15 Generalized Linear Models

Due originally to Nelder and Wedderburn (1972), generalized linear models are a remarkable synthesis and extension of familiar regression models such as the linear models described in Part II of this text and the logit and probit models described in the preceding chapter. The current chapter begins with a consideration of the general structure and range of application of generalized linear models; proceeds to examine in greater detail generalized linear models for count data, including contingency tables; briefly sketches the statistical theory underlying generalized linear models; and concludes with the extension of regression diagnostics to generalized linear models.

The unstarred sections of this chapter are perhaps more difficult than the unstarred material in preceding chapters. Generalized linear models have become so central to effective statistical data analysis, however, that it is worth the additional effort required to acquire a basic understanding of the subject.

15.1 The Structure of Generalized Linear Models

A *generalized linear model* (or GLM[1]) consists of three components:

1. A *random component*, specifying the conditional distribution of the response variable, Y_i (for the ith of n independently sampled observations), given the values of the explanatory variables in the model. In Nelder and Wedderburn's original formulation, the distribution of Y_i is a member of an *exponential family*, such as the Gaussian (normal), binomial, Poisson, gamma, or inverse-Gaussian families of distributions. Subsequent work, however, has extended GLMs to multivariate exponential families (such as the multinomial distribution), to certain non-exponential families (such as the two-parameter negative-binomial distribution), and to some situations in which the distribution of Y_i is not specified completely. Most of these ideas are developed later in the chapter.

2. A *linear predictor*—that is a linear function of regressors

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

   As in the linear model, and in the logit and probit models of Chapter 14, the regressors X_{ij} are prespecified functions of the explanatory variables and therefore may include quantitative explanatory variables, transformations of quantitative explanatory variables, polynomial regressors, dummy regressors, interactions, and so on. Indeed, one of the advantages of GLMs is that the structure of the linear predictor is the familiar structure of a linear model.

3. A smooth and invertible linearizing *link function* $g(\cdot)$, which transforms the expectation of the response variable, $\mu_i \equiv E(Y_i)$, to the linear predictor:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

[1] Some authors use the acronym "GLM" to refer to the "*general* linear model"—that is, the linear regression model with normal errors described in Part II of the text—and instead employ "GLIM" to denote *generalized* linear models (which is also the name of a computer program used to fit GLMs).

Table 15.1 Some Common Link Functions and Their Inverses

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Identity	μ_i	η_i
Log	$\log_e \mu_i$	e^{η_i}
Inverse	μ_i^{-1}	η_i^{-1}
Inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
Square-root	$\sqrt{\mu_i}$	η_i^2
Logit	$\log_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
Complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

NOTE: μ_i is the expected value of the response; η_i is the linear predictor; and $\Phi(\cdot)$ is the cumulative distribution function of the standard-normal distribution.

Because the link function is invertible, we can also write

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})$$

and, thus, the GLM may be thought of as a linear model for a transformation of the expected response or as a nonlinear regression model for the response. The inverse link $g^{-1}(\cdot)$ is also called the *mean function*. Commonly employed link functions and their inverses are shown in Table 15.1. Note that the *identity link* simply returns its argument unaltered, $\eta_i = g(\mu_i) = \mu_i$, and thus $\mu_i = g^{-1}(\eta_i) = \eta_i$.

The last four link functions in Table 15.1 are for binomial data, where Y_i represents the observed proportion of "successes" in n_i independent binary trials; thus, Y_i can take on any of the values $0, 1/n_i, 2/n_i, \ldots, (n_i - 1)/n_i, 1$. Recall from Chapter 15 that binomial data also encompass binary data, where all the observations represent $n_i = 1$ trial, and consequently Y_i is either 0 or 1. The expectation of the response $\mu_i = E(Y_i)$ is then the probability of success, which we symbolized by π_i in the previous chapter. The logit, probit, log-log, and complementary log-log links are graphed in Figure 15.1. In contrast to the logit and probit links (which, as we noted previously, are nearly indistinguishable once the variances of the underlying normal and logistic distributions are equated), the log-log and complementary log-log links approach the asymptotes of 0 and 1 asymmetrically.[2]

Beyond the general desire to select a link function that renders the regression of Y on the Xs linear, a promising link will remove restrictions on the range of the expected response. This is a familiar idea from the logit and probit models discussed in Chapter 14, where the object was to model the probability of "success," represented by μ_i in our current general notation. As a probability, μ_i is confined to the unit interval [0,1]. The logit and probit links map this interval to the entire real line, from $-\infty$ to $+\infty$. Similarly, if the response Y is a count, taking on only non-negative integer values, 0, 1, 2, ..., and consequently μ_i is an expected count, which (though not necessarily an integer) is also non-negative, the log link maps μ_i to the whole real line. This is not to say that the choice of link function is entirely determined by the range of the response variable.

[2]Because the log-log link can be obtained from the complementary log-log link by exchanging the definitions of "success" and "failure," it is common for statistical software to provide only one of the two—typically, the complementary log-log link.

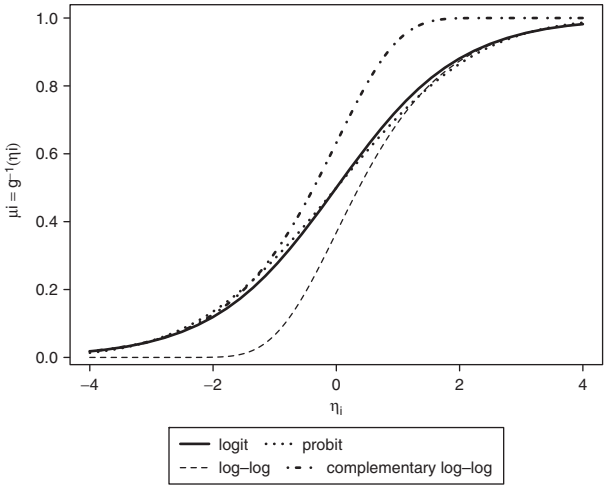

Figure 15.1 Logit, probit, log-log, and complementary log-log links for binomial data. The variances of the normal and logistic distributions have been equated to facilitate the comparison of the logit and probit links [by graphing the cumulative distribution function of $N(0, \pi^2/3)$ for the probit link].

