

UNIVERSITY OF WOLLONGONG
STAT902. Advanced Data Analysis

Generalised Regression Models

These notes are an excerpt (Chapter 10) from the book *Semiparametric Regression* by D. Ruppert, M.P. Wand & R.J. Carroll. They include material on both generalised linear models (GLM) and generalised linear mixed models (GLMM); both of which are components of STAT902.

©David Ruppert, M.P. Wand, R.J. Carroll 2003

Generalized Parametric Regression

1.1 Introduction

The data that we have dealt with in the preceding chapters has the feature that the response variable is *continuous*. This usually means that, possibly with the help of a transformation, the data can be modeled to be normal, and linear regression techniques such as least squares and best linear unbiased prediction can be used for fitting. However it is often the case that the response variable is not continuous but rather categorical or perhaps a count. Examples include: tumor present or absent; customer prefers green, pink, orange or yellow packaging; number of emergency asthma admissions on a given day. Such response variables cannot be handled through the normal regression framework. In many fields, e.g., medicine and marketing, categorical response variables are more the exception than the rule. Some continuous response data cannot be handled satisfactorily within the normal errors framework; e.g., if they are heavily skewed. Skewed data often can be transformed to near symmetry, but an alternative is to apply a Gamma model (Section 1.4.3) to the untransformed data.

Regression models that aim to handle non-Gaussian response variables such as these are usually termed *generalized linear models* (GLM). The first part of this chapter gives a brief overview of GLM. A reader with plenty of experience in this topic could skip this part of the chapter. The second part of the chapter deals with generalized linear mixed models (GLMM), and is recommended for all readers.

1.2 Binary Response Data

The data depicted in Figure 1.1 correspond to 223 birthweight measurements (grams) and occurrence of *bronchopulmonary dysplasia* (BPD) for a set of babies. The BPD data are coded as

$$\text{BPD}_i = \begin{cases} 1, & \text{if } i\text{th baby has BPD} \\ 0, & \text{otherwise} \end{cases}$$

Such 0-1 data are usually referred to as *binary*. It is of interest to

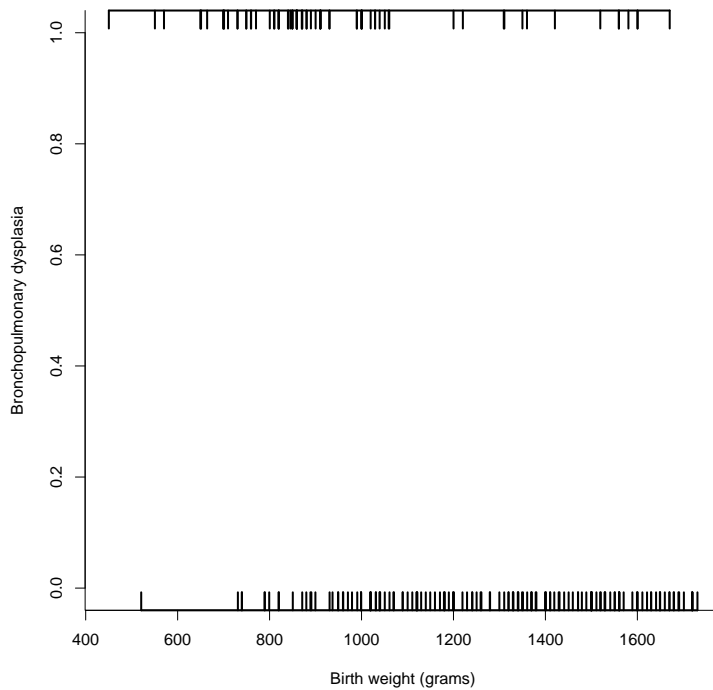


Figure 1.1: Plot of the occurrence of the coded bronchopulmonary dysplasia data against birthweight (grams) for 223 babies.

measure the effect of birthweight on the occurrence of BPD. Consider, for the moment, the model

$$\text{BPD}_i = \beta_0 + \beta_1 \text{birthweight}_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2). \quad (1.1)$$

This is easily seen to be inappropriate. First of all, the right hand side of (1.1) is not guaranteed to be binary. Secondly, it implies that the BPD_i are normally distributed and with homoscedastic errors. Such an assumption is easily refuted for binary data. Finally, the expected values of the fit need not be probabilities, and there is a danger that,

The source of the BPD data is *Principles of Biostatistics* by M. Pagano and K. Gauvreau (Duxbury, 1993). We are grateful to Professor Kim Gauvreau for sharing the data.

when fit, the model will report estimated probabilities of the occurrence of BPD for certain values of birthweight that are negative, or exceed one!

1.3 Logistic Regression

A remedy for the problems raised above is to change the model to

$$P(\text{BPD}_i = 1 | \text{birthweight}) = F(\beta_0 + \beta_1 \text{birthweight}_i) \quad (1.2)$$

where F is a function that maps any real number to a number between 0 and 1. To retain interpretability of the coefficients F should also be strictly increasing. There are many functions that have these properties. In fact, the cumulative distribution function of any continuous distribution with an everywhere positive density, e.g., the standard normal distribution, must meet these requirements. Some examples are shown in Table 1.1 They are each plotted in Figure 1.2, where

Table 1.1: Some cumulative distribution functions used with a binary response.

$F(x)$	distribution
$H(x) = \exp(x) / \{1 + \exp(x)\}$	logistic
$\Phi(x)$	normal
$1 - \exp(-e^x)$	complimentary log-log

it is seen that they have the properties mentioned above. The functions are each shifted and scaled to have the same value and slope at the origin, to allow easier comparison.

Logistic regression uses the logistic probability distribution function, while probit regression uses the normal probability distribution function. Paradoxically, although the normal distribution is used in almost all branches of statistics, for binary data the logistic distribution is used in most applications. The reason for this choice goes back many years and is both philosophical and computational. Unlike the probit model, the logistic model possesses nontrivial sufficient statistics, thus allowing data compression and exact (finite sample) inference. In addition, although this is no longer a major issue, logistic regression requires only the exponential function, something hard-coded into computers, while probit regression requires the calculation of the normal distribution function. Finally, the logistic regression model leads to particularly simple expressions when it comes to fitting the model. The complementary log-log distribution function is

more often used for ordered categorical data. It should be mentioned, however, that Bayesians are abandoning the logistic model in favor of the probit model since the probit is much simpler to implement with MCMC than the logistic.

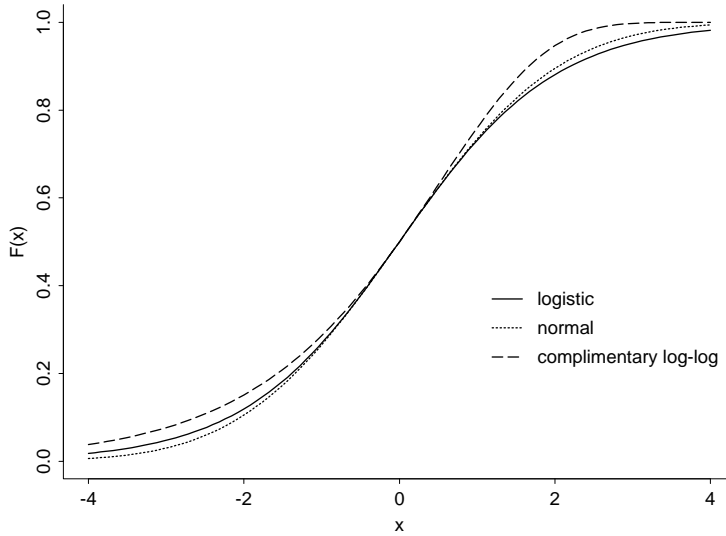


Figure 1.2: Plots of the 3 cumulative distribution functions listed in Table 1.1; but shifted and scaled to have the same value and slope at the origin.

