CHAPTER 4

Generalized Linear Models: Model Fitting and Inference

We now extend our scope from the linear model to the *generalized linear model* (GLM). This extension encompasses (1) non-normal response distributions and (2) link functions of the mean equated to the linear predictor. Section 1.1.5 introduced examples of GLMs: *Loglinear models* using the log-link function for a Poisson (count) response and *logistic models* using the logit-link function for a binomial (binary) response.

Section 4.1 provides more details about exponential family distributions for the random component of a GLM. In Section 4.2 we derive likelihood equations for the maximum likelihood (ML) estimators of model parameters and show their large-sample normal distribution. Section 4.3 summarizes the likelihood ratio, score, and Wald inference methods for the model parameters. Then in Section 4.4 we introduce the *deviance*, a generalization of the residual sum of squares used in inference, such as to compare nested GLMs. That section also presents residuals for GLMs and ways of checking the model. Section 4.5 presents two standard methods, *Newton–Raphson* and *Fisher scoring*, for solving the likelihood equations to fit GLMs. Section 4.6 discusses the selection of explanatory variables for a model, followed by an example. A chapter appendix shows that fundamental results for linear models about orthogonality of fitted values and residuals do not hold exactly for GLMs, but analogs hold for an adjusted, weighted version of the response variable that satisfies a linear model with approximately constant variance.

4.1 EXPONENTIAL DISPERSION FAMILY DISTRIBUTIONS FOR A GLM

In Section 1.1 we introduced the three components of a GLM: (1) random component, (2) linear predictor, (3) link function. We now take a closer look at the random

Foundations of Linear and Generalized Linear Models, First Edition. Alan Agresti.

component, showing an exponential family form that encompasses standard distributions such as the normal, Poisson, and binomial and that has general expressions for moments and for likelihood equations.

4.1.1 Exponential Dispersion Family for a Random Component

The *random component* of a GLM consists of a response variable y with independent observations $(y_1, \ldots, y_n)$ from a distribution having probability density or mass function for y_i of the form

$$f(y_i; \theta_i, \phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\}. \tag{4.1}$$

This is called the *exponential dispersion family*. The parameter θ_i is called the *natural parameter*, and ϕ is called the *dispersion parameter*. Often $a(\phi) = 1$ and $c(y_i, \phi) = c(y_i)$, giving the *natural exponential family* of the form $f(y_i; \theta_i) = h(y_i)\exp[y_i\theta_i - b(\theta_i)]$. Otherwise, usually $a(\phi)$ has the form $a(\phi) = \phi$ or $a(\phi) = \phi/\omega_i$ for $\phi > 0$ and a known weight ω_i. For instance, when y_i is a mean of n_i independent readings, $\omega_i = n_i$. Various choices for the functions $b(\cdot)$ and $a(\cdot)$ give rise to different distributions.

Expressions for $E(y_i)$ and $\text{var}(y_i)$ use quantities in (4.1). Let $L_i = \log f(y_i; \theta_i, \phi)$ denote the contribution of y_i to the log-likelihood function, $L = \sum_i L_i$. Since

$$L_i = [y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi), \tag{4.2}$$

$$\partial L_i/\partial\theta_i = \left[y_i - b'(\theta_i)\right]/a(\phi), \quad \partial^2 L_i/\partial\theta_i^2 = -b''(\theta_i)/a(\phi),$$

where $b'(\theta_i)$ and $b''(\theta_i)$ denote the first two derivatives of $b(\cdot)$ evaluated at θ_i. We now apply the general likelihood results

$$E\left(\frac{\partial L}{\partial\theta}\right) = 0 \quad \text{and} \quad -E\left(\frac{\partial^2 L}{\partial\theta^2}\right) = E\left(\frac{\partial L}{\partial\theta}\right)^2,$$

which hold under regularity conditions satisfied by the exponential dispersion family. From the first formula applied with a single observation,

$$E\left[y_i - b'(\theta_i)\right]/a(\phi) = 0, \quad \text{so that} \quad \mu_i = E(y_i) = b'(\theta_i). \tag{4.3}$$

From the second formula,

$$b''(\theta_i)/a(\phi) = E\left[(y_i - b'(\theta_i))/a(\phi)\right]^2 = \text{var}(y_i)/[a(\phi)]^2,$$

so that

$$\text{var}(y_i) = b''(\theta_i)a(\phi). \tag{4.4}$$

In summary, the function $b(\cdot)$ in (4.1) determines moments of y_i. This function is called the *cumulant function*, because when $a(\phi) = 1$ its derivatives yield the cumulants[1] of the distribution.

4.1.2 Poisson, Binomial, and Normal in Exponential Dispersion Family

We illustrate the exponential dispersion family by showing its representations for Poisson, binomial, and normal distributions. We then evaluate the mean and variance expressions for these cases.

When y_i has a Poisson distribution, the probability mass function is

$$\begin{aligned} f(y_i;\mu_i) &= \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \exp[y_i \log \mu_i - \mu_i - \log(y_i!)] \\ &= \exp[y_i\theta_i - \exp(\theta_i) - \log(y_i!)], \quad y_i = 0, 1, 2, \ldots, \end{aligned} \tag{4.5}$$

where the natural parameter $\theta_i = \log \mu_i$. This has exponential dispersion form (4.1) with $b(\theta_i) = \exp(\theta_i)$, $a(\phi) = 1$, and $c(y_i, \phi) = -\log(y_i!)$. By (4.3) and (4.4),

$$\begin{aligned} E(y_i) &= b'(\theta_i) = \exp(\theta_i) = \mu_i, \\ \text{var}(y_i) &= b''(\theta_i) = \exp(\theta_i) = \mu_i. \end{aligned}$$

Next, suppose that $n_i y_i$ has a bin(n_i, π_i) distribution; that is, here y_i is the sample *proportion* (rather than *number*) of successes, so $E(y_i) = \pi_i$ does not depend on n_i. Let $\theta_i = \log[\pi_i/(1-\pi_i)]$. Then $\pi_i = \exp(\theta_i)/[1+\exp(\theta_i)]$ and $\log(1-\pi_i) = -\log[1+\exp(\theta_i)]$. We can express

$$\begin{aligned} f(y_i;\pi_i,n_i) &= \binom{n_i}{n_i y_i}\pi_i^{n_i y_i}(1-\pi_i)^{n_i - n_i y_i}, \quad y_i = 0, \frac{1}{n_i}, \frac{2}{n_i}, \ldots, 1, \\ &= \exp\left[\frac{y_i\theta_i - \log[1+\exp(\theta_i)]}{1/n_i} + \log\binom{n_i}{n_i y_i}\right]. \end{aligned} \tag{4.6}$$

This has exponential dispersion form (4.1) with $b(\theta_i) = \log[1+\exp(\theta_i)]$, $a(\phi) = 1/n_i$, and $c(y_i, \phi) = \log\binom{n_i}{n_i y_i}$. The natural parameter is $\theta_i = \log[\pi_i/(1-\pi_i)]$, the *logit*. By (4.3) and (4.4),

$$\begin{aligned} E(y_i) &= b'(\theta_i) = \exp(\theta_i)/[1+\exp(\theta_i)] = \pi_i, \\ \text{var}(y_i) &= b''(\theta_i)a(\phi) = \exp(\theta_i)/\{[1+\exp(\theta_i)]^2 n_i\} = \pi_i(1-\pi_i)/n_i. \end{aligned}$$

[1]Recall that cumulants $\{\kappa_n\}$ are coefficients in a power series expansion of the log mgf, $\log[E(e^{ty})] = \sum_{n=1}^{\infty} \kappa_n t^n/n!$. The moments determine the cumulants, and vice versa.

For the normal distribution, observation i has probability density function

$$f(y_i;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right]$$
$$= \exp\left[\frac{y_i\mu_i - \frac{1}{2}\mu_i^2}{\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{y_i^2}{2\sigma^2}\right].$$

This satisfies the exponential dispersion family (4.1) with natural parameter $\theta_i = \mu_i$ and

$$b(\theta_i) = \frac{1}{2}\mu_i^2 = \frac{1}{2}\theta_i^2, \quad a(\phi) = \sigma^2, \quad c(y_i;\phi) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{y_i^2}{2\sigma^2}.$$

Then

$$E(y_i) = b'(\theta_i) = \theta_i = \mu_i \quad \text{and} \quad \text{var}\left(y_i\right) = b''(\theta_i)a(\phi) = \sigma^2.$$

4.1.3 The Canonical Link Function of a Generalized Linear Model

The *link function* of a GLM connects the random component and the linear predictor. That is, a GLM states that a linear predictor $\eta_i = \sum_{j=1}^{p}\beta_j x_{ij}$ relates to μ_i by $\eta_i = g(\mu_i)$, for a link function g. Equivalently, the *response function* g^{-1} maps linear predictor values to the mean.

The link function g that transforms the mean μ_i to the natural parameter θ_i in (4.1) is called the *canonical link*. For it, the direct relationship

$$\theta_i = \sum_{j=1}^{p}\beta_j x_{ij}$$

equates the natural parameter to the linear predictor. From the exponential dispersion family expressions just derived, the canonical link functions are the log link for the Poisson distribution, the logit link for the binomial distribution, and the identity link for the normal distribution. Section 4.5.5 shows special results that apply for GLMs that use the canonical link function.

4.2 LIKELIHOOD AND ASYMPTOTIC DISTRIBUTIONS FOR GLMS

We next obtain general expressions for likelihood equations and asymptotic distributions of ML parameter estimators for GLMs. For n independent observations, from (4.2) the log likelihood is

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} L_i = \sum_{i=1}^{n}\log f(y_i;\theta_i,\phi) = \sum_{i=1}^{n}\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i,\phi). \qquad (4.7)$$

The notation $L(\boldsymbol{\beta})$ reflects the dependence of $\boldsymbol{\theta}$ on the model parameters $\boldsymbol{\beta}$. For the canonical link function, $\theta_i = \sum_j \beta_j x_{ij}$, so when $a(\phi)$ is a fixed constant, the part of the log likelihood involving both the data and the model parameters is

$$\sum_{i=1}^{n} y_i \left(\sum_{j=1}^{p} \beta_j x_{ij} \right) = \sum_{j=1}^{p} \beta_j \left(\sum_{i=1}^{n} y_i x_{ij} \right).$$

Then the sufficient statistics for $\{\beta_j\}$ are $\{\sum_{i=1}^{n} y_i x_{ij},\ j = 1, \ldots, p\}$.

4.2.1 Likelihood Equations for a GLM

For a GLM $\eta_i = \sum_j \beta_j x_{ij} = g(\mu_i)$ with link function g, the likelihood equations are

$$\partial L(\boldsymbol{\beta})/\partial \beta_j = \sum_{i=1}^{n} \partial L_i/\partial \beta_j = 0, \quad \text{for all } j.$$

To differentiate the log likelihood (4.7), we use the chain rule,

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}. \tag{4.8}$$

Since $\partial L_i/\partial \theta_i = [y_i - b'(\theta_i)]/a(\phi)$, and since $\mu_i = b'(\theta_i)$ and $\text{var}(y_i) = b''(\theta_i)a(\phi)$ from (4.3) and (4.4),

$$\partial L_i/\partial \theta_i = (y_i - \mu_i)/a(\phi), \quad \partial \mu_i/\partial \theta_i = b''(\theta_i) = \text{var}(y_i)/a(\phi).$$

Also, since $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$, $\partial \eta_i/\partial \beta_j = x_{ij}$. Finally, since $\eta_i = g(\mu_i)$, $\partial \mu_i/\partial \eta_i$ depends on the link function for the model. In summary, substituting into (4.8) gives us

$$\begin{aligned} \frac{\partial L_i}{\partial \beta_j} &= \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\ &= \frac{(y_i - \mu_i)}{a(\phi)} \frac{a(\phi)}{\text{var}(y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = \frac{(y_i - \mu_i) x_{ij}}{\text{var}(y_i)} \frac{\partial \mu_i}{\partial \eta_i}. \end{aligned} \tag{4.9}$$

Summing over the n observations yields the likelihood equations.

Likelihood equations for a GLM:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{\text{var}(y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, 2, \ldots, p, \tag{4.10}$$

where $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij} = g(\mu_i)$ for link function g.

Let $\boldsymbol{V}$ denote the diagonal matrix of variances of the observations, and let $\boldsymbol{D}$ denote the diagonal matrix with elements $\partial\mu_i/\partial\eta_i$. For the GLM expression $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ with a model matrix $\boldsymbol{X}$, these likelihood equations have the form

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0}. \tag{4.11}$$

Although $\boldsymbol{\beta}$ does not appear in these equations, it is there implicitly through $\boldsymbol{\mu}$, since $\mu_i = g^{-1}\big(\sum_{j=1}^{p}\beta_j x_{ij}\big)$. Different link functions yield different sets of equations. The likelihood equations are nonlinear functions of $\boldsymbol{\beta}$ that must be solved iteratively. We defer details to Section 4.5.

4.2.2 Likelihood Equations for Poisson Loglinear Model

For count data, one possible GLM assumes a Poisson random component and uses the log-link function. The *Poisson loglinear model* is $\log(\mu_i) = \sum_{j=1}^{p}\beta_j x_{ij}$. For the log link, $\eta_i = \log\mu_i$, so $\mu_i = \exp(\eta_i)$ and $\partial\mu_i/\partial\eta_i = \exp(\eta_i) = \mu_i$. Since $\mathrm{var}(y_i) = \mu_i$, the likelihood equations (4.10) simplify to

$$\sum_{i=1}^{n}(y_i - \mu_i)x_{ij} = 0, \quad j = 1, 2, \ldots, p. \tag{4.12}$$

These equate the sufficient statistics $\{\sum_i y_i x_{ij}\}$ for $\boldsymbol{\beta}$ to their expected values. Section 4.5.5 shows that these equations occur for GLMs that use the canonical link function.

4.2.3 The Key Role of the Mean–Variance Relation

Interestingly, the likelihood equations (4.10) depend on the distribution of y_i only through μ_i and $\mathrm{var}(y_i)$. The variance itself depends on the mean through a functional form[2]

$$\mathrm{var}(y_i) = v(\mu_i),$$

for some function v. For example, $v(\mu_i) = \mu_i$ for the Poisson, $v(\mu_i) = \mu_i(1 - \mu_i)/n_i$ for the binomial proportion, and $v(\mu_i) = \sigma^2$ (i.e., constant) for the normal.

When the distribution of y_i is in the exponential dispersion family, the relation between the mean and the variance characterizes[3] the distribution. For instance, if y_i has distribution in the exponential dispersion family and if $v(\mu_i) = \mu_i$, then necessarily y_i has the Poisson distribution.

4.2.4 Large-Sample Normal Distribution of Model Parameter Estimators

From a fundamental property of maximum likelihood, under standard regularity conditions[4], for large n the ML estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ for a GLM is efficient and has an

[2]We express the variance of y as $v(\mu)$ to emphasize that it is a function of the mean.

[3]See Jørgensen (1987), Tweedie (1947), and Wedderburn (1974).

