to compute the confidence intervals ignores the clustering in the data so are smaller than they should be; however, they give a sense of the dispersion around the means. The pattern of the interaction suggests that black boys have on average lower values of segregation than white students and black girls. In other words, the critical variable that might help explain the appearance of two

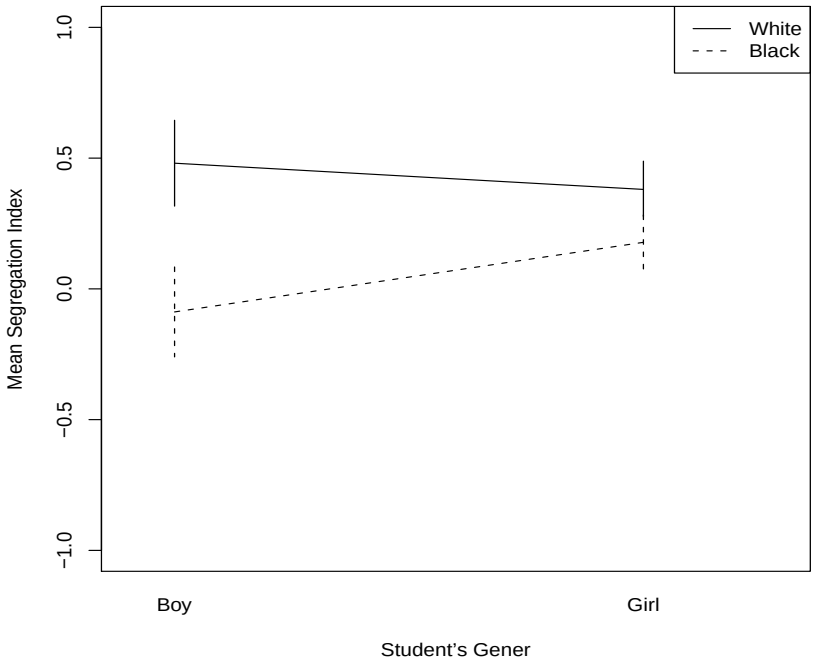

Fig. 7.5 Mean of the segregation index from Rodkin et al. (2007) plotted versus student's gender with a separate curve for each student ethnicity.

groups of boys in Figure 7.1.1.2 is ethnicity (i.e., an interaction between gender and ethnicity).

7.1.2 Graphing Potential Between Cluster Predictors

Similar to the previous section, graphs for numerical/conitnuous between cluster predictors are illustrated followed by discrete ones. Detecting effects is similar to the previous section; however, between cluster (Level 2) predictors of coefficients of the within (Level 1) intercepts and predictors variables. Plots of regression lines of the response variable by a possible Level 1 predictor can include information regarding possible Level 2 predictors. Means over different values of the between

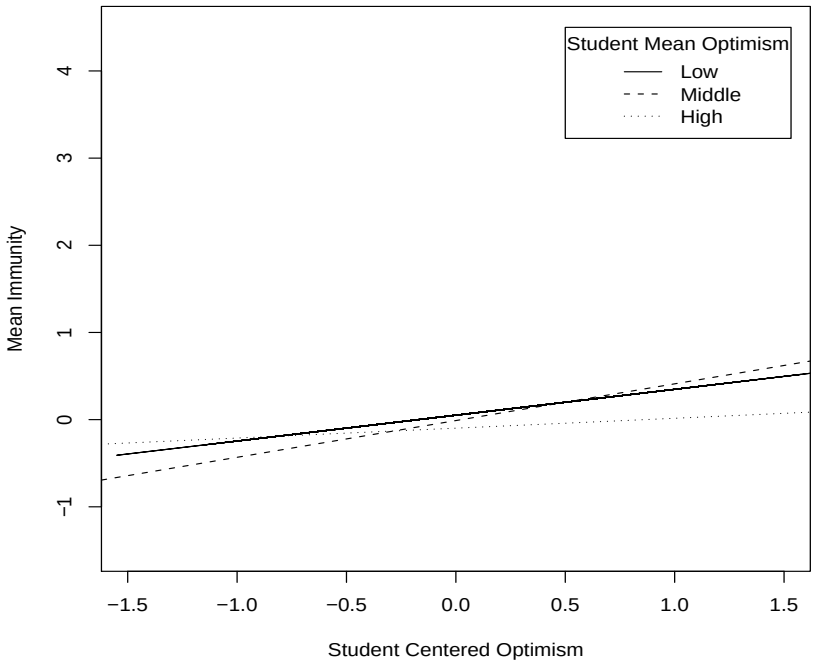

Fig. 7.6 Regressions of immunity as the response and student mean centered optimism as the predictors with different regressions for students that have low, medium and high levels of mean optimism.

cluster variables can be plotted rather than raw data. Such plots can help identify predictors of a random intercept and random slopes. If such lines show vertical separation, then the Level 2 variable may be a good predictor of the intercept. If the lines are not parallel (i.e., cross, diverge or converge), then Level 2 variable might help account for between class variability in the effects of the Level 1 variables (i.e., a cross-level interaction). Lastly if the lines are basically on top of each other, the Level 2 effect is most likely not a useful predictor of between cluster differences of either the intercept or slope.

7.1.2.1 Numerical or Continuous Level 2 Predictors

Plotting separate regressions for each level of a continuous Level 2 predictor is not practical; however, a continuous variable can be grouped into a few levels (e.g., low, middle, high) and regressions fit to each of these levels[3]. For example, in the immunity and optimism example from **?**, a student's average optimism is a possible predictor of differences between student's in terms of their immunity. The data were split into three groups according to students' mean optimism score. Figure 7.1.2.1 is a plot of the regressions for each of these three groups with the measure of immunity regressed onto students' centered optimism score. The regression lines are basically on top of each other and will likely not help account for variability between students. The slope of these lines is slightly positive and indicates that the fixed effect for student centered optimism may be present in the data. Although mean optimism does not look like a good predictor of differences between students either in terms of the intercept or random effect of optimism, for substantive reasons, initial modeling of data should include both the student centered and student mean optimism.

7.1.2.2 Discrete Level 2 Predictors

Discrete Level 2 predictors can be graphed in the same way as continuous predictors; however, the first step of breaking data into groups in not necessary. The different levels of a discrete predictor provide natural grouping of clusters and lines connecting means for each cluster can be plotted. For example, consider the classroom segregation data from Rodkin et al. (2007). Figure 7.1.2.2 consists of a plot of the mean segregation index by student ethnicity for each classroom type. The lines cross and suggest racial composition of a class (i.e., class type) may be a good predictor of the effect of ethnicity on segregation. Contrast Figure 7.1.2.2 to Figure 7.1.2.2 where in the latter has gender along the horizonal axis. The lines for each classroom type are nearly indistinguishable and suggest that the effect of gender will not be mediated by classroom type.

7.1.3 Preliminary Level 1 Model

Based on the goals of an analysis or as suggested by an exploratory analysis, a sub-set of preliminary cluster-specific fixed effects models may exist. If cluster

[3] Grouping a continuous variable is only for the purpose of graphing. When the data are modeled, the variable should not be be collapsed into levels because this throws information away and can lead to spurious results.

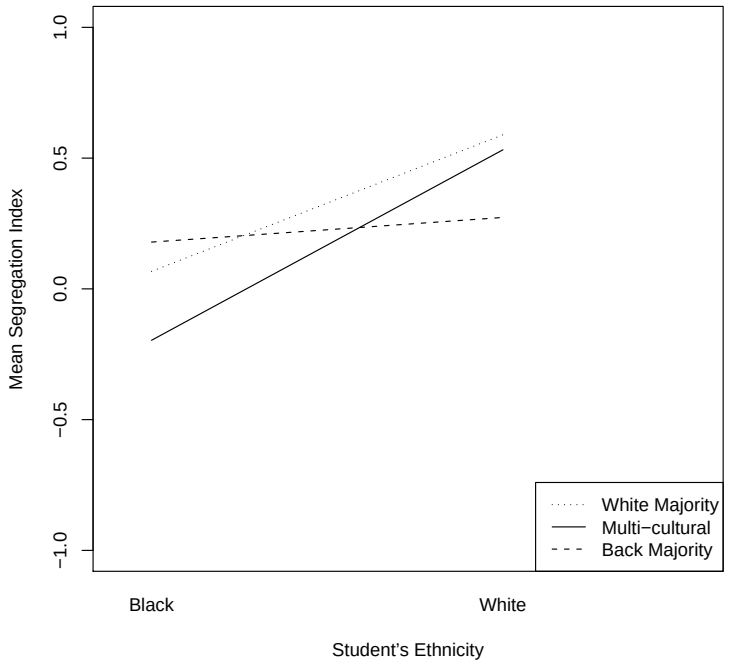

Fig. 7.7 Mean integration index from Rodkin et al. (2007) for ethnicity with a separate line for each classroom type.

sizes are large enough relative to the number of Level 1 predictor variables, a regression model can be fit to each cluster thus allowing each cluster to have its own intercept and coefficients for predictor variables. The multiple R^2s from the regressions fit to each clusters' data measure of the best achievable goodness-of-fit of the model to each clusters' data. Since R^2 may depend on the cluster size n_j, the R^2 can be plotted versus n_j (Verbeke & Molenberghs 2000).

As an example, the segregation index for each cluster (classroom) was regressed onto three different sets of possible predictors. An overly simplistic model was fit that only included gender as a predictor, one that included gender and ethnicity as main effects, and the most complex model included gender, ethnicity and an interaction between gender and ethnicity. The most complex model is the one suggested from the graphs presented above and the simple model was fit for illustration. The R^2s for these models for each classroom are plotted in Figure 7.1.3

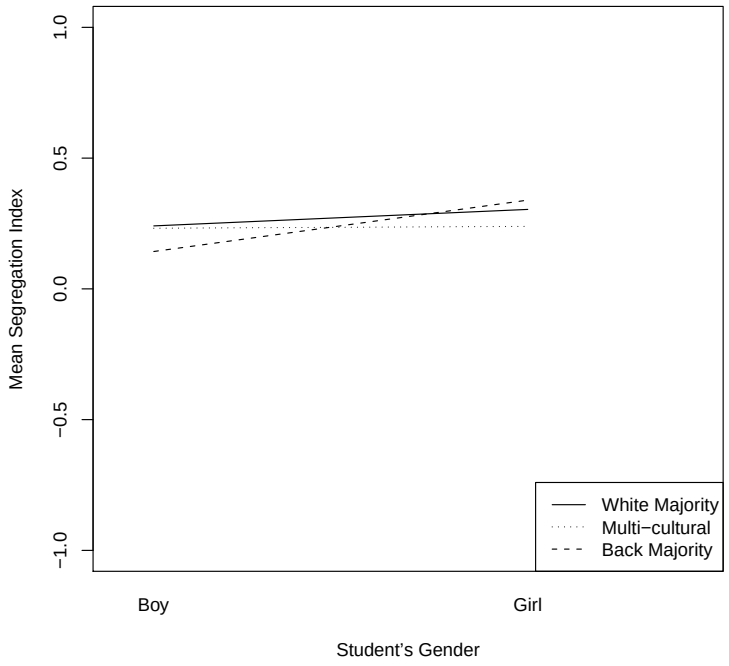

Fig. 7.8 Mean integration index Rodkin et al. (2007) for gender with a separate line for each classroom type.

where the upper left plot is for the simplest model, the upper right is the main effects model, and the lower left is the gender by ethnicity interaction model. The improvement by adding ethnicity to the simple model is readily apparent and the subsequent addition of the interaction noticeably improves the fit of the models to each classroom's data. The improvement between the main effects and the interaction model is less obvious; therefore, the R^2s from the main effects and the interaction models are plotted against each other in the lower left part of Figure 7.1.3. Points that lie above the identity line are those classrooms whose data were better fit by the regression that included the interaction between gender and ethnicity. The classrooms on the identity line are those whose data were equally well fit by the two models.