Because $H^{-1}(y) = \log\{y/(1-y)\}$, the logistic regression model can be rewritten as

$$\log \left\{ \frac{P(\text{BPD}_i = 1 | \text{birthweight})}{1 - P(\text{BPD}_i = 1 | \text{birthweight})} \right\} = \beta_0 + \beta_1 \text{birthweight}_i \quad (1.3)$$

The left hand side is the logarithm of the *odds* of BPD for a given birthweight, sometimes called the *log odds* for short. A convenient shorthand is to define the logit function as H^{-1} so that

$$\text{logit}(u) = \log \left(\frac{u}{1-u} \right),$$

which leads to

$$\text{logit}\{P(\text{BPD}_i = 1 | \text{birthweight})\} = \beta_0 + \beta_1 \text{birthweight}_i$$

Notice that logit maps numbers in $(0, 1)$ to the real line. The logit transformation of the probability of BPD is an example of what is commonly called a *link* transformation. In particular, the logit transformation, being the one that results in the simplest likelihood, is called the *canonical link*. A more precise definition of canonical link will be given in Section 1.4.

1.4 Other Generalized Linear Models

While the logistic regression model

$$y_i \sim \text{Bernoulli} \left[\frac{\exp\{(\mathbf{X}\boldsymbol{\beta})_i\}}{1 + \exp\{(\mathbf{X}\boldsymbol{\beta})_i\}} \right]$$

is the most common generalized linear model, there are others that are commonly used in practice. These include the Poisson regression model

$$y_i \sim \text{Poisson} [\exp\{(\mathbf{X}\boldsymbol{\beta})_i\}],$$

appropriate for count data, and the Gamma regression models such as

$$y_i \sim \text{Gamma} [\{1/(\mathbf{X}\boldsymbol{\beta})_i\}, \phi] \quad (1.4)$$

and

$$y_i \sim \text{Gamma} [\exp\{(\mathbf{X}\boldsymbol{\beta})_i\}, \phi], \quad (1.5)$$

that are appropriate for right-skewed continuous data. Although (1.4) has the canonical link, (1.5) is more commonly used because the log-link guarantees a positive mean. Here Gamma (μ, ϕ) means a gamma distribution with mean μ and coefficient of variation $\sqrt{\phi}$. The gamma and the Gaussian families are examples of GLMs with dispersion parameters, the standard deviation for the Gaussian family and coefficient of variation for the gamma family.

The coefficient of variation of a distribution is the ratio of the standard deviation to the mean.

A GLM begins with a one-parameter exponential family of distribution for the response with density of the form

$$f(y; \eta) = \exp \left\{ \frac{y\eta - b(\eta)}{\phi} + c(y, \phi) \right\}, \quad (1.6)$$

for some functions $b(\eta)$ and $c(y, \phi)$; see Table 1.2. Here ϕ is a dispersion parameter; the Bernoulli and Poisson distributions have no dispersion parameters so for these distributions one takes $\phi \equiv 1$. The parameter η is called the natural parameter. It can be shown that $E(y) = b'(\eta)$ and $\text{var}(y) = \phi b''(\eta)$, where $b'(\eta)$ and $b''(\eta)$ are the first and second derivatives of b . In a GLM, it is assumed that the natural parameter for y_i , η_i , depends on a vector of predictor variables, \mathbf{x}_i . More explicitly, it is assumed that for some function, ψ , $\eta_i = \psi(\mathbf{x}_i^T \boldsymbol{\beta})$. The canonical link occurs when ψ is the identity function and hence $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$.

More generally, the link function \mathcal{L} is defined by the equation $\mathcal{L}\{E(y_i)\} = \mathbf{x}_i^T \boldsymbol{\beta}$. Later we will need the notation $\mu(\cdot) = \mathcal{L}(\cdot)^{-1}$.

Note that the inverse link, $\mu(\cdot)$, converts the linear predictor $\mathbf{x}_i^T \boldsymbol{\beta}$ to the expectation of y_i : $\mu(\mathbf{x}_i^T \boldsymbol{\beta}) = \mathbb{E}(y_i) = \mu_i$. For logistic regression, $\mathcal{L}(u) = \text{logit}(u)$ and $\mu(u) = H(u)$ where H is the logistic function.

The dispersion parameter ϕ is assumed to not depend on i ; this is a generalization of the constant variance assumption of the linear model. With these assumptions, the density of \mathbf{y} is

$$f(\mathbf{y}; \boldsymbol{\beta}) = \exp \left[\frac{\mathbf{y}^T \psi(\mathbf{X} \boldsymbol{\beta}) - \mathbf{1}^T b\{\psi(\mathbf{X} \boldsymbol{\beta})\}}{\phi} + \mathbf{1}^T c(\mathbf{y}, \phi) \right]. \quad (1.7)$$

name	$b(\eta)$	$b'(\eta)$	canon. link	$c(y, \phi)$
Bernoulli	$\log(1 + e^\eta)$	$e^\eta / (1 + e^\eta)$	$\text{logit}(\mu)$	0
Poisson	e^η	e^η	$\log(\mu)$	$-\log(y!)$
Gamma	$\log(\eta)$	$1/\eta$	$1/\mu$	see text
Gaussian	$\eta^2/2$	η	μ	see text

Table 1.2: Functions b and c for some common one-parameter exponential family.

1.4.1 Poisson regression and overdispersion

The Poisson GLM, often called *Poisson regression*, uses the Poisson density

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, \dots$$

The logarithm of this density is $y \log(\mu) - \mu - \log(y!)$. Letting $\eta = \log(\mu)$, the log density is $y\eta - e^\eta - \log(y!)$. Therefore, $b'(\eta) = e^\eta$ and $c(y, \phi) = -\log(y!)$.

The Poisson regression model is often used when the response is a count. However, the assumption that a count is Poisson distributed should not be taken lightly, since the variance of a Poisson regression equals its mean. Often, a response that is a count has a variance that is larger than its mean, sometimes much larger. This is “overdispersion” relative to the Poisson model (McCullagh and Nelder, 1989). In such a case, the Poisson assumption can lead to an underestimation of variability. This would, for example, cause confidence intervals to be too small with coverage probability smaller than the nominal value. With serious overdispersion, which occurs fairly frequently, the size of the undercoverage could be substantial. See Section 1.7 for discussion of overdispersion and other models for the variance.

There are various models for overdispersion, the most well-known being the negative binomial distribution. Another way to achieve

Poisson regression is often appropriate for count data, but overdispersion often occurs and can affect inference.

overdispersion is to introduce random effects, see Section 1.8 and following for details. Let u be a normal random variable with mean $-\sigma_u^2/2$ and variance σ_u^2 ; these parameters are chosen so that $E\{\exp(u)\} = 1$. Suppose that given u and the covariate \mathbf{x} , y is Poisson with mean $\exp(\mathbf{x}^T \boldsymbol{\beta} + u)$. Then unconditionally, y has mean $\exp(\mathbf{x}^T \boldsymbol{\beta})$ but its variance is

$$\exp(\mathbf{x}^T \boldsymbol{\beta}) + \exp(2\mathbf{x}^T \boldsymbol{\beta}) \{ \exp(\sigma_u^2) - 1 \} \geq \exp(\mathbf{x}^T \boldsymbol{\beta}).$$

Thus, the variance exceeds the mean (overdispersion) unless the random effects are all zero. With this mixed model formulation, overdispersion is a natural consequence of nonzero random effects.

1.4.2 The Gaussian GLM: a model for symmetrically distributed and homoscedastic responses

Ordinary multiple linear regression is a Gaussian GLM.

The Gaussian GLM is just the ordinary linear model, which shows that generalized linear models are, in fact, a generalization of linear models. The Gaussian family of densities, when parametrized by the mean μ and variance ϕ , is

$$\frac{1}{\sqrt{2\pi\phi}} \exp \left\{ -\frac{1}{2\phi}(y - \mu)^2 \right\}.$$

The log density is

$$\frac{y\mu - \mu^2/2}{\phi} - \frac{1}{2} \left\{ \frac{y^2}{\phi} + \log(2\pi\phi) \right\}.$$

Therefore, $\eta = \mu$, $b(\eta) = \eta^2/2$, and $c(y, \phi) = -\frac{1}{2} \left\{ \frac{y^2}{\phi} + \log(2\pi\phi) \right\}$.