[4]See Cox and Hinkley (1974, p. 281). Mainly, $\boldsymbol{\beta}$ falls in the interior of the parameter space and p is fixed as n increases.

approximate normal distribution. We next use the log-likelihood function for a GLM to find the covariance matrix of that distribution. The covariance matrix is the inverse of the information matrix $\boldsymbol{\mathcal{J}}$, which has elements $E[-\partial^2 L(\boldsymbol{\beta})/\partial\beta_h\,\partial\beta_j]$. The estimator $\hat{\boldsymbol{\beta}}$ is more precise when the log-likelihood function has greater curvature at $\boldsymbol{\beta}$. To find the covariance matrix, for the contribution L_i to the log likelihood we use the helpful result

$$E\left(\frac{-\partial^2 L_i}{\partial\beta_h\,\partial\beta_j}\right) = E\left[\left(\frac{\partial L_i}{\partial\beta_h}\right)\left(\frac{\partial L_i}{\partial\beta_j}\right)\right],$$

which holds for distributions in the exponential dispersion family. Thus, using (4.9),

$$E\left(\frac{-\partial^2 L_i}{\partial\beta_h\,\partial\beta_j}\right) = E\left[\frac{(y_i-\mu_i)x_{ih}}{\text{var}(y_i)}\frac{\partial\mu_i}{\partial\eta_i}\frac{(y_i-\mu_i)x_{ij}}{\text{var}(y_i)}\frac{\partial\mu_i}{\partial\eta_i}\right]$$
$$= \frac{x_{ih}x_{ij}}{\text{var}(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2.$$

Since $L(\boldsymbol{\beta}) = \sum_{i=1}^n L_i$,

$$E\left(-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_h\,\partial\beta_j}\right) = \sum_{i=1}^n \frac{x_{ih}x_{ij}}{\text{var}(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2.$$

Let $\boldsymbol{W}$ be the diagonal matrix with main-diagonal elements

$$w_i = \frac{(\partial\mu_i/\partial\eta_i)^2}{\text{var}(y_i)}.$$

Then, generalizing from the typical element of the information matrix to the entire matrix, with the model matrix $\boldsymbol{X}$,

$$\boldsymbol{\mathcal{J}} = \boldsymbol{X}^{\text{T}}\boldsymbol{W}\boldsymbol{X}. \tag{4.13}$$

The form of $\boldsymbol{W}$, and hence $\boldsymbol{\mathcal{J}}$, depends on the link function g, since $\partial\eta_i/\partial\mu_i = g'(\mu_i)$. In summary,

Asymptotic distribution of $\hat{\boldsymbol{\beta}}$ for GLM $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} \text{ has an approximate } N[\boldsymbol{\beta}, (\boldsymbol{X}^{\text{T}}\boldsymbol{W}\boldsymbol{X})^{-1}] \text{ distribution,} \tag{4.14}$$

where $\boldsymbol{W}$ is the diagonal matrix with elements $w_i = (\partial\mu_i/\partial\eta_i)^2/\text{var}(y_i)$.

The asymptotic covariance matrix is estimated by $\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^{\text{T}}\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}$, where $\hat{\boldsymbol{W}}$ is $\boldsymbol{W}$ evaluated at $\hat{\boldsymbol{\beta}}$.

For example, the Poisson loglinear model has the GLM form

$$\log \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}.$$

For this case, $\eta_i = \log(\mu_i)$, so $\partial\eta_i/\partial\mu_i = 1/\mu_i$. Thus,$w_i = (\partial\mu_i/\partial\eta_i)^2/\text{var}(y_i) = \mu_i$, and in the asymptotic covariance matrix (4.14) of $\hat{\boldsymbol{\beta}}$, $\boldsymbol{W}$ is the diagonal matrix with the elements of $\boldsymbol{\mu}$ on the main diagonal.

For some GLMs, the parameter vector partitions into the parameters $\boldsymbol{\beta}$ for the linear predictor and other parameters $\boldsymbol{\phi}$ (such as a dispersion parameter) needed to specify the model completely. Sometimes[5], $E(\partial^2 L/\partial\beta_j\partial\phi_k) = 0$ for each j and k. Similarly, the inverse of the expected information matrix has 0 elements connecting each β_j with each ϕ_k. Because this inverse is the asymptotic covariance matrix, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\phi}}$ are then asymptotically independent. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are said to be *orthogonal*. This is the generalization to GLMs of the notion of orthogonal parameters for linear models (Cox and Reid 1987). For the exponential dispersion family (4.1), θ and ϕ are orthogonal parameters.

4.2.5 Delta Method Yields Covariance Matrix for Fitted Values

The estimated linear predictor relates to $\hat{\boldsymbol{\beta}}$ by $\hat{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Thus, for large samples, its covariance matrix

$$\text{var}(\hat{\boldsymbol{\eta}}) = \boldsymbol{X}\text{var}(\hat{\boldsymbol{\beta}})\boldsymbol{X}^{\text{T}} \approx \boldsymbol{X}(\boldsymbol{X}^{\text{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\text{T}}.$$

We can obtain the asymptotic $\text{var}(\hat{\boldsymbol{\mu}})$ from $\text{var}(\hat{\boldsymbol{\eta}})$ by the *delta method*, which gives approximate variances using linearizations from a Taylor-series expansion. For example, in the univariate case with a smooth function h, the linearization $h(y) - h(\mu) \approx (y - \mu)h'(\mu)$, which holds for y near μ, implies that $\text{var}[h(y)] \approx [h'(\mu)]^2\text{var}(y)$ when $\text{var}(y)$ is small. For a vector $\boldsymbol{y}$ with covariance matrix $\boldsymbol{V}$ and a vector $\boldsymbol{h}(\boldsymbol{y}) = (h_1(\boldsymbol{y}), \ldots, h_n(\boldsymbol{y}))^{\text{T}}$, let $(\partial\boldsymbol{h}/\partial\boldsymbol{\mu})$ denote the Jacobian matrix with entry in row i and column j equal to $\partial h_i(\boldsymbol{y})/\partial y_j$ evaluated at $\boldsymbol{y} = \boldsymbol{\mu}$. Then the delta method yields $\text{var}[\boldsymbol{h}(\boldsymbol{y})] \approx (\partial\boldsymbol{h}/\partial\boldsymbol{\mu})\boldsymbol{V}(\partial\boldsymbol{h}/\partial\boldsymbol{\mu})^{\text{T}}$. So, by the delta method, using the diagonal matrix $\boldsymbol{D}$ with elements $\partial\mu_i/\partial\eta_i$, for large samples the covariance matrix of the fitted values

$$\text{var}(\hat{\boldsymbol{\mu}}) \approx \boldsymbol{D}\text{var}(\hat{\boldsymbol{\eta}})\boldsymbol{D} \approx \boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}^{\text{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\text{T}}\boldsymbol{D}.$$

However, to obtain a confidence interval for μ_i when g is not the identity link, it is preferable to construct one for η_i and then apply the response function g^{-1} to the endpoints, thus avoiding the further delta method approximation.

[5] An example is the negative binomial GLM for counts in Section 7.3.3.

These results for $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\mu}}$ are based on those for $\hat{\boldsymbol{\beta}}$, for which the asymptotics refer to $n \to \infty$. However, $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\mu}}$ have length n. Asymptotics make more sense for them when n is fixed and each component is based on an increasing number of subunits, such that the observations themselves become approximately normal. One such example is a fixed number of binomial observations, in which the asymptotics refer to each binomial sample size $n_i \to \infty$. In another example, each observation is a Poisson cell count in a contingency table with fixed dimensions, and the asymptotics refer to each expected cell count growing. Such cases can be expressed as exponential dispersion families in which the dispersion parameter $a(\phi) = \phi/\omega_i$ has weight ω_i growing. This component-specific large-sample theory is called *small-dispersion asymptotics* (Jørgensen 1987). The covariance matrix formulas are also used in an approximate sense in the more standard asymptotic cases with large n.

4.2.6 Model Misspecification: Robustness of GLMs with Correct Mean

Like other ML estimators of a fixed-length parameter vector, $\hat{\boldsymbol{\beta}}$ is consistent (i.e., $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ as $n \to \infty$). As n increases, $\boldsymbol{X}$ has more rows, the diagonal elements of the asymptotic covariance matrix $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}$ of $\hat{\boldsymbol{\beta}}$ tend to be smaller, and $\hat{\boldsymbol{\beta}}$ tends to fall closer to $\boldsymbol{\beta}$.

But what if we have misspecified the probability distribution for $\boldsymbol{y}$? Models, such as GLMs, that assume a response distribution from an exponential family have a certain robustness property. If the model for the mean is correct, that is, if we have specified the link function and linear predictor correctly, then $\hat{\boldsymbol{\beta}}$ is still consistent[6] for $\boldsymbol{\beta}$. However, if the assumed variance function is incorrect (which is likely when the assumed distribution for $\boldsymbol{y}$ is incorrect), then so is the formula for var$(\hat{\boldsymbol{\beta}})$. Moreover, not knowing the actual distribution for $\boldsymbol{y}$, we would not know the correct expression for var$(\hat{\boldsymbol{\beta}})$. Section 8.3 discusses model misspecification issues and ways of dealing with it, including using the sample variability to help obtain a consistent estimator of the appropriate covariance matrix.

4.3 LIKELIHOOD-RATIO/WALD/SCORE METHODS OF INFERENCE FOR GLM PARAMETERS

Inference about GLMs has three standard ways to use the likelihood function. For a generic scalar model parameter β, we focus on tests[7] of H_0: $\beta = \beta_0$ against H_1: $\beta \neq \beta_0$. We then explain how to construct confidence intervals using those tests.

4.3.1 Likelihood-Ratio Tests

A general purpose significance test method uses the likelihood function through the ratio of (1) its value ℓ_0 at β_0, and (2) its maximum ℓ_1 over β values permitting H_0

[6]Gourieroux et al. (1984) proved this and showed the key role of the natural exponential family and a generalization that includes the exponential dispersion family.

[7]Here, β_0 denotes a particular null value, typically 0, not the intercept parameter.

or H_1 to be true. The ratio $\Lambda = \ell_0/\ell_1 \leq 1$, since ℓ_0 results from maximizing at a restricted β value. The *likelihood-ratio test statistic* is[8]

$$-2\log\Lambda = -2\log(\ell_0/\ell_1) = -2(L_0 - L_1),$$

where L_0 and L_1 denote the maximized log-likelihood functions. Under regularity conditions, it has a limiting null chi-squared distribution as $n \to \infty$, with $df = 1$. The P-value is the chi-squared probability above the observed test statistic value.

This test extends directly to multiple parameters. For instance, for $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$, consider H_0: $\boldsymbol{\beta}_0 = \mathbf{0}$. Then ℓ_1 is the likelihood function calculated at the $\boldsymbol{\beta}$ value for which the data would have been most likely, and ℓ_0 is the likelihood function calculated at the $\boldsymbol{\beta}_1$ value for which the data would have been most likely when $\boldsymbol{\beta}_0 = \mathbf{0}$. The chi-squared *df* equal the difference in the dimensions of the parameter spaces under $H_0 \cup H_1$ and under H_0, which is $\dim(\boldsymbol{\beta}_0)$ when the model is parameterized to achieve identifiability. The test also extends to the general linear hypothesis H_0: $\boldsymbol{\Lambda\beta} = \mathbf{0}$, since the linear constraints imply a new model that is a special case of the original one.

4.3.2 Wald Tests

Standard errors obtained from the inverse of the information matrix depend on the unknown parameter values. When we substitute the unrestricted ML estimates (i.e., not assuming the null hypothesis), we obtain an *estimated* standard error (*SE*) of $\hat{\beta}$. For H_0: $\beta = \beta_0$, the test statistic using this non-null estimated standard error,

$$z = (\hat{\beta} - \beta_0)/SE,$$

is called[9] a *Wald statistic*. It has an approximate standard normal distribution when $\beta = \beta_0$, and z^2 has an approximate chi-squared distribution with $df = 1$.

For multiple parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$, to test H_0: $\boldsymbol{\beta}_0 = \mathbf{0}$, the Wald chi-squared statistic is

$$\hat{\boldsymbol{\beta}}_0^{\mathrm{T}}[\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_0)]^{-1}\hat{\boldsymbol{\beta}}_0,$$

where $\hat{\boldsymbol{\beta}}_0$ is the unrestricted ML estimate of $\boldsymbol{\beta}_0$ and $\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_0)$ is a block of the unrestricted estimated covariance matrix of $\hat{\boldsymbol{\beta}}$.

4.3.3 Score Tests

A third inference method uses the *score statistic*. The score test, referred to in some literature as the *Lagrange multiplier test*, uses the slope (i.e., the *score function*) and

[8]The general form was proposed by Samuel S. Wilks in 1938; see Cox and Hinkley (1974, pp. 313, 314, 322, 323) for a derivation of the chi-squared limit.
[9]The general form was proposed by Abraham Wald in 1943.

expected curvature of the log-likelihood function, evaluated at the null value β_0. The chi-squared form[10] of the score statistic is

$$\frac{[\partial L(\beta)/\partial\beta_0]^2}{-E[\partial^2 L(\beta)/\partial\beta_0^2]},$$

where the notation reflects derivatives with respect to β that are evaluated at β_0. In the multiparameter case, the score statistic is a quadratic form based on the vector of partial derivatives of the log likelihood and the inverse information matrix, both evaluated at the H_0 estimates.

4.3.4 Illustrating the Likelihood-Ratio, Wald, and Score Tests

Figure 4.1 plots a generic log-likelihood function $L(\beta)$ and illustrates the three tests of H_0: $\beta = \beta_0$, at $\beta_0 = 0$. The Wald test uses $L(\beta)$ at the ML estimate $\hat{\beta}$, having chi-squared form $(\hat{\beta}/SE)^2$ with SE of $\hat{\beta}$ based on the curvature of $L(\beta)$ at $\hat{\beta}$. The score test uses the slope and curvature of $L(\beta)$ at $\beta_0 = 0$. The likelihood-ratio test combines information about $L(\beta)$ at $\hat{\beta}$ and at $\beta_0 = 0$. In Figure 4.1, this statistic is twice the vertical distance between values of $L(\beta)$ at $\beta = \hat{\beta}$ and at $\beta = 0$.

To illustrate, consider a binomial parameter π and testing H_0: $\pi = \pi_0$. With sample proportion $\hat{\pi} = y$ for n observations, you can show that the chi-squared forms of the

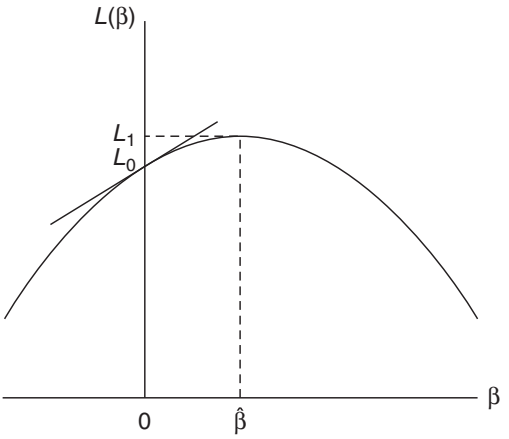

Figure 4.1 Log-likelihood function and information used in likelihood-ratio, score, and Wald tests of H_0: $\beta = 0$.

[10]The general form was proposed by C. R. Rao in 1948.

test statistics are

$$\text{Likelihood-ratio:} \quad -2(L_0 - L_1) = -2\log\left[\frac{\pi_0^{ny}(1-\pi_0)^{n(1-y)}}{y^{ny}(1-y)^{n(1-y)}}\right];$$

$$\text{Wald:} \quad z^2 = \frac{(y-\pi_0)^2}{[y(1-y)/n]};$$

$$\text{Score:} \quad z^2 = \frac{(y-\pi_0)^2}{[\pi_0(1-\pi_0)/n]}.$$

As $n \to \infty$, the three tests have certain asymptotic equivalences[11]. For the best-known GLM, the normal linear model, the three types of inference provide identical results. Unlike the other methods, though, we show in Section 5.3.3 that the results of the Wald test depend on the scale for the parameterization. Also, Wald inference is useless when an estimate or H_0 value is on the boundary of the parameter space. Examples are $\hat{\pi} = 0$ for a binomial and $\hat{\beta} = \infty$ in a GLM (not unusual in logistic regression).

4.3.5 Constructing Confidence Intervals by Inverting Tests

For any of the three test methods, we can construct a confidence interval by inverting the test. For instance, in the single-parameter case a 95% confidence interval for β is the set of β_0 for which the test of H_0: $\beta = \beta_0$ has P-value exceeding 0.05.