A generalized linear model (or GLM) consists of three components:

1. A random component, specifying the conditional distribution of the response variable, Y_i (for the ith of n independently sampled observations), given the values of the explanatory variables in the model. In the initial formulation of GLMs, the distribution of Y_i was a member of an exponential family, such as the Gaussian, binomial, Poisson, gamma, or inverse-Gaussian families of distributions.
2. A linear predictor—that is a linear function of regressors,

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

3. A smooth and invertible linearizing link function $g(\cdot)$, which transforms the expectation of the response variable, $\mu_i = E(Y_i)$, to the linear predictor:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

A convenient property of distributions in the exponential families is that the conditional variance of Y_i is a function of its mean μ_i [say, $v(\mu_i)$] and, possibly, a *dispersion parameter* ϕ. The variance functions for the commonly used exponential families appear in Table 15.2. The conditional variance of the response in the Gaussian family is a constant, ϕ, which is simply alternative notation for what we previously termed the error variance, σ_ε^2. In the binomial and Poisson families, the dispersion parameter is set to the fixed value $\phi = 1$.

Table 15.2 also shows the range of variation of the response variable in each family, and the so-called *canonical* (or *"natural"*) *link function* associated with each family. The canonical link

Table 15.2 Canonical Link, Response Range, and Conditional Variance Function for Exponential Families

Family	*Canonical Link*	*Range of* Y_i	$V(Y_i\|\eta_i)$
Gaussian	Identity	$(-\infty, +\infty)$	ϕ
Binomial	Logit	$\frac{0,1,\ldots,n_i}{n_i}$	$\frac{\mu_i(1-\mu_i)}{n_i}$
Poisson	Log	0,1,2,...	μ_i
Gamma	Inverse	$(0,\infty)$	$\phi\mu_i^2$
Inverse-Gaussian	Inverse-square	$(0,\infty)$	$\phi\mu_i^3$

NOTE: ϕ is the dispersion parameter, η_i is the linear predictor, and μ_i is the expectation of Y_i (the response). In the binomial family, n_i is the number of trials.

simplifies the GLM,[3] but other link functions may be used as well. Indeed, one of the strengths of the GLM paradigm—in contrast to transformations of the response variable in linear regression—is that the choice of linearizing transformation is partly separated from the distribution of the response, and the same transformation does not have to both normalize the distribution of Y and make its regression on the Xs linear.[4] The specific links that may be used vary from one family to another and also—to a certain extent—from one software implementation of GLMs to another. For example, it would not be promising to use the identity, log, inverse, inverse-square, or square-root links with binomial data, nor would it be sensible to use the logit, probit, log-log, or complementary log-log link with nonbinomial data.

I assume that the reader is generally familiar with the Gaussian and binomial families and simply give their distributions here for reference. The Poisson, gamma, and inverse-Gaussian distributions are perhaps less familiar, and so I provide some more detail:[5]

- The Gaussian distribution with mean μ and variance σ^2 has density function

$$p(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{(y-\mu)^2}{2\sigma^2}\right] \tag{15.1}$$

- The binomial distribution for the proportion Y of successes in n independent binary trials with probability of success μ has probability function

$$p(y) = \binom{n}{ny} \mu^{ny}(1-\mu)^{n(1-y)} \tag{15.2}$$

[3] This point is pursued in Section 15.3.

[4] There is also this more subtle difference: When we transform Y and regress the transformed response on the Xs, we are modeling the expectation of the transformed response,

$$E[g(Y_i)] = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

In a GLM, in contrast, we model the transformed expectation of the response,

$$g[E(Y_i)] = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

While similar in spirit, this is not quite the same thing when (as is true except for the identity link) the link function $g(\cdot)$ is nonlinear.

[5] The various distributions used in this chapter are described in a general context in Appendix D on probability and estimation.

Here, ny is the observed *number* of successes in the n trials, and $n(1-y)$ is the number of failures; and

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

is the binomial coefficient.

- The Poisson distributions are a discrete family with probability function indexed by the *rate parameter* $\mu > 0$:

$$p(y) = \mu^y \times \frac{e^{-\mu}}{y!} \text{ for } y = 0, 1, 2, \ldots$$

The expectation and variance of a Poisson random variable are both equal to μ. Poisson distributions for several values of the parameter μ are graphed in Figure 15.2. As we will see in Section 15.2, the Poisson distribution is useful for modeling count data. As μ increases, the Poisson distribution grows more symmetric and is eventually well approximated by a normal distribution.

- The gamma distributions are a continuous family with probability-density function indexed by the *scale parameter* $\omega > 0$ and *shape parameter* $\psi > 0$:

$$p(y) = \left(\frac{y}{\omega}\right)^{\psi-1} \times \frac{\exp\left(\frac{-y}{\omega}\right)}{\omega\Gamma(\psi)} \text{ for } y > 0 \tag{15.3}$$

where $\Gamma(\cdot)$ is the gamma function.[6] The expectation and variance of the gamma distribution are, respectively, $E(Y) = \omega\psi$ and $V(Y) = \omega^2\psi$. In the context of a generalized linear model, where, for the gamma family, $V(Y) = \phi\mu^2$ (recall Table 15.2 on page 382), the dispersion parameter is simply the inverse of the shape parameter, $\phi = 1/\psi$. As the names of the parameters suggest, the scale parameter in the gamma family influences the spread (and, incidentally, the location) but not the shape of the distribution, while the shape parameter controls the skewness of the distribution. Figure 15.3 shows gamma distributions for scale $\omega = 1$ and several values of the shape parameter ψ. (Altering the scale parameter would change only the labelling of the horizontal axis in the graph.) As the shape parameter gets larger, the distribution grows more symmetric. The gamma distribution is useful for modeling a positive continuous response variable, where the conditional variance of the response grows with its mean but where the *coefficient of variation* of the response, $\text{SD}(Y)/\mu$, is constant.

- The inverse-Gaussian distributions are another continuous family indexed by two parameters, μ and λ, with density function

$$p(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left[-\frac{\lambda(y-\mu)^2}{2y\mu^2}\right] \text{ for } y > 0$$

The expectation and variance of Y are $E(Y) = \mu$ and $V(Y) = \mu^3/\lambda$. In the context of a GLM, where, for the inverse-Gaussian family, $V(Y) = \phi\mu^3$ (as recorded in Table 15.2

[6] * The gamma function is defined as

$$\Gamma(x) = \int_0^\infty e^{-z} z^{x-1} dz$$

and may be thought of as a continuous generalization of the factorial function in that when x is a non-negative integer, $x! = \Gamma(x+1)$.