The horizonal lines in Figure 7.1.3 is a global measure or meta-R^2 of how well the model fits the classrooms' data. Meta-R^2 equals the ratio of the sum of model

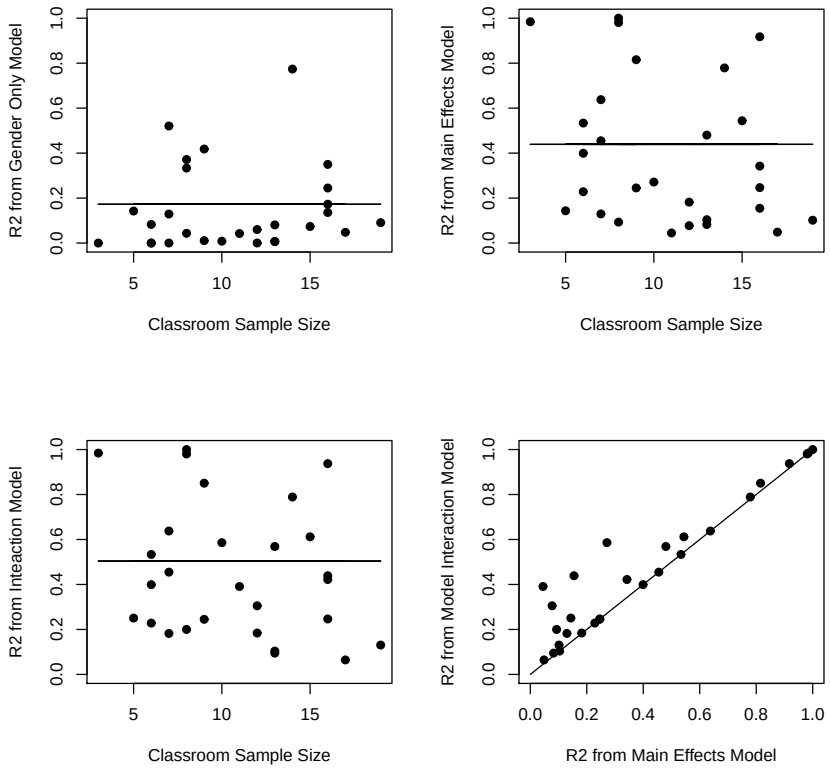

Fig. 7.9 Multiple R^2s for models fit to each classroom's data from plotted versus classroom size where gender is the only predictor (upper left), main effects for gender and ethnicity are predictors (upper right), and main effects and an interaction between gender and ethnicity are included (lower left). The horizonal lines corresponds to the meta-R^2 (i.e., $R^2 = .17$, .44, and .50, respectively). The lower right figure is a plot of the R^2 comparing the main effects and interaction models.

sum of squares divided by the sum of total sum of squares; that is,

$$\text{meta-}R^2 = \frac{\sum_{j=1}^{N}(\text{Model Sum of Squares})_j}{\sum_{j=1}^{N}(\text{Total Sum of Squares})_j},$$

(Verbeke & Molenberghs 2000). In the classroom segregation example, meta-$R^2 =$ .17 for the simple model, meta-$R^2 = .44$ for the main effects model, and meta-$R^2 = .50$ for the most complex model. There is overall a clear improvement by adding ethnicity to the model and some improvement appears to have been gained by adding an interaction between gender and ethnicity. At this point the interaction appears to be important; however, the final decision regarding the interaction will come by modeling the data.

7.2 Modeling the Data

Modeling data is a process that is guided by substantive theory, results of exploratory analysis, and results from fitting various models to the data. The two most common approaches advocated in the literature on GLMMs are a "step-up" and a "top-down" method (**?**). The step-up method starts with a simple within structure with a random intercept to which fixed effects are successively added after which random effects are then added to the model (**??**). The top-down approach starts with the most complex polynomial model for within (Level 1) effects with a random model intercept where the first step is to determines the correct order of the polynomial and in the second step to build the random part of the model (**?**Verbeke & Molenberghs 2000). The top-down approach lends itself better than the step-up approach to longitudinal data where change is not always linear. The little research that exists on the subject of modeling approaches with LMMs finds that the step-up approach tends to identify the true model in simulations more often than the bottom-down method (**?**).

In this section a version of the step-up approach is used where we start with simple models and work toward more complex ones. Besides the work by **?**, a reason for preferring this approach is that a model with all potentially interesting fixed and random effects or that suggested by the exploratory analysis may fail as a starting model because complex models may not be supported by the data; that is, such models may fail to converge or yield improper solutions (e.g., $\hat{\boldsymbol{\Psi}}$ that is not a proper covariance matrix).

Section 7.2.1 starts the modeling process by considering models with a random intercept and adds fixed effects that potentially can account for within cluster variation. In Section 7.2.2, complexity is increased by adding covariates that potentially account for between cluster variation. In Section 7.2.3, more random effects are added to the model to account for unsystematic differences between clusters and to further build a model for the variance of $\underline{y}_{ij}$. The modeling process is not linear because what is in the random part of the model affects what in the fixed part and visa versa. In Section 7.2.4, both the fixed and random effects are re-examined. Throughout this section, the classroom segregation data from Rod-

kin et al. (2007) is used to illustrate the process and properties of the models, including the interpretation of the results in Section 7.2.4.

7.2.1 Systematic Within Cluster Variation

Systematic differences in the response variables due to within cluster variation and between cluster differences are modeled by fixed effects. In this section, the fixed effects that describe characteristics or attributes of the units within clusters will be entered into the model.

7.2.1.1 The Unconditional Means Model & the ICC

In many published applications of multilevel random effects models, the first model reported is the *unconditional means* model, also known as the *empty* or *null HLM*. This model is a random intercept model with no predictor variables and is just a simple random effects ANOVA. The LMM of the null HLM is

$$\underline{y}_{ij} = \beta_{00} + \underline{\gamma}_{0j} + \underline{\varepsilon}_{ij},$$

where

$$\begin{pmatrix} \underline{\varepsilon}_{ij} \\ \underline{\gamma}_{0j} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \psi_{00} \end{pmatrix}\right) \; i.i.d. \tag{7.2}$$

For cluster j, the conditional mean equals $\mathrm{E}(\underline{y}_{ij}|\gamma_{0j}) = \beta_{00} + \gamma_{0j}$, and the unconditional mean or overall average is $\mathrm{E}(\underline{y}_{ij}) = \beta_{00}$.

Although null model is not particularly interesting, it provides a baseline against which to compare more complex models. The parameter estimates from the null or empty model are used in the computation of various statistics, including a measure of within cluster dependency, a measure of the proportion of variance due to differences between clusters, and extensions of multiple R^2 to LMMs. From Chapter 3, it was shown that the variance of the response variable based on a random intercept model equals

$$\mathrm{var}(\underline{y}_{ij}) = \psi_{00} + \sigma^2,$$

and the covariance between two observations within a cluster equals

$$\mathrm{cov}(\underline{y}_{ij}, \underline{y}_{i'j}) = \psi_{00}.$$

Using these two pieces of information, the *interclass correlation* (ICC) can be computed,

$$\text{ICC} = \frac{\text{cov}(\underline{y}_{ij}, \underline{y}_{i'j})}{\sqrt{\text{var}(\underline{y}_{ij})}\sqrt{\text{var}(\underline{y}_{i'j})}} = \frac{\psi_{00}}{\psi_{00} + \sigma^2}.$$

The ICC is the correlation between any two randomly sampled observations within the same cluster and is assumed to be the same over clusters. This is also the proportion of variance of the response accounted for by differences between clusters.

For the classroom segregation data from Rodkin et al. (2007), the MLE parameter estimates and various fit statistics for the null model are reported in Table 7.2.1.2 under the column labeled "Model 1". Recall from Chapter **??** that if the random structure is too simple, the estimated model based standard errors can be very poor and leading to invalid t and F tests of β parameters. Since there is only a random intercept and our exploratory analysis indicate that there might be random effects for gender, if anything, the structure is too simple. The reported standard errors in Table 7.2.1.2 and Table **??** are the robust or sandwich estimates that may still yield a reasonable statistical tests of fixed effects.

cja: I'm not sure why Table number is wrong.

For Model 1, the $ICC = .0633/(.0633 + .3253) = 0.16$ indicating that 16% of the variation of the response is due to differences between the level of segregation over classrooms, and that the correlation between observations within the same classroom equal .16. Statistically testing whether $H_0 : \psi_{00} = 0$ versus $H_1 : \psi_{00} \neq 0$, the likelihood ratio statistic is $LR = 20.59$, and using a mixture chi-square distribution, $p < .01$. This result corroborates what is seen in Figures 7.1.1.2 and **??**; namely, that there are differences between the classrooms' intercepts.

Although reporting the ICC has been standard practice in applications, obtaining a $\psi_{00} = 0$ does not necessarily imply within cluster independence, nor does it imply the absence of between cluster random variation. For example, consider Figure 7.2.1.1 where a response variable y_{ij} is regressed onto a predictor x_{ij} for each cluster. The intercept is the same for all clusters (i.e., when $x_{ij} = 0$, $\gamma_{0j} = 0$), but there is clearly variation in the slopes. In this case, if variation in slopes is random, then there will be dependence of observations and differences between clusters. A pattern as in Figure 7.2.1.1 might occur in an experiment on some performance measure where all subjects start at the same point, but over the course of the experiment subjects increase their performance at different rates as function of x_{ij}.