Let z_a denote the $(1-a)$ quantile of the standard normal distribution. A $100(1-\alpha)\%$ confidence interval based on asymptotic normality uses $z_{\alpha/2}$, for instance, $z_{0.025} = 1.96$ for 95% confidence. The Wald confidence interval is the set of β_0 for which $|\hat{\beta} - \beta_0|/SE < z_{\alpha/2}$. This gives the interval $\hat{\beta} \pm z_{\alpha/2}(SE)$. The score-test-based confidence interval often simplifies to the set of β_0 for which $|\hat{\beta} - \beta_0|/SE_0 < z_{\alpha/2}$, where SE_0 is the standard error estimated under the restriction that $\beta = \beta_0$. Let $\chi^2_d(a)$ denote the $(1-a)$ quantile of the chi-squared distribution with $df = d$. The likelihood-ratio-based confidence interval is the set of β_0 for which $-2[L(\beta_0) - L(\hat{\beta})] < \chi^2_1(\alpha)$. [Note that $\chi^2_1(\alpha) = z^2_{\alpha/2}$.]

When $\hat{\beta}$ has a normal distribution, the log-likelihood function is a second-degree polynomial and thus has a parabolic shape. For small samples of highly non-normal data or when β falls near the boundary of the parameter space, $\hat{\beta}$ may have distribution far from normality, and the log-likelihood function can be far from a symmetric, parabolic curve. A marked divergence in the results of Wald and likelihood-ratio inference indicates that the distribution of $\hat{\beta}$ may not be close to normality. It is then preferable to use the likelihood-ratio inference or higher order asymptotic methods[12].

[11] See, for example, Cox and Hinkley (1974, Section 9.3).

[12] For an introduction to higher-order asymptotics, see Brazzale et al. (2007).

4.3.6 Profile Likelihood Confidence Intervals

For confidence intervals for multiparameter models, especially useful is the *profile likelihood* approach. It is based on inverting likelihood-ratio tests for the various possible null values of β, regarding the other parameters $\boldsymbol{\psi}$ in the model as *nuisance parameters*. In inverting a likelihood-ratio test of H_0: $\beta = \beta_0$ to check whether β_0 belongs in the confidence interval, the ML estimate $\hat{\boldsymbol{\psi}}(\beta_0)$ of $\boldsymbol{\psi}$ that maximizes the likelihood under the null varies as β_0 does. The *profile log-likelihood function* is $L(\beta_0, \hat{\boldsymbol{\psi}}(\beta_0))$, viewed as a function of β_0. For each β_0 this function gives the maximum of the ordinary log-likelihood subject to the constraint $\beta = \beta_0$. Evaluated at $\beta_0 = \hat{\beta}$, this is the maximized log likelihood $L(\hat{\beta}, \hat{\boldsymbol{\psi}})$, which occurs at the unrestricted ML estimates. The *profile likelihood confidence interval* for β is the set of β_0 for which

$$-2[L(\beta_0, \hat{\boldsymbol{\psi}}(\beta_0)) - L(\hat{\beta}, \hat{\boldsymbol{\psi}})] < \chi_1^2(\alpha).$$

The interval contains all β_0 not rejected in likelihood-ratio tests of nominal size α. The profile likelihood interval is more complex to calculate than the Wald interval, but it is available in software[13].

4.4 DEVIANCE OF A GLM, MODEL COMPARISON, AND MODEL CHECKING

For a particular GLM with observations $\boldsymbol{y} = (y_1, \ldots, y_n)$, let $L(\boldsymbol{\mu}; \boldsymbol{y})$ denote the log-likelihood function expressed in terms of the means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. Let $L(\hat{\boldsymbol{\mu}}; \boldsymbol{y})$ denote the maximum of the log likelihood for the model. Considered for all possible models, the maximum achievable log likelihood is $L(\boldsymbol{y}; \boldsymbol{y})$. This occurs for the most general model, having a separate parameter for each observation and the perfect fit $\hat{\boldsymbol{\mu}} = \boldsymbol{y}$. This model is called the *saturated model*. It explains all variation by the linear predictor of the model. A perfect fit sounds good, but the saturated model is not a helpful one. It does not smooth the data or have the advantages that a simpler model has because of its parsimony, such as better estimation of the true relation. However, it often serves as a baseline for comparison with other model fits, such as for checking goodness of fit.

4.4.1 Deviance Compares Chosen Model with Saturated Model

For a chosen model, for all i denote the ML estimate of the natural parameter θ_i by $\hat{\theta}_i$, corresponding to the estimated mean $\hat{\mu}_i$. Let $\tilde{\theta}_i$ denote the estimate of θ_i for the saturated model, with corresponding $\tilde{\mu}_i = y_i$. For maximized log likelihoods $L(\hat{\boldsymbol{\mu}}; \boldsymbol{y})$ for the chosen model and $L(\boldsymbol{y}; \boldsymbol{y})$ for the saturated model,

$$-2 \log \left[\frac{\text{maximum likelihood for model}}{\text{maximum likelihood for saturated model}} \right] = -2[L(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) - L(\boldsymbol{y}; \boldsymbol{y})]$$

[13]Examples are the `confint` function and `ProfileLikelihood` and `cond` packages in R.

is the likelihood-ratio statistic for testing H_0 that the model holds against H_1 that a more general model holds. It describes lack of fit. From (4.7),

$$\begin{aligned}&-2[L(\hat{\boldsymbol{\mu}};\boldsymbol{y}) - L(\boldsymbol{y};\boldsymbol{y})]\\&= 2\sum_i [y_i\tilde{\theta}_i - b(\tilde{\theta}_i)]/a(\phi) - 2\sum_i [y_i\hat{\theta}_i - b(\hat{\theta}_i)]/a(\phi).\end{aligned}$$

Usually $a(\phi) = \phi/\omega_i$, in which case this difference equals

$$2\sum_i \omega_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]/\phi = D(\boldsymbol{y};\hat{\boldsymbol{\mu}})/\phi, \tag{4.15}$$

called the *scaled deviance*. The statistic $D(\boldsymbol{y};\hat{\boldsymbol{\mu}})$ is called the *deviance*.

Since $L(\hat{\boldsymbol{\mu}};\boldsymbol{y}) \le L(\boldsymbol{y};\boldsymbol{y})$, $D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) \ge 0$. The greater the deviance, the poorer the fit. For some GLMs, such as binomial and Poisson GLMs under small-dispersion asymptotics in which the number of observations n is fixed and the individual observations converge to normality, the scaled deviance has an approximate chi-squared distribution. The *df* equal the difference between the numbers of parameters in the saturated model and in the chosen model. When ϕ is known, we use the scaled deviance for model checking. The main use of the deviance is for inferential comparisons of models (Section 4.4.3).

4.4.2 The Deviance for Poisson GLMs and Normal GLMs

For Poisson GLMs, from Section 4.1.2, $\hat{\theta}_i = \log\ \hat{\mu}_i$ and $b(\hat{\theta}_i) = \exp(\hat{\theta}_i) = \hat{\mu}_i$. Similarly, $\tilde{\theta}_i = \log\ y_i$ and $b(\tilde{\theta}_i) = y_i$ for the saturated model. Also $a(\phi) = 1$, so the deviance and scaled deviance (4.15) equal

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\sum_i [y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i].$$

When a model with log link contains an intercept term, the likelihood equation (4.12) implied by that parameter is $\sum_i y_i = \sum_i \hat{\mu}_i$. Then the deviance simplifies to

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\sum_i y_i \log(y_i/\hat{\mu}_i). \tag{4.16}$$

For some applications with Poisson GLMs, such as modeling cell counts in contingency tables, the number n of counts is fixed. With p model parameters, as the expected counts grow the deviance converges in distribution to chi-squared with $df = n - p$. Chapter 7 shows that the deviance then provides a test of model fit.

For normal GLMs, by Section 4.1.2, $\hat{\theta}_i = \hat{\mu}_i$ and $b(\hat{\theta}_i) = \hat{\theta}_i^2/2$. Similarly, $\tilde{\theta}_i = y_i$ and $b(\tilde{\theta}_i) = y_i^2/2$ for the saturated model. So the deviance equals

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\sum_i \left[y_i(y_i - \hat{\mu}_i) - \frac{y_i^2}{2} + \frac{\hat{\mu}_i^2}{2}\right] = \sum_i (y_i - \hat{\mu}_i)^2.$$

For linear models, this is the residual sum of squares, which we have denoted by SSE. Also $\phi = \sigma^2$, so the scaled deviance is $[\sum_i (y_i - \hat{\mu}_i)^2]/\sigma^2$. When the model holds, we have seen (Section 3.2.2, by Cochran's theorem) that this has a χ^2_{n-p} distribution.

For a particular GLM, *maximizing the likelihood corresponds to minimizing the deviance*. Using least squares to minimize SSE for a linear model generalizes to using ML to minimize a deviance for a GLM.

4.4.3 Likelihood-Ratio Model Comparison Uses Deviance Difference

Methods for comparing deviances generalize methods for normal linear models that compare residual sums of squares. When $\phi = 1$, such as for a Poisson or binomial model, the deviance (4.15) equals

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = -2[L(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) - L(\boldsymbol{y}; \boldsymbol{y})].$$

Consider two nested models, M_0 with p_0 parameters and fitted values $\hat{\boldsymbol{\mu}}_0$ and M_1 with p_1 parameters and fitted values $\hat{\boldsymbol{\mu}}_1$, with M_0 a special case of M_1. Section 3.2.2 showed how to compare nested linear models. Since the parameter space for M_0 is contained in that for M_1, $L(\hat{\boldsymbol{\mu}}_0; \boldsymbol{y}) \leq L(\hat{\boldsymbol{\mu}}_1; \boldsymbol{y})$. Since $L(\boldsymbol{y}; \boldsymbol{y})$ is identical for each model,

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1) \leq D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0).$$

Simpler models have larger deviances.

Assuming that model M_1 holds, the likelihood-ratio test of the hypothesis that M_0 holds uses the test statistic

$$\begin{aligned} -2[L(\hat{\boldsymbol{\mu}}_0; \boldsymbol{y}) - L(\hat{\boldsymbol{\mu}}_1; \boldsymbol{y})] &= -2[L(\hat{\boldsymbol{\mu}}_0; \boldsymbol{y}) - L(\boldsymbol{y}; \boldsymbol{y})] - \{-2[L(\hat{\boldsymbol{\mu}}_1; \boldsymbol{y}) - L(\boldsymbol{y}; \boldsymbol{y})]\} \\ &= D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1), \end{aligned}$$

when $\phi = 1$. This statistic is large when M_0 fits poorly compared with M_1. In expression (4.15) for the deviance, since the terms involving the saturated model cancel,

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1) = 2 \sum_i \omega_i [y_i(\hat{\theta}_{1i} - \hat{\theta}_{0i}) - b(\hat{\theta}_{1i}) + b(\hat{\theta}_{0i})].$$

This also has the form of the deviance. Under standard regularity conditions for which likelihood-ratio statistics have large-sample chi-squared distributions, this difference has approximately a chi-squared null distribution with $df = p_1 - p_0$.

For example, for a Poisson loglinear model with an intercept term, from expression (4.16) for the deviance, the difference in deviances uses the observed counts and the two sets of fitted values in the form

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1) = 2 \sum_i y_i \log(\hat{\mu}_{1i}/\hat{\mu}_{0i}).$$

We denote the likelihood-ratio statistic for comparing nested models by $G^2(M_0 \mid M_1)$.

4.4.4 Score Tests and Pearson Statistics for Model Comparison

For GLMs having variance function $\text{var}(y_i) = v(\mu_i)$ with $\phi = 1$, the score statistic for comparing a chosen model with the saturated model is[14]

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}. \tag{4.17}$$

For Poisson y_i, for which $v(\hat{\mu}_i) = \hat{\mu}_i$, this has the form

$$\sum (\text{observed} - \text{fitted})^2/\text{fitted}.$$

This is known as the *Pearson chi-squared statistic*, because Karl Pearson introduced it in 1900 for testing various hypotheses using the chi-squared distribution, such as the hypothesis of independence in a two-way contingency table (Section 7.2.2). The generalized Pearson statistic (4.17) is an alternative to the deviance for testing the fit of certain GLMs.

For two nested models, a generalized Pearson statistic for comparing nested models is

$$X^2(M_0 \mid M_1) = \sum_i (\hat{\mu}_{1i} - \hat{\mu}_{0i})^2/v(\hat{\mu}_{0i}). \tag{4.18}$$

This is a quadratic approximation for $G^2(M_0 \mid M_1)$, with the same null asymptotic behavior. However, this is not the score statistic for comparing the models, which is more complex. See Note 4.4.

4.4.5 Residuals and Fitted Values Asymptotically Uncorrelated

Examining residuals helps us find where the fit of a GLM is poor or where unusual observations occur. As in ordinary linear models, we would like to exploit the decomposition

$$\boldsymbol{y} = \hat{\boldsymbol{\mu}} + (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \quad \text{(i.e., data = fit + residuals)}.$$

With GLMs, however, $\hat{\boldsymbol{\mu}}$ and $(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ are not orthogonal when we leave the simple linear model case of identity link with constant variance. Pythagoras's theorem does not apply, because maximizing the likelihood does not correspond to minimizing $\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|$. With a nonlinear link function, although the space of linear predictor values $\boldsymbol{\eta}$ that satisfy a particular GLM is a linear vector space, the corresponding set of $\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta})$ values is not. Fundamental results for ordinary linear models about projections and orthogonality of fitted values and residuals do not hold exactly for GLMs.

[14] See Lovison (2005, 2014), Pregibon (1982), and Smyth (2003).

We next obtain an asymptotic covariance matrix for the residuals. From Section 4.2.4, $\boldsymbol{W} = \text{diag}\{(\partial\mu_i/\partial\eta_i)^2/\text{var}(y_i)\}$ and $\boldsymbol{D} = \text{diag}\{\partial\mu_i/\partial\eta_i\}$, so we can express the diagonal matrix $\boldsymbol{V} = \text{var}(\boldsymbol{y})$ as $\boldsymbol{V} = \boldsymbol{D}\boldsymbol{W}^{-1}\boldsymbol{D}$. For large n, if $\hat{\boldsymbol{\mu}}$ is approximately uncorrelated with $(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$, then $\boldsymbol{V} \approx \text{var}(\hat{\boldsymbol{\mu}}) + \text{var}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$. Then, using the approximate expression for $\text{var}(\hat{\boldsymbol{\mu}})$ from Section 4.2.5 and $\boldsymbol{V}^{1/2} = \boldsymbol{D}\boldsymbol{W}^{-1/2}$,

$$\begin{aligned}\text{var}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) &\approx \boldsymbol{V} - \text{var}(\hat{\boldsymbol{\mu}}) \approx \boldsymbol{D}\boldsymbol{W}^{-1}\boldsymbol{D} - \boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D} \\ &= \boldsymbol{D}\boldsymbol{W}^{-1/2}[\boldsymbol{I} - \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{1/2}]\boldsymbol{W}^{-1/2}\boldsymbol{D}.\end{aligned}$$

This has the form $\boldsymbol{V}^{1/2}[\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{W}}]\boldsymbol{V}^{1/2}$, where $\boldsymbol{I}$ is the identity matrix and

$$\boldsymbol{H}_{\boldsymbol{W}} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{1/2}. \tag{4.19}$$

You can verify that $\boldsymbol{H}_{\boldsymbol{W}}$ is a projection matrix by showing it is symmetric and idempotent. McCullagh and Nelder (1989, p. 397) noted that it is approximately a hat matrix for standardized units of $\boldsymbol{y}$, with

$$\boldsymbol{H}_{\boldsymbol{W}}\boldsymbol{V}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu}) \approx \boldsymbol{V}^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}).$$

The chapter appendix shows that the estimate of $\boldsymbol{H}_{\boldsymbol{W}}$ is also a type of hat matrix, applying to weighted versions of the response and the linear predictor.