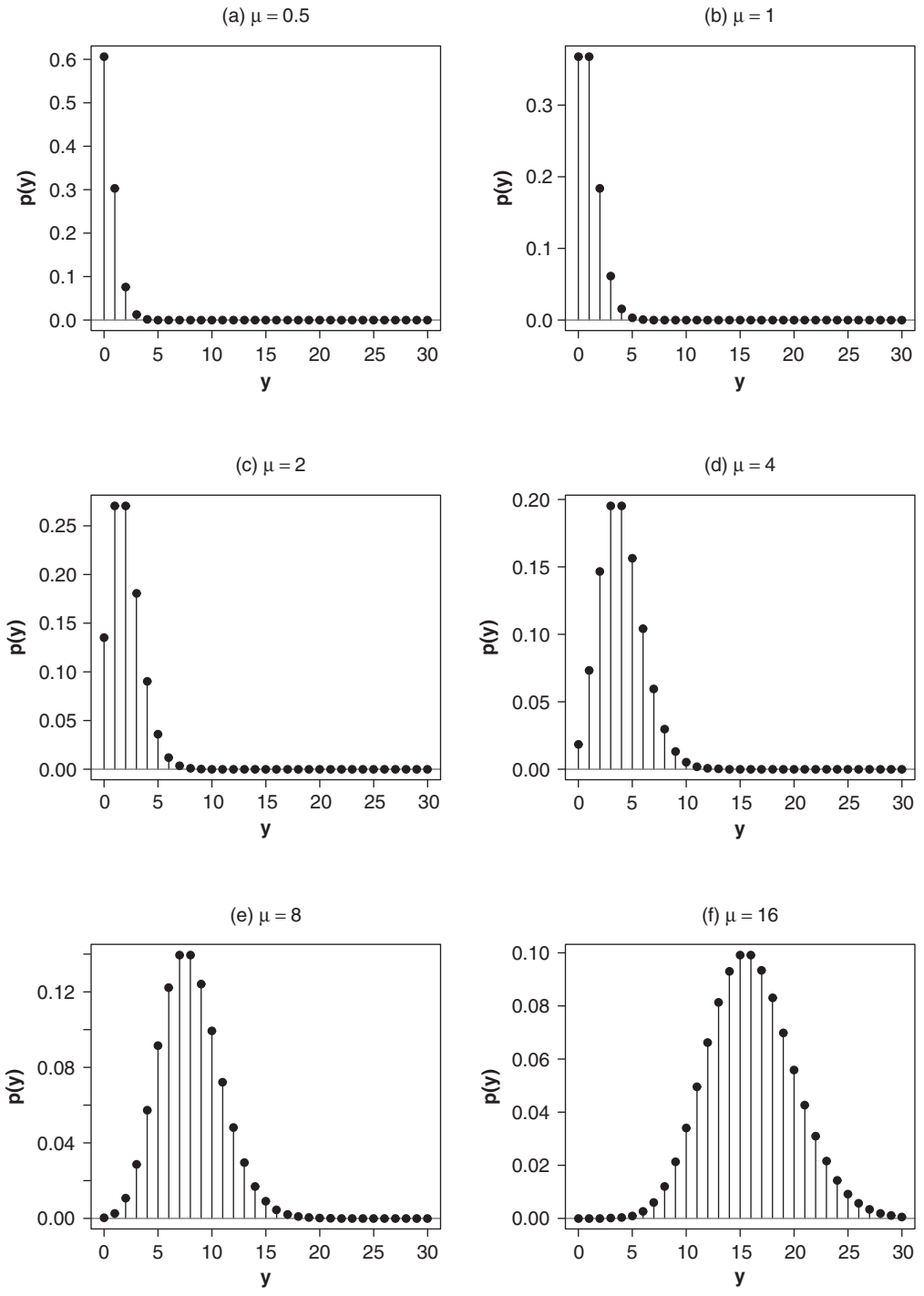

Figure 15.2 Poisson distributions for various values of the rate parameter μ.

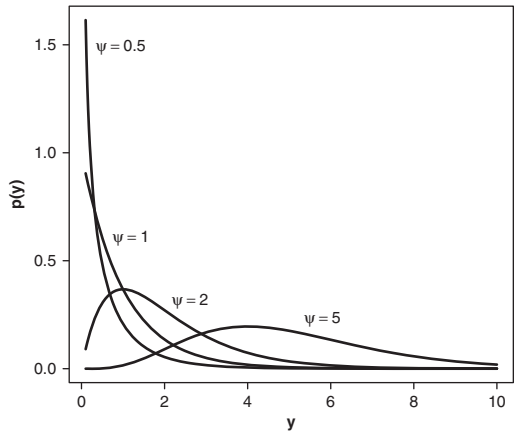

Figure 15.3 Several gamma distributions for scale $\omega = 1$ and various values of the shape parameter ψ.

on page 382), λ is the inverse of the dispersion parameter ϕ. Like the gamma distribution, therefore, the variance of the inverse-Gaussian distribution increases with its mean, but at a more rapid rate. Skewness also increases with the value of μ and decreases with λ. Figure 15.4 shows several inverse-Gaussian distributions.

A convenient property of distributions in the exponential families is that the conditional variance of Y_i is a function of its mean μ_i and, possibly, a dispersion parameter ϕ. In addition to the familiar Gaussian and binomial families (the latter for proportions), the Poisson family is useful for modeling count data, and the gamma and inverse-Gaussian families for modeling positive continuous data, where the conditional variance of Y increases with its expectation.