1.4.3 The Gamma GLM: a model with a constant coefficient of variation

There are many ways to parameterize the gamma family. Following McCullagh and Nelder (1989), we will use the mean, μ , and squared

coefficient of variation (variance over squared mean), denoted by ϕ , as the parameters. Then the gamma density with parameters (μ, ϕ) is

$$\frac{1}{y \Gamma(\phi^{-1})} \left(\frac{y}{\phi \mu} \right)^{\phi^{-1}} \exp \left(-\frac{y}{\phi \mu} \right). \quad (1.8)$$

Define $\eta = -1/\mu$. Then the log of the density in (1.8) is

$$\frac{y \eta + \log(-\eta)}{\phi} - \log \{y \Gamma(\phi^{-1})\} + \phi^{-1} \log(y/\eta),$$

which is in exponential family form with $b(\eta) = -\log(-\eta)$ and $c(y, \phi) = -\log\{y \Gamma(\phi^{-1})\} + \phi^{-1} \log(y/\eta)$.

The gamma model can be used when the responses have a right-skewed distribution and are heteroscedastic. However, only a special type of heteroscedasticity can be modeled by the gamma family; $\text{var}(y_i|\mathbf{x}_i)$ must be proportional to $\{E(y_i|\mathbf{x}_i)\}^2$. Other types of heteroscedasticity should be modeled by variance function models or transformation models; see Carroll and Ruppert (1988), which also has an extensive discussion of how to detect the presence and functional form of heteroscedasticity. Ruppert, Carroll, and Cressie (1989, 1991) compare the gamma/GLM approach to modeling skewness and heteroscedasticity with the more flexible variance function/transformation approach. A brief discussion of variance function estimation in non-parametric regression is given in Chapter ???. Section 1.7 below contains details of estimation in overdispersion and variance function models.

1.5 Iteratively Reweighted Least Squares

In the GLM family, we have seen that if the mean $E(y|\mathbf{x}) = \mu(\mathbf{x}^T \boldsymbol{\beta})$, then the variance is $\text{Var}(y|\mathbf{x}) = \phi V(\mathbf{x}^T \boldsymbol{\beta})$ for some function V . In canonical exponential families, it can be shown that the first derivative of μ is V , i.e., $\mu' = V$.

Computing parameter estimates in GLMs is particularly simple, and uses a method called iteratively reweighted least squares. This method also turns out to be equivalent to *Fisher's method of scoring*, which is simply the Newton–Raphson method with the Hessian replaced by its expected value.

Iteratively reweighted least squares is the main computational engine for GLM's and also for some implementations of generalized linear mixed models

Suppose now that y_1, \dots, y_n denote the response variables and $\mathbf{x}_1, \dots, \mathbf{x}_n$ are vectors of predictor variables. For the BPD example

$$y_i = \text{BPD}_i \quad \text{and} \quad \mathbf{x}_i = \begin{bmatrix} 1 \\ \text{birthweight}_i \end{bmatrix}.$$

Define

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}.$$

In the case of the birthweight data

$$\mathbf{X} = \begin{bmatrix} 1 & \text{birthweight}_1 \\ \vdots & \vdots \\ 1 & \text{birthweight}_{223} \end{bmatrix}.$$

Notice that \mathbf{X} is precisely the model matrix in a linear regression model.

In iteratively reweighted least squares, the basic idea is as follows. Suppose the current estimate is $\boldsymbol{\beta}^{(t)}$. Form the weights $w = 1/V(\mathbf{x}^T \boldsymbol{\beta}^{(t)})$. Then in iteratively reweighted least squares, we update the current estimate by performing *one step* of Fisher's method of scoring for weighted least squares. When implemented, the algorithm takes the following form. Let

$$\begin{aligned} \mathbf{W}_{1,\boldsymbol{\beta}} &\equiv \text{diag} \{ \mu'(\mathbf{x}_i^T \boldsymbol{\beta}) \}; \\ \mathbf{W}_{2,\boldsymbol{\beta}} &\equiv \text{diag} \{ V(\mathbf{x}_i^T \boldsymbol{\beta}) \}. \end{aligned}$$

Then the updating step is

$$\hat{\boldsymbol{\beta}} \leftarrow \hat{\boldsymbol{\beta}} + (\mathbf{X}^T \mathbf{W}_{1,\hat{\boldsymbol{\beta}}} \mathbf{W}_{2,\hat{\boldsymbol{\beta}}}^{-1} \mathbf{W}_{1,\hat{\boldsymbol{\beta}}} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_{1,\hat{\boldsymbol{\beta}}} \mathbf{W}_{2,\hat{\boldsymbol{\beta}}}^{-1} \{\mathbf{y} - \mu(\mathbf{X} \hat{\boldsymbol{\beta}})\}. \quad (1.9)$$

For canonical links, $\mathbf{W}_{1,\boldsymbol{\beta}} = \mathbf{W}_{2,\boldsymbol{\beta}}$, and the algorithm takes the usual generalized least squares form

$$\hat{\boldsymbol{\beta}} \leftarrow \hat{\boldsymbol{\beta}} + (\mathbf{X}^T \mathbf{W}_{2,\hat{\boldsymbol{\beta}}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \{\mathbf{y} - \mu(\mathbf{X} \hat{\boldsymbol{\beta}})\}. \quad (1.10)$$

In the logistic regression case, the algorithm takes the simple form

$$\hat{\boldsymbol{\beta}} \leftarrow \hat{\boldsymbol{\beta}} + \left[\mathbf{X}^T \text{diag} \left\{ \frac{e^{\mathbf{x} \hat{\boldsymbol{\beta}}}}{(1+e^{\mathbf{x} \hat{\boldsymbol{\beta}}})^2} \right\} \mathbf{X} \right]^{-1} \mathbf{X}^T \left(\mathbf{y} - \frac{e^{\mathbf{x} \hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x} \hat{\boldsymbol{\beta}}}} \right). \quad (1.11)$$

1.6 Hat Matrix, Degrees of Freedom, and Standard Errors

It is relatively easy to define analogues to the hat matrix and degrees of freedom, as well as to obtain standard errors for the parameter estimates. The calculations necessary are outlined for the logistic case at the end of this chapter in Section 1.10.

Using those type of calculations, the hat matrix is defined as

$$\mathbf{H}_\beta = \mathbf{W}_{1,\beta} \mathbf{X} \left(\mathbf{X}^T \mathbf{W}_{1,\beta} \mathbf{W}_{2,\beta}^{-1} \mathbf{W}_{1,\beta} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W}_{1,\beta} \mathbf{W}_{2,\beta}^{-1} \quad (1.12)$$

and the degrees of freedom of the fit is $\text{tr}(\mathbf{H}_\beta)$ which equals

$$\text{tr} \left\{ \left(\mathbf{X}^T \mathbf{W}_{1,\beta} \mathbf{W}_{2,\beta}^{-1} \mathbf{W}_{1,\beta} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W}_{1,\beta} \mathbf{W}_{2,\beta}^{-1} \mathbf{W}_{1,\beta} \mathbf{X} \right\} = p. \quad (1.13)$$

Moreover, the estimated variance matrix of $\hat{\beta}$ is

$$\widehat{\text{Cov}}(\hat{\beta}) = \left(\mathbf{X}^T \mathbf{W}_{1,\hat{\beta}} \mathbf{W}_{2,\hat{\beta}}^{-1} \mathbf{W}_{1,\hat{\beta}} \mathbf{X} \right)^{-1}. \quad (1.14)$$

Standard errors for each component of the estimate of β are formed by taking the square root of the diagonal of the matrix in (1.14).

The fit to the BPD data based on this estimation strategy is shown in Figure 1.3. One can see that the estimated probability of BPD decreases from approximately 0.80 to approximately 0.05, indicating a strong dependence of BPD on birth weight. Of course, to assess this dependence statistically, we need inference procedures, and for this we use standard errors.