7.2.1.2 Within Cluster Predictors

In normal linear regression, part of the process of modeling data is choosing predictor or explanatory variables, considering interactions, and possible transfor-

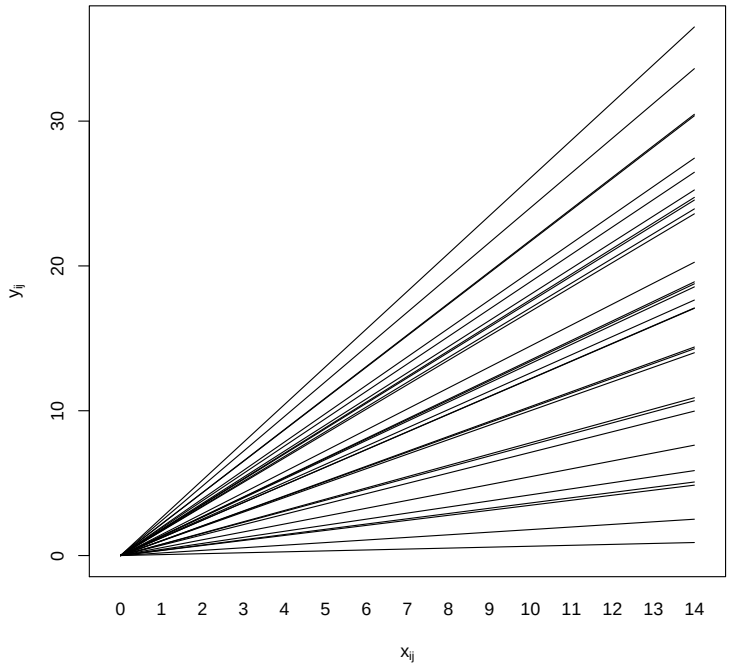

Fig. 7.10 A hypothetical example where the ICC $= 0$; however, there are random slopes.. . . maybe re-do this figure so that figure have as lines with negative slopes so that on average line might be flat.

mations of them. The same is true for LMM. If the predictor is important in a statistical sense, adding predictors to a cluster-specific regression model for observations within clusters should lead to decreases in the within cluster variance, σ^2. The between cluster variance for the intercepts, ψ_{00}, should not be affected unless a predictor carries both within and between cluster information.

For example, in the classroom segregation data, an important issue whether a child's ethnicity has an effect on their level of segregation. Coding dummy coding ethnicity as $x_{1ij} = 1$ for White and 0 for Black students, the model with ethnicity is

$$\begin{aligned} \text{Level 1:} \quad \underline{y}_{ij} &= \underline{\omega}_{0j} + \omega_{1j} x_{1ij} + \underline{\varepsilon}_{ij} \\ \text{Level 2:} \quad \underline{\omega}_{0j} &= \beta_{00} + \underline{\gamma}_{0j} \\ \omega_{10} &= \beta_{10}, \end{aligned}$$

and the LMM or combined model is

$$\underline{y}_{ij} = \beta_{00} + \beta_{10} x_{1ij} + \underline{\gamma}_{0j} + \underline{\varepsilon}_{ij}, \tag{7.3}$$

where the distributional assumptions are the same as in (7.2). The estimated parameters for this model are reported in Table 7.2.1.2 under the column labeled "Model 2". All effects of models reported in Table 7.2.1.2 that are 0 (due to dummy coding) are not reported in the table. For example, in Model 2, only one parameter is estimated for ethnicity due to the dummy coding of ethnicity. The fitted mean value for Black students is 0, and in this simple model, it equals the intercept. According to the information criteria adding ethnicity results in a better model. Furthermore, the likelihood ratio test of $H_o : \beta_{10} = 0$ supports this conclusion (i.e., $LR = 548.5106 - 517.4178 = 31.10$, $df = 1$, $p < .01$). The estimated variance $\hat{\sigma}^2$ has decreased from 0.33 to 0.29, and the variance of the intercepts has remained essentially the same.

In a random intercept model that includes fixed effects for Level 1 explanatory variables, the variance of $\underline{y}_{ij}$ equals $\psi_{00} + \sigma^2$. A *residual ICC* can be computed that measures the proportion of the remaining variance of the response variable due to between cluster heterogeneity. The formula is the same as that for the ICC. Since the value of σ^2 becomes smaller as fixed effects that are good predictors are added, the value of a residual ICC should become larger. For example, Model 2 has a residual ICC value of $.0650/(.0650 + .2910) = .1826$ and is larger than the ICC from Model 1 (i.e., .1629). The residual ICC may decrease when fixed effects accounting for between cluster heterogeneity (i.e., Level 2 predictors) are included in the model, because adding such effects can decrease ψ_{00}, as well as effect the estimate of σ^2 (Snijders reference that adding Level 2 fixed effect changes estimate of ψ_{00} and σ^2–).

On average no relationship between x_{ij} and y_{ij} may exist even though the effect of x_{ij} randomly differs over clusters. In such a case, β for the x_{ij} equals 0, but the explanatory variable x_{ij} should still be included in the model. For example, although gender is less important than ethnicity from a substantive point of view, the exploratory analysis indicated that there might a random effect due to gender. In Model 3, gender is added to the model and leads to nearly identical results as Model 2 that does not include gender in the model. The likelihood ratio test for gender $\beta_{20} = 0$ is not significant ($LR = 1.13$, $df = 1$, $p = .29$). However, given the possibility of a random effect for gender (see Figure **??**), it will be kept in the model for now.

cja: The estimated standard error differ quite a bit here. Might be good for statistical inference chapter where robust/sandwich estimators will be introduced.

When developing the cluster-specific fixed effects model, interactions and transformations of predictors should be considered. predictors. For example, from the EDA of the classroom segregation data, a potential gender by ethnicity interaction was seen in Figure **??** and including the interaction in models fit to each clusters' data yields a noticeable increase in the R^2 for nearly all classroom and a sizable increase in meta-R^2. Model 4 in Table **??** includes the interaction and according the likelihood ratio test, the interaction is significant ($LR = 4.05$, $df = 1$, $p = 0.04$). This is a case where the likelihood ratio test and the t-test using the model based standard errors yield a different conclusion than the t-test using robust standard errors. The t test using empirical standard errors is not significant at the $\alpha = .05$ level (i.e., $t = 1.83$, $df = 271$, $p = 0.07$), but the t-test using model based standard errors is significant (i.e., $t = 2.03$, $df = 271$, $p = 0.04$). The interaction will be retained for now, but will be reassessed after developing a more complex model for the mean and the covariance structure of the data.

Table 7.2 Statistics for random intercept models fit to classroom segregation data (Rodkin et al. 2007) by MLE and robust (empirical, sandwich) standard errors. If a level of a predictor is not given (e.g., black for ethnicity, girl for gender, or no majority for class type), then the corresponding parameter is 0.

		Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
Effect	Parameter	est.	(s.e.)	est.	(s.e.)	est.	(s.e.)	est.	(s.e.)	est.	(s.e.)	est.	(s.e.)
Fixed Effects													
Intercept	β_{00}	.2370	(.0589)	.0365	(.0756)	.0644	(.0735)	.1289	(.0730)	.1791	(.0922)	.0931	(.0952)
Ethnicity: White	β_{10}			.3683	(.0865)	.3713	(.0861)	.2552	(.0867)	.2604	(.0912)	.4552	(.1531)
Gender: Boy	β_{20}					−.0699	(.0935)	−.2103	(.1197)	.2121	(.1194)	−.1851	(.1127)
Ethnicity × Gender:													
White × Boy	β_{30}							.2656	(.1449)	.2661	(.1413)	.2389	(.1354)
Class Type:													
Black majority	β_{01}									−.1060	(.1319)	.1435	(.1506)
Multiclutural	β_{02}									−.0466	(.1642)	−.1498	(.2280)
Ethnicity × Class type:													
White × Black majority	β_{11}											−.4675	(.1749)
White × Multicultural	β_{12}											.0935	(.2230)
Covariance Parameters													
Intercept	ψ_{00}	.0633	(.0268)	.0650	(.0260)	.0648	(.0259)	.0602	(.0247)	.0575	(.0241)	.0592	(.0241)
Residual	σ^2	.3253	(.0279)	.2910	(.0249)	.2899	(.0248)	.2871	(.0246)	.2873	(.0246)	.2718	(.0233)
Fit Statistics													
−2ln(Like)		548.5106		517.4178		516.2854		512.2332		511.5749		496.3719	
AIC		554.5		525.4		526.3		524.2		527.6		516.4	
BIC		559.0		531.4		533.8		533.2		539.5		531.3	
(residual) ICC		.1629		.1826		.1827		.1733		.1950		.1789	
R_1^2		—		.08		.09		.11		.11		.15	
R_2^2		—		.02		.02		.07		.10		.10	

Table 7.3 Statistics for random intercept and slope effect models fit to classroom segregation data (Rodkin et al. 2007) by MLE and robust (empirical, sandwich) standard errors. If a level of a predictor is not given (e.g., black for ethnicity, girl for gender, or no majority for class type), then the corresponding parameter is 0. Note that the harmonic mean using to compute $R_2^2 = 8.90$.

		Model 7		Model 8		Model 9		Model 10		Model 11		Model 12	
Effect	Parameter	est.	(s.e.)	est.	(s.e.)	est.	(s.e.)	est.	(s.e.)	est.	(s.e.)	est.	(s.e.)
Fixed Effects													
Intercept	β_{00}	.0989	(.0962)	.1315	(.0854)	.1276	(.0932)	.1205	(.0885)	.0762	(.0873)	.0750	(.0870)
Ethnicity: White	β_{10}	.4602	(.1544)	.4187	(.1450)	.4279	(.1444)	.4093	(.1493)	.4405	(.1245)	.4374	(.1124)
Gender: Boy	β_{20}	−.1988	(.1115)	−.1885	(.1109)	−.2055	(.1061)	−.1814	(.1117)	−.1928	(.1175)	−.1909	(.1176)
Ethnicity × Gender:													
White × Boy	β_{30}	.2523	(.1295)	.2245	(.1451)	.2563	(.1369)	.2311	(.1450)	.2363	(.1491)	.2367	(.1250)
Class Type:													
Black majority	β_{01}	.1568	(.1614)	.1231	(.1388)	.1560	(.1575)	.1327	(.1408)	.1789	(.1442)	.1790	(.1436)
Multiculutural	β_{02}	−.1394	(.2399)	−.1942	(.2103)	−.1878	(.2317)	−.1714	(.2124)				
Ethnicity × Class type:													
White × Black majority	β_{11}	−.4922	(.1861)	−.4494	(.1706)	−.4784	(.1839)	−.4394	(.1748)	−.4718	(.1508)	−.4663	(.1499)
White × Multicultural	β_{12}	.0862	(.2404)	.1305	(.2215)	.1112	(.2387)	−.1424	(.2262)				
Random Effects													
VARIANCES													
Intercept, $\underline{\gamma}_{0j}$	ψ_{00}	.1080	(.0545)	.0216	(.0178)	.0902	(.0623)	.0295	(.0196)	.0325	(.0206)	.0313	(.0191)
White, $\underline{\gamma}_{1j}$	ψ_{11}	.0705	(.0584)			.0681	(.0588)	.0096	(.0258)	.0094	(.0260)		
Boy, $\underline{\gamma}_{2j}$	ψ_{22}			.0890	(.0522)	.0915	(.0553)	.1096	(.0534)	.1077	(.0528)	.1088	(.0527)
Residual, $\underline{\varepsilon}_{ij}$	σ^2	.2534	(.0239)	.2522	(.0225)	.2347	(.0234)	.2465	(.0235)	.2467	(.0235)	.2502	(.0220)
COVARIANCES													
Intercept × White	ψ_{10}	−.0545	(.0495)			−.0709	(.0564)						
Intercept × Boy	ψ_{20}			.0220	(.0220)	.0087	(.0389)						
Black × Boy	ψ_{21}					.0178	(.0384)						
Fit Statistics													
$-2\ln(\text{Like})$		493.6396		483.1040		479.5496		483.7761		484.8546		485.0076	
AIC		517.6		507.1		509.9		507.8		504.9		503.0	
BIC		535.6		525.1		532.0		525.7		519.8		516.5	
R_1^2		.12		.15		.12		.16		.15		.16	
R_2^2		−.02		.10		−.01		.13		.11		.14	

7.2.2 Systematic Between Cluster Variation

In this section, the focus is on modeling systematic variation due to differences between clusters. To account for differences between clusters, predictors that describe attributes or characteristic of the cluster can be added to the within cluster model. In some cases, these are aggregates of Level 1 predictors (e.g., means, $\overline{x}_{\bullet j}$), but in all cases they describe some aspect of the cluster. These predictors provide information about the about context within which Level 1 units are observed. A common starting point is to add fixed effects to the regression model for the intercept and then consider fixed effects for the other Level 1 regression coefficients.