15.1.1 Estimating and Testing GLMs

GLMs are fit to data by the method of maximum likelihood, providing not only estimates of the regression coefficients but also estimated asymptotic (i.e., large-sample) standard errors of the coefficients.[7] To test the null hypothesis H_0: $\beta_j = \beta_j^{(0)}$ we can compute the Wald statistic $Z_0 = \left(B_j - \beta_j^{(0)}\right)/\text{SE}(B_j)$, where $\text{SE}(B_j)$ is the asymptotic standard error of the estimated coefficient B_j. Under the null hypothesis, Z_0 follows a standard normal distribution.[8]

As explained, some of the exponential families on which GLMs are based include an unknown dispersion parameter ϕ. Although this parameter can, in principle, be estimated by maximum likelihood as well, it is more common to use a "method of moments" estimator, which I will denote $\widetilde{\phi}$.[9]

[7] Details are provided in Section 15.3.2. The method of maximum likelihood is introduced in Appendix D on probability and estimation.

[8] Wald tests and F-tests of more general linear hypotheses are described in Section 15.3.3.

[9] Again, see Section 15.3.2.

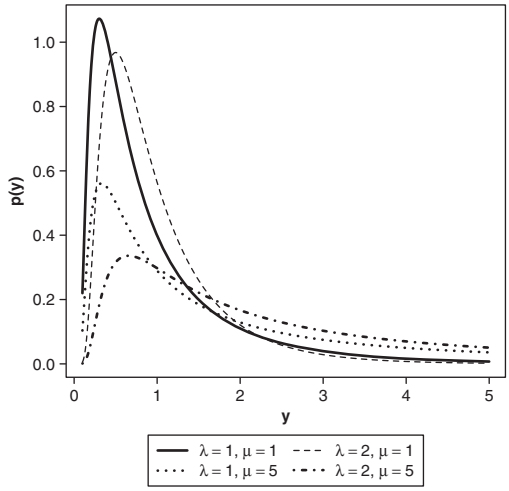

Figure 15.4 Inverse-Gaussian distributions for several combinations of values of the mean μ and inverse-dispersion λ.

As is familiar from the preceding chapter on logit and probit models, the ANOVA for linear models has a close analog in the *analysis of deviance* for GLMs. In the current more general context, the *residual deviance* for a GLM is

$$D_m \equiv 2(\log_e L_s - \log_e L_m)$$

where L_m is the maximized likelihood under the model in question and L_s is the maximized likelihood under a *saturated model*, which dedicates one parameter to each observation and consequently fits the data as closely as possible. The residual deviance is analogous to (and, indeed, is a generalization of) the residual sum of squares for a linear model.

In GLMs for which the dispersion parameter is fixed to 1 (i.e., binomial and Poisson GLMs), the likelihood-ratio test statistic is simply the difference in the residual deviances for nested models. Suppose that Model 0, with $k_0 + 1$ coefficients, is nested within Model 1, with $k_1 + 1$ coefficients (where, then, $k_0 < k_1$); most commonly, Model 0 would simply omit some of the regressors in model 1. We test the null hypothesis that the restrictions on Model 1 represented by Model 0 are correct by computing the likelihood-ratio test statistic

$$G_0^2 = D_0 - D_1$$

Under the hypothesis, G_0^2 is asymptotically distributed as chi-square with $k_1 - k_0$ degrees of freedom.

Likelihood-ratio tests can be turned around to provide confidence intervals for coefficients; as mentioned in Section 14.1.4 in connection with logit and probit models, tests and intervals based on the likelihood-ratio statistic tend to be more reliable than those based on the Wald statistic. For example, the 95% confidence interval for β_j includes all values β_j' for which the hypothesis H_0: $\beta_j = \beta_j'$ is acceptable at the .05 level—that is, all values of β_j' for which $2(\log_e L_1 - \log_e L_0) \leq \chi^2_{.05,1} = 3.84$, where $\log_e L_1$ is the maximized log likelihood for the full model, and $\log_e L_0$ is the maximized log likelihood for a model in which β_j is constrained to the value β_j'. This procedure is computationally intensive because it required "profiling" the likelihood—refitting the model for various fixed values β_j' of β_j.

For GLMs in which there is a dispersion parameter to estimate (Gaussian, gamma, and inverse-Gaussian GLMs), we can instead compare nested models by an F-test,

$$F_0 = \frac{\dfrac{D_0 - D_1}{k_1 - k_0}}{\widetilde{\phi}}$$

where the estimated dispersion $\widetilde{\phi}$, analogous to the estimated error variance for a linear model, is taken from the *largest* model fit to the data (which is not necessarily Model 1). If the largest model has $k + 1$ coefficients, then, under the hypothesis that the restrictions on Model 1 represented by Model 0 are correct, F_0 follows an F-distribution with $k_1 - k_0$ and $n - k - 1$ degrees of freedom. Applied to a Gaussian GLM, this is simply the familiar incremental F-test. The residual deviance divided by the estimated dispersion, $D^* \equiv D/\widetilde{\phi}$, is called the *scaled deviance*.[10]

As we did for logit and probit models,[11] we can base a GLM analog of the squared multiple correlation on the residual deviance: Let D_0 be the residual deviance for the model including only the regression constant α—termed the *null deviance*—and D_1 the residual deviance for the model in question. Then,

$$R^2 \equiv 1 - \frac{D_1}{D_0}$$

represents the proportion of the null deviance accounted for by the model.

GLMs are fit to data by the method of maximum likelihood, providing not only estimates of the regression coefficients but also estimated asymptotic standard errors of the coefficients.

The ANOVA for linear models has an analog in the analysis of deviance for GLMs. The residual deviance for a GLM is $D_m = 2(\log_e L_s - \log_e L_m)$, where L_m is the maximized likelihood under the model in question and L_s is the maximized likelihood under a saturated model. The residual deviance is analogous to the residual sum of squares for a linear model.

In GLMs for which the dispersion parameter is fixed to 1 (binomial and Poisson GLMs), the likelihood-ratio test statistic is the difference in the residual deviances for nested models. For GLMs in which there is a dispersion parameter to estimate (Gaussian, gamma, and inverse-Gaussian GLMs), we can instead compare nested models by an incremental F-test.

15.2 Generalized Linear Models for Counts

The basic GLM for count data is the Poisson model with log link. Consider, by way of example, Michael Ornstein's data on interlocking directorates among 248 dominant Canadian firms, previously discussed in Chapters 3 and 4. The number of interlocks for each firm is the number of ties

[10] Usage is not entirely uniform here, and either of the residual deviance or the scaled deviance is often simply termed "the deviance."