Application of the standard error formulae to the BPD example leads to the standard errors and t-values in Table 1.3. Note that, as expected, birth weight is a statistically significant predictor of BPD, with higher birth weights associated with lower risk of BPD.

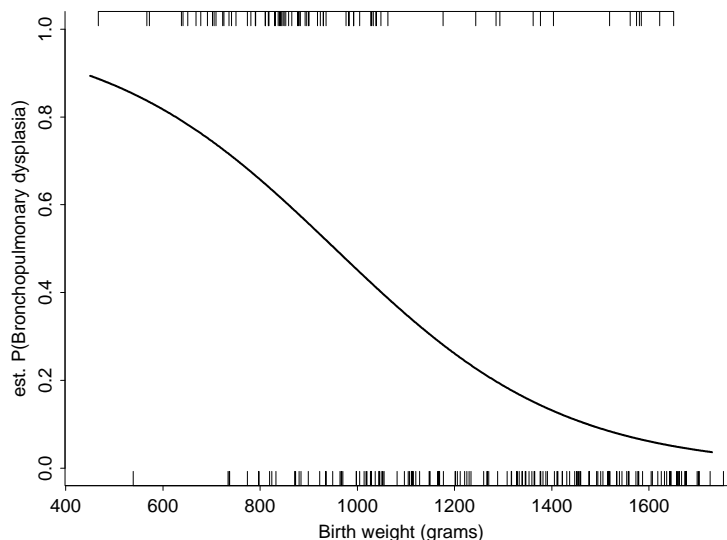
coefficient	value	st.dev	t-value
Intercept	1.21	1.06×10^{-1}	11.5
birth.weight	-7.46×10^{-4}	8.70×10^{-5}	-8.58

These quantities are the natural generalizations from ordinary linear regression to GLMs.

Table 1.3: Results of the logistic regression analysis of the BPD data.

1.7 Overdispersion and Variance Functions: Pseudolikelihood

Figure 1.3: Plot of the occurrence of the coded bronchopulmonary dysplasia data against birthweight (grams) for 223 babies.



In general regression problems, suppose that the mean is $f(\mathbf{x}^T \boldsymbol{\beta})$ and the variance is $V(\mathbf{x}^T \boldsymbol{\beta}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is an unknown parameter. For example, consider Poisson data. We discussed in Section 1.4.1 a model for count data in which

$$\begin{aligned} f(\mathbf{x}^T \boldsymbol{\beta}) &= \exp(\mathbf{x}^T \boldsymbol{\beta}); \\ V(\mathbf{x}^T \boldsymbol{\beta}, \boldsymbol{\theta}) &= f(\mathbf{x}^T \boldsymbol{\beta}) + \boldsymbol{\theta} f^2(\mathbf{x}^T \boldsymbol{\beta}). \end{aligned}$$

Variance function estimation and the pseudolikelihood algorithm form a general approach to the problem of non-constant variances in regression, and are applicable to nonlinear least squares problems as well.

Other common models include the power of the mean model, so that with $\boldsymbol{\theta} = (\theta_0, \theta_1)$,

$$V(\mathbf{x}^T \boldsymbol{\beta}, \boldsymbol{\theta}) = \theta_0 f(\mathbf{x}^T \boldsymbol{\beta})^{\theta_1}.$$

Carroll and Ruppert (1988) describe the pseudolikelihood method for estimating $(\boldsymbol{\beta}, \boldsymbol{\theta})$. The algorithm is given as follows. The method refers to iteratively reweighted least squares, see Section 1.5 for details. We assume a sample of size n .

- (1) Set $t = 0$ and estimate $\boldsymbol{\beta}$ by (unweighted) iteratively reweighted least squares with a constant variance function. Call the estimate $\boldsymbol{\beta}^{(t)}$.
- (2) Estimate $\boldsymbol{\theta}$ by maximizing in $\boldsymbol{\theta}$ only the pseudolikelihood

$$-\sum_{i=1}^n \log \left\{ V(\mathbf{x}_i^T \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}) \right\} - \sum_{i=1}^n \left\{ y_i - f(\mathbf{x}_i^T \boldsymbol{\beta}^{(t)}) \right\}^2 / V(\mathbf{x}_i^T \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}).$$

Call the estimate $\boldsymbol{\theta}^{(t)}$.

- (3) Update $\boldsymbol{\beta}^{(t)}$ to $\boldsymbol{\beta}^{(t+1)}$ using iteratively reweighted least squares with the variance function $V(\mathbf{x}^T \boldsymbol{\beta}, \boldsymbol{\theta}^{(t)})$.
- (4) Set $t = t + 1$ and return to step 2.
- (5) Iterate until convergence.

1.7.1 Quasilikelihood and overdispersion parameters

One common approach to quasilikelihood starts with a full parametric model, but then relaxes the assumptions by specifying that the variance function is ϕ times the variance function specified by that model, where ϕ is an unknown overdispersion parameter. For example, one might specify that the variance of count data is ϕ times the mean. Quasilikelihood estimation is used in the GLIMMIX macro of SAS that is discussed in Section 1.8.3. GLIMMIX will provide an estimate of ϕ . If $\hat{\phi} > 1$, then there is an indication of overdispersion.

1.8 Generalized Linear Mixed Models

As in the case of linear models, it is sometimes useful to incorporate random effects into a generalized linear model. The resultant models are known as *generalized linear mixed models*. However, their fitting presents some computational challenges. This has led to a large amount of recent research aimed at overcoming these challenges.

Each of the generalized linear models in Section 1.4 can be extended to allow for some effects to be random. We will denote such random effects by \mathbf{u} , and we will assume that they are normally distributed with mean zero and a covariance matrix \mathbf{G}_θ , where \mathbf{G}_θ is a positive definite matrix that depends on a parameter vector $\boldsymbol{\theta}$ usually called the variance components.

The most common such models are the logistic-normal mixed model

$$y_i | \mathbf{u} \sim \text{Bernoulli} \left[\frac{\exp \{(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i\}}{1 + \exp \{(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i\}} \right], \quad \mathbf{u} \sim N(\mathbf{0}, \mathbf{G}_\theta)$$

Generalized linear mixed models include hierarchical models, longitudinal models and cluster variation models. This area is one of the most rapidly expanding and vigorously researched fields in Statistics.

and the general Poisson-normal mixed model

$$y_i | \mathbf{u} \sim \text{Poisson}[\exp\{(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i\}], \quad \mathbf{u} \sim N(\mathbf{0}, \mathbf{G}_\theta).$$

In what follows, for purposes of explication we will assume that the dispersion parameter ϕ is known and equal to 1, e.g., logistic or Poisson regression. We will also work entirely within the context of the canonical exponential family. We can treat both the logistic-normal and Poisson-normal models with the one-parameter exponential family notation:

$$\begin{aligned} f(\mathbf{y}|\mathbf{u}) &= \exp\left\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})\right. \\ &\quad \left. - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) + \mathbf{1}^T c(\mathbf{y})\right\}, \\ f(\mathbf{u}) &= (2\pi)^{-q/2} |\mathbf{G}_\theta|^{-1/2} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{G}_\theta^{-1} \mathbf{u}\right), \end{aligned} \quad (1.15)$$

where q is the dimension of \mathbf{u} . The last formula is the probability density function of the random effects.

1.8.1 Estimation of model parameters

The parameters in the model are $(\boldsymbol{\beta}, \boldsymbol{\theta})$, and the corresponding likelihood is

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}) &= f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\theta}) \\ &= \int_{\mathbb{R}^q} f(\mathbf{y}|\mathbf{u}) f(\mathbf{u}) d\mathbf{u} \\ &= (2\pi)^{-q/2} |\mathbf{G}_\theta|^{-1/2} \exp\{\mathbf{1}^T c(\mathbf{y})\} J(\boldsymbol{\beta}, \boldsymbol{\theta}) \end{aligned}$$

The computational issues in GLMMs are nontrivial and require special tools. There is considerable research interest in the computational methods themselves, as well as the modeling.

where

$$J(\boldsymbol{\beta}, \boldsymbol{\theta}) = \int_{\mathbb{R}^q} \exp\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \frac{1}{2} \mathbf{u}^T \mathbf{G}_\theta^{-1} \mathbf{u}\} d\mathbf{u}. \quad (1.16)$$

Maximum likelihood estimation of $\ell(\boldsymbol{\beta}, \boldsymbol{\theta})$ is hindered by the presence of this q -dimensional integral. As in Section ??, if we want to use GLMMs to fit penalized splines, q is the number of knots: even for 5 knots the integral becomes essentially intractable by direct calculation.