7.2.2.1 Predicting the Intercept Variance

To incorporate cluster-level predictors into the model as predictors of intercept variance, consider the following Level 1 model of a multilevel model:

$$\text{Level 1:} \qquad \underline{y}_{ij} = \underline{\omega}_{0j} + \sum_{p=1}^{P} \omega_{pj} x_{pij} + \underline{\varepsilon}_{ij}. \tag{7.4}$$

The heterogeneity between clusters in terms of their intercepts are modeled as a function of cluster level explanatory variables, but for now, the effects of other variables in the Level 1 models are assumed to be the same over clusters; namely,

$$\begin{aligned} \text{Level 2:} \qquad \underline{\omega}_{0j} &= \beta_{00} + \sum_{q=1}^{Q} \beta_{0q} z_{qj} + \underline{\gamma}_{0j} \\ \omega_{1j} &= \beta_{10} \\ &\vdots \\ \omega_{pj} &= \beta_{p0}. \end{aligned}$$

Using the Level 2 models in the Level 1 model yield the following LMM or combined model,

$$\underline{y}_{ij} = \underbrace{\beta_{00} + \sum_{q=1}^{Q} \beta_{q0} z_{qj} + \underline{\gamma}_{00}}_{\text{Intercept}} + \sum_{p=1}^{P} \beta_{p0} x_{pij} + \underline{\varepsilon}_{ij} \tag{7.5}$$

The degree to which the z_{qj}'s predict systematic variation in the intercepts leads to decreases in the variance of γ_{00}, and can alter the value σ^2 (—Snijder reference?—).

For example, in the study on classroom segregation, studying the effect that the racial composition of the classroom has on segregation is one of the main goals of the study. Since The most complex Level 1 model from the previous section (i.e., Model 4) that was also the best preliminary models found in the EDA is used here; that is,

$$\begin{aligned}\underline{\texttt{segregation}}_{ij} = \underline{\omega}_{0j} + \omega_{1j}\texttt{Ethnicity}_{ij} + \omega_{2j}\texttt{Gender}_{ij} \\ +\omega_{3j}(\texttt{Ethnicity}_{ij})(\texttt{Gender}_{ij}) + \underline{\varepsilon}_{ij}.\end{aligned}$$

The Level 2 model for the random intercept includes predictors describing the classroom type. The original coding of classroom type was designed to test substantive hypotheses regarding differences between classrooms that had either an African American or a European American majority, and between multi-cultural classrooms and those with a majority from one ethnicity[4]. However, in this chapter, a different coding is used make it easier to examine all pairs of possible configurations of classroom types. The change in coding was motivated, in part, by Figure 7.1.2.2 that shows an interaction between the three types of classrooms and a student's ethnicity where the effect of a multi-cultural classroom or a European American majority classroom appear the same but differ from classrooms with an African American majority. In this analysis, two dummy variables for classroom type are defined as

$$\texttt{Bmaj} = \begin{cases} 1 \text{ if majority is Black} \\ 0 \text{ otherwise} \end{cases} \qquad \texttt{MultiC} = \begin{cases} 1 \text{ if classroom is multiculutural} \\ 0 \text{ otherwise} \end{cases}$$

When the classroom majority is White, $\texttt{Bmaj} = \texttt{MultiC} = 0$. When classroom type is included in the model, the original coding and the dummy coding defined above lead to equivalent models (e.g., same model goodness-of-fit to data, predictions, etc.). What differs is the values of the βs for classroom type and the value β_{00} (the intercept).

The Level 2 model in this example is

$$\begin{aligned}\underline{\omega}_{0j} &= \beta_{00} + \beta_{01}\texttt{Bmaj}_j + \beta_{02}\texttt{MultiC}_j + \underline{\gamma}_{0j} \\ \omega_{pj} &= \beta_{p0} \qquad \text{for } p = 1,2,3.\end{aligned}$$

Combining the Level 1 and 2 models, yields the following LMM:

[4] This was achieved by coding majority= -1 for African American, 1 for European American, and 0 for multi-cultural. The other code was multi-cultural= 1 for a multi-cultural classroom and 0 for either African or European American majority classroom.

$$\begin{aligned}\underline{\texttt{segregation}}_{ij} = {} & \beta_{00} + \beta_{01}\texttt{Bmaj}_j + \beta_{02}\texttt{MultiC}_j + \beta_{10}\texttt{Ethnicity}_{ij} \\ & + \beta_{20}\texttt{Gender}_{ij} + \beta_{30}(\texttt{Ethnicity}_{ij})(\texttt{Gender}_{ij}) \\ & + \underline{\gamma}_{0j} + \underline{\varepsilon}_{ij},\end{aligned}$$

where the distributional assumption for $\underline{\varepsilon}_{ij}$ and $\underline{\psi}_{00}$ are given in (7.2). The results for this model, Model 5, are reported in Table **??**. The addition of classroom type into the model has negligible effect on estimated parameters, standard errors and fit statistics. Classroom type does not appear to be a significant predictor of the intercept ($LR = 0.66$, $df = 2$, $p = .72$); however, if there is an interaction between classroom type and ethnicity (as suggested by the EDA) and is of substantive interest, then it may be best to retain classroom type in the model. The predictors of the intercept at Level 2 become main effects in the LMM. Not including main effects comprising interactions can cause interpretational difficulties.

Models for other Level 1 regression coefficients can also include predictors even though they are not random. Such predictors may account for systematic variation and lead to *cross-level interactions* in the LMM (i.e., interactions between Level 1 predictors and Level 2 predictors). For example, classroom type is added to the Level 2 model for coefficient for ethnicity,

$$\begin{aligned}\underline{\omega}_{0j} &= \beta_{00} + \beta_{01}\texttt{Bmaj}_j + \beta_{02}\texttt{MultiC}_j\underline{\gamma}_{0j} \\ \omega_{1j} &= \beta_{10} + \beta_{11}\texttt{Bmaj}_j + \beta_{12}\texttt{MultiC}_j \\ \omega_{2j} &= \beta_{20} \\ \omega_{3j} &= \beta_{30}\end{aligned}$$

Since there is not a random component in the models for ω_{1j}, the coefficient for ethnicity, classroom type explains all the differences between the classrooms in terms of the effect of ethnicity on segregation. The LMM is

$$\begin{aligned}\underline{\texttt{segregation}}_{ij} = {} & \beta_{00} + \beta_{01}\texttt{BMaj}_j + \beta_{02}\texttt{MultiC}_j \\ & + \beta_{10}\texttt{Ethnicity}_{ij} + \beta_{20}\texttt{Gender}_{ij} + \beta_{30}(\texttt{Ethnicity}_{ij})(\texttt{Gender}_{ij}) \\ & + \beta_{11}(\texttt{BMaj}_j)(\texttt{Ethnicity}_{ij}) + \beta_{12}(\texttt{MultiC}_j)(\texttt{Ethnicity}_{ij}) \\ & + \underline{\gamma}_{0j} + \underline{\varepsilon}_{ij}.\end{aligned}$$

In this model, $\texttt{BMaj}_j$ and $\texttt{MultiC}_j$ interact with $\texttt{Ethnicity}_{ij}$; that is, classroom type mediates the effect of ethnicity on segregation. The statistics for this model, Model 6, are reported in Table **??**.

If predictors z_{qj} are in the models for regression coefficients ω_{pj} ($p > 0$), then it is generally best to also include them in the model for the intercept regardless of whether the intercept is random or not. Consider what happens when the predictors are not in the model for the intercept but they are in the model for other

effects, then the β_{pq}'s may be difficult to interpret. For example, suppose the following simple HLM is holds in the population:

$$\begin{aligned} \text{Level 1:} \quad \underline{y}_{ij} &= \underline{\omega}_{0j} + \omega_{1j} x_{ij} + \underline{\varepsilon}_{ij} \\ \text{Level 2:} \quad \underline{\omega}_{0j} &= \beta_{00} + \beta_{01} z_{ij} + \underline{\gamma}_{0j} \\ \omega_{1j} &= \beta_{10} + \beta_{11} z_{ij}, \end{aligned}$$

leads to the LMM

$$\underline{y}_{ij} = \beta_{00} + \beta_{01} z_{ij} + \beta_{10} x_{ij} + \beta_{11} z_{ij} x_{ij} + \underline{\gamma}_{0j} + \underline{\varepsilon}_{ij}. \tag{7.6}$$

This LMM has main effects for both z_{ij} and x_{ij} and β_{11} represents the interaction effect holding the main effects constant. Now suppose that z_{ij} is dropped from model for the intercept, then the LMM would be

$$\underline{y}_{ij} = \beta^*_{00} + \beta^*_{10} x_{ij} + \beta^*_{11} z_j x_{ij} + \underline{\gamma}^*_{0j} + \underline{\varepsilon}_{ij}. \tag{7.7}$$

In the second LMM, the main effect become part of error, in this case $\underline{\gamma}^*_{0j} = \underline{\gamma}_{0j} + \beta_{01} z_j$, and the variance $\underline{\gamma}^*_{0j}$ will be larger than that for $\underline{\gamma}_{0j}$. Models (7.6) and (7.7) may be empirically different (i.e., lead to different fitted values and fit statistics). Furthermore, since $x_{ij} z_j$ is generally correlated with z_j, the fixed part of the model and the random part are correlated. This violates an assumption of the model that the predictor (fixed part of the model) are uncorrelated with the (random) errors.

7.2.3 Random Effects

The models considered so far have only allowed for unexplained variance in the intercepts; however, clusters may differ in terms of the effects of the Level 1 predictors. This is dealt with by adding random terms to the Level 2 models for these other effects. Essentially, the variance of the responses due to unsystematic heterogeneity between clusters is accounted from by proposing unobserved random variables. The distribution of these unobserved variables is assumed to be normal with mean arbitrarily set to zero and their variances and covariances are estimated.