[11] See Section 14.1.4.

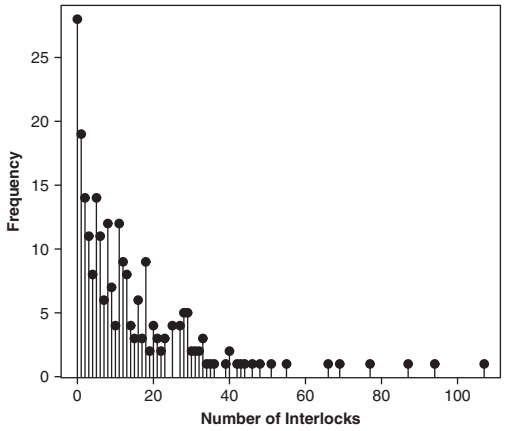

Figure 15.5 The distribution of number of interlocks among 248 dominant Canadian corporations.

that a firm maintained by virtue of its board members and top executives also serving as board members or executives of other firms in the data set. Ornstein was interested in the regression of number of interlocks on other characteristics of the firms—specifically, on their assets (measured in billions of dollars), nation of control (Canada, the United States, the United Kingdom, or another country), and the principal sector of operation of the firm (10 categories, including banking, other financial institutions, heavy manufacturing, etc.).

Examining the distribution of number of interlocks (Figure 15.5) reveals that the variable is highly positively skewed, and that there are many zero counts. Although the conditional distribution of interlocks given the explanatory variables could differ from its marginal distribution, the extent to which the marginal distribution of interlocks departs from symmetry bodes ill for least-squares regression. Moreover, no transformation will spread out the zeroes.[12]

The results of the Poisson regression of number of interlocks on assets, nation of control, and sector are summarized in Table 15.3. I set the *United States* as the baseline category for nation of control, and *Construction* as the baseline category for sector—these are the categories with the smallest fitted numbers of interlocks controlling for the other variables in the regression, and the dummy-regressor coefficients are therefore all positive.

The residual deviance for this model is $D(\text{Assets, Nation, Sector}) = 1887.402$ on $n - k - 1 = 248 - 13 - 1 = 234$ degrees of freedom. Deleting each explanatory variable in turn from the model produces the following residual deviances and degrees of freedom:

Explanatory Variables	*Residual Deviance*	*df*
Nation, Sector	2278.298	235
Assets, Sector	2216.345	237
Assets, Nation	2248.861	243

[12]Ornstein (1976) in fact performed a linear least-squares regression for these data, though one with a slightly different specification from that given here. He cannot be faulted for having done so, however, inasmuch as Poisson regression models—and, with the exception of loglinear models for contingency tables, other specialized models for counts—were not typically in sociologists' statistical toolkit at the time.

Table 15.3 Estimated Coefficients for the Poisson Regression of Number of Interlocks on Assets, Nation of Control, and Sector, for Ornstein's Canadian Interlocking-Directorate Data

Coefficient	*Estimate*	*Standard Error*
Constant	0.8791	0.2101
Assets	0.02085	0.00120
Nation of Control (baseline: United States)		
Canada	0.8259	0.0490
Other	0.6627	0.0755
United Kingdom	0.2488	0.0919
Sector (Baseline: Construction)		
Wood and paper	1.331	0.213
Transport	1.297	0.214
Other financial	1.297	0.211
Mining, metals	1.241	0.209
Holding companies	0.8280	0.2329
Merchandising	0.7973	0.2182
Heavy manufacturing	0.6722	0.2133
Agriculture, food, light industry	0.6196	0.2120
Banking	0.2104	0.2537

Taking differences between these deviances and the residual deviance for the full model yields the following analysis-of-deviance table:

Source	G_0^2	*df*	*p*
Assets	390.90	1	$\ll .0001$
Nation	328.94	3	$\ll .0001$
Sector	361.46	9	$\ll .0001$

All the terms in the model are therefore highly statistically significant.

Because the model uses the log link, we can interpret the exponentiated coefficients (i.e., the e^{B_j}) as multiplicative effects on the expected number of interlocks. Thus, for example, holding nation of control and sector constant, increasing assets by 1 billion dollars (the unit of the assets variable) multiplies the estimated expected number of interlocks by $e^{0.02085} = 1.021$—that is, an increase of just over 2%. Similarly, the estimated expected number of interlocks is $e^{0.8259} = 2.283$ times as high in a Canadian-controlled firm as in a comparable U.S.-controlled firm.

As mentioned, the residual deviance for the full model fit to Ornstein's data is $D_1 = 1887.402$; the deviance for a model fitting only the constant (i.e., the null deviance) is $D_0 = 3737.010$. Consequently, $R^2 = 1 - 1887.402/3737.010 = .495$, revealing that the model accounts for nearly half the deviance in number of interlocks.

The Poisson-regression model is a nonlinear model for the expected response, and I therefore find it generally simpler to interpret the model graphically using effect displays than to examine the estimated coefficients directly. The principles of construction of effect displays for GLMs are essentially the same as for linear models and for logit and probit models:[13] We usually construct one display for each high-order term in the model, allowing the explanatory variables in that

[13] See Section 15.3.4 for details.