There has been a great deal of research, accelerating in the 1990s, on remedies to the computational problem. There are also a variety of software options, see for example the web site multilevel.ioe.ac.uk for various links. These remedies may be divided into four distinct categories:

(1) Laplace approximation of (1.16): PQL.

Laplace's method is a classical approximation technique for handling intractable multivariate integrals. Application to (1.16) reduces the problem to one that is akin to fitting a generalized linear model (among many others see Schall, 1991; McGilchrist and Aisbett, 1991; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993). The only difference is that the coefficients are subject to a penalty, and nowadays the name *penalized quasi-likelihood* (PQL) is usually associated with the method. Interestingly, PQL is essentially the same as maximizing the joint likelihood of the observed data and random effects (Gilmour, Anderson and Rae, 1985; Harville and Mee, 1984; Schall, 1991), simultaneously, see below for more details. This corresponds to Henderson's (1950) justification for Gaussian mixed models; see equation (?). Improved Laplace approximation through higher-order expansion has been investigated by Shun and McCullagh (1995), Shun (1997) and Raudenbush, Yang and Yosef (2000).

(2) Bias corrections to PQL.

The approximations used by PQL induce bias in the estimates. This has resulted in a stream of research (Breslow and Lin, 1995; Goldstein and Rasbash, 1996; Lin and Breslow, 1996) that uses asymptotic arguments to devise bias corrections to the PQL estimates.

(3) Fitting via Expectation Maximization (EM).

The Expectation Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) can be used to fit mixed models by treating the random effects as missing. However, the E-step involves intractable integrals, so Laplace integration (Steele, 1996) or Monte Carlo methods (McCulloch, 1997; Booth and Hobert, 1999) need to be employed.

(4) Bayesian fitting via Markov Chain Monte Carlo (MCMC).

This involves a Bayesian formulation of the generalized linear mixed model in which $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is treated as randomly distributed according to some prior distribution. The posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is intractable, so Markov Chain Monte Carlo algorithms (see e.g. Robert and Casella, 1999) are used to generate samples from this distribution and allow estimation and inference for these parameters (Zeger and Karim,

1991; Clayton, 1996; Diggle, Tawn and Moyeed, 1998). See Chapter ?? for a discussion of Bayesian fitting by MCMC.

Generalized estimating equations (GEE) (Gourieroux, Monfort and Trognon, 1984, Liang and Zeger, 1986) are also, in some sense, a remedy to the maximum likelihood problem conveyed by (1.16). However, it is specific to the longitudinal data setting rather than the type of mixed models that arise in smoothing. Hence, we will forego outlining details of this methodology.

We now discuss each of (1)–(4).

1.8.2 Penalized quasilikelihood (PQL)

Penalized quasi-likelihood is a relatively simple method for fitting generalized linear mixed models. The fits from PQL also serve as useful starting values for the other fitting approaches. As we will see in Section ??, its application to penalized spline fitting is equivalent to the penalized likelihood approach traditionally used there. In this subsection, we state the necessary formulae required to implement PQL. Derivation of these equations is given in Section 1.10.4.

Recall that in what follows, for purposes of explication we will assume that the data come from a canonical exponential family and that the dispersion parameter ϕ is known and equal to 1. Write $\mu = b'$ and $V = b''$ as the mean and variance functions. Also write $\boldsymbol{\mu} = \mu(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) = b'(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) = E(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{u})$, and $\mathbf{W} = \text{diag}\{b''(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})\} = \text{Var}(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{u})$.

Recall from (1.15) that $f(\mathbf{y}|\mathbf{u})$ is the notation we use for the likelihood of the data given the random effects \mathbf{u} . PQL estimates of $(\boldsymbol{\beta}, \mathbf{u})$ are obtained by treating the random effects \mathbf{u} as fixed parameters, but the likelihood is penalized according to the distribution of \mathbf{u} . Thus, for given $\boldsymbol{\theta}$, $(\boldsymbol{\beta}, \mathbf{u})$ is obtained by maximizing the penalized log-likelihood

$$\log\{f(\mathbf{y}|\mathbf{u})\} - \frac{1}{2}\mathbf{u}^T \mathbf{G}_{\boldsymbol{\theta}}^{-1} \mathbf{u}.$$

The notion that the likelihood is penalized leads to the name *penalized likelihood*. The nomenclature penalized quasilikelihood (PQL) results from a technical extension of likelihood called quasilikelihood.

Given $\boldsymbol{\theta}$, direct differentiation of the penalized likelihood leads to the score equations for $(\boldsymbol{\beta}, \mathbf{u})$:

$$\begin{bmatrix} \mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu}) \\ \mathbf{Z}^T(\mathbf{y} - \boldsymbol{\mu}) - \mathbf{G}_{\boldsymbol{\theta}}^{-1} \mathbf{u} \end{bmatrix} = \mathbf{0}. \tag{1.17}$$

PQL is only an approximation to a full likelihood analysis, except in the Gaussian GLMM, i.e., an ordinary LMM, where it is exact. Sometimes the approximation works remarkably well, but in some problems such as logistic regression the variance components may not be well estimated.

The Hessian of (1.17) is independent of \mathbf{y} and is given by

$$-\begin{bmatrix} \mathbf{X}^T \mathbf{W} \mathbf{X} & \mathbf{X}^T \mathbf{W} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{W} \mathbf{X} & \mathbf{Z}^T \mathbf{W} \mathbf{Z} + \mathbf{G}_{\boldsymbol{\theta}}^{-1} \end{bmatrix}.$$

In this case, Newton–Raphson and Fisher’s method of scoring are identical.

There is an identical formulation of Fisher’s method of scoring that leads to the PQL estimates of $\boldsymbol{\theta}$. Consider the pseudodata

$$\mathbf{y}_{\text{pseudo}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{W}^{-1}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}_{\text{pseudo}}.$$

This is in the form of a *linear* mixed model (LMM) where in the notation of Section ??, the covariance matrix of the pseudoerrors $\boldsymbol{\varepsilon}_{\text{pseudo}}$ is $\mathbf{R} = \mathbf{W}^{-1}$. Fisher’s method of scoring turns out to be nothing more than iterative updating of the LMM formula (??) and (??), using the pseudodata $\mathbf{y}_{\text{pseudo}}$ as the response.

1.8.2.1 Estimation of Variance Components via Mixed Models

This formulation of the GLMM as an interactively updated form of the LMM led Breslow and Clayton (1993) to suggest a PQL method for estimating $\boldsymbol{\theta}$. Specifically, fixing $\boldsymbol{\beta}$ and \mathbf{u} at their current values, they suggest updating $\boldsymbol{\theta}$ at each stage of the iteration by using the ML or REML estimates of Section ?? applied to the pseudodata and with $\mathbf{R} = \mathbf{W}^{-1}$.

1.8.2.2 Estimation of Variance Components via Cross-Validation

An alternative method for estimating the variance components appropriate for smoothing is cross-validation, see Chapter ??.

1.8.3 GLIMMIX

The SAS macro GLIMMIX implements a refinement of PQL due to Wolfinger and O’Connell (1993) which they call pseudo-likelihood. Pseudo-likelihood incorporates an overdispersion parameter, ϕ . The GLIMMIX macro makes PL available on a standard statistical package, and values of $\hat{\phi}$ substantially larger than 1 are an indication of overdispersion.