7.2.3.1 Additional Random Effects

When adding a random effect to the model, the variance of this random effect is estimated and if desired the the covariance with the intercept are estimated. Suppose that the coefficient of x_{ij} is random as well as the intercept. The random part of the LLM would become $\underline{\gamma}_{0j} + \underline{\gamma}_{1j} x_{ij} + \underline{\varepsilon}_{ij}$. Since it's the stochastic part of the LMM that determines the variance of the response variable $\underline{y}_{ij}$, the variance now depends on x_{ij}. The variance in this simple example is

$$\begin{aligned}
\text{var}(\underline{y}_{ij}) &= \boldsymbol{z}_{ij}' \boldsymbol{\Psi} \boldsymbol{z}_{ij} + \sigma^2 \\
&= (1, x_{ij}) \begin{pmatrix} \psi_{00} & \psi_{10} \\ \psi_{10} & \psi_{11} \end{pmatrix} \begin{pmatrix} 1 \\ x_{ij} \end{pmatrix} + \sigma^2 \\
&= \psi_{00} + 2\psi_{10} + \psi_{11} x_{ij}^2 + \sigma^2.
\end{aligned}$$

The variance is a function of x_{ij}; that is, the data illustrate heteroscedasticity. If the clustering in the data is ignored, the constant variance assumption of normal linear regression would be violated.

Due to the heteroscedasticity, the correlation between two randomly sampled units within the same cluster is not computed for models with random effects[5], because the variance and the covariance between units depend on the value of z_{ij}. For example, consider two observations i and $i*$ in cluster j,

$$\text{cor}(\underline{y}_{ij}, \underline{y}_{i*j}) = \frac{\boldsymbol{z}_{ij} \boldsymbol{\Psi} \boldsymbol{z}_{i*j}}{\sqrt{\boldsymbol{z}_{ij}' \boldsymbol{\Psi} \boldsymbol{z}_{ij} + \sigma^2} \sqrt{\boldsymbol{z}_{i*j}' \boldsymbol{\Psi} \boldsymbol{z}_{i*j} + \sigma^2}}.$$

For the simple example of one random effect x_{ij}, this correlation equals

$$\text{cor}(\underline{y}_{ij}, \underline{y}_{i*j}) = \frac{\psi_{00} + \psi_{10}(x_{ij} + xi * j) + \psi_{11} x_{ij} x_{i*j}}{\sqrt{\psi_{00} + 2\psi_{10} + \psi_{11} x_{ij}^2 + \sigma^2} \sqrt{\psi_{00} + 2\psi_{10} + \psi_{11} x_{i*j}^2 + \sigma^2}}.$$

There are as many correlations as there are unique pairs of units within a cluster.

The covariance between the random intercept and effect, say x_{ij}, indicates whether and how the random effects $\underline{\gamma}_{0j}$ and $\underline{\gamma}_{1j}$ co-vary. Suppose that students are nested within classes and the response variable is a test score that has a maximum value and x_{ij} is aptitude. A negative ψ_{10} might be expected because those classes with large values of γ_{0j}) have high test scores that could be near the maximum and thus would tend to have smaller slopes. Conversely a positive ψ_{10} might

[5] An exception is for longitudinal data where correlations between time points are meaningful. As the time between observations increases, such correlations often decrease and modeling this change is important.

be expected whenever large intercepts tend to occur with large slopes as might be expected in a study on the effectiveness of a drug to treat depression. In other cases, it may be expected that $\psi_{10} = 0$. One example of this is the segregation study that is further analyzed in the following sections. Other examples can be found in Chapters 11 and Chapter **??**.

7.2.3.2 Example: Segregation Data

To illustrate the addition of a random effect, a random term for a student's ethnicity is added to the Level 2 regression of Model 6. Recall that the variable $\texttt{Ethnicity} = 1$ if a student is African American, and 0 if European American. Treating this as a numerical variable,

$$\begin{aligned}
\underline{\omega}_{0j} &= \beta_{00} + \underline{\gamma}_{0j} \\
\underline{\omega}_{1j} &= \beta_{10} + \beta_{11}\texttt{Bmaj}_j + \beta_{12}\texttt{MultiC}_j + \underline{\gamma}_{1j} \\
\omega_{2j} &= \beta_{20} \\
\omega_{3j} &= \beta_{30}.
\end{aligned}$$

The resulting LMM is

$$\begin{aligned}
\underline{\texttt{segregation}}_{ij} = {} & \beta_{00} + \beta_{01}\texttt{BMaj}_j + \beta_{02}\texttt{MultiC}_j + \beta_{10}\texttt{Ethnicity}_{ij} \\
& + \beta_{20}\texttt{Gender}_{ij} + \beta_{30}(\texttt{Ethnicity}_{ij})(\texttt{Gender}_{ij}) \\
& + \underline{\gamma}_{0j} + \underline{\gamma}_{1j}\texttt{Ethnicity} + \underline{\varepsilon}_{ij},
\end{aligned}$$

where

$$\begin{pmatrix} \varepsilon_{ij} \\ \gamma_{0j} \\ \gamma_{1j} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \psi_{00} & \psi_{10} \\ 0 & \psi_{10} & \psi_{11} \end{pmatrix}\right) \; i.i.d. \tag{7.8}$$

Whereas the random terms in the Level 2 model are often thought of as residual terms in the regressions that account for unobserved predictors, the random terms in the LMM can also be thought of as random coefficients for 1 (the intercept) and $\texttt{Ethnicity}_j$. The statistics for the model given above, Model 7, are reported in Table **??**. The mixture chi-square test for $H_o : \psi_{11} = \psi_{01} = 0$ suggests that a random effect for ethnicity is not needed (i.e., $LR = 2.73$, $df = 1$ & 2, $p = .18$). This is not too surprising given Figure 7.1.1.2 where most classrooms lines are flat and roughly parallel as is consistent with a non-random effect for ethnicity. As further evidence of including a random effect for ethnicity, note that the measure $R_2^2 = -.02$. As discussed later in this chapter, a negative value indicates a possible miss-specification of the model.

Rather than a random effect for ethnicity, in Model 8, a random effect for gender is included in the Level 2 model for ω_{2j} such that the new Level 2 model is

$$\begin{aligned}
\underline{\omega_{0j}} &= \beta_{00} + \underline{\gamma_{0j}} \\
\omega_{1j} &= \beta_{10} + \beta_{11}\texttt{Bmaj}_j + \beta_{12}\texttt{MultiC}_j \\
\omega_{2j} &= \beta_{20} + \underline{\gamma_{2j}} \\
\omega_{3j} &= \beta_{30},
\end{aligned}$$

and the corresponding LMM is

$$\begin{aligned}
\underline{\texttt{segregation}}_{ij} = \beta_{00} &+ \beta_{01}\texttt{BMaj}_j + \beta_{02}\texttt{MultiC}_j + \beta_{10}\texttt{Ethnicity}_{ij} \\
&+ \beta_{20}\texttt{Gender}_{ij} + \beta_{30}(\texttt{Ethnicity}_{ij})(\texttt{Gender}_{ij}) \\
&+ \underline{\gamma_{0j}} + \underline{\gamma_{1j}}\texttt{Boy}_{ij} + \underline{\varepsilon_{ij}}.
\end{aligned}$$

This model, Model 8, has the smallest AIC and BIC among those fit to the data so far. The test for a gender random effect, $H_0 : \psi_{22} = \psi_{02} = 0$, supports the conclusion that a random effect for gender is needed (i.e., $LR = 13.27$, $df = 1$ & 2, $p < .01$). Note that the effect for gender in the random part of the model was switched to $\texttt{Boy}_{ij} = 1$ for boys and $= 0$ for girl. The reason for this is explained in the following section.

7.2.3.3 Discrete Level 2 Predictors

Add references to bib:

Steven W. Raudenbush and R.T Brennan and R. Barnet (1995) A multilevel hierarchial model for studying change within married couples. Journal of Family Psychology, 9, 161-174.

David A. Kenny and Deborah A. Kashy (2011). Dyadic data analysis using multilevel models. In Joop J. Hox and J. Kyle Roberts (Eds) Handbook of Advanced Multilevel Analysis, pp 335–370. ISBN: 978-1-84169-722-2. Routledge, NY, NY

Care must be taken with discrete Level 2 predictors. Suppose that in addition to a random intercept, a dichotomous predictor x_{ij} is specified as being a random effect. How such predictors are dealt and interpreted is a parametrization (coding) issue. A discrete predictor basically shifts the intercept and a random discrete predictor permits the variance in intercepts to differ over levels of the discrete variable. Problems are encounter when the intercept and a discrete Level 1 predictor are both specified as random effects. **?** illustrate a two random intercept model for

the case of a dichotomous variable in the context of distinguishable dyadic data (i.e., husband and wife) (see also (**?**)). The underlying issue pertains to the general case of random discrete Level 1 predictors. Suppose that a dichotomous variable x_{ij} is dummy or effect coded and is entered x_{ij} in a model is just like it would be if x_{ij} were any other numerical or continuous predictor (i.e. specify a random intercept and a random x_{ij}). Assuming that the data support a random x_{ij} (i.e., yields a proper solution), then there are no problems with the model converging or yielding an improper solution. An equivalent model with a different parametrization is also possible where the number of random effects entered for a discrete variable is the same as the number of categories of the predictor. In the latter case, the intercept must be specified as fixed otherwise the model will fail to converge or yield an improper solution.

To illustrate consider the following simple model that has a random intercept and x_{1ij} is either dummy or effect coded:

$$\begin{aligned} \text{Level 1:} \quad & \underline{y}_{ij} = \underline{\omega}_{0j} + \underline{\omega}_{1j} x_{ij} + \underline{\varepsilon}_{ij} \\ \text{Level 2:} \quad & \underline{\omega}_{0j} = \beta_{00} + \underline{\gamma}_{0j} \\ & \underline{\omega}_{1j} = \beta_{10} + \underline{\gamma}_{1j}, \end{aligned}$$

and the distribution of the random effects is

$$\begin{pmatrix} \varepsilon_{ij} \\ \underline{\gamma}_{0j} \\ \underline{\gamma}_{1j} \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \psi_{00} & \psi_{10} \\ 0 & \psi_{10} & \psi_{11} \end{pmatrix} \right).$$

An equivalent model with the same Level 1 model but different parametrization of the Level 2 model is

$$\begin{aligned} \omega_{0j} &= \beta_{00} \\ \underline{\omega}_{1j} &= \begin{cases} \beta_{10} + \underline{\gamma}_{1j} & \text{if } x_{ij} = 0 \\ \beta_{10} + \underline{\gamma}_{2j} & \text{if } x_{ij} = 1, \end{cases} \end{aligned}$$

where the distribution of the random effects is basically as above, except the labeling of the elements of $\boldsymbol{\Psi}$. A model that would yield an improper solution has the same Level 1 model above but the Level 2 model is

$$\begin{aligned} \underline{\omega}_{0j} &= \beta_{00} + \underline{\gamma}_{0j} \\ \underline{\omega}_{1j} &= \begin{cases} \beta_{10} + \underline{\gamma}_{1j} \text{ when } x_{ij} = 0 \\ \beta_{10} + \underline{\gamma}_{2j} \text{ when } x_{ij} = 1 \end{cases} \end{aligned}$$

The problems with this model is that the random part of the cluster-specific effect when $x_{ij} = 0$ equals $\underline{\gamma}_{0j} + \underline{\gamma}_{1j}$ and when $x_{ij} = 1$, the random part equals $\underline{\gamma}_{0j} + \underline{\gamma}_{1j}$. There not enough data to estimate three variances ψ_{kk} and three $\psi_{k\ell}$s. The matrix $\boldsymbol{\Psi}$ will not be positive definite (i.e., a proper covariance matrix).