So why is $(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ asymptotically uncorrelated with $\hat{\boldsymbol{\mu}}$, thus generalizing the exact orthogonal decomposition for linear models? Lovison (2014) gave an argument that seems relevant for small-dispersion asymptotic cases in which "large samples" refer to the individual components, such as binomial indices. If $(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ and $\hat{\boldsymbol{\mu}}$ were not approximately uncorrelated, one could construct an asymptotically unbiased estimator of $\boldsymbol{\mu}$ that is asymptotically more efficient than $\hat{\boldsymbol{\mu}}$ using $\hat{\boldsymbol{\mu}}^* = [\hat{\boldsymbol{\mu}} + \boldsymbol{L}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})]$ for a matrix of constants $\boldsymbol{L}$. But this would contradict the ML estimator $\hat{\boldsymbol{\mu}}$ being asymptotically efficient. Such an argument is an asymptotic version for ML estimators of the one in the Gauss–Markov theorem (Section 2.7.1) that unbiased estimators other than the least squares estimator have difference from that estimator that is uncorrelated with it. The small-dispersion asymptotic setting applies for the discrete-data models we will present in the next three chapters for situations in which residuals are mainly useful, in which individual y_i have approximate normal distributions. Then $(\boldsymbol{y} - \boldsymbol{\mu})$ and $(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ jointly have an approximate normal distribution, as does their difference.

4.4.6 Pearson, Deviance, and Standardized Residuals for GLMs

For a particular model with variance function $v(\mu)$, the *Pearson residual* for observation y_i and its fitted value $\hat{\mu}_i$ is

$$\text{Pearson residual:} \quad e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}}. \tag{4.20}$$

Their squared values sum to the generalized Pearson statistic (4.17). For instance, consider a Poisson GLM. The Pearson residual is

$$e_i = (y_i - \hat{\mu}_i)/\sqrt{\hat{\mu}_i},$$

and when $\{\mu_i\}$ are large and the model holds, e_i has an approximate normal distribution and $X^2 = \sum_i e_i^2$ has an approximate chi-squared distribution (Chapter 7). For a binomial GLM in which $n_i y_i$ has a bin(n_i, π_i) distribution, the Pearson residual is

$$e_i = (y_i - \hat{\pi}_i)/\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)/n_i},$$

and when $\{n_i\}$ are large, $X^2 = \sum_i e_i^2$ also has an approximate chi-squared distribution (Chapter 5). In these cases, such statistics are used in model goodness-of-fit tests.

In expression (4.15) for the deviance, let $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \sum_i d_i$, where

$$d_i = 2\omega_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)].$$

The *deviance residual* is

$$\text{Deviance residual:} \quad \sqrt{d_i} \times \text{sign}(y_i - \hat{\mu}_i). \tag{4.21}$$

The sum of squares of these residuals equals the deviance.

To judge when a residual is "large" it is helpful to have residual values that, when the model holds, have means of 0 and variances of 1. However, Pearson and deviance residuals tend to have variance less than 1 because they compare y_i with the fitted mean $\hat{\mu}_i$ rather than the true mean μ_i. For example, the denominator of the Pearson residual estimates $[v(\mu_i)]^{1/2} = [\text{var}(y_i - \mu_i)]^{1/2}$ rather than $[\text{var}(y_i - \hat{\mu}_i)]^{1/2}$. The *standardized residual* divides each raw residual $(y_i - \hat{\mu}_i)$ by its standard error. From Section 4.4.5, $\text{var}(y_i - \hat{\mu}_i) \approx v(\mu_i)(1 - h_{ii})$, where h_{ii} is the diagonal element of the generalized hat matrix $\boldsymbol{H}_{\boldsymbol{W}}$ for observation i, its *leverage*. Let $\hat{h}_{ii}$ denote the estimate of h_{ii}. Then, standardizing by dividing $y_i - \hat{\mu}_i$ by its estimated *SE* yields

$$\text{Standardized residual:} \quad r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)(1-\hat{h}_{ii})}} = \frac{e_i}{\sqrt{1-\hat{h}_{ii}}}. \tag{4.22}$$

For Poisson GLMs, for instance, $r_i = (y_i - \hat{\mu}_i)/\sqrt{\hat{\mu}_i(1-\hat{h}_{ii})}$. Likewise, deviance residuals have standardized versions. They are most useful for small-dispersion asymptotic cases, such as for relatively large Poisson means and relatively large binomial indices. In such cases their model-based distribution is approximately standard normal.

To detect a model's lack of fit, any particular type of residual can be plotted against the component fitted values in $\hat{\boldsymbol{\mu}}$ and against each explanatory variable. As with the linear model, the fit could be quite different when we delete an observation that has a

large standardized residual and a large leverage. The estimated leverages fall between 0 and 1 and sum to p. Unlike in ordinary linear models, the generalized hat matrix depends on the fit as well as on the model matrix, and points that have extreme values for the explanatory variables need not have high estimated leverage. To gauge influence, an analog of Cook's distance (2.11) uses both the standardized residuals and the estimated leverages, by $r_i^2[\hat{h}_{ii}/p(1-\hat{h}_{ii})]$.

4.5 FITTING GENERALIZED LINEAR MODELS

How do we find the ML estimator $\hat{\boldsymbol{\beta}}$ of GLM parameters? The likelihood equations (4.10) are usually nonlinear in $\hat{\boldsymbol{\beta}}$. We next describe a general purpose iterative method for solving nonlinear equations and apply it in two ways to determine the maximum of the likelihood function.

4.5.1 Newton–Raphson Method

The *Newton–Raphson method* iteratively solves nonlinear equations, for example, to determine the point at which a function takes its maximum. It begins with an initial approximation for the solution. It obtains a second approximation by approximating the function in a neighborhood of the initial approximation by a second-degree polynomial and then finding the location of that polynomial's maximum value. It then repeats this step to generate a sequence of approximations. These converge to the location of the maximum when the function is suitable and/or the initial approximation is good.

Mathematically, here is how the Newton–Raphson method determines the value $\hat{\boldsymbol{\beta}}$ at which a function $L(\boldsymbol{\beta})$ is maximized. Let

$$\boldsymbol{u} = \left(\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_1}, \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_2}, \ldots, \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_p}\right)^{\mathrm{T}}.$$

Let $\boldsymbol{H}$ denote[15] the matrix having entries $h_{ab} = \partial^2 L(\boldsymbol{\beta})/\partial\beta_a\partial\beta_b$, called the *Hessian matrix*. Let $\boldsymbol{u}^{(t)}$ and $\boldsymbol{H}^{(t)}$ be $\boldsymbol{u}$ and $\boldsymbol{H}$ evaluated at $\boldsymbol{\beta}^{(t)}$, approximation t for $\hat{\boldsymbol{\beta}}$. Step t in the iterative process ($t = 0, 1, 2, \ldots$) approximates $L(\boldsymbol{\beta})$ near $\boldsymbol{\beta}^{(t)}$ by the terms up to the second order in its Taylor series expansion,

$$L(\boldsymbol{\beta}) \approx L(\boldsymbol{\beta}^{(t)}) + \boldsymbol{u}^{(t)T}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) + \left(\frac{1}{2}\right)(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^{\mathrm{T}}\boldsymbol{H}^{(t)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}).$$

Solving $\partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta} \approx \boldsymbol{u}^{(t)} + \boldsymbol{H}^{(t)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) = \mathbf{0}$ for $\boldsymbol{\beta}$ yields the next approximation,

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - (\boldsymbol{H}^{(t)})^{-1}\boldsymbol{u}^{(t)}, \tag{4.23}$$

assuming that $\boldsymbol{H}^{(t)}$ is nonsingular.

[15]Here, $\boldsymbol{H}$ is *not* the hat matrix; it is conventional to use $\boldsymbol{H}$ for a Hessian matrix.

Iterations proceed until changes in $L(\boldsymbol{\beta}^{(t)})$ between successive cycles are sufficiently small. The ML estimator is the limit of $\boldsymbol{\beta}^{(t)}$ as $t \to \infty$; however, this need not happen if $L(\boldsymbol{\beta})$ has other local maxima at which $\boldsymbol{u}(\boldsymbol{\beta}) = \mathbf{0}$. In that case, a good initial approximation is crucial. Figure 4.2 illustrates a cycle of the method, showing the parabolic (second-order) approximation at a given step.

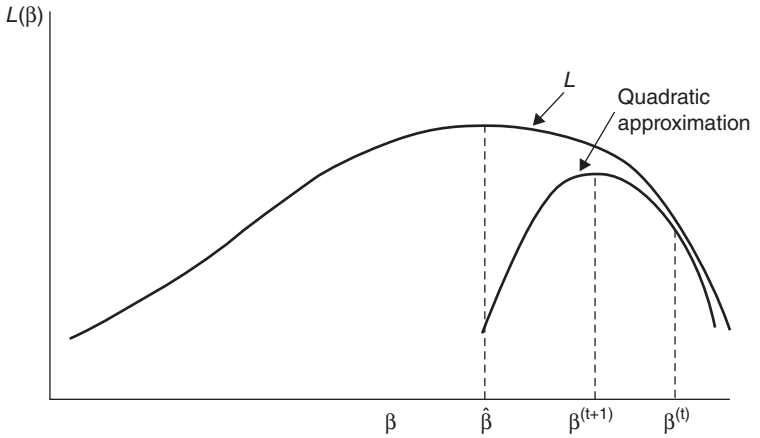

Figure 4.2 Illustration of a cycle of the Newton–Raphson method.

For many GLMs, including Poisson loglinear models and binomial logistic models, with full-rank model matrix the Hessian is negative definite, and the log likelihood is a strictly concave function. Then ML estimates of model parameters exist and are unique under quite general conditions[16]. The convergence of $\boldsymbol{\beta}^{(t)}$ to $\hat{\boldsymbol{\beta}}$ in the neighborhood of $\hat{\boldsymbol{\beta}}$ is then usually fast.

4.5.2 Fisher Scoring Method

Fisher scoring is an alternative iterative method for solving likelihood equations. The difference from Newton–Raphson is in the way it uses the Hessian matrix. Fisher scoring uses the *expected value* of this matrix, called the *expected information*, whereas Newton–Raphson uses the Hessian matrix itself, called the *observed information*.

Let $\boldsymbol{\mathcal{J}}^{(t)}$ denote approximation t for the ML estimate of the expected information matrix; that is, $\boldsymbol{\mathcal{J}}^{(t)}$ has elements $-E\,(\partial^2 L(\boldsymbol{\beta})/\partial\beta_a\,\partial\beta_b)$, evaluated at $\boldsymbol{\beta}^{(t)}$. The formula for Fisher scoring is

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + (\boldsymbol{\mathcal{J}}^{(t)})^{-1}\boldsymbol{u}^{(t)}, \quad \text{or} \quad \boldsymbol{\mathcal{J}}^{(t)}\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\mathcal{J}}^{(t)}\boldsymbol{\beta}^{(t)} + \boldsymbol{u}^{(t)}. \tag{4.24}$$

Formula (4.13) showed that $\boldsymbol{\mathcal{J}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X}$, where $\boldsymbol{W}$ is diagonal with elements $w_i = (\partial\mu_i/\partial\eta_i)^2/\mathrm{var}(y_i)$. Similarly, $\boldsymbol{\mathcal{J}}^{(t)} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(t)}\boldsymbol{X}$, where $\boldsymbol{W}^{(t)}$ is $\boldsymbol{W}$ evaluated at $\boldsymbol{\beta}^{(t)}$. The estimated asymptotic covariance matrix $\boldsymbol{\mathcal{J}}^{-1}$ of $\hat{\boldsymbol{\beta}}$ [see (4.14)] occurs as

[16]See, for example, Wedderburn (1976).

a by-product of this algorithm as $(\mathcal{J}^{(t)})^{-1}$ for t at which convergence is adequate. For GLMs with a canonical link function, Section 4.5.5 shows that the observed and expected information are the same.

A simple way to begin either iterative process takes the initial estimate of $\boldsymbol{\mu}$ to be the data $\boldsymbol{y}$, smoothed to avoid boundary values. This determines the initial estimate of the weight matrix $\boldsymbol{W}$ and hence the initial approximation for $\hat{\boldsymbol{\beta}}$.

4.5.3 Newton–Raphson and Fisher Scoring for a Binomial Parameter

In the next three chapters we use the Newton–Raphson and Fisher scoring methods for models for categorical data and count data. We illustrate them here with a simpler problem for which we know the answer, maximizing the log likelihood with a sample proportion y from a bin(n, π) distribution. The log likelihood to be maximized is then $L(\pi) = \log[\pi^{ny}(1-\pi)^{n-ny}] = ny\log\pi + (n-ny)\log(1-\pi)$.

The first two derivatives of $L(\pi)$ are

$$u = (ny - n\pi)/\pi(1-\pi), \quad H = -[ny/\pi^2 + (n-ny)/(1-\pi)^2].$$

Each Newton–Raphson step has the form

$$\pi^{(t+1)} = \pi^{(t)} + \left[\frac{ny}{(\pi^{(t)})^2} + \frac{n-ny}{(1-\pi^{(t)})^2}\right]^{-1} \frac{ny - n\pi^{(t)}}{\pi^{(t)}(1-\pi^{(t)})}.$$

This adjusts $\pi^{(t)}$ up if $y > \pi^{(t)}$ and down if $y < \pi^{(t)}$. For instance, with $\pi^{(0)} = \frac{1}{2}$, you can check that $\pi^{(1)} = y$. When $\pi^{(t)} = y$, no adjustment occurs and $\pi^{(t+1)} = y$, which is the correct answer for $\hat{\pi}$. From the expectation of H above, the information is $n/[\pi(1-\pi)]$. A step of Fisher scoring gives

$$\begin{aligned}\pi^{(t+1)} &= \pi^{(t)} + \left[\frac{n}{\pi^{(t)}(1-\pi^{(t)})}\right]^{-1} \frac{ny - n\pi^{(t)}}{\pi^{(t)}(1-\pi^{(t)})} \\ &= \pi^{(t)} + (y - \pi^{(t)}) = y.\end{aligned}$$

This gives the correct answer for $\hat{\pi}$ after a single iteration and stays at that value for successive iterations.

4.5.4 ML as Iteratively Reweighted Least Squares

A relation exists between using Fisher scoring to find ML estimates and *weighted least squares estimation*. We refer here to the general linear model

$$\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

When the covariance matrix of $\boldsymbol{\epsilon}$ is $\boldsymbol{V}$, from Section 2.7.2 the generalized least squares estimator of $\boldsymbol{\beta}$ is

$$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{z}.$$

When $\boldsymbol{V}$ is diagonal, this is referred to as a *weighted least squares* estimator.