1.8.4 Bias correction to PQL

PQL is based only on an approximate likelihood, and thus estimates of the variance component $\boldsymbol{\theta}$ are asymptotically biased, as are estimates of $\boldsymbol{\beta}$. This has led Breslow and Lin (1995) and Lin and Breslow (1996) to derive corrections to the PQL estimates based on small $\boldsymbol{\theta}$ asymptotics.

1.8.5 Fitting via expectation maximization

The Expectation Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) is a general purpose method for maximum likelihood estimation in the presence of missing data (see e.g. McLachlan and Krishnan, 1997). It can be used for fitting mixed models by treating the random effects as missing data. For the Gaussian mixed model it provides an alternative to BLUP/REML for estimation of the model parameters (Laird and Ware, 1982). It can also be used to guide the choice of the parameters in the generalized context. For the generalized linear mixed model

The EM algorithm in GLMMs often requires simulation, and is sometimes referred to as *Monte Carlo EM*.

$$f(\mathbf{y}|\mathbf{u}) = \exp\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) + \mathbf{1}^T c(\mathbf{y})\},$$

$$f(\mathbf{u}) = (2\pi)^{-q/2} |\mathbf{G}_{\boldsymbol{\theta}}|^{-1/2} \exp(-\frac{1}{2} \mathbf{u}^T \mathbf{G}_{\boldsymbol{\theta}}^{-1} \mathbf{u}),$$

let $\boldsymbol{\psi} \equiv (\boldsymbol{\beta}, \boldsymbol{\theta})$ be the parameter vector. The EM algorithm iterates between the *E-step* (Expectation) and the *M-step* (Maximization) until convergence. The E-step requires computation of an expectation:

$$Q(\boldsymbol{\psi}' | \boldsymbol{\psi}) \equiv E_{\mathbf{u}|\mathbf{y}; \boldsymbol{\psi}} \{\log f(\mathbf{y}, \mathbf{u}; \boldsymbol{\psi}')\}$$

while the M-step involves an update of parameter estimates through maximization:

$$\boldsymbol{\psi}_{\text{new}} = \underset{\boldsymbol{\psi}}{\operatorname{argmax}} Q(\boldsymbol{\psi} | \boldsymbol{\psi}_{\text{old}}). \quad (1.18)$$

Since, from Bayes Rule,

$$f(\mathbf{u}|\mathbf{y}; \boldsymbol{\psi}) = \frac{f(\mathbf{y}|\mathbf{u}; \boldsymbol{\psi})f(\mathbf{u})}{\int_{\mathbb{R}^q} f(\mathbf{y}|\mathbf{u}; \boldsymbol{\psi})f(\mathbf{u}; \boldsymbol{\psi}) d\mathbf{u}} \quad (1.19)$$

we have the representation

$$Q(\boldsymbol{\psi}' | \boldsymbol{\psi}) = \frac{\int_{\mathbb{R}^q} \log f(\mathbf{y}, \mathbf{u}; \boldsymbol{\psi}') f(\mathbf{y}, \mathbf{u}; \boldsymbol{\psi}) d\mathbf{u}}{\int_{\mathbb{R}^q} f(\mathbf{y}, \mathbf{u}; \boldsymbol{\psi}) d\mathbf{u}}. \quad (1.20)$$

However, computation of (1.20) is at least as difficult as computation of the log-likelihood $\ell(\boldsymbol{\psi})$.

One solution is to use Laplace's approximation to handle the integrals in (1.20) (Steele, 1996). Techniques designed for approximating ratios of integrals (e.g. Tierney, Kass and Kadane, 1989) are appropriate in this case. Alternatively one can use a *Monte Carlo EM*:

$$\widehat{Q}(\boldsymbol{\psi}' | \boldsymbol{\psi}) = \frac{1}{m} \sum_{i=1}^m \log f(\mathbf{y}, \mathbf{u}_i; \boldsymbol{\psi}'), \quad (1.21)$$

where $\mathbf{u}_1, \dots, \mathbf{u}_m$ is a Monte Carlo generated sample from $[\mathbf{u} | \mathbf{y}; \boldsymbol{\psi}]$ (Wei and Tanner, 1990). Inspection of (1.19) shows $[\mathbf{u} | \mathbf{y}; \boldsymbol{\psi}]$ to have a complicated distribution from which sampling is difficult. One remedy is to use the Metropolis-Hastings (MH) algorithm (McCulloch, 1997). Another is to replace (1.21) by

$$\frac{1}{m} \sum_{i=1}^m \frac{\log f(\mathbf{y}, \mathbf{u}_i^h; \boldsymbol{\psi}') f(\mathbf{y} | \mathbf{u}_i^h) f(\mathbf{u}_i^h)}{h(\mathbf{u}_i^h)} \quad (1.22)$$

where h is a standard density such as that of the multivariate t-distribution and $\mathbf{u}_1^h, \dots, \mathbf{u}_m^h$ are a random sample from h . This is known as *importance sampling* (e.g. Rubinstein, 1981; Booth and Hobert, 1999). Note that (1.22) estimates the numerator of (1.20) rather than $Q(\boldsymbol{\psi}' | \boldsymbol{\psi})$ itself. However, since the denominator does not involve $\boldsymbol{\psi}'$ the M-step (1.18) is unaffected.

1.8.6 Bayesian fitting via Markov Chain Monte Carlo

Another approach to fitting an GLMM is to put priors on all parameters and to simulate from the posterior by Markov Chain Monte Carlo (MCMC). This, in effect, integrates out the unobserved random effects; it also imputes the values of the random effects so that we get the equivalent of their BLUPs. The parameters are estimated by their posterior means that are approximated by their sample averages from the MCMC output. This method is introduced in Chapter ?? after MCMC methods are discussed. See Section ??.

1.8.7 Prediction of random effects

For the generalized extension of semiparametric models such as those described in Chapters ?? and ??, prediction of \mathbf{u} is required. For known $(\boldsymbol{\beta}, \boldsymbol{\theta})$, the best predictor of \mathbf{u} is

$$\tilde{\mathbf{u}} = E_{(\boldsymbol{\beta}, \boldsymbol{\theta})}(\mathbf{u} | \mathbf{y})$$

which suggests the predictor

$$\hat{\mathbf{u}} = E_{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})}(\mathbf{u}|\mathbf{y}).$$

Note that

$$\tilde{\mathbf{u}} = \frac{\int_{\mathbb{R}^q} \mathbf{u} \exp\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \frac{1}{2}\mathbf{u}^T \mathbf{G}_{\boldsymbol{\theta}}^{-1} \mathbf{u}\} d\mathbf{u}}{\int_{\mathbb{R}^q} \exp\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \frac{1}{2}\mathbf{u}^T \mathbf{G}_{\boldsymbol{\theta}}^{-1} \mathbf{u}\} d\mathbf{u}}$$

so computation is hindered by the presence of higher dimensional integrals.

Prediction of the random effects is crucial in smoothing and semiparametric fitting.

PQL of course directly estimates \mathbf{u} . Methods that provide corrected estimates of $\boldsymbol{\theta}$ can then provide an estimate of \mathbf{u} through the solving (1.17). Monte–Carlo EM produces an estimate of $\boldsymbol{\beta}$, and since one generates a sample of \mathbf{u} 's from the distribution of \mathbf{u} given \mathbf{y} , the mean of these samples provides an estimate of \mathbf{u} . Similarly, in the Bayesian formulation, MCMC also provides a sample of the \mathbf{u} 's, and the mean of this sample yields an estimate of \mathbf{u} .