Different parameterizations yield different values for the estimated ψs, but they have the same $\hat{\sigma}^2$, $\widehat{\text{var}}(\underline{y}_{ij})$, the same estimated fixed effects, and the same goodness-of-fit statistics. To illustrate this Model 8 was fit to the segregation data using a dummy code for gender, an effect code for gender, and a fixed intercept with two values of the random effect for $\underline{\omega}_{2j}$. Since $\text{var}(\underline{y}_{ij}) = \mathbf{z}'_{ij}\boldsymbol{\Psi}\mathbf{z}_{ij}$ and $\texttt{Boy}_{ij}$ dummy coding, $\boldsymbol{z}' = (1, \texttt{Boy}_{ij})$ and

$$\text{var}(\underline{y}_{ij}) = \begin{cases} \psi_{00} + 2\psi_{10} + \psi_{11} + \sigma^2 & \text{for boys} \\ \psi_{00} + \sigma^2 & \text{for girls} \end{cases}$$

The variance of $\underline{y}_{ij}$ for boys is greater than that for girls.

Table 7.4 Three difference estimated covariance matrices of equivalent models from different ways coding gender.

	Random Intercept				No Random Intercept	
	Effect code		One Dummy Code		Two Dummy Codes	
Parameter	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)
ψ_{00}	0.0659	(0.0264)	0.0216	(0.0178)		
ψ_{01}	0.0333	(0.0152)	0.0220	(0.0220)		
ψ_{11}	0.0223	(0.0130)	0.0890	(0.0522)	0.1546	(0.0607)
ψ_{12}					0.0436	(0.0256)
ψ_{22}					0.0216	(0.0178)
σ^2	0.2522	(0.0225)	0.2522	(0.0225)	0.2522	(0.0225)

To illustrate that the effect of different parameterizations, the estimated variances and covariances using three different codings of gender are given in Table 7.4. Note that $\hat{\sigma}^2 = 0.2552$ (0.0225) for all the models. The estimated variances for boys and girls are computed below using each of the different parameterizations. The first line for each of the parameterizations is the estimated variance for boys (i.e., 0.41) and the second is the variance for girls (i.e., 0.28):

$$\text{Effect code: } \widehat{\text{var}}(\underset{\sim}{y}_{ij}) = \begin{cases} 0.0659 + 2(0.0333) + 0.0223 + 0.2552 = 0.41 \\ 0.0659 - 2(0.0333) + 0.0223 + 0.2552 = 0.28 \end{cases}$$

$$\text{Dummy code: } \widehat{\text{var}}(\underset{\sim}{y}_{ij}) = \begin{cases} 0.0216 + 2(0.0220) + 0.0890) + 0.2552 = 0.41 \\ 0.0216 + 0.2552 = 0.28 \end{cases}$$

$$\text{Two Dummy codes: } \widehat{\text{var}}(\underset{\sim}{y}_{ij}) = \begin{cases} 0.1546 + 0.2552 = 0.41 \\ 0.0216 + 0.2552 = 0.28 \end{cases}.$$

With two dummy codes, the covariance ψ_{12} "drops" out of the equation for $\widehat{\text{var}}(\underset{\sim}{y}_{ij})$; however, the covariance can be and was estimated. The estimated covariance with dummy coding equals $\hat{\psi}_{12} = .0416$ (0.0256) and indicates that classrooms with larger effects for girls (i.e., γ_{2j}) tend to occur with larger ones for boys (i.e., γ_{1j}).

If a discrete predictor has three or more categories, then there would be as many variances of the intercept as there are levels of the discrete predictor, one for each category.

7.2.3.4 Two or More Random Effects

As the number of random effects increases, the risk of having a model yield an improper solution or fail to converge increases. If a computer takes an unusually long time to reach convergence, this may be a sign that the model is too complex for the data[6]. For example, for the segregation data, making all the Level 1 variables (i.e., $\texttt{Ethnicity}_{ij}$, $\texttt{Gender}_{ij}$, and $(\texttt{Ethnicity}_{ij})(\texttt{Gender}_{ij})$) yields a non-positive definite $\hat{\boldsymbol{\Psi}}$. Simplifying the model by setting all covariances equal to 0 converges and yields an acceptable solution, as well as dropping the random interaction. The latter case is reported in Table 7.2.1.2 as Model 9.

If covariances are not expected between random effects and are not significant as indicated by z-tests for each (or a sub-set) of the ψ_{kl}s, these covariances can be set equal to zero. For example, the individual tests for the ψ_{kl}, $k \neq l$, are all not significant in Model 9, and setting them all equal to zero, Model 10 in Table 7.2.1.2, leads to a non-significant change in the fit of the model to the data (i.e., $LR = 4.23$, $df = 3$, $p = .24$).

[6] A computer program may also take a long time to convergence when models are fit to very large data sets.

7.2.4 Reassessing Fixed and Random Effects

Since the random part of the model and fixed part of a LMM are interdependent, after initial decisions regarding fixed effects and subsequently random ones, it is worthwhile to re-assess and perhaps refine the fixed effects. After re-examining the fixed effects, the random part of the models can also be re-examined. This process may require multiple iterations until a final model is selected. Given such a final model or a sub-set of reasonable ones, further detailed assessment of the mode's adequacy and validity of its assumptions should be conducted. This latter assessment is the subject of the following chapter because it pertains to all GLMMs and not just LMMs.

Continuing the classroom segregation example using Model 10, we start by examining whether the interactions are needed. Recall that the EDA suggested an interaction between classroom type and ethnicity where the effect for multicultural and White majority classrooms appeared have the same (positive) slopes when plotted against student ethnicity. Furthermore, the effect for Black majority classrooms appeared to be 0 (i.e., flat). In Model 10, the cross-level interaction $\texttt{Ethnicity}_{ij})(\texttt{Classtyp}_j)$ is significant (i.e., $F_{2,269} = 5.02$, $p = .01$); however, of the two parameters for this effect, the β for $\texttt{White}_{ij})\texttt{Multic}_j$ is not significant ($t = 0.63$, $df = 269$, $p = .53$). In other words, the effects for multi-cultural and white majority classes are not statistically different from each other. Furthermore, since the β for the main effect for $\texttt{Multic}_j$ is not significant ($t = -0.81$, $df = 269$, $p = .43$), the β's' for multi=cultural classrooms can be set equal to 0 (i.e., the value for white majority classrooms) by only including $\texttt{Bmaj}_j$ in the model. The resulting model, Model 11 in Table 7.2.1.2, has the smallest value for AIC and BIC, and the restriction on the βs for multi-cultural classroom did not lead to a significant decrease in fit (i.e., $LR = 0.08$, $df = 2$, $p = .58$).

Given the $\hat{\psi}_{11}$ for ethnicity is smaller than its standard error in Model 11, in the next model fit to data, Model 12, the random effect for ethnicity is dropped. This leads to smaller values for both *AIC* and *BIC*, a negligible change in the log of the likelihood, and nearly the same estimates for all other parameters in the model.

Whether to include the the interaction between ethnicity and gender is a judgment call on the part of the researcher. If it is removed from Model 11, $LR = 1.079$, $df = 1$ and $p = .30$); however, if it is removed from Model 12, $LR = 3.54$, $df = 1$, $p = .06$. The information criteria for the model without the interaction are $AIC = 504.5$ and $BIC = 516.5$ and are nearly the same as those for Model 12 (i.e., $AIC = 503.0$ and $BIC = 516.5$). As would be expected if the model is correctly specified, the model based standard error estimates for the fixed effects are very similar to the robust (sandwich) estimates, and the tests for fixed effects yield the same conclusions regarding the significance of the fixed effects. In all models the ethnicity by gender interaction is never clearly significant; therefore,

as a final model we drop the interaction. The parameters estimates and various goodness-of-fit statistics for this model at in Table 7.5.

Table 7.5 Parameter estimates and fit statistics for final model fit to the classroom segregation data.

			Standard Error				
Effect	Parameter	Estimate	Model	Robust	df	t	p
Intercept	β_{00}	0.0187	0.0779	0.0840	26	0.22	0.83
Ethnicity: White	β_{10}	0.5420	0.0805	0.0996	271	5.44	$<.01$
Gender: Boy	β_{20}	-0.0671	0.0916	0.0872	271	-0.77	0.44
Class Type:							
Black majority	β_{01}	0.1843	0.1244	0.1422	26	1.30	0.21
Ethnicity $\times$ Class type:							
White $\times$ Black majority	β_{11}	-0.4803	0.1246	0.1444	271	-3.33	$<.01$
Random Effects (Variances)							
Intercept, $\underline{\gamma}_{0j}$	ψ_{00}	0.0346	0.0199				
Boy, $\underline{\gamma}_{1j}$	ψ_{11}	0.1123	0.0545				
Residual, $\underline{\varepsilon}_{ij}$	σ^2	0.2518	0.0222				
Fit Statistics							
$-2\ln(\text{Like})$		488.5435					
AIC		504.5					
BIC		516.5					
R_1^2		.14					
R_2^2		.10					

Given some interpretation for the final model: fixed effects, random

7.3 R^2 Type Measures

A standard way to assess goodness of model fit to data in normal OLS regression is to use multiple R^2. With HLMs there are complications and problems for the use of R^2 to assess how well a model represents the data. One complication is that there are models at multiple levels that represent different aspects of the data, cluster-specific model and between-cluster models; therefore, R^2 type measures are needed for each level, R_1^2 for Level 1 model and R_2^2 for the Level 2 models.

In normal linear regression, R^2 has a number of different interpretations, including the squared correlation between observed and predicted values of the response, the proportion decrease in modeled variance of the response given the predictor variables, and the proportional reduction of prediction error variance. The proportional reduction in explained variance at different levels is used by **?**

and **?**. This definition is problematic because a null or empty HLM (no predictor variables) and one with predictors variables can theoretically and empirically lead to negative values. The problem with this definition stems from the fact that estimates of Level 2 variances can increase or decrease as predictors are added to a model. For example, using this as a definition in the classroom segregation example, the proportional amount of variance between groups that is explain by Model 2 is

$$R_2^2 = 1 - \frac{.0650}{.0663} = -.03,$$

where $\hat{\psi}_{00} = .0650$ is from Model 2 and $\hat{\psi}_{00} = .0663$ is from Model 1, the empty HLM with no predictor variables.