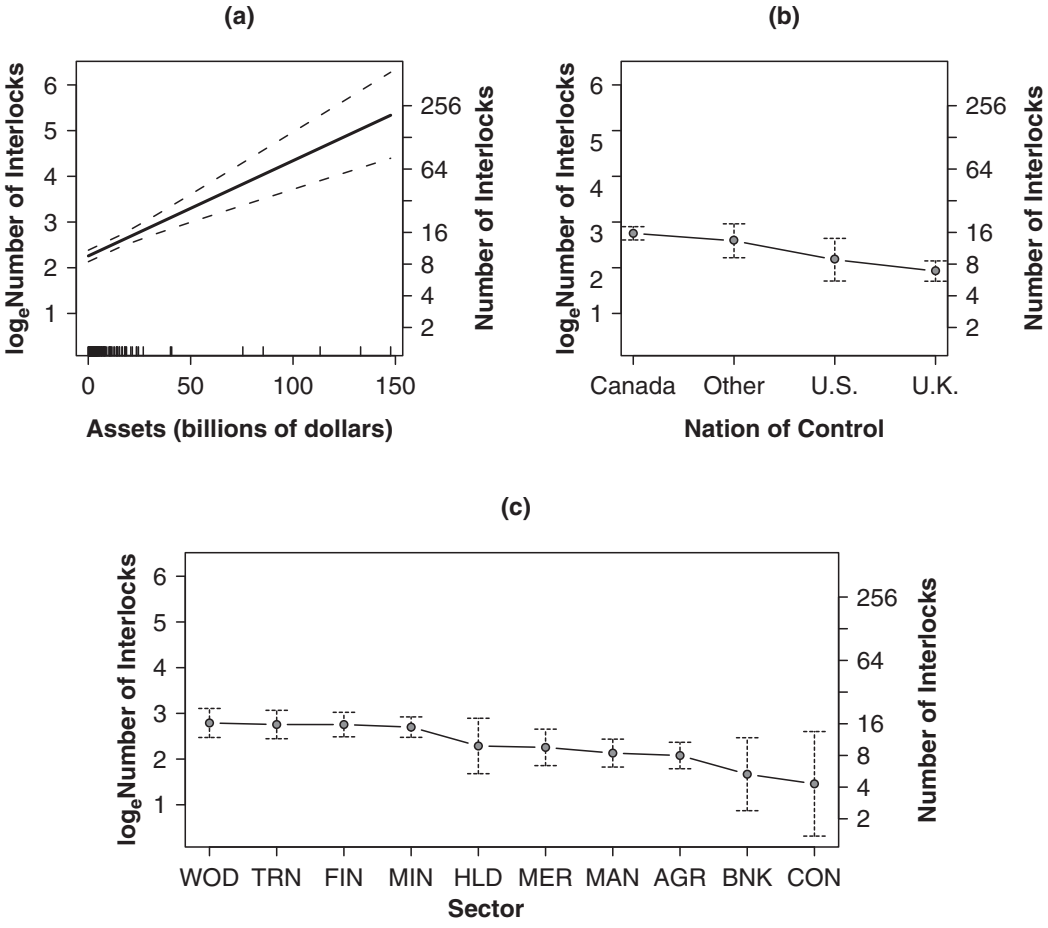

Figure 15.6 Effect displays for (a) assets, (b) nation of control, and (c) sector in the Poisson regression for Ornstein's interlocking-directorate data. The broken lines and error bars give 95% confidence intervals around the fitted effects (computed using the quasi-Poisson model described below). A "rug-plot" at the bottom of panel (a) shows the distribution of assets.

term to range over their values, while holding other explanatory variables in the model to typical values. In a GLM, it is advantageous to plot effects on the scale of the estimated linear predictor, $\widehat{\eta}$, a procedure that preserves the linear structure of the model. In a Poisson model with the log link, the linear predictor is on the log-count scale. We can, however, make the display easier to interpret by relabeling the vertical axis in the scale of the expected response, $\widehat{\mu}$, most informatively by providing a second vertical axis on the right-hand side of the plot. For a Poisson model, the expected response is a count.

Effect displays for the terms in Ornstein's Poisson regression are shown in Figure 15.6. This model has an especially simple structure because each high-order term is a main effect—there are no interactions in the model. The effect display for assets shows a one-dimensional scatterplot (a "rug-plot") for this variable at the bottom of the graph, revealing that the distribution of assets

is highly skewed to the right. Skewness produces some high-leverage observations and suggests the possibility of a nonlinear effect for assets, points that I pursue later in the chapter.[14]

15.2.1 Models for Overdispersed Count Data

The residual deviance for the Poisson regression model fit to the interlocking-directorate data, $D = 1887.4$, is much larger than the 234 residual degrees of freedom for the model. If the Poisson model fits the data reasonably, we would expect the residual deviance to be roughly equal to the residual degrees of freedom.[15] That the residual deviance is so large suggests that the conditional variation of the expected number of interlocks exceeds the variation of a Poisson-distributed variable, for which the variance equals the mean. This common occurrence in the analysis of count data is termed *overdispersion*.[16] Indeed, overdispersion is so common in regression models for count data, and its consequences are potentially so severe, that models such as the quasi-Poisson and negative-binomial GLMs discussed in this section should be employed as a matter of course.

The Quasi-Poisson Model

A simple remedy for overdispersed count data is to introduce a dispersion parameter into the Poisson model, so that the conditional variance of the response is now $V(Y_i|\eta_i) = \phi\mu_i$. If $\phi > 1$, therefore, the conditional variance of Y increases more rapidly than its mean. There is no exponential family corresponding to this specification, and the resulting GLM does not imply a specific probability distribution for the response variable. Rather, the model specifies the conditional mean and variance of Y_i directly. Because the model does not give a probability distribution for Y_i, it cannot be estimated by maximum likelihood. Nevertheless, the usual procedure for maximum-likelihood estimation of a GLM yields the so-called *quasi-likelihood* estimators of the regression coefficients, which share many of the properties of maximum-likelihood estimators.[17]

As it turns out, the quasi-likelihood estimates of the regression coefficients are identical to the ML estimates for the Poisson model. The estimated coefficient standard errors differ, however: If $\widetilde{\phi}$ is the estimated dispersion for the model, then the coefficient standard errors for the *quasi-Poisson model* are $\widetilde{\phi}^{1/2}$ times those for the Poisson model. In the event of overdispersion, therefore, where $\widetilde{\phi} > 1$, the effect of introducing a dispersion parameter and obtaining quasi-likelihood estimates is (realistically) to inflate the coefficient standard errors. Likewise, F-tests for terms in the model will reflect the estimated dispersion parameter, producing smaller test statistics and larger p-values.

As explained in the following section, we use a method-of-moments estimator for the dispersion parameter. In the quasi-Poisson model, the dispersion estimator takes the form

$$\widetilde{\phi} = \frac{1}{n-k-1}\sum\frac{(Y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}$$

[14] See Section 15.4 on diagnostics for GLMs.

[15] That is, the ratio of the residual deviance to degrees of freedom can be taken as an estimate of the dispersion parameter ϕ, which, in a Poisson model, is fixed to 1. It should be noted, however, that this deviance-based estimator of the dispersion can perform poorly. A generally preferable "method of moments" estimator is given in Section 15.3.