1.8.8 Standard error estimation

Standard error estimates for the estimates of $(\boldsymbol{\beta}, \mathbf{u})$ can be constructed in each case. For the EM algorithm, consult Louis (1982). Bayesian methods yield standard error estimates and posterior confidence intervals as part of the MCMC calculations. Both methods account for the estimation of the variance component $\boldsymbol{\theta}$. On the other hand, known standard error estimates for PQL do not account for estimation of $\boldsymbol{\theta}$. However, using the identification of PQL as an iterative updating of the linear mixed model (LMM), we have a result similar to (??) for the LMM:

$$\text{Cov} \left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} \middle| \mathbf{u} \right) \simeq (\mathbf{C}^T \mathbf{W} \mathbf{C} + \mathbf{B})^{-1} \mathbf{C}^T \mathbf{W} \mathbf{C} (\mathbf{C}^T \mathbf{W} \mathbf{C} + \mathbf{B})^{-1}, \quad (1.23)$$

where

$$\mathbf{C} = [\mathbf{X} \ \mathbf{Z}], \quad \mathbf{W} = \text{diag}\{b''(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})\} = \text{Var}(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{u}) \quad (1.24)$$

and $\mathbf{B} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{\boldsymbol{\theta}}^{-1} \end{bmatrix}.$

These formulae for standard errors are useful for smoothing and semiparametric modeling.

Evaluation of (1.23) allows for standard error bars for the fitted function and its linear part. Suppose that interest focuses on estimating at a given (\mathbf{x}, \mathbf{z}) , so that the linear predictor is $\mathbf{x}^T \hat{\boldsymbol{\beta}} + \mathbf{z}^T \hat{\mathbf{u}}$. The

estimated standard error of this prediction is

$$\sqrt{\begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix}^T \text{Cov} \left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} \mid \mathbf{u} \right) \begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix}}$$

where the $\boldsymbol{\beta}$ and \mathbf{u} appearing in \mathbf{W} are replaced by their estimates. Often more interesting is a standard error estimate for the estimated mean $\mu(\mathbf{x}^T \hat{\boldsymbol{\beta}} + \mathbf{z}^T \hat{\mathbf{u}})$. This standard error estimate is

$$\sqrt{\left\{ \mu'(\mathbf{x}^T \hat{\boldsymbol{\beta}} + \mathbf{z}^T \hat{\mathbf{u}}) \right\}^2 \begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix}^T \text{Cov} \left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} \mid \mathbf{u} \right) \begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix}}.$$

Since, $\mu' = b'' = V$, there are several ways to re-express this standard error.

1.8.9 Bias adjustment

In Section ?? we argued in favor of confidence bands based on the unconditional covariance matrix

$$\text{Cov} \left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} - \mathbf{u} \end{bmatrix} \right)$$

because they adjust for the additional uncertainty in the fitted curve due to bias whereas confidence regions based on the conditional covariance

$$\text{Cov} \left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} \mid \mathbf{u} \right)$$

make no such adjustment. Such confidence bands have not yet been studied for generalized regression, but we believe this would be a useful area for further inquiry.

1.9 Deviance

In Chapter ?? we saw that sums of squared deviations of y_i from its fitted value \hat{y}_i played an important role in inference for Gaussian regression models. In a GLM, *deviance* is the analog of the residual sum of squares. The deviance of a model compares the fit for that model with the fit for the so-called *saturated* model where there is a separate parameter for each observation. More specifically, the deviance of any model is twice the difference in log-likelihoods between the saturated model and the given model.

Let \hat{y}_i be the fitted value for a given model. This means that \hat{y}_i is the expected value of y_i , given the covariates for the i case, evaluated at $\hat{\boldsymbol{\beta}}$. We will assume that $\hat{y}_i = y_i$ for the saturated model, which is true for any of the GLMs we are considering. For logistic regression the deviance is

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

while for Poisson regression the deviance is

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

where $0 \log(0) = 0$; see McCullagh and Nelder (1989) who give the deviance for other GLMs as well.

For Gaussian linear models, the deviance is just the residual sum of squares. As discussed in detail in McCullagh and Nelder (1989), the analysis of variance for Gaussian linear models can be generalized to the analysis of deviance for GLMs.

1.10 Technical Details

In this section we collect some technical details that may be of use to those who want to understand the methods of this chapter algebraically. Those readers whose interests are mainly in applications and an intuitive understanding of the theory should skip this section.

1.10.1 Fitting a logistic regression

The general logistic regression model can be written as

$$\text{logit}\{P(y_i = 1 | \mathbf{x}_i)\} = \boldsymbol{\beta}^T \mathbf{x}_i, \quad i = 1, \dots, n \quad (1.25)$$

The log-likelihood for this problem is

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= \sum_{i=1}^n \{y_i(\boldsymbol{\beta}^T \mathbf{x}_i) - \log(1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i})\} \\ &= \mathbf{y}^T \mathbf{X} \boldsymbol{\beta} - \mathbf{1}^T \log(\mathbf{1} + e^{\mathbf{X} \boldsymbol{\beta}}). \end{aligned} \quad (1.26)$$

Differentiation with respect to $\boldsymbol{\beta}$ leads to the *score equations*

$$\mathbf{S}(\boldsymbol{\beta}) \equiv \mathbf{X}^T \left(\mathbf{y} - \frac{e^{\mathbf{X} \boldsymbol{\beta}}}{1 + e^{\mathbf{X} \boldsymbol{\beta}}} \right) = \mathbf{0}.$$

The prototype of tools for solving a vector equation of the form

$$\mathbf{S}(\boldsymbol{\beta}) = \mathbf{0} \quad (1.27)$$

is the *Newton-Raphson* technique. It involves the updating step

$$\hat{\boldsymbol{\beta}} \leftarrow \hat{\boldsymbol{\beta}} - \{\mathbf{DS}(\hat{\boldsymbol{\beta}})\}^{-1} \mathbf{S}(\hat{\boldsymbol{\beta}}) \quad (1.28)$$

where $\mathbf{DS}(\boldsymbol{\beta})$ is called the Hessian, and is the square matrix with (i, j) entry equal to

$$\frac{\partial}{\partial \beta_j} \mathbf{S}(\boldsymbol{\beta})_i.$$

Provided $\mathbf{S}(\hat{\boldsymbol{\beta}})$ is well-behaved, the iteration defined by (1.28) leads to rapid convergence to the solution of (1.27)

Fisher's method of scoring uses the same basic algorithm as in (1.28), except that if the Hessian depends on \mathbf{y} , it is replaced by its expected value. We call this the *scoring Hessian*.

For logistic regression, the Hessian and the scoring Hessian are the same and equal

$$\begin{aligned} \mathbf{DS}(\boldsymbol{\beta}) &= \mathbf{D} \mathbf{X}^T \left(\mathbf{y} - \frac{e^{\mathbf{X}\boldsymbol{\beta}}}{1 + e^{\mathbf{X}\boldsymbol{\beta}}} \right) \\ &= -\mathbf{X}^T \text{diag} \left\{ \frac{e^{\mathbf{X}\boldsymbol{\beta}}}{(1 + e^{\mathbf{X}\boldsymbol{\beta}})^2} \right\} \mathbf{X} = \mathbf{X}^T \mathbf{W}_{\boldsymbol{\beta}} \mathbf{X}, \end{aligned}$$

so the updating step is as given in (1.11).

Convergence is usually very rapid, with 5–10 iterations being sufficient in most circumstances.

In the logistic regression model, and more generally for canonical exponential models, the Newton-Raphson algorithm is identical to Fisher's method of scoring and to iteratively reweighted least-squares (see Section 1.5). For other GLMs, the Newton-Raphson differs from the other two algorithms, and is generally not used.

1.10.2 Standard error estimation in logistic regression

There are two ways to make inference about the regression parameter $\boldsymbol{\beta}$: likelihood ratio tests and confidence intervals and using standard t-test/interval methods after having obtained standard error estimates. In this section, we discuss how to obtain standard error estimates for logistic regression.