From (4.11), the score vector for a GLM is $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$. Since $\boldsymbol{D} = \text{diag}\{\partial\mu_i/\partial\eta_i\}$ and $\boldsymbol{W} = \text{diag}\{(\partial\mu_i/\partial\eta_i)^2/\text{var}(y_i)\}$, we have $\boldsymbol{D}\boldsymbol{V}^{-1} = \boldsymbol{W}\boldsymbol{D}^{-1}$ and we can express the score function as

$$\boldsymbol{u} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{D}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}).$$

Since $\boldsymbol{\mathcal{J}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X}$, it follows that in the Fisher scoring formula (4.24),

$$\begin{aligned}\boldsymbol{\mathcal{J}}^{(t)}\boldsymbol{\beta}^{(t)} + \boldsymbol{u}^{(t)} &= (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(t)}\boldsymbol{X})\boldsymbol{\beta}^{(t)} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(t)}(\boldsymbol{D}^{(t)})^{-1}(\boldsymbol{y}-\boldsymbol{\mu}^{(t)}) \\ &= \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(t)}[\boldsymbol{X}\boldsymbol{\beta}^{(t)} + (\boldsymbol{D}^{(t)})^{-1}(\boldsymbol{y}-\boldsymbol{\mu}^{(t)})] = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)},\end{aligned}$$

where $\boldsymbol{z}^{(t)}$ has elements

$$z_i^{(t)} = \sum_j x_{ij}\,\beta_j^{(t)} + \left(y_i - \mu_i^{(t)}\right)\frac{\partial\eta_i^{(t)}}{\partial\mu_i^{(t)}} = \eta_i^{(t)} + \left(y_i - \mu_i^{(t)}\right)\frac{\partial\eta_i^{(t)}}{\partial\mu_i^{(t)}}.$$

The Fisher scoring equations then have the form

$$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(t)}\boldsymbol{X})\boldsymbol{\beta}^{(t+1)} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)}.$$

These are the normal equations for using weighted least squares to fit a linear model for a response variable $\boldsymbol{z}^{(t)}$, when the model matrix is $\boldsymbol{X}$ and the inverse of the covariance matrix is $\boldsymbol{W}^{(t)}$. The equations have the solution

$$\boldsymbol{\beta}^{(t+1)} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)}.$$

The vector $\boldsymbol{z}^{(t)}$ in this formulation is an estimated linearized form of the link function g, evaluated at $\boldsymbol{y}$,

$$g(y_i) \approx g\left(\mu_i^{(t)}\right) + \left(y_i - \mu_i^{(t)}\right)g'\left(\mu_i^{(t)}\right) = \eta_i^{(t)} + \left(y_i - \mu_i^{(t)}\right)\frac{\partial\eta_i^{(t)}}{\partial\mu_i^{(t)}} = z_i^{(t)}. \qquad (4.25)$$

The *adjusted response variable* $\boldsymbol{z}$ has element i approximated by $z_i^{(t)}$ for cycle t of the iterative scheme. That cycle regresses $\boldsymbol{z}^{(t)}$ on $\boldsymbol{X}$ with weight (i.e., inverse covariance) $\boldsymbol{W}^{(t)}$ to obtain a new approximation $\boldsymbol{\beta}^{(t+1)}$. This estimate yields a new linear predictor value $\boldsymbol{\eta}^{(t+1)} = \boldsymbol{X}\boldsymbol{\beta}^{(t+1)}$ and a new approximation $\boldsymbol{z}^{(t+1)}$ for the adjusted response for the next cycle. The ML estimator results from iterative use of weighted least squares,

in which the weight matrix changes at each cycle. The process is called *iteratively reweighted least squares* (IRLS). The weight matrix $\boldsymbol{W}$ used in $\text{var}(\hat{\boldsymbol{\beta}}) \approx (\boldsymbol{X}^{\text{T}}\boldsymbol{W}\boldsymbol{X})^{-1}$, in the generalized hat matrix (4.19), and in Fisher scoring is the inverse covariance matrix of the linearized form $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$ of $g(\boldsymbol{y})$. At convergence,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\text{T}}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\text{T}}\widehat{\boldsymbol{W}}\hat{\boldsymbol{z}},$$

for the estimated adjusted response $\hat{\boldsymbol{z}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{D}}^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$.

4.5.5 Simplifications for Canonical Link Functions

Certain simplifications result for GLMs that use the canonical link function. For that link,

$$\eta_i = \theta_i = \sum_{j=1}^{p} \beta_j x_{ij},$$

and

$$\partial\mu_i/\partial\eta_i = \partial\mu_i/\partial\theta_i = \partial b'(\theta_i)/\partial\theta_i = b''(\theta_i).$$

Since $\text{var}(y_i) = b''(\theta_i)a(\phi)$, the contribution (4.9) to the likelihood equation for β_j simplifies to

$$\frac{\partial L_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{\text{var}(y_i)} b''(\theta_i) x_{ij} = \frac{(y_i - \mu_i)x_{ij}}{a(\phi)}. \tag{4.26}$$

Often $a(\phi)$ is identical for all observations, such as for Poisson GLMs [$a(\phi) = 1$] and for binomial GLMs with each $n_i = 1$ [for which $a(\phi) = 1$]. Then, the likelihood equations are

$$\sum_{i=1}^{n} x_{ij} y_i = \sum_{i=1}^{n} x_{ij}\mu_i, \quad j = 1, 2, \ldots, p. \tag{4.27}$$

We noted at the beginning of Section 4.2 that $\{\sum_{i=1}^{n} x_{ij} y_i\}$ are the sufficient statistics for $\{\beta_j\}$. So equation (4.27) illustrates a fundamental result:

- For GLMs with canonical link function, the likelihood equations equate the sufficient statistics for the model parameters to their expected values.

For a normal distribution with identity link, these are the *normal equations*. We obtained them for Poisson loglinear models in (4.12).

From expression (4.26) for $\partial L_i/\partial\beta_j$, with the canonical link function the second partial derivatives of the log likelihood are

$$\frac{\partial^2 L_i}{\partial\beta_h\partial\beta_j} = -\frac{x_{ij}}{a(\phi)}\left(\frac{\partial\mu_i}{\partial\beta_h}\right).$$

This does not depend on y_i, so

$$\partial^2 L(\boldsymbol{\beta})/\partial\beta_h\partial\beta_j = E[\partial^2 L(\boldsymbol{\beta})/\partial\beta_h\partial\beta_j].$$

That is, $\boldsymbol{H} = -\boldsymbol{\mathcal{J}}$, so the Newton–Raphson and Fisher scoring algorithms are identical for GLMs that use the canonical link function (Nelder and Wedderburn 1972).

Finally, in the canonical link case the log likelihood is necessarily a concave function, because the log likelihood for an exponential family distribution is concave in the natural parameter. In using iterative methods to find the ML estimates, we do not need to worry about the possibility of multiple maxima for the log likelihood.

4.6 SELECTING EXPLANATORY VARIABLES FOR A GLM

Model selection for GLMs faces the same issues as for ordinary linear models. The selection process becomes more difficult as the number of explanatory variables increases, because of the rapid increase in possible effects and interactions. The selection process has two competing goals. The model should be complex enough to fit the data well. On the other hand, it should smooth rather than overfit the data and ideally be relatively simple to interpret.

Most research studies are designed to answer certain questions. Those questions guide the choice of model terms. Confirmatory analyses then use a restricted set of models. For instance, a study hypothesis about an effect may be tested by comparing models with and without that effect. For studies that are exploratory rather than confirmatory, a search among possible models may provide clues about the structure of effects and raise questions for future research. In either case, it is helpful first to study the marginal effect of each predictor by itself with descriptive statistics and a scatterplot matrix, to get a feel for those effects.

This section discusses some model-selection procedures and issues that affect the selection process. Section 4.7 presents an example and illustrates that the variables selected, and the influence of individual observations, can be highly sensitive to the assumed distribution for y.

4.6.1 Stepwise Procedures: Forward Selection and Backward Elimination

With p explanatory variables, the number of potential models is 2^p, as each variable either is or is not in the chosen model. The *best subset selection* identifies the model that performs best according to a criterion such as maximizing the adjusted R^2 value. This is computationally intensive when p is large. Alternative algorithmic methods

can search among the models. In exploratory studies, such methods can be informative if we use the results cautiously.

Forward selection adds terms sequentially. At each stage it selects the term giving the greatest improvement in fit. A point of diminishing returns occurs in adding explanatory variables when new ones added are themselves so well predicted by ones already used that they do not provide a substantive improvement in R^2. The process stops when further additions do not improve the fit, according to statistical significance or a criterion for judging the model fit (such as the AIC, introduced below in Section 4.6.3). A stepwise variation of this procedure rechecks, at each stage, whether terms added at previous stages are still needed. *Backward elimination* begins with a complex model and sequentially removes terms. At each stage, it selects the term whose removal has the least damaging effect on the model, such as the largest P-value in a test of its significance or the least deterioration in a criterion for judging the model fit. The process stops when any further deletion leads to a poorer fit.

With either approach, an interaction term should not be in a model without its component main effects. Also, for qualitative predictors with more than two categories, the process should consider the entire variable at any stage rather than just individual indicator variables. Add or drop the entire variable rather than only one of its indicators. Otherwise, the result depends on the choice of reference category for the indicator coding.

Some statisticians prefer backward elimination over forward selection, feeling it safer to delete terms from an overly complex model than to add terms to an overly simple one. Forward selection based on significance testing can stop prematurely because a particular test in the sequence has low power. It also has the theoretical disadvantage that in early stages both models being compared are likely to be inadequate, making the basis for a significance test dubious. Neither strategy necessarily yields a meaningful model. When you evaluate many terms, some that are not truly important may seem so merely because of chance. For instance, when all the true effects are weak, the largest sample effect is likely to overestimate substantially its true effect. Also, the use of standard significance tests in the process lacks theoretical justification, because the distribution of the minimum or maximum P-value evaluated over a set of explanatory variables is not the same as that of a P-value for a preselected variable. Use variable-selection algorithms in an informal manner and with caution. Backward and forward selection procedures yielding quite different models is an indication that such results are of dubious value.

For any method, since statistical significance is not the same as practical significance, a significance test should not be the sole criterion for including a term in a model. It is sensible to include a variable that is central to the purposes of the study and report its estimated effect even if it is not statistically significant. Keeping it in the model may make it possible to compare results with other studies where the effect is significant, perhaps because of a larger sample size. If the variable is a potential confounder, including it in the model may help to reduce bias in estimating relevant effects of key explanatory variables. But also a variable should not be kept merely because it is statistically significant. For example, if a selection method results in a model having adjusted $R^2 = 0.39$ but a simpler model without the interaction

terms has adjusted $R^2 = 0.38$, for ease of interpretation it may be preferable to drop the interaction terms. Algorithmic selection procedures are no substitute for careful thought in guiding the formulation of models.

Some variable-selection methods adapt stepwise procedures to take such issues into account. For example, Hosmer et al. (2013, Chapter 4) recommended a *purposeful selection* model-building process that also pays attention to potential confounding variables. In outline, they suggest constructing an initial main-effects model by (1) choosing a set of explanatory variables that include the known clinically important variables and others that show *any* evidence of being relevant predictors in a univariable analysis (e.g., having P-value <0.25), (2) conducting backward elimination with the full set from (1), keeping a variable if it is either significant at a somewhat more stringent level or shows evidence of being a relevant confounder, in the sense that the estimated effect of a key variable changes by at least 20% when it is removed, (3) checking whether any variables not included in (1) are significant when adjusting for the variables in the model after Step (2). One then checks for plausible interactions among variables in the model after Step (3), using significance tests at conventional levels such as 0.05, followed by the usual diagnostic investigations presented in Section 4.4.

4.6.2 Model Selection: The Bias–Variance Tradeoff

In selecting a model from a set of candidates, we are mistaken if we think that there is a "correct" one. Any model is a simplification of reality. For instance, an explanatory variable will not have exactly a linear effect, no matter which link function we use. And it is not always a good idea to choose a more complex model in order to obtain a better fit. A simple model that fits adequately has the advantages of model parsimony, including a tendency to provide more accurate estimates of the quantities of interest. The choice of how complex a model to use is at the heart of the basic statistical tradeoff between the variance of an estimator and its bias. Here, bias occurs when the true $\{E(y_i)\}$ values differ from the values $\{\mu_{Mi}\}$ corresponding to fitting model M to the population. Using a simpler model has the disadvantage of increasing the bias; that is, the differences $\{|\mu_{Mi} - E(y_i)|\}$ between the model-based means and the true means tend to be larger. But a simpler model has the advantage that the decrease in the number of model parameters results in decreased variance in the estimators. This can result in overall lower mean squared error[17] in estimating characteristics such as the true $\{E(y_i)\}$ values.

In practice, many models can be consistent with the data. If not one of them is "correct," it is logically inconsistent to choose one model based on its fitting the data well and then make subsequent inferences as if the model had been chosen before seeing the data. Although this is common practice, it results in a tendency to underestimate uncertainty and to exaggerate significance. Keep in mind the selection uncertainty in making inferences based on a model, because those inferences use the same data that helped you to select the model. Although selection procedures are

[17]Recall that MSE = variance + (bias)^2.

helpful tools, results of an exploratory study are highly tentative and useful mainly for suggesting effects and hypotheses to analyze in future studies. The model-building process should also be based on theory and common sense.

Other criteria besides significance tests comparing models can help you to select a sensible model. We next introduce the best known of such criteria.

4.6.3 AIC: Minimizing Distance of the Fit from the Truth

The *Akaike information criterion* (AIC) judges a model by how close we can expect its sample fit to be to the true model fit. In the population of interest, even though a simple model is farther from the true relationship than is a more complex model, for a sample it may tend to provide a closer fit because of the advantages of model parsimony. In a set of potential models, the optimal model is the one that tends to have sample fit closest to the true model fit.

Here "closeness" is defined in terms of the *Kullback–Leibler divergence* of a model M from the unknown true model. Let $p(\boldsymbol{y})$ denote the density (or probability, in the discrete case) of the data under the true model, and let $p_M(\boldsymbol{y}; \boldsymbol{\beta}_M)$ be the density under model M with parameters $\boldsymbol{\beta}_M$. For a given value of the ML estimator $\hat{\boldsymbol{\beta}}_M$ of $\boldsymbol{\beta}_M$ and for a future sample $\boldsymbol{y}^*$ from $p(\cdot)$, the Kullback–Leibler divergence between the true and fitted distributions is

$$KL[p, p_M(\hat{\boldsymbol{\beta}}_M)] = E\left[\log \frac{p(\boldsymbol{y}^*)}{p_M(\boldsymbol{y}^*; \hat{\boldsymbol{\beta}}_M)}\right],$$

where the expectation is taken relative to the true distribution $p(\cdot)$. The goal of AIC is to choose the model to minimize $E[KL(p, p_M(\hat{\boldsymbol{\beta}}_M))]$ for a set of potential models, where this expectation also is taken relative to $p(\cdot)$, now with $\hat{\boldsymbol{\beta}}_M$ as the random variable for another sample. To do this, it is sufficient to minimize $E\{-E\log[p_M(\boldsymbol{y}^*; \hat{\boldsymbol{\beta}}_M)]\}$ over the set of models. The true distribution $p(\cdot)$ needed to evaluate this expectation is unknown, but the expectation can be estimated consistently. Akaike (1973) showed that when M is reasonably close to the true model, the maximized log likelihood $L(\hat{\boldsymbol{\beta}}_M)$ for M is a biased estimator of $E\{E\log[p_M(\boldsymbol{y}^*; \hat{\boldsymbol{\beta}}_M)]\}$, and for large sample sizes the bias is reduced by subtracting the number of parameters in M. This implies that out of a set of reasonably fitting models, the optimal model minimizes[18]

$$\text{AIC} = -2\left[L(\hat{\boldsymbol{\beta}}_M) - \text{number of parameters in } M\right].$$

Although the role of subtracting the number of parameters in M is to adjust for bias, the AIC essentially penalizes a model for having many parameters. With many potential explanatory variables, using AIC can aid in variable selection. Out of a set of candidate models, we identify the one with smallest AIC or identify parsimonious

[18]Akaike introduced the multiple of 2 merely for convenience, to link the AIC formula with likelihood-ratio chi-squared statistics.

models that have AIC near the minimum value. The candidate models need not be nested or even based on the same family of distributions for the random component.

An alternative to AIC, a *Bayesian information criterion* (BIC), penalizes more severely for the number of model parameters. It replaces 2 by $\log(n)$ as its multiple. Compared with AIC, BIC gravitates less quickly toward more complex models as n increases. It is based on a Bayesian argument for determining which of a set of models has highest posterior probability (Schwarz 1978). Because of selection bias, however, model-selection criteria such as minimizing AIC or minimizing BIC can result in inclusion of irrelevant variables (George 2000). This can be especially problematic when p is large and few variables truly have an effect[19].