[16] Although it is much less common, it is also possible for count data to be *underdispersed*—that is, for the conditional variation of the response to be *less than* the mean. The remedy for underdispsered count data is the same as for overdispersed data; for example, we can fit a quasi-Poisson model with a dispersion parameter, as described immediately below.

[17] See Section 15.3.2.

where $\widehat{\mu}_i = g^{-1}(\widehat{\eta}_i)$ is the fitted expectation of Y_i. Applied to Ornstein's interlocking-directorate regression, for example, we get $\widetilde{\phi} = 7.9435$, and, therefore, the standard errors of the regression coefficients for the Poisson model in Table 15.3 are each multiplied by $\sqrt{7.9435} = 2.818$.

I note in passing that there is a similar *quasi-binomial model* for over-dispersed proportions, replacing the fixed dispersion parameter of 1 in the binomial distribution with a dispersion parameter ϕ to be estimated from the data. Overdispersed binomial data can arise, for example, when different individuals who share the same values of the explanatory variables nevertheless differ in their probability μ of success, a situation that is termed *unmodelled heterogeneity*. Similarly, overdispersion can occur when binomial observations are not independent, as required by the binomial distribution—for example, when each binomial observation is for related individuals, such as members of a family.

The Negative-Binomial Model

There are several routes to models for counts based on the negative-binomial distribution (see, e.g., Long, 1997, sect. 8.3; McCullagh & Nelder, 1989, sect. 6.2.3). One approach (following McCullagh & Nelder, 1989, p. 233) is to adopt a Poisson model for the count Y_i but to suppose that the expected count μ_i^* is itself an unobservable random variable that is gamma-distributed with mean μ_i and constant scale parameter ω (implying that the the gamma shape parameter is $\psi_i = \mu_i/\omega$ [18]). Then the observed count Y_i follows a *negative-binomial distribution*,[19]

$$p(y_i) = \frac{\Gamma(y_i + \omega)}{y!\Gamma(\omega)} \times \frac{\mu_i^{y_i}\omega^{\omega}}{(\mu_i + \omega)^{\mu_i + \omega}} \tag{15.4}$$

with expected value $E(Y_i) = \mu_i$ and variance $V(Y_i) = \mu_i + \mu_i^2/\omega$. Unless the parameter ω is large, therefore, the variance of Y increases more rapidly with the mean than the variance of a Poisson variable. Making the expected value of Y_i a random variable incorporates additional variation among observed counts for observations that share the same values of the explanatory variables and consequently have the same linear predictor η_i.

With the gamma scale parameter ω fixed to a known value, the negative-binomial distribution is an exponential family (in the sense of Equation 15.15 in Section 15.3.1), and a GLM based on this distribution can be fit by iterated weighted least squares (as developed in the next section). If instead—and is typically the case—the value of ω is unknown, and must therefore be estimated from the data, standard methods for GLMs based on exponential families do not apply. We can, however, obtain estimates of both the regression coefficients and ω by the method of maximum likelihood. Applied to Ornstein's interlocking-directorate regression, and using the log link, the negative-binomial GLM produces results very similar to those of the quasi-Poisson model (as the reader may wish to verify). The estimated scale parameter for the negative-binomial model is $\widehat{\omega} = 1.312$, with standard error $\text{SE}(\widehat{\omega}) = 0.143$; we have, therefore, strong evidence that the conditional variance of the number of interlocks increases more rapidly than its expected value.[20]

Zero-Inflated Poisson Regression

A particular kind of overdispersion obtains when there are more zeroes in the data than is consistent with a Poisson (or negative-binomial) distribution, a situation that can arise when only certain members of the population are "at risk" of a nonzero count. Imagine, for example, that

[18] See Equation 15.3 on page 383.
[19] A simpler form of the negative-binomial distribution is given in Appendix D on probability and estimation.
[20] See Exercise 15.1 for a test of overdispersion based on the negative-binomial GLM.

we are interested in modeling the number of children born to a woman. We might expect that this number is a partial function of such explanatory variables as marital status, age, ethnicity, religion, and contraceptive use. It is also likely, however, that some women (or their partners) are infertile and are distinct from fertile women who, though at risk for bearing children, happen to have none. If we knew which women are infertile, we could simply exclude them from the analysis, but let us suppose that this is not the case. To reiterate, there are two sources of zeroes in the data that cannot be perfectly distinguished: women who cannot bear children and those who can but have none.

Several statistical models have been proposed for count data with an excess of zeroes, including the *zero-inflated Poisson regression* (or ZIP) *model*, due to Lambert (1992). The ZIP model consists of two components: (1) A binary logistic-regression model for membership in the *latent class* of individuals for whom the response variable is necessarily 0 (e.g., infertile individuals)[21] and (2) a Poisson-regression model for the latent class of individuals for whom the response may be 0 or a positive count (e.g., fertile women).[22]

Let π_i represent the probability that the response Y_i for the ith individual is necessarily 0. Then

$$\log_e \frac{\pi_i}{1-\pi_i} = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_p z_{ip} \tag{15.5}$$

where the z_{ij} are regressors for predicting membership in the first latent class; and

$$\log_e \mu_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \tag{15.6}$$

$$p(y_i | x_1, \ldots, x_k) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \text{ for } y_i = 0, 1, 2, \ldots$$

where $\mu_i \equiv E(Y_i)$ is the expected count for an individual in the second latent class, and the x_{ij} are regressors for the Poisson submodel. In applications, the two sets of regressors—the Xs and the Zs—are often the same, but this is not necessarily the case. Indeed, a particularly simple special case arises when the logistic submodel is $\log_e \pi_i/(1-\pi_i) = \gamma_0$, a constant, implying that the probability of membership in the first latent class is identical for all observations.

The probability of observing a 0 count is

$$p(0) \equiv \Pr(Y_i = 0) = \pi_i + (1-\pi_i)e^{-\mu_i}$$

and the probability of observing any particular nonzero count y_i is

$$p(y_i) = (1-\pi_i) \times \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

The conditional expectation and variance of Y_i are

$$E(Y_i) = (1-\pi_i)\mu_i$$
$$V(Y_i) = (1-\pi_i)\mu_i(1+\pi_i\mu_i)$$

with $V(Y_i) > E(Y_i)$ for $\pi_i > 0$ [unlike a pure Poisson distribution, for which $V(Y_i) = E(Y_i) = \mu_i$].[23]

[21] See Section 14.1 for a discussion of logistic regression.