Following **?**, the extensions presented here use the *proportional decrease in prediction error variance* (i.e., the mean squared error of prediction) definition to develop measures for within and between cluster models. These R^2 measures presented are only for two-level models and are appropriate when data come from an observational study. Observational data is required because they depend on the predictor or explanatory variables being random as well as the response variable. Although these extensions of R^2 serve not only as measures of fit, they also have diagnostic uses.

7.3.1 Level 1: R_1^2

Consider the following general LMM where the predictor variables are random,

$$\underline{y}_{ij} = \underline{\boldsymbol{x}}'_{ij}\boldsymbol{\beta} + \underline{\boldsymbol{z}}'_{ij}\underline{\boldsymbol{\gamma}}_j + \underline{\varepsilon}_{ij}, \tag{7.9}$$

and the usual assumption for $\underline{\boldsymbol{\gamma}}_j$ and $\underline{\varepsilon}_{ij}$, but now $\underline{\boldsymbol{x}}_{ij}$ and $\underline{\boldsymbol{z}}_{ij}$ are assumed to be random. Independence is assumed between $\underline{\boldsymbol{x}}_{ij}$ and $\underline{\boldsymbol{\gamma}}_j$ and $\underline{\varepsilon}_{ij}$[7]. When data are from an observational study, $\underline{\boldsymbol{x}}_{ij}$ are sampled from some population.

Without predictors, the best prediction of $\underline{y}_{ij}$ is $\bar{y}$, in which case the mean squared error of prediction equals the variance, $\text{var}(\underline{y}_{ij})$. Given predictor variables, the best linear predictor of $\underline{y}_{ij}$ is $\underline{\boldsymbol{x}}'_{ij}\boldsymbol{\beta}$, in which case the mean square prediction error equals $\text{var}((\underline{y}_{ij} - \underline{\boldsymbol{x}}'_{ij}\boldsymbol{\beta})|\boldsymbol{z}_{ij})$. Note the this variance is conditional on the Level 2 predictors. Defining R_1^2 as the proportional reduction in the mean squared predictor error gives leads to

[7] The explanatory variables $\underline{\boldsymbol{x}}_{ij}$ and $\underline{\boldsymbol{z}}_{ij}$ are necessarily dependent since $\underline{\boldsymbol{z}}_{ij}$ is a sub-set of $\underline{\boldsymbol{x}}_{ij}$.

$$R_1^2 = \frac{\text{var}(\underline{y}_{ij}) - \text{var}(\underline{y}_{ij} - \underline{\boldsymbol{x}}'_{ij}\boldsymbol{\beta})}{\text{var}(\underline{y}_{ij})} \tag{7.10}$$

$$= 1 - \frac{\text{var}(\underline{y}_{ij} - \underline{\boldsymbol{x}}'_{ij}\boldsymbol{\beta})}{\text{var}(\underline{y}_{ij})}, \tag{7.11}$$

where the numerator in (7.10) equals the reduction in square prediction error.

Computing R_1^2 as given in (7.11) requires variance estimates. Rather than the sample variance of $\underline{y}_{ij}$, the variance from the unconditional means model is used for $\text{var}(\underline{y}_{ij})$, because the empy/null HLM is used the baseline model; that is,

$$\widehat{\text{var}}(\underline{y}_{ij}) = \hat{\psi}^*_{00} + \hat{\sigma}^{2*}.$$

The estimator for the numerator in (7.11) equals

$$\widehat{\text{var}}(\underline{y}_{ij} - \underline{\boldsymbol{x}}'_{ij}\boldsymbol{\beta}) = \underline{\bar{\boldsymbol{z}}}'\hat{\boldsymbol{\Psi}}\underline{\bar{\boldsymbol{z}}} + \text{trace}(\boldsymbol{\Psi}(\boldsymbol{S}_z^B + \boldsymbol{S}_z^W)) + \hat{\sigma}^2, \tag{7.12}$$

where $\bar{\boldsymbol{z}}$ is the vector of means of the Level 1 variables that have random effects, $\hat{\boldsymbol{\Psi}}$ is the estimate covariance matrix of the γ_js, $\boldsymbol{S}_z^B$ is the estimated between cluster covariance matrix for the predictors $\boldsymbol{z}_{ij}$ that are random, and $\boldsymbol{S}_z^W$ equals within cluster or pooled covariance matrix of the z variables.

To illustrate the computation of R_1^2, consider a simple random intercept model where the vector $\boldsymbol{z}_{ij}$ is a scalar (i.e., $\boldsymbol{z}_{ij} = (1)$), $s_{00}^B = s_{00}^W = 0$, and (7.12) simplifies to

$$\begin{aligned}\widehat{\text{var}}(\underline{y}_{ij} - \boldsymbol{x}'_{ij}\boldsymbol{\beta}) &= 1(\hat{\psi}_{00})1 + \hat{\psi}_0(0+0) + \hat{\sigma}^2 \\ &= \hat{\psi}_{00} + \hat{\sigma}^2.\end{aligned}$$

For example, consider Model 6 in Table **??** where $\hat{\psi}_{00} = .0592$, $\hat{\sigma}^2 = .2718$, and for the empty model (Model 1) $\hat{\psi}^*_{00} = .0633$ and $\hat{\sigma}_{2*} = .3253$,

$$R_1^2 = 1 - \frac{.0592 + .2718}{.0633 + .3253} = .15.$$

For Model 6, there is a 15% reduction in the mean square prediction error within clusters.

For a slightly more complex model, with a random intercept and one other random effect x_{ij}, $\boldsymbol{z}'_{ij} = (1, x_{ij})$, the estimated variance equals

$$\begin{aligned}\widehat{\text{var}}(\underline{y}_{ij} - \boldsymbol{x}'_{ij}\boldsymbol{\beta}) &= (1, \bar{x}_1)\begin{pmatrix}\hat{\psi}_{00} & \hat{\psi}_{10}\\ \hat{\psi}_{10} & \hat{\psi}_{11}\end{pmatrix}\begin{pmatrix}1\\ \bar{x}\end{pmatrix}\\ &\quad + \text{trace}\left(\begin{pmatrix}\hat{\psi}_{00} & \hat{\psi}_{10}\\ \hat{\psi}_{10} & \hat{\psi}_{11}\end{pmatrix}\left(\begin{pmatrix}0 & 0\\ 0 & s^B_{11}\end{pmatrix} + \begin{pmatrix}0 & 0\\ 0 & s^W_{11}\end{pmatrix}\right)\right) + \hat{\sigma}^2\\ &= \hat{\psi}_{00} + 2\bar{x}_1\hat{\psi}_{10} + \hat{\psi}_{11}(\bar{x}_1^2 + s^B_{11} + s^W_{11}) + \hat{\sigma}^2\end{aligned}$$

For example, consider Model 8 in Table **??** where the intercept and $\texttt{boy}_{ij}$ both have random effects. Using standard formulas for s^B_{11} and s^W_{11}, i.e.,

$$s^B_{11} = \frac{\sum_j^J n_j(\bar{x}_j - \bar{x})^2}{(\sum_{j=1}^J n_j) - 1} \qquad \text{and} \qquad s^W_{11} = \frac{\sum_{i=1}^{nj}(n_j - 1)s_j^2}{(\sum_{j=1}^J n_j) - 1},$$

the values of statistics for $\texttt{boy}_{ij}$ that are needed to computer R^2_1 equal $\bar{x} = 0.4272$, $s^B_{11} = 0.0329$, $s^W_{11} = 0.2126$, and the $\hat{\psi}_{jk}$ and $\hat{\sigma}^2$ values from Table **??**,

$$\begin{aligned}R^2_1 &= 1 - \frac{.0216 + 2(.4272)(.0220) + .0890((.4272)^2 + .0329 + .2126) + .2522}{.06326 + .3253}\\ &= .15.\end{aligned}$$

7.3.2 Level 2: R^2_2

The development of an R^2 at Level 2 proceeds in much the same way; however, it is the cluster means for which a measure of the proportional reduction in mean squared prediction error is desired. When there are no predictors, the best predication for the cluster mean is the sample mean, $\bar{y}_j$. Given predictors variables, the best linear prediction is $\bar{\boldsymbol{x}}'_{ij}\boldsymbol{\beta}$. The reduction in prediction error equals $(\bar{y}_{ij} - \bar{\boldsymbol{x}}_{ij}\boldsymbol{\beta})$ and the proportional reduction in mean square prediction error is

$$R^2_2 = \frac{\text{var}(\underline{\bar{y}}_j) - \text{var}(\underline{\bar{y}}_j - \underline{\bar{\boldsymbol{x}}}'_{ij}\boldsymbol{\beta})}{\text{var}(\underline{\bar{y}}_j)} \tag{7.13}$$

$$= 1 - \frac{\underline{\bar{\boldsymbol{z}}}'\boldsymbol{\Psi}\underline{\bar{\boldsymbol{z}}} + \text{trace}(\boldsymbol{\Psi}(\boldsymbol{S}^B + \boldsymbol{S}^W/n) + \sigma^2/n)}{\psi^*_{00} + \sigma^{2*}/n}, \tag{7.14}$$

where n is a typical or representative cluster size. When cluster sizes differ, **?** suggest using the harmonic mean for n; that is, $n = N/(\sum_{j=1}^J 1/n_j)$.

In the special case of the random intercept model, the formula for R^2_2 simplifies to

$$R_2^2 = 1 - \frac{\hat{\psi}_{00} + \hat{\sigma}^2/n}{\hat{\psi}^*_{00} + \hat{\sigma}^{2*}/n}.$$

For example, in the classroom segregation data, the harmonic mean equals 8.90, and for Model 6 in Table **??**,

$$R_2^2 = 1 - \frac{.0592 + .2718/8.90}{.0633 + .3253/8.90} = .10$$

Another special case where there is only a random intercept and one random effect, the formula for R_2^2 simplifies in a similar manner as that for R_1^2, except for the division of $\boldsymbol{S}^W$ and σ^2 by n,

$$R_2^2 = 1 - \frac{\psi_{00} + 2\psi_{10}\bar{x} + \psi_{11}(\bar{x^2} + S^B + S^2/n) + \sigma^2/n}{\psi^*_{00} + \sigma^{2*}/n}.$$

For example, substituting the estimated values for Model 8 in Table **??** and sample statistics for $\texttt{boy}_{ij}$, $R_2^2 = .10$.

7.3.3 Properties and Uses R_1^2 and R_2^2

In the population, R_1^2 and R_2^2 do not decrease as predictor variables are added to the model and will be non-negative . However, when dealing with data sampled from the population, the R^2 may not behave ideally. Negative values or moderate decreases in R^2s are signs of model miss-specification (**?**). For example, consider Model 7 in Table **??** where $R_2^2 = -.01$. The fact that it is negative for Model 7 that include a random effect for ethnicity suggests that the data do not support ethnicity as a random effect. Also note that R_1^2 and R_2^2 are smaller for Model 9 than they are for Model 8 even though Model 9 have more parameters.