4.6.4 Summarizing Predictive Power: R-Squared and Other Measures

In ordinary linear models, R^2 and the multiple correlation R describe how well the explanatory variables predict the sample response values, with $R = 1$ for perfect prediction. For any GLM, the correlation between the fitted values $\{\hat{\mu}_i\}$ and the observed responses $\{y_i\}$ measures predictive power. It is also useful for comparing fits of different models for the same data. For the ordinary linear model, $\text{corr}(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ is the multiple correlation. An advantage of the correlation, relative to its square, is the appeal of working on the original scale and its approximate proportionality to effect size: For a small effect with a single explanatory variable, doubling the slope corresponds approximately to doubling the correlation. For GLMs, unlike linear models, $\text{corr}(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ need not be nondecreasing as the model gets more complex, although it usually is.

Other measures of predictive power directly use the likelihood function. Denote the maximized log likelihood by L_M for a given model, L_S for the saturated model, and L_0 for the null model containing only an intercept term. Then, $L_0 \leq L_M \leq L_S$, and

$$\frac{L_M - L_0}{L_S - L_0} \tag{4.28}$$

falls between 0 and 1. It equals 0 when the model provides no improvement in fit over the null model, and it equals 1 when the model fits as well as the saturated model. A weakness is that the scale for the log likelihood may not be as easy to interpret as the scale for the response variable itself. The measure is mainly useful for comparing models.

With any such measure, with many explanatory variables, the sample estimators can be biased upward in estimating the true population value. It can be misleading to compare sample values for models with quite different numbers of parameters. Bias corrections are possible, for example, by using cross-validation (Stone 1974) or the jackknife (Zheng and Agresti 2000).

[19]For example, when no variables truly have an effect, for t tests of the individual partial effects, $E(t^2_{\max}) \approx 2 \log p$ (George 2000).

4.6.5 Effects of Collinearity

In an observational study with many explanatory variables, relations among them may suggest that not one variable is important when all the others are in the model. A variable may have little partial effect because it is predicted well by the others. Deleting a nearly redundant predictor can be helpful, for instance, to reduce standard errors of other estimated effects.

In a linear model, the variance of $\hat{\beta}_j$ is

$$\text{var}(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \left[\frac{\sigma^2}{\sum_i (x_{ij} - \bar{x}_j)^2} \right],$$

where R_j^2 denotes the value of R^2 for predicting x_j as a response using the other explanatory variables in the model. One can derive this formula from an expression of $\hat{\beta}_j$ for a regression using two sets of residuals, as in Section 2.5.6 (e.g., see Greene 2011, p. 90). The ratio $VIF_j = 1/(1 - R_j^2)$ is called the *variance inflation factor* for predictor x_j. It is the multiple by which the variance increases because the other predictors are correlated with x_j. As R_j^2 increases, $\text{var}(\hat{\beta}_j)$ increases. If $R_j^2 = 1$, there is extrinsic aliasing (Section 1.3.2): The model matrix has less than full rank, and there are infinitely many solutions for $\hat{\boldsymbol{\beta}}$. When R_j^2 is near 1, $\hat{\beta}_j$ can be unstable. When $R_j^2 = 0$, $\hat{\beta}_j$ and its variance are identical to their values when x_j is the sole explanatory variable in the model.

To illustrate, for the horseshoe crab data (Section 1.5.1), the width of the carapace shell is highly statistically significant as a predictor of a female crab's number of satellites. What happens if we add the crab's weight as a predictor? Here is the result of fitting Poisson loglinear models:

```
------------------------------------------------------------------------
> attach(Crabs) # y is number of satellites
> summary(glm(y ~ width, family=poisson(link=log)))
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.30476    0.54224  -6.095  1.1e-09
width        0.16405    0.01997   8.216  < 2e-16
----
> summary(glm(y ~ weight + width, family=poisson(link=log)))
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.29521    0.89890  -1.441  0.14962
weight       0.44697    0.15862   2.818  0.00483
width        0.04608    0.04675   0.986  0.32433
----
> cor(weight, width)
[1] 0.8868715
------------------------------------------------------------------------
```

Width loses its significance. The loss also happens with normal linear models and with a more appropriate two-parameter distribution for count data that Chapter 7

uses. The dramatic reduction in the significance of the crab's shell width when its weight is added to the model reflects the correlation of 0.887 between weight and width. The variance inflation factor for the effect of either predictor in a linear model is $1/[1-(0.887)^2] = 4.685$. The *SE* for the effect of width more than doubles when weight is added to the model, and the estimate itself is much smaller, reflecting also the strong correlation.

This example illustrates a general phenomenon in modeling. When an explanatory variable x_j is highly correlated with a linear combination of other explanatory variables in the model, the relation is said to exhibit[20] *collinearity* (also referred to as *multicollinearity*).

When collinearity exists, one approach chooses a subset of the explanatory variables, removing those variables that explain a small portion of the remaining unexplained variation in y. When several predictors are highly correlated and are indicators of a common feature, another approach constructs a summary index by combining responses on those variables. Also, methods such as *principal components analysis* create artificial variables from the original ones in such a way that the new variables are uncorrelated. In most applications, though, it is more advisable from an interpretive standpoint to use a subset of the variables or create some new variables directly. The effect of interaction terms on collinearity is diminished if we center the explanatory variables before entering them in the model. Section 11.1.2 introduces alternative methods, such as *ridge regression*, that produce estimates that are biased but less severely affected by collinearity.

Collinearity does not adversely affect all aspects of regression. Although collinearity makes it difficult to assess partial effects of explanatory variables, it does not hinder the assessment of their joint effects. If newly added explanatory variables overlap substantially with ones already in the model, R^2 will not increase much, but the presence of collinearity has little effect on the global test of significance.

4.7 EXAMPLE: BUILDING A GLM

Section 3.4 introduced a dataset on home selling prices. The response variable is selling *price* in thousands of dollars. The explanatory variables are *size* of the home in square feet, whether it is *new* (1 = yes, 0 = no), annual *tax* bill in dollars, number of *bedrooms*, and number of *bathrooms*. A scatterplot matrix has limited use for highly discrete variables such as new, beds, and baths, but Figure 4.3 does reveal the strong positive correlation for each pair of price, size, and taxes.

```
------------------------------------------------------------------------------
> attach(Houses) # data at www.stat.ufl.edu/~aa/glm/data
> pairs(cbind(price,size,taxes)) # scatterplot matrix for pairs of var's
```

[20]Technically, collinearity refers to an *exact* linear dependence, but the term is used in practice when there is a *near* dependence.

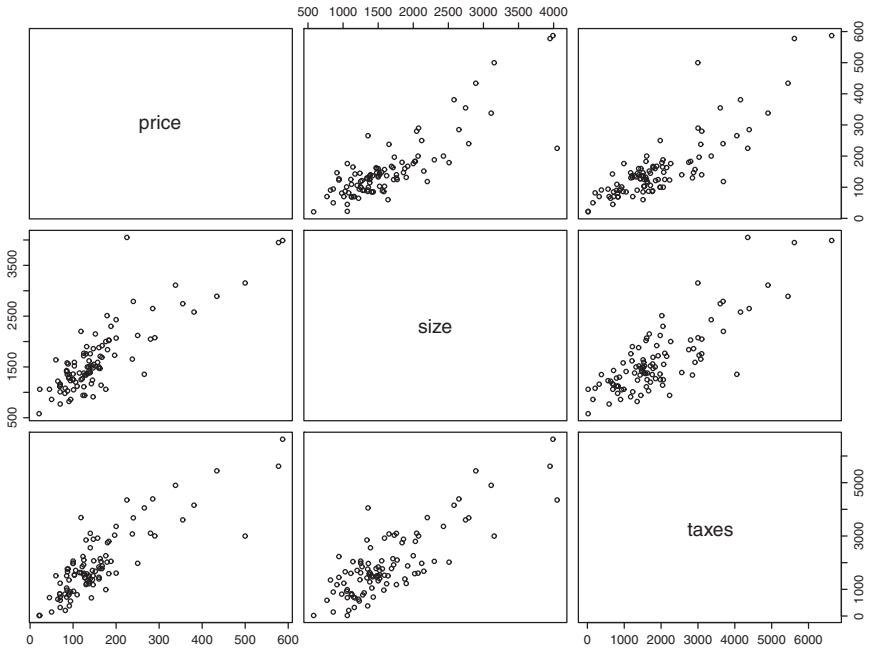

Figure 4.3 Scatterplot matrix for price, size, and taxes in dataset on house selling prices.

```
> cor(cbind(price,size,taxes,beds,baths)) # correlation matrix
        price    size   taxes    beds   baths
price  1.0000  0.8338  0.8420  0.3940  0.5583
size   0.8338  1.0000  0.8188  0.5448  0.6582
taxes  0.8420  0.8188  1.0000  0.4739  0.5949
beds   0.3940  0.5448  0.4739  1.0000  0.4922
baths  0.5583  0.6582  0.5949  0.4922  1.0000
------------------------------------------------------------------------------
```

4.7.1 Backward Elimination with House Selling Price Data

We illustrate a backward elimination process for selecting a model, using all the variables except taxes. (A chapter exercise uses all the variables.) Rather than relying solely on significance tests, we combine a backward process with judgments about practical significance.

To gauge how complex a model may be needed, we begin by comparing models containing the main effects only, also the second-order interactions, and also the third-order interactions. The `anova` function in R executes the F test comparing nested normal linear models (Section 3.2.2).

```
------------------------------------------------------------------------------
> fit1 <- lm(price ~ size + new + baths + beds)
> fit2 <- lm(price ~ (size + new + baths + beds)^2)
```

```
> fit3 <- lm(price ~ (size + new + baths + beds)^3)
> anova(fit1, fit2)
Analysis of Variance Table
Model 1: price ~ size + new + baths + beds
Model 2: price ~ (size + new + baths + beds)^2
  Res.Df     RSS  Df  Sum of Sq       F     Pr(>F)
1     95  279624
2     89  217916   6      61708  4.2004  0.0009128
-------------------------------------------------------------------------
```

A statistically significant improvement results from adding six pairwise interactions to the main effects model, with a drop in SSE of 61,708. A similar analysis (not shown here) indicates that we do not need three-way interactions. The R^2 values for the three models are 0.724, 0.785, and 0.804. In this process we compare models with quite different numbers of parameters, so we instead focus on the adjusted R^2 values: 0.713, 0.761, and 0.771. So we search for a model that fits adequately but is simpler than the model with all the two-way interactions.

In *fit2* (not shown), the least significant two-way interaction is baths × beds. Removing that interaction yields *fit4* with adjusted $R^2 = 0.764$. Then the least significant remaining two-way interaction is size × baths. With *fit5* we remove it, obtaining adjusted $R^2 = 0.766$. At that stage, the new × beds interaction is least significant, and we remove it, yielding adjusted $R^2 = 0.769$. The result is *fit6*:

```
-------------------------------------------------------------------------
> summary(fit6)
              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    135.6459     54.1902    2.503    0.0141
size            -0.0032      0.0323   -0.098    0.9219
new             90.7242     77.5413    1.170    0.2450
baths           12.2813     12.1814    1.008    0.3160
beds           -55.0541     17.6201   -3.125    0.0024
size:new         0.1040      0.0286    3.630    0.0005
size:beds        0.0309      0.0091    3.406    0.0010
new:baths     -111.5444     45.3086   -2.462    0.0157
---
Multiple R-squared: 0.7851,     Adjusted R-squared: 0.7688
-------------------------------------------------------------------------
```

The three remaining two-way interactions are statistically significant at the 0.02 level. However, the *P*-values are only rough guidelines, and dropping the new × baths interaction (*fit7*, not shown) has only a slight effect, adjusted R^2 dropping to 0.756. At this stage we could drop baths from the model, as it is not in the remaining interaction terms and its $t = 0.40$.

```
-------------------------------------------------------------------------
> fit8 <- update(fit7, .~. - baths)
> summary(fit8)
```

```
               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   143.47098     54.1412    2.650    0.0094
size            0.00684      0.0326    0.210    0.8345
new           -56.68578     49.3006   -1.150    0.2531
beds          -53.63734     17.9848   -2.982    0.0036
size:new        0.05441      0.0210    2.588    0.0112
size:beds       0.03002      0.0092    3.254    0.0016
---
Multiple R-squared: 0.7706,     Adjusted R-squared: 0.7584
---
> plot(fit8)
-------------------------------------------------------------------------
```

Both interactions are highly statistically significant, and adjusted R^2 drops to 0.716 if we drop them both. Viewing this as a provisional model, let us interpret the effects in *fit8*:

- For an older two-bedroom home, the effect on the predicted selling price of a 100 square foot increase in size is 100[0.00684 + 2(0.03002), or \$6688. For an older three-bedroom home, it is 100[0.00684 + 3(0.03002)], or \$9690, and for an older four-bedroom home, it is 100[0.00684 + 4(0.03002)], or \$12,692. For a new home, \$5441 is added to each of these three effects.
- Adjusted for the number of bedrooms, the effect on the predicted selling price of a home's being new (instead of older) is −56.686 + 1000(0.0544), or −\$2277, for a 1000-square-foot home, −56.686 + 2000(0.0544), or \$52,132, for a 2000-square-foot home, and −56.686 + 3000(0.0544), or \$106,541 for a 3000-square-foot home.
- Adjusted for whether a house is new, the effect on the predicted selling price of an extra bedroom is −53.637 + 1000(0.0300), or −\$23, 616, for a 1000-square-foot home, −53.637 + 2000(0.0300), or \$6405, for a 2000-square-foot home, and −53.637 + 3000(0.0300), or \$36,426, for a 3000-square-foot home.

For many purposes in an exploratory study, a simple model is adequate. We obtain a reasonably effective fit by removing the beds effects from *fit8*, yielding adjusted R^2 = 0.736 and very simple interpretations from the fit $\hat{\mu} = -22.228 + 0.1044(\text{size}) - 78.5275(\text{new}) + 0.0619(\text{size} \times \text{new})$. For example, the estimated effect of a 100 square-foot increase in size is \$10,440 for an older home and \$16,630 for a new home. In fact, this is the model having minimum BIC. The model having minimum AIC is[21] slightly more complex, the same as *fit6* above.

```
-------------------------------------------------------------------------
> step(lm(price ~ (size + new + beds + baths)^2))
Start:  AIC=790.67 # AIC for initial model with two-factor interactions
...
```

[21]The AIC value reported by the `step` and `extractAIC` functions in R ignores certain constants, which the `AIC` function in R includes.

```
Step:  AIC=784.78 # lowest AIC for special cases of starting model
price ~ size + new + beds + baths + size:new + size:beds + new:baths
> AIC(lm(price ~ size+new+beds+baths+size:new+size:beds+new:baths))
[1] 1070.565 # correct value using AIC formula for normal linear model
> BIC(lm(price ~ size+new+size:new)) # this is model with lowest BIC
[1] 1092.973
```

4.7.2 Gamma GLM Has Standard Deviation Proportional to Mean

We ignored an important detail in the above model selection process. Section 3.4.2 noted that observation 64 in the dataset is an outlier that is highly influential in least squares fitting. Repeating the backward elimination process without it yields a different final model. This makes any conclusions even more highly tentative.

Section 3.4 noted some evidence of greater variability when mean selling prices are greater. This seems plausible and often happens for positive-valued response variables. At settings of explanatory variables for which $E(y)$ is low, we would not expect much variability in y (partly because y cannot be < 0), whereas when $E(y)$ is high, we would expect considerable variability. In each case, we would also expect some skew to the right in the response distribution, which could partly account for relatively large values. For such data, ordinary least squares is not optimal. One approach instead uses weighted least squares, by weighting observations according to how the variance depends on the mean. An alternative GLM approach assumes a distribution for y for which the variance increases as the mean increases. The family of *gamma distributions* has this property.