The maximum likelihood estimate of $\boldsymbol{\beta}$ satisfies

$$S(\hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

Thus we can make the informal Taylor's Theorem argument

$$\begin{aligned} \mathbf{0} &= S(\hat{\boldsymbol{\beta}}) \\ &= S(\boldsymbol{\beta} + \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\simeq S(\boldsymbol{\beta}) + \text{D} S(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \end{aligned}$$

Rearranging we get

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \simeq -\{\text{D}S(\boldsymbol{\beta})\}^{-1} \mathbf{X}^T \left(\mathbf{y} - \frac{e^{\mathbf{X}\boldsymbol{\beta}}}{1 + e^{\mathbf{X}\boldsymbol{\beta}}} \right).$$

This means that

$$\text{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \simeq \mathbf{0} \quad \text{and} \quad \text{Cov}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \simeq (\mathbf{X}^T \mathbf{W}_{\boldsymbol{\beta}} \mathbf{X})^{-1}.$$

From this it follows that

$$\widehat{\text{st.dev}}(\hat{\boldsymbol{\beta}})_i = \sqrt{i\text{th diagonal entry of } (\mathbf{X}^T \mathbf{W}_{\hat{\boldsymbol{\beta}}} \mathbf{X})^{-1}}$$

1.10.3 The hat matrix and degrees of freedom

For linear regression, we defined the hat matrix $\hat{\mathbf{y}}$ in Section ??, noting that multiplying it by the response \mathbf{y} led to the predicted values. There is an analogue of the hat matrix for GLMs, one that reflects both leverage and degrees of freedom.

In a generalized linear model, $\hat{\mathbf{y}}$ is a nonlinear function of \mathbf{y} ; there is no matrix \mathbf{H} such that $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, and hence the usual definition of a hat matrix does not apply. However, we can define a hat matrix by a linearization using the following analogy with a linear model. For a linear model

$$\hat{\mathbf{y}} - \text{E}(\mathbf{y}) = \mathbf{H}\{\mathbf{y} - \text{E}(\mathbf{y})\} = \mathbf{H}\boldsymbol{\varepsilon};$$

the left hand side of this equation is the error in estimating $\text{E}(\mathbf{y})$ by $\hat{\mathbf{y}}$. We will define the hat matrix $\mathbf{H}_{\boldsymbol{\beta}}$ to be the matrix such that, with μ being the inverse link function,

$$\mu(\mathbf{X}\hat{\boldsymbol{\beta}}) - \mu(\mathbf{X}\boldsymbol{\beta}) \simeq \mathbf{H}_{\boldsymbol{\beta}}\{\mathbf{y} - \text{E}(\mathbf{y})\}.$$

It is a general fact that

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \simeq (\mathbf{X}^T \mathbf{W}_{\boldsymbol{\beta}} \mathbf{X})^{-1} \mathbf{X}^T \{\mathbf{y} - \text{E}(\mathbf{y})\}.$$

Then by a Taylor approximation and using $\mathbf{W}_\beta = \text{diag}\{H'(\mathbf{X}\beta)\}$, we have

$$\mu(\mathbf{X}\hat{\beta}) - \mu(\mathbf{X}\beta) \approx \mathbf{W}_\beta \mathbf{X}(\mathbf{X}^T \mathbf{W}_\beta \mathbf{X})^{-1} \mathbf{X}^T \{\mathbf{y} - \mathbb{E}(\mathbf{y})\}.$$

Thus, an appropriate definition for the hat matrix is

$$\mathbf{H}_\beta \equiv \mathbf{W}_\beta \mathbf{X}(\mathbf{X}^T \mathbf{W}_\beta \mathbf{X})^{-1} \mathbf{X}^T.$$

Notice that

$$\text{tr}(\mathbf{H}_\beta) = \text{tr}\{(\mathbf{X}^T \mathbf{W}_\beta \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_\beta \mathbf{X}\} = \text{tr}(\mathbf{I}_p) = p$$

where

$$p = \text{number of parameters in model.}$$

This argument shows that

$$df_{\text{fit}} \equiv \text{tr}(\mathbf{H}_{\hat{\beta}})$$

is a reasonable definition for degrees of freedom in generalized parametric regression models. Later we will see that the same definition can be used to quantify the effective number of parameters in generalized semiparametric models.

1.10.4 Derivation of PQL

Here we show how PQL can be derived as an approximation to the maximum likelihood solution. We hold $\boldsymbol{\theta}$ fixed, and as before consider only the canonical model with $\phi = 1$ known.

Write

$$J(\boldsymbol{\beta}, \boldsymbol{\theta}) = \int_{\mathbb{R}^q} \exp\{h(\mathbf{u})\} d\mathbf{u} \quad (1.29)$$

where

$$h(\mathbf{u}) = \mathbf{y}^T (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \frac{1}{2} \mathbf{u}^T \mathbf{G}_\theta^{-1} \mathbf{u}. \quad (1.30)$$

Recall that \mathbf{D} refers to first order differentiation, while \mathbf{H} refers to second order differentiation. Laplace approximation of $J(\boldsymbol{\beta}, \boldsymbol{\theta})$ starts with the Taylor series approximation

$$h(\mathbf{u}) \simeq h(\mathbf{u}^0) + \mathbf{D} h(\mathbf{u}^0)(\mathbf{u} - \mathbf{u}^0) + \frac{1}{2}(\mathbf{u} - \mathbf{u}^0)^T \mathbf{H} h(\mathbf{u}^0)(\mathbf{u} - \mathbf{u}^0).$$

Choose \mathbf{u}^0 to solve

$$\mathbf{D} h(\mathbf{u}^0) = \mathbf{0}.$$

This leads to the approximation

$$h(\mathbf{u}) \simeq h(\mathbf{u}^0) + \frac{1}{2}(\mathbf{u} - \mathbf{u}^0)^T \mathbf{H} h(\mathbf{u}^0)(\mathbf{u} - \mathbf{u}^0).$$

Using the expression for the density of a $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random vector we have

$$\int_{\mathbb{R}^q} (2\pi)^{-q/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\} d\mathbf{x} = 1.$$

Combination of this result with (1.29) and (1.30) results in the approximation

$$J(\boldsymbol{\beta}, \boldsymbol{\theta}) \simeq (2\pi)^{q/2} |\mathbf{H} h(\mathbf{u}^0)|^{-1/2} \exp\{h(\mathbf{u}^0)\}. \quad (1.31)$$

Vector differential calculus (Appendix ??) shows that

$$\begin{aligned} \mathbf{D} h(\mathbf{u}) &= \{\mathbf{y} - b'(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})\}^T \mathbf{Z} - \mathbf{u}^T \mathbf{G}_{\boldsymbol{\theta}}^{-1} \\ \text{and } \mathbf{H} h(\mathbf{u}) &= -\mathbf{Z}^T \text{diag}\{b''(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})\} \mathbf{Z} - \mathbf{G}_{\boldsymbol{\theta}}^{-1}. \end{aligned}$$

The resulting loglikelihood approximation is then

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}) &\simeq \mathbf{y}^T (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}^0) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}^0) + \mathbf{1}^T c(\mathbf{y}) - \frac{1}{2} \mathbf{u}^{0T} \mathbf{G}_{\boldsymbol{\theta}}^{-1} \mathbf{u}^0 \\ &\quad - \frac{1}{2} \log |\mathbf{I} + \mathbf{G}_{\boldsymbol{\theta}} \mathbf{Z}^T \text{diag}\{b''(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}^0)\} \mathbf{Z}|. \end{aligned}$$

However, for ease of fitting, PQL uses one final approximation, based on the assumption that $b''(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}^0)$ is relatively constant as a function of $\boldsymbol{\beta}$. For the purpose of maximizing $\ell(\boldsymbol{\beta}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\beta}$, this gives some justification for its omission from the log-likelihood to yield

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) \simeq \mathbf{y}^T (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}^0) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}^0) + \mathbf{1}^T c(\mathbf{y}) - \frac{1}{2} \mathbf{u}^{0T} \mathbf{G}_{\boldsymbol{\theta}}^{-1} \mathbf{u}^0.$$

Maximizing this leads to the solutions described in Section 1.8.2.

1.11 Bibliographic Notes

Chapter 9 of McCullagh and Nelder (1989) provides a thorough overview of quasilielihood estimation. McCulloch and Searle (2000) is an excellent introduction to generalized linear and mixed linear models. Other good sources of information on generalized linear models include Aitkin *et al.* (1989), Green and Silverman (1994), Lindsey (1997), Fahrmeir and Tutz (2001), Gill (2001), Hardin and Hilbe (2001), Myers *et al.* (2001), and Dobson (2002).