The values of R_1^2 and R_2^2 for random intercept and random intercept and slope models will tend to be comparable when they have the same fixed effects. Since computing R_1^2 and R_2^2 is easier for random intercept models than it is for more complex models, one can generally compute R_1^2 and R_2^2 for the random intercept models (**?**).

7.4 Centering Level 1 Predictors

Centering or setting the location of Level 1 predictor variables can have an effect on parameter stability, statistical equivalence of models, and interpretation of pa-

rameters of a model. This is only an issue only for Level 1 predictor variables. For example, consider Figure 7.2.1.1 where the *ICC* was equal to zero; however, the slopes were different over clusters. If the scale for x_{ij} was shifted, thus changing the 0 point, between cluster differences in both the intercept and slopes over clusters would be found.

The material presented in this chapter also pertains to other types of GLMMS. After discussing different type of centering, the effect of centering on stability, equivalence, interpretation and recommendations are discussed.

7.4.1 Types of Centering

Level 1 predictors can be centered around the grand mean of all units in a study or experiment. Let x_{ij} be the measured value or raw score, then grand mean centering corresponds to $x^*_{ij} = x_{ij} - \bar{x}$. When grand mean centering is used, the 0 point of x^*_{ij} corresponds to the overall mean of the data. This impacts the interpretation of the intercept. With grand mean centering all observations are shifted by the same amount.

A second type of centering is cluster or group mean centering where the Level 1 variable is centered around the cluster's mean; that is, $\tilde{x}_{ij} = x_{ij} - \bar{x}_j$. With group mean centering, data within clusters are shifted by the same amount, but by different amounts over clusters. The 0 point in the $\tilde{x}_{ij}$ scale corresponds to cluster means. The intercept is then interpreted as the value of the response for an average or typical member of a cluster.

Difference processes can be at work within clusters and between clusters. If a predictor variable carries information about units and processes within clusters but also information about differences between clusters, this can create difficulties for interpretation. This situation can be detected if in the addition of a within cluster predictor leads to a smaller value of both σ^2 and ψ_{00} in a random intercept model. Centering can around the cluster means can alleviate this problem.

Centering around the cluster mean and using the mean as a Level 2 predictor may be called based on a researcher's hypotheses. For example, in the study by **?** where the focus was on whether optimism is related to immunity. Immunity and optimism were measure multiple times during a year. An important issue was whether a student who is generally optimistic tends to have greater immunity than those who are less optimistic or whether the deviation from a person's general level of optimism more important. In other words, is optimism as a trait or the current level of optimism is more predictor of immunity? To test this hypothesis requires cluster (student) mean centering for optimism at as a Level 1 predictor and the cluster mean as a Level 2 predictor.

7.4.2 Statistical Equivalence

Models are statistically equivalent if they have the same fitted values and same value for the log-likelihood. The type of clustering can lead to statistically equivalent models but have different parameters estimates. If models are statistically equivalent, a transformation of the parameters (and standard errors) of one model can be found that equal the parameters of the statistically equivalent one (and vice versa).

Whether models are equivalent depends on the complexity of the model. For random intercept models the following models are statistically equivalent but may have differences in terms of parameter estimates:

- Raw scores and grand mean centering.
- Raw scores and cluster mean centering if the cluster mean is also a predictor of the intercept.

For random intercept and slope models, the only ones that will yield statistically equivalent models are the raw score and grand mean centering.

To illustrate this, five random intercept and five random intercept and slope models are fit to the immunity of optimism data and their log-likelihood valued are compared.

Table 7.6 Put on here

		Random Intercept	and Slope
Type of Centering	Variables in model	$-2lnlike$	$-2lnlike$
None (raw score)	x_{ij}		
Grand mean centering	$x^* = x_{ij} - \bar{x}$		
Raw score & cluster mean			
Cluster mean centering			
& cluster mean			

7.4.3 Parameter Stability

7.4.4 Interpretational Considerations

7.4.5 Recommendations

Best centering for comparing group differences (see course slides when I talk about centering).

Basis for test of dependency between $\boldsymbol{x}$ and ε.

- Types of centering and effects on random intercept & slope models.
- Putting cluster mean back in at Level 2 when cluster center.
- When do need to or should center.

7.5 Three-level model

So far we have restricted attention to two level models. In many cases, there may be more than two levels of nesting. For example, students could be nested within schools and schools within districts, or observations at different time points nested within individuals nested within treatments. The two level models are extend to three levels; however, following the same process discussed here, they can be extended to four and high level models.

Let i index Level 1 units, j index Level 2 units and k index Level 3 units. The first two levels of a three Level HLM are

$$\begin{aligned}
\text{Level 1 } \underline{y}_{ijk} &= \underline{\omega}_{0jk} + \underline{\omega}_{1jk}x_{1ijk} + \ldots + \underline{\omega}_{pjk}x_{pijk} + \underline{\varepsilon}_{ijk} \\
\text{Level 2: } \underline{\omega}_{0jk} &= \underline{\beta}_{00k} + \underline{\beta}_{01k}z_{1jk} + \ldots + \underline{\beta}_{0qk}z_{qjk} + \underline{\gamma}_{0k} \\
\underline{\omega}_{1jk} &= \underline{\beta}_{10k} + \underline{\beta}_{11k}z_{1jk} + \ldots + \underline{\beta}_{1qk}z_{qjk} + \underline{\gamma}_{1k} \\
&\vdots \quad \vdots \\
\underline{\omega}_{pjk} &= \underline{\beta}_{q0k} + \underline{\beta}_{q1k}z_{1jk} + \ldots + \underline{\beta}_{pqk}z_{qjk} + \underline{\gamma}_{pk}.
\end{aligned}$$

Or more succulently,

$$\begin{aligned}
\text{Level 1: } \underline{y}_{ijk} &= \underline{\boldsymbol{\omega}}'_{jk}\boldsymbol{x}_{ijk} + \underline{\varepsilon}_{ijk} \\
\text{Level 2: } \boldsymbol{\omega}_{jk} &= \underline{\boldsymbol{\beta}}_{k}\boldsymbol{z}_{jk} + \underline{\boldsymbol{\gamma}}_{k},
\end{aligned}$$

where

$$\begin{pmatrix} \underline{\varepsilon}_{ijk} \\ \underline{\gamma}_{0k} \\ \underline{\gamma}_{1k} \\ \vdots \\ \underline{\gamma}_{qk} \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \psi_{00} & \psi_{01} & \dots & \psi_{0p} \\ 0 & \psi_{01} & \psi_{11} & \dots & \psi_{1p} \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & \psi_{0p} & \psi_{1p} & \dots & \psi_{pp} \end{pmatrix} \right) \; i.i.d.$$

Note that the Level 2 regression coefficients may be random.

Fluid Intelligence Example (a 3-level model)

To analyze the data from this experiment we considered each year as essentially a replication of the base study, but under slightly different circumstances. That is, we assume that the studies are exchangeable. The primary goal of each study was to measure the effect of the intervention. However, context changed slightly over time. We also include two additional variables, whether the student was male or female, and whether the intervention had passed or not. Thus there are three dummy variables. To make our results as interpretable as possible, all variables were standardized. The Cattell F scores were standardized according to population norms by subtracting 100 and dividing by 16. We also standardized the other variables in the model by z-scoring them. This may seem unusual because they were dummies, but it is really just a computational trick to obtain Cohen's d effect sizes from the regression output. Nothing in the structure of the model changes thereby.

Essentially this becomes a three-level hierarchical linear model (e.g., Raudenbush and Bryk, 2002), with year being level 3 and student being level 2, with three dummy variable predictors. Because of the relatively modest size of each experiment, we fit only random intercepts. Stata 12's xtmixed was used to fit the model, using a restricted maximum likelihood (REML) specification due to the relatively small N (StataCorp, 2011). No problems were observed in estimation.

Of primary interest are two coefficients, the standardized effect for the intervention (0.29, SE 0.020, p-value ¡ 0.0001, 95These numbers are well within line with what would be expected from the raw data and, more importantly, what could be expected from experiments of this type. The increase in Cattell fluid intelligence score is modest, but notable. In addition, the between-student heterogeneity and residual error are also quite reasonable.

Stata output:

. xtmixed cfz zinterv zpost zmale, —— year:, covariance(independent) —— id:, covariance(independent) reml Note: single-variable random-effects specification; covariance structure set to identity Note: single-variable random-effects specification; covariance structure set to identity Performing EM optimization: Performing gradient-based optimization: Iteration 0: log restricted-likelihood = -154.80083 Iteration 1: log restricted-likelihood = -154.80083 Computing standard errors:

Problems & Exercises

7.1. Use other response variables in classroom segregation data set.

7.2. Add in those data sets that had normal response variables from GLM chapter. Compare results. Ask to do various things illustrated and discussed in the chapter

7.3. Model bully scale score from Espelage et al. (2003) where peer groups are the clusters. Use empathy, gender, age, fight self-report measure as predictors. Readers should be given some guidance so they discover:

- Plot similar to Figure7.1.1.1 does suggest non-normality (re-visit this example at exercise for Chapter **??**—a non-normal fits better).
- A close look at the data will reveal that gender is actually a level two variable (girl groups, boy groups and a few mixed gender groups).
- Mean centering with mean at Level 2 ends up with non-significant mean (I think).

7.4. Use the data that requires spline. Need to put in more guidance on how to use the spline.

7.5. In the text R_1^2 and R_2^2 were computed for Model 8 in Table **??**. Using the classroom segregation data,

- Verify that the values for Model 12 are correct.
- Verify that the values for Model 7 are correct. (the mean for $\texttt{white}_{ij}$ and the within and between classroom covariances will need to be computed).
- Verify that the values for Model 9 are correct. (a bit more complex here, have to do a little linear algebra).

References

Agresti, A. (2002), *Categorical Data Analysis*, second edn, Wiley, NY.

Agresti, A. (2007), *An Introduction to Categorical Data Analysis*, second edn, Wiley, NY.

Aitkin, M., Anderson, D., Francis, B. & Hinde, J. (1989), *Statistical Modelling in GLIM*, Oxford University Press, Oxford.

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.

Davis, M. H. (1996), *Empathy: A Social Psychological Approach*, Brown and Benchmark Publishers, Madison, WI.

Demidenko, E. (2004), *Mixed Models: Theory and Applications*, Wiley, ?not in book?

Dobson, A. J. & Barnett, A. (2008), *An Introduction to Generalized Linear Models*, third edn, 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.

Ferrari, S. L. P. & Cribari-Neto, F. (2004), 'Beta regression for modelling rates and proportions', *Journal of Applied Statistics* **31**, 799–815.

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.

Javdani, S., Allen, N. E., Todd, N. R. & Anderson, C. J. (2011), 'Examining systems change in the responses to domestic violence: Innovative applications of multilevel modeling', *Violence against women* **17**, 359–375.

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.

Kreft, I. & de Leeuw, J. (1998), 'Introducing multilevel modeling'.

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 sggressive 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.

Smithson, M. & Verkuilen, J. (2006), 'A better lemon squeeser? maximimum-likelihood estimation with beta-distributed dependent variables', *Psychological Methods* **11**, 54–71.

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 Statistican* **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*