The two-parameter gamma probability density function for y, parameterized in terms of its mean μ and the shape parameter $k > 0$, is

$$f(y; k, \mu) = \frac{(k/\mu)^k}{\Gamma(k)} e^{-ky/\mu} y^{k-1}, \quad y \geq 0, \tag{4.29}$$

$$\text{for which } E(y) = \mu, \quad \text{var}(y) = \mu^2/k.$$

Gamma GLMs usually assume k to be constant but unknown, like σ^2 in ordinary linear models. Then the coefficient of variation, $\sqrt{\text{var}(y)}/\mu = 1/\sqrt{k}$, is constant as μ varies, and the standard deviation increases proportionally with the mean. The density is skewed to the right, but the degree of skewness (which equals $2/\sqrt{k}$) decreases as k increases. The mode is 0 when $k \leq 1$ and $\mu(k-1)/k$ when $k > 1$, with $k = 1$ giving the exponential distribution. The chi-squared distribution is the special case with $\mu = df$ and $k = df/2$.

The gamma distribution is in the exponential dispersion family with natural parameter $\theta = -1/\mu$, $b(\theta) = -\log(-\theta)$, and dispersion parameter $\phi = 1/k$. The scaled deviance for a gamma GLM has approximately a chi-squared distribution. However, the dispersion parameter is usually treated as unknown. We can mimic how we eliminate it in ordinary linear models by constructing an F statistic. For example, consider

testing M_0 against M_1 for nested GLMs M_0 and M_1 with $p_0 < p_1$ parameters. Using the model deviances, the test statistic

$$\frac{[D(M_0) - D(M_1)]/(p_1 - p_0)}{D(M_1)/(n - p_1)},$$

has an approximate $F_{p_1-p_0,n-p_1}$ distribution, if the numerator and denominator are approximately independent[22]. Or, we can explicitly estimate ϕ for the more complex model and use the approximation

$$\frac{[D(M_0) - D(M_1)]/(p_1 - p_0)}{\hat{\phi}} \sim F_{p_1-p_0,n-p_1}.$$

Some software (e.g., SAS) uses ML to estimate ϕ. However, the ML estimator is inconsistent if the variance function is correct but the distribution is not truly the assumed one (McCullagh and Nelder 1989, p. 295). Other software (e.g., R) uses[23] the scaling $\hat{\phi} = X^2/(n-p)$ of the Pearson statistic (4.17), which is based on equating the average squared Pearson residual to 1, adjusted by using the dimension of the error space $n - p$ instead of n in the denominator (Wedderburn 1974). It is consistent when $\hat{\boldsymbol{\beta}}$ is. In the gamma context, this estimate is

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

When k is large, a gamma variate y has distribution close to normal. However, the gamma GLM fit is more appropriate than the least squares fit because the standard deviation increases as the mean does. Sometimes the identity link function is inadequate, because y must be nonnegative. It is then more common to use the log link. With that link, results are similar to least squares with a log-normal assumption for the response, that is, applying least squares to a linear model expressed in terms of $\log(y)$ (Exercise 4.27).

4.7.3 Gamma GLMs for House Selling Price Data

For the house selling price data, perhaps observation 64 is *not* especially unusual if we assume a gamma distribution for price. Using the same linear predictor as in the model (with *fit8*) interpreted in Section 4.7.1, we obtain:

```
------------------------------------------------------------------------
> fit.gamma <- glm(price ~ size + new + beds + size:new + size:beds,
                   family = Gamma(link = identity))
> summary(fit.gamma)$coef
              Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)    44.3759     48.5978    0.9131    0.3635
```

[22]This holds when the dispersion parameter is small, so the gamma distribution is approximately normal. See Jørgensen (1987) for the general case using the F.

[23]But ML is available in R with the `gamma.dispersion` function in the MASS package.

```
size             0.0740      0.0400    1.8495    0.0675
new            -60.0290     65.7655   -0.9128    0.3637
beds           -22.7131     17.6312   -1.2882    0.2008
size:new         0.0538      0.0376    1.4325    0.1553
size:beds        0.0100      0.0126    0.7962    0.4279
-------------------------------------------------------------------------
```

Now, neither interaction is significant! This also happens if we fit the model without observation 64. Including that observation, its standardized residual is now only -1.63, not at all unusual, because this model expects more variability in the data when the mean is larger. In fact, we may not need any interaction terms:

```
-------------------------------------------------------------------------
> fit.g1 <- glm(price ~ size+new+baths+beds, family=Gamma(link=identity))
> fit.g2 <- glm(price~(size+new+baths+beds)^2,family=Gamma(link=identity))
> anova(fit.g1, fit.g2, test="F")
Analysis of Deviance Table
  Resid. Df  Resid. Dev  Df  Deviance       F  Pr(>F)
1        95     10.4417
2        89      9.8728   6    0.5689  0.8438  0.5396
-------------------------------------------------------------------------
```

Further investigation using various model-building strategies reveals that according to AIC the model with size alone does well (AIC = 1050.7), as does the model with size and beds (AIC = 1048.3) and the model with size and new (AIC = 1049.5), with a slight improvement from adding the size × new interaction (AIC = 1047.9). Here is the output for the latter gamma model and for the corresponding normal linear model that we summarized near the end of Section 4.7.1:

```
-------------------------------------------------------------------------
> summary(glm(price ~ size+new+size:new, family=Gamma(link=identity)))
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   -7.4522     12.9738   -0.574    0.5670
size           0.0945      0.0100    9.396  2.95e-15
new          -77.9033     64.5827   -1.206    0.2307
size:new       0.0649      0.0367    1.769    0.0801 .
---
(Dispersion parameter for Gamma family taken to be 0.11021)
Residual deviance: 10.563  on 96  degrees of freedom
AIC: 1047.9
> plot(glm(price ~ size + new + size:new, family=Gamma(link=identity)))
> summary(lm(price ~ size + new + size:new))
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  -22.2278     15.5211   -1.432    0.1554
size           0.1044      0.0094   11.082   < 2e-16
new          -78.5275     51.0076   -1.540    0.1270
size:new       0.0619      0.0217    2.855    0.0053
---
Residual standard error: 52 on 96 degrees of freedom
Multiple R-squared:  0.7443,    Adjusted R-squared:  0.7363
-------------------------------------------------------------------------
```

Effects are similar, but the interaction term in the gamma model has larger *SE*. For this gamma model, $\hat{\phi} = 0.11021$, so the estimated shape parameter is $\hat{k} = 1/\hat{\phi} = 9.07$, which corresponds to a bell shape with some skew to the right. The estimated standard deviation $\hat{\sigma}$ of the conditional distribution of y relates to the estimated mean $\hat{\mu}$ by

$$\hat{\sigma} = \sqrt{\hat{\phi}}\hat{\mu} = \hat{\mu}/\sqrt{\hat{k}} = 0.33197\hat{\mu}.$$

For example, at predictor values having estimated mean selling price $\hat{\mu} = \$100{,}000$, the estimated standard deviation is \$33,197, whereas at $\hat{\mu} = \$400{,}000$, $\hat{\sigma}$ is four times as large.

The reported AIC value of 1047.9 for this gamma model is much better than the AIC for the normal linear model with the same explanatory variables, or for the normal linear model (*fit6*) in Section 4.7.1 that minimized AIC, of the models with main effects and two-way interactions.

```
------------------------------------------------------------------------------
> AIC(lm(price ~ size + new + size:new))
[1] 1079.9
> AIC(lm(price ~ size +new +beds +baths +size:new +size:beds +new:baths))
[1] 1070.6
------------------------------------------------------------------------------
```

We learn an important lesson from this example:

- In modeling, it is not sufficient to focus on how $E(y_i)$ depends on $\boldsymbol{x}_i$ for all i. The assumption about how $\text{var}(y_i)$ depends on $E(y_i)$ can have a significant impact on conclusions about the effects.

Other approaches, such as using the log link instead of the identity link, yield other plausible models. Analyses that are beyond our scope here (such as Q–Q plots) indicate that selling prices may have a somewhat longer right tail than gamma and log-normal models permit. An alternative response distribution having this property is the *inverse Gaussian*, which has variance proportional to μ^3 (Seshadri 1994).

APPENDIX: GLM ANALOGS OF ORTHOGONALITY RESULTS FOR LINEAR MODELS

This appendix presents approximate analogs of linear model orthogonality results. Lovison (2014) showed that a weighted version of the estimated adjusted responses that has approximately constant variance has the same orthogonality of fitted values and residuals as occurs in ordinary linear models.

Recall that $\boldsymbol{D} = \text{diag}\{\partial\mu_i/\partial\eta_i\}$ and $\boldsymbol{W} = \text{diag}\{(\partial\mu_i/\partial\eta_i)^2/\text{var}(y_i)\}$. From Section 4.5.4, the IRLS fitting process is naturally expressed in terms of the estimate $\hat{z} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{D}}^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ of an *adjusted* response variable $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$. Since

$$\hat{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^{\text{T}}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\text{T}}\widehat{\boldsymbol{W}}\hat{\boldsymbol{z}}$$

for the fitted linear predictor values, $\boldsymbol{X}(\boldsymbol{X}^{\text{T}}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\text{T}}\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{W}}^{-1/2}\widehat{\boldsymbol{H}}_{\boldsymbol{W}}\widehat{\boldsymbol{W}}^{1/2}$ is a sort of asymmetric projection adaptation of the estimate of the generalized hat matrix (4.19), namely,

$$\widehat{\boldsymbol{H}}_{\boldsymbol{W}} = \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}(\boldsymbol{X}^{\text{T}}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\text{T}}\widehat{\boldsymbol{W}}^{1/2}.$$

Consider the weighted adjusted responses and linear predictor, $\boldsymbol{z}_0 = \boldsymbol{W}^{1/2}\boldsymbol{z}$ and $\boldsymbol{\eta}_0 = \boldsymbol{W}^{1/2}\boldsymbol{\eta}$. For $\boldsymbol{V} = \text{var}(\boldsymbol{y})$, $\boldsymbol{W} = \boldsymbol{D}\boldsymbol{V}^{-1}\boldsymbol{D}$ and $\boldsymbol{W}^{-1} = \boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1}$. Since $\text{var}(\boldsymbol{z}) = \boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1} = \boldsymbol{W}^{-1}$, it follows that $\text{var}(\boldsymbol{z}_0) = \boldsymbol{I}$. Likewise, let $\hat{\boldsymbol{z}}_0 = \widehat{\boldsymbol{W}}^{1/2}\hat{\boldsymbol{z}}$ and $\hat{\boldsymbol{\eta}}_0 = \widehat{\boldsymbol{W}}^{1/2}\hat{\boldsymbol{\eta}}$. Then

$$\hat{\boldsymbol{\eta}}_0 = \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}\hat{\boldsymbol{\beta}} = \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}(\boldsymbol{X}^{\text{T}}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\text{T}}\widehat{\boldsymbol{W}}\hat{\boldsymbol{z}} = \widehat{\boldsymbol{H}}_{\boldsymbol{W}}\hat{\boldsymbol{z}}_0.$$

So the weighted fitted linear predictor values are the orthogonal projection of the estimated weighted adjusted response variable onto the vector space spanned by the columns of the weighted model matrix $\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}$. The estimated generalized hat matrix $\widehat{\boldsymbol{H}}_{\boldsymbol{W}}$ equals $\boldsymbol{X}_0(\boldsymbol{X}_0^{\text{T}}\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^{\text{T}}$ for the weighted model matrix $\boldsymbol{X}_0 = \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}$.

For the estimated weighted adjusted response, the raw residual is

$$\boldsymbol{e}_0 = \hat{\boldsymbol{z}}_0 - \hat{\boldsymbol{\eta}}_0 = (\boldsymbol{I} - \widehat{\boldsymbol{H}}_{\boldsymbol{W}})\hat{\boldsymbol{z}}_0,$$

so these residuals are orthogonal to the weighted fitted linear predictor values. Also, these residuals equal

$$\boldsymbol{e}_0 = \widehat{\boldsymbol{W}}^{1/2}(\hat{\boldsymbol{z}} - \hat{\boldsymbol{\eta}}) = \widehat{\boldsymbol{W}}^{1/2}\widehat{\boldsymbol{D}}^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = \widehat{\boldsymbol{V}}^{-1/2}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}),$$

which are the Pearson residuals defined in (4.20).

A corresponding approximate version of Pythagoras's theorem states that

$$\|\hat{\boldsymbol{z}}_0 - \boldsymbol{\eta}_0\|^2 \approx \|\hat{\boldsymbol{z}}_0 - \hat{\boldsymbol{\eta}}_0\|^2 + \|\hat{\boldsymbol{\eta}}_0 - \boldsymbol{\eta}_0\|^2 = \|\boldsymbol{e}_0\|^2 + \|\hat{\boldsymbol{\eta}}_0 - \boldsymbol{\eta}_0\|^2.$$

The relation is not exact, because $\boldsymbol{\eta}_0 = \boldsymbol{W}^{1/2}\boldsymbol{X}\boldsymbol{\beta}$ lies in $C(\boldsymbol{W}^{1/2}\boldsymbol{X})$, not $C(\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X})$. Likewise, other decompositions for linear models occur only in an approximate manner for GLMs. For example, Firth (1991) noted that orthogonality of columns of $\boldsymbol{X}$ does not imply orthogonality of corresponding model parameters, except when the link function is such that $\boldsymbol{W}$ is a constant multiple of the identity matrix.

CHAPTER NOTES

Section 4.1: Exponential Dispersion Family Distributions for a GLM

4.1 **Exponential dispersion**: Jørgensen (1987, 1997) developed properties of the exponential dispersion family, including showing a convolution result and approximate normality for small values of the dispersion parameter. Davison (2003, Section 5.2), Morris (1982, 1983a), and Pace and Salvan (1997, Chapters 5 and 6) surveyed properties of exponential family models and their extensions.

4.2 **GLMs**: For more on GLMs, see Davison (2003), Fahrmeir and Tutz (2001), Faraway (2006), Firth (1991), Hastie and Pregibon (1991), Lee et al. (2006), Lovison (2014), Madsen and Thyregod (2011), McCullagh and Nelder (1989), McCulloch et al. (2008), and Nelder and Wedderburn (1972). For asymptotic theory, including conditions for consistency of $\hat{\boldsymbol{\beta}}$, see Fahrmeir and Kaufmann (1985).

Section 4.4: Deviance of a GLM, Model Comparison, and Model Checking

4.3 **Diagnostics**: Cox and Snell (1968) generalized residuals from ordinary linear models, including standardizations. Haberman (1974, Chapter 4) proposed standardized residuals for Poisson models, and Gilchrist (1981) proposed them for GLMs. For other justification for them, see Davison and Snell (1991). Pierce and Schafer (1986) and Williams (1984) evaluated residuals and presented standardized deviance residuals. Lovison (2014) proposed other adjusted residuals and showed their relations with test statistics for comparing nested models. See also Fahrmeir and Tutz (2001, pp. 147–148) and Tutz (2011, Section 3.10). Atkinson and Riani (2000), Davison and Tsai (1992), and Williams (1987) proposed other diagnostic measures for GLMs. Since residuals have limited usefulness for assessing GLMs, Cook and Weisberg (1997) proposed marginal model plots that compare nonparametric smoothings of the data to the model fit, both plotted as a function of characteristics such as individual predictors and the linear predictor values.

4.4 **Score statistics**: For comparing nested models M_0 and M_1, let $\boldsymbol{X}$ be the model matrix for M_1 and let $\boldsymbol{V}(\hat{\boldsymbol{\mu}}_0)$ be the estimated variances of $\boldsymbol{y}$ under M_0. With the canonical link, Lovison (2005) showed that the score statistic is

$$(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^{\mathrm{T}}\boldsymbol{X}[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}(\hat{\boldsymbol{\mu}}_0)\boldsymbol{X}]^{-1}\boldsymbol{X}^{\mathrm{T}}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$$

and this statistic bounds below the $X^2(M_0 \mid M_1)$ statistic in (4.18). Pregibon (1982) showed that the score statistic equals $X^2(M_0) - X^2(M_1)$ when $X^2(M_1)$ uses a one-step approximation to $\hat{\boldsymbol{\mu}}_1$. Pregibon (1982) and Williams (1984) showed that the squared standardized residual is a score statistic for testing whether the observation is an outlier.

Section 4.5: Fitting Generalized Linear Models

4.5 **IRLS**: For more on iteratively reweighted least squares and ML, see Bradley (1973), Green (1984), and Jørgensen (1983). Wood (2006, Chapter 2) illustrated the geometry of GLMs and IRLS.

4.6 **Observed versus expected information**: Fisher scoring has the advantages that it produces the asymptotic covariance matrix as a by-product, the expected information

is necessarily nonnegative-definite, and the method relates to weighted least squares for ordinary linear models. For complex models, the observed information is often simpler to calculate. Efron and Hinkley (1978) argued that observed information has variance estimates that better approximate a relevant conditional variance (conditional on ancillary statistics not relevant to the parameter being estimated), it is "close to the data" rather than averaged over data that could have occurred but did not, and it tends to agree more closely with variances from Bayesian analyses.

Section 4.6: Selecting Explanatory Variables for a GLM

4.7 **Bias–variance tradeoff**: See Davison (2003, p. 405) and James et al. (2013, Section 2.2) for informative discussions of the bias–variance tradeoff.

4.8 **AIC and BIC**: Burnham and Anderson (2010) and Davison (2003, Sections 4.7 and 8.7) justified and illustrated the use of AIC for model comparison and suggested adjustments when n/p is not large. Raftery (1995) showed that differences between BIC values for two models relate to a Bayes factor comparing them. George (2000) presented a brief survey of variable selection methods and cautioned against using a criterion such as minimizing AIC or BIC to select a model.

4.9 **Collinearity**: Other measures besides *VIF* summarize the severity of collinearity and detect the variables involved. A *condition number* is the ratio of largest to smallest eigenvalues of $\boldsymbol{X}$, with large values (e.g., above 30) being problematic. See Belsley et al. (1980) and Rawlings et al. (1998, Chapter 13) for details.

EXERCISES

4.1 Suppose that y_i has a $N(\mu_i, \sigma^2)$ distribution, $i = 1, \ldots, n$. Formulate the normal linear model as a GLM, specifying the random component, linear predictor, and link function.

4.2 Show the exponential dispersion family representation for the gamma distribution (4.29). When do you expect it to be a useful distribution for GLMs?

4.3 Show that the t distribution is not in the exponential dispersion family. (Although GLM theory works out neatly for family (4.1), in practice it is sometimes useful to use other distributions, such as the Cauchy special case of the t.)

4.4 Show that an alternative expression for the GLM likelihood equations is

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\text{var}(y_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0, \quad j = 1, 2, \ldots, p.$$

Show that these equations result from the generalized least squares problem of minimizing $\sum_i [(y_i - \mu_i)^2 / \text{var}(y_i)]$, treating the variances as known constants.

4.5 For a GLM with canonical link function, explain how the likelihood equations imply that the residual vector $\boldsymbol{e} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ is orthogonal with $C(\boldsymbol{X})$.

4.6 Suppose y_i has a Poisson distribution with $g(\mu_i) = \beta_0 + \beta_1 x_i$, where $x_i = 1$ for $i = 1, \ldots, n_A$ from group A and $x_i = 0$ for $i = n_A + 1, ..., n_A + n_B$ from group B, and with all observations being independent. Show that for the log-link function, the GLM likelihood equations imply that the fitted means $\hat{\mu}_A$ and $\hat{\mu}_B$ equal the sample means.

4.7 Refer to the previous exercise. Using the likelihood equations, show that the same result holds for (**a**) *any* link function for this Poisson model, (**b**) any GLM of the form $g(\mu_i) = \beta_0 + \beta_1 x_i$ with a binary indicator predictor.

4.8 For the two-way layout with one observation per cell, consider the model whereby $y_{ij} \sim N(\mu_{ij}, \sigma^2)$ with

$$\mu_{ij} = \beta_0 + \beta_i + \gamma_j + \lambda\beta_i\gamma_j.$$

For independent observations, is this a GLM? Why or why not? (Tukey (1949) proposed a test of H_0: $\lambda = 0$ as a way of testing for interaction; in this setting, after we form the usual interaction SS, the residual SS is 0, so the ordinary test that applies with multiple observations degenerates.)

4.9 Consider the expression for the weight matrix $\boldsymbol{W}$ in $\text{var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^{\text{T}}\boldsymbol{W}\boldsymbol{X})^{-1}$ for a GLM. Find $\boldsymbol{W}$ for the ordinary normal linear model, and show how $\text{var}(\hat{\boldsymbol{\beta}})$ follows from the GLM formula.

4.10 For the normal bivariate linear model, the asymptotic variance of the correlation r is $(1 - \rho^2)^2/n$. Using the delta method, show that the transform $\frac{1}{2}\log[(1 + r)/(1 - r)]$ is variance stabilizing. (Fisher (1921) noted this, showing that $1/(n - 3)$ is an improved variance for the transform.) Explain how to use this result to construct a confidence interval for ρ.

4.11 For a binomial random variable ny with parameter π, consider the null model.

- **a.** Explain how to invert the Wald, likelihood-ratio, and score tests of H_0: $\pi = \pi_0$ against H_1: $\pi \neq \pi_0$ to obtain 95% confidence intervals for π.
- **b.** In teaching an introductory statistics class, one year I collected data from the students to use for lecture examples. One question in the survey asked whether the student was a vegetarian. Of 25 students, 0 said "yes." Treating this as a random sample from some population, find the 95% confidence interval for π using each method in (**a**).
- **c.** Do you trust the Wald interval in (**b**)? (Your answer may depend on whether you regard the standard error estimate for the interval to be credible.) Explain why the Wald method may behave poorly when a parameter takes value near the parameter space boundary.

4.12 For the normal linear model, Section 3.3.2 showed how to construct a confidence interval for $E(y)$ at a fixed $\boldsymbol{x}_0$. Explain how to do this for a GLM.

4.13 For a GLM assuming $y_i \sim N(\mu_i, \sigma^2)$, show that the Pearson chi-squared statistic is the same as the deviance. Find the form of the difference between the deviances for nested models M_0 and M_1.

4.14 In a GLM that uses a noncanonical link function, explain why it need not be true that $\sum_i \hat{\mu}_i = \sum_i y_i$. Hence, the residuals need not have a mean of 0. Explain why a canonical link GLM needs an intercept term in order to ensure that this happens.

4.15 For a binomial GLM, explain why the Pearson residual for observation i, $e_i = (y_i - \hat{\pi}_i)/\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)/n_i}$, does not have an approximate standard normal distribution, even for a large n_i.

4.16 Find the form of the deviance residual (4.21) for an observation in (**a**) a binomial GLM, (**b**) a Poisson GLM.

4.17 Suppose x is uniformly distributed between 0 and 100, and y is binary with $\log[\pi_i/(1-\pi_i)] = -2.0 + 0.04x_i$. Randomly generate $n = 25$ independent observations from this model. Fit the model, and find $\text{corr}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}})$. Do the same for $n = 100$, $n = 1000$, and $n = 10,000$, and summarize how the correlation seems to depend on n.

4.18 Derive the formula $\text{var}(\hat{\beta}_j) = \sigma^2/\{(1-R_j^2)[\sum_i (x_{ij} - \bar{x}_j)^2]\}$.

4.19 Consider the value $\hat{\beta}$ that maximizes a function $L(\beta)$. This exercise motivates the Newton–Raphson method by focusing on the single-parameter case.

a. Using $L'(\hat{\beta}) = L'(\beta^{(0)}) + (\hat{\beta} - \beta^{(0)})L''(\beta^{(0)}) + \cdots$, argue that for an initial approximation $\beta^{(0)}$ close to $\hat{\beta}$, approximately $0 = L'(\beta^{(0)}) + (\hat{\beta} - \beta^{(0)})L''(\beta^{(0)})$. Solve this equation to obtain an approximation $\beta^{(1)}$ for $\hat{\beta}$.

b. Let $\beta^{(t)}$ denote approximation t for $\hat{\beta}$, $t = 0, 1, 2, \ldots$. Justify that the next approximation is

$$\beta^{(t+1)} = \beta^{(t)} - L'(\beta^{(t)})/L''(\beta^{(t)}).$$

4.20 For n independent observations from a Poisson distribution with parameter μ, show that Fisher scoring gives $\mu^{(t+1)} = \bar{y}$ for all $t > 0$. By contrast, what happens with the Newton–Raphson method?

4.21 For an observation y from a Poisson distribution, write a short computer program to use the Newton–Raphson method to maximize the likelihood. With $y = 0$, summarize the effects of the starting value on speed of convergence.

4.22 For noncanonical link functions in a GLM, show that the observed information matrix may depend on the data and hence differs from the expected information

matrix. Thus, the Newton–Raphson method and Fisher scoring may provide different standard errors.

4.23 The bias–variance tradeoff: Before an election, a polling agency randomly samples $n = 100$ people to estimate π = population proportion who prefer candidate A over candidate B. You estimate π by the sample proportion $\hat{\pi}$. I estimate it by $\frac{1}{2}\hat{\pi} + \frac{1}{2}(0.50)$. Which estimator is biased? Which estimator has smaller variance? For what range of π values does my estimator have smaller mean squared error?

4.24 In selecting explanatory variables for a linear model, what is inadequate about the strategy of selecting the model with largest R^2 value?

4.25 For discrete probability distributions of $\{p_j\}$ for the "true" model and $\{p_{Mj}\}$ for a model M, prove that the Kullback–Leibler divergence $E\{\log[p(\boldsymbol{y})/p_M(\boldsymbol{y})]\} \geq 0$.

4.26 For a normal linear model M_1 with $p + 1$ parameters, namely, $\{\beta_j\}$ and σ^2, which has ML estimator $\hat{\sigma}^2 = [\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2]/n$, show that

$$\text{AIC} = n[\log(2\pi\hat{\sigma}^2) + 1] + 2(p + 1).$$

Using this, when M_2 has q additional terms, show that M_2 has smaller AIC value if $\text{SSE}_2/\text{SSE}_1 < e^{-2q/n}$.

4.27 Section 4.7.2 mentioned that using a gamma GLM with log-link function gives similar results to applying a normal linear model to $\log(\boldsymbol{y})$.

a. Use the delta method to show that when y has standard deviation σ proportional to μ (as does the gamma GLM), $\log(y)$ has approximately constant variance for small σ.

b. The gamma GLM with log link refers to $\log[E(y_i)]$, whereas the ordinary linear model for the transformed response refers to $E[\log(y_i)]$. Show that if $\log(y_i) \sim N(\mu_i, \sigma^2)$, then $\log[E(y_i)] = E[\log(y_i)] + \sigma^2/2$.

c. For the lognormal fitted mean L_i for the linear model for $\log(y_i)$, explain why $\exp(L_i)$ is the fitted median for the conditional distribution of y_i. Explain why the fitted median would often be more relevant than the fitted mean of that distribution.

4.28 Download the `Houses.dat` data file from `www.stat.ufl.edu/~aa/glm/data`. Summarize the data with descriptive statistics and plots. Using a forward selection procedure with all five predictors together with judgments about practical significance, select and interpret a linear model for selling price. Check whether results depend on any influential observations.

4.29 Refer to the previous exercise. Use backward elimination to select a model.

a. Use an initial model containing the two-factor interactions. When you reach the stage at which all terms are statistically significant, adjusted R^2 should still be about 0.87. See whether you can simplify further without serious loss of practical significance. Interpret your final model.

b. A simple model for these data has only main effects for size, new, and taxes. Compare your model with this model in terms of adjusted R^2, AIC, and the summaries of effects.

c. If any observations seem to be influential, redo the analyses to analyze their impact.

4.30 Refer to the previous two exercises. Conduct a model-selection process assuming a gamma distribution for y, using (**a**) identity link, (**b**) log link. For each, interpret the final model.

4.31 For the Scottish races data of Section 2.6, the Bens of Jura Fell Race was an outlier for an ordinary linear model with main effects of climb and distance in predicting record times. Alternatively the residual plots might merely suggest increasing variability at higher record times. Fit this model and the corresponding interaction model, assuming a gamma response instead of normal. Interpret results. According to AIC, what is your preferred model for these data?

4.32 Exercise 1.21 presented a study comparing forced expiratory volume after 1 hour of treatment for three drugs (a, b, and p = placebo), adjusting for a baseline measurement x_1. Table 4.1 shows the results of fitting some normal GLMs (with identity link, except one with log link) and a GLM assuming a gamma response. Interpret results.

Table 4.1 Results of Fitting GLMs for Exercise 4.32

Explanatory Variables	R^2	AIC	Fitted Linear Predictor
base	0.393	134.4	$0.95 + .90x_1$
drug	0.242	152.4	$3.49 + .20b - .67p$
base + drug	0.627	103.4	$1.11 + .89x_1 + .22b - .64p$
base + drug (gamma)	0.626	106.2	$0.93 + .97x_1 + .20b - .66p$
base + drug (log link)	0.609	106.8	$0.55 + .25x_1 + .06b - .20p$
base + drug + base:drug	0.628	107.1	$1.33 + .81x_1 - .17b - .91p + .15x_1b + .10x_1p$

4.33 Refer to Exercise 2.45 and the study for comparing instruction methods. Write a report summarizing a model-building process. Include instruction type in the chosen model, because of the study goals and the small n, which results in little power for finding significance for that effect. Check and interpret the final model.

4.34 The horseshoe crab dataset `Crabs2.dat` at the text website comes from a study of factors that affect sperm traits of males. One response variable is ejaculate size, measured as the log of the amount of ejaculate (microliters) measured after 10 seconds of stimulation. Explanatory variables are the location of the observation, carapace width (centimeters), mass (grams), color (1 = dark, 2 = medium, 3 = light), the operational sex ratio (OSR, the number of males per females on the beach), and a subjective condition number that takes into account mucus, pitting on the prosoma, and eye condition (the higher the better). Prepare a report (maximum 4 pages) describing a model-building process for these data. Attach edited software output as an appendix to your report.

4.35 The MASS package of R contains the `Boston` data file, which has several predictors of the median value of owner-occupied homes, for 506 neighborhoods in the suburbs near Boston. Describe a model-building process for these data, using the first 253 observations. Fit your chosen model to the other 253 observations. Compare how well the model fits in the two cases. Attach edited software output in your report.

4.36 For x between 0 and 100, suppose the normal linear model holds with

$$E(y) = 45 + 0.1x + 0.0005x^2 + 0.0000005x^3 + 0.0000000005x^4 + 0.0000000000005x^5$$

and $\sigma = 10.0$. Randomly generate 25 observations from the model, with x having a uniform distribution between 0 and 100. Fit the simple model $E(y) = \beta_0 + \beta_1 x$ and the "correct" model $E(y) = \beta_0 + \beta_1 x + \cdots + \beta_5 x^5$. Construct plots, showing the data, the true relationship, and the model fits. For each model, summarize the quality of the fit by the mean of $|\hat{\mu}_i - \mu_i|$. Summarize, and explain what this exercise illustrates about model parsimony.

4.37 What does the fit of the "correct" model in the previous exercise illustrate about collinearity?

4.38 Randomly generate 100 observations (x_i, y_i) that are independent uniform random variables over [0, 100]. Fit a sequence of successively more complex polynomial models for using x to predict y, of degree 1, 2, 3, ... In principle, even though the true model is $E(y) = 50$ with population $R^2 = 0$, you should be able to fit a polynomial of degree 99 to the data and achieve $R^2 = 1$. Note that when you get to $p \approx 15$, $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})$ is effectively singular and effects of collinearity appear. As p increases, monitor R^2, adjusted R^2, and the P-value for testing significance of the intercept term. Summarize your results.