[22] Although this form of the zero-inflated count model is the most common, Lambert (1992) also suggested the use of other binary GLMs for membership in the zero latent class (i.e., probit, log-log, and complementary log-log models) and the alternative use of the negative-binomial distribution for the count submodel (see Exercise 15.2).

[23] See Exercise 15.2.

*Estimation of the ZIP model would be simple if we knew to which latent class each observation belongs, but, as I have pointed out, that is not true. Instead, we must maximize the somewhat more complex combined log likelihood for the two components of the ZIP model:[24]

$$\log_e L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{y_i=0} \log_e \left\{ \exp\left(\mathbf{z}_i'\boldsymbol{\gamma}\right) + \exp\left[-\exp(\mathbf{x}_i'\boldsymbol{\beta})\right] \right\} + \sum_{y_i>0} \left[y_i \mathbf{x}_i'\boldsymbol{\beta} - \exp(\mathbf{x}_i'\boldsymbol{\beta}) \right] \quad (15.7)$$
$$- \sum_{i=1}^{n} \log_e \left[1 + \exp(\mathbf{z}_i'\boldsymbol{\gamma})\right] - \sum_{y_i>0} \log_e(y_i!)$$

where $\mathbf{z}_i' \equiv [1, z_{i1}, \ldots, z_{ip}]$, $\mathbf{x}_i' \equiv [1, x_{i1}, \ldots, x_{ik}]$, $\boldsymbol{\gamma} \equiv [\gamma_0, \gamma_1, \ldots, \gamma_p]'$, and $\boldsymbol{\beta} \equiv [\alpha, \beta_1, \ldots, \beta_k]'$.

The basic GLM for count data is the Poisson model with log link. Frequently, however, when the response variable is a count, its conditional variance increases more rapidly than its mean, producing a condition termed overdispersion, and invalidating the use of the Poisson distribution. The quasi-Poisson GLM adds a dispersion parameter to handle overdispersed count data; this model can be estimated by the method of quasi-likelihood. A similar model is based on the negative-binomial distribution, which is not an exponential family. Negative-binomial GLMs can nevertheless be estimated by maximum likelihood. The zero-inflated Poisson regression model may be appropriate when there are more zeroes in the data than is consistent with a Poisson distribution.

15.2.2 Loglinear Models for Contingency Tables

The joint distribution of several categorical variables defines a *contingency table*. As discussed in the preceding chapter,[25] if one of the variables in a contingency table is treated as the response variable, we can fit a logit or probit model (that is, for a dichotomous response, a binomial GLM) to the table. *Loglinear models*, in contrast, which are models for the associations among the variables in a contingency table, treat the variables symmetrically—they do not distinguish one variable as the response. There is, however, a relationship between loglinear models and logit models that I will develop later in this section. As we will see as well, loglinear models have the formal structure of two-way and higher-way ANOVA models[26] and can be fit to data by Poisson regression.

Loglinear models for contingency tables have many specialized applications in the social sciences—for example to "square" tables, such as mobility tables, where the variables in the table have the same categories. The treatment of loglinear models in this section merely scratches the surface.[27]

[24] See Exercise 15.2.
[25] See Section 14.3.
[26] See Sections 8.2 and 8.3.
[27] More extensive accounts are available in many sources, including Agresti (2002), Fienberg (1980), and Powers and Xie (2000).

Table 15.4 Voter Turnout by Intensity of Partisan Preference, for the 1956 U.S. Presidential Election

	Voter Turnout		
Intensity of Preference	*Voted*	*Did Not Vote*	*Total*
Weak	305	126	431
Medium	405	125	530
Strong	265	49	314
Total	975	300	1275

Table 15.5 General Two-Way Frequency Table

	Variable C				
Variable R	1	2	$\cdots$	c	*Total*
1	Y_{11}	Y_{12}	$\cdots$	Y_{1c}	Y_{1+}
2	Y_{21}	Y_{22}	$\cdots$	Y_{2c}	Y_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	Y_{r1}	Y_{r2}	$\cdots$	Y_{rc}	Y_{r+}
Total	Y_{+1}	Y_{+2}	$\cdots$	Y_{+c}	n

Two-Way Tables

I will examine contingency tables for two variables in some detail, for this is the simplest case, and the key results that I establish here extend straightforwardly to tables of higher dimension. Consider the illustrative *two-way table* shown in Table 15.4, constructed from data reported in the *American Voter* (Campbell, Converse, Miller, & Stokes, 1960), introduced in the previous chapter.[28] The table relates intensity of partisan preference to voting turnout in the 1956 U.S. presidential election. To anticipate my analysis, the data indicate that voting turnout is positively associated with intensity of partisan preference.

More generally, two categorical variables with r and c categories, respectively, define an $r \times c$ contingency table, as shown in Table 15.5, where Y_{ij} is the *observed frequency count* in the i, j th cell of the table. I use a "+" to represent summation over a subscript; thus $Y_{i+} \equiv \sum_{j=1}^{c} Y_{ij}$ is the *marginal frequency* in the ith row; $Y_{+j} \equiv \sum_{i=1}^{r} Y_{ij}$ is the marginal frequency in the jth column; and $n = Y_{++} \equiv \sum_{i=1}^{r} \sum_{j=1}^{c} Y_{ij}$ is the number of observations in the sample.

I assume that the n observations in Table 15.5 are independently sampled from a population with proportion π_{ij} in cell i, j, and therefore that the probability of sampling an individual observation in this cell is π_{ij}. Marginal probability distributions π_{i+} and π_{+j} may be defined as above; note that $\pi_{++} = 1$. If the row and column variables are statistically independent in the population, then the joint probability π_{ij} is the product of the marginal probabilities for all i and j: $\pi_{ij} = \pi_{i+}\pi_{+j}$.

Because the observed frequencies Y_{ij} result from drawing a random sample, they are random variables that generally take on different values in different samples. The *expected frequency* in

[28]Table 14.9 (page 371) examined the relationship of voter turnout to intensity of partisan preference *and* perceived closeness of the election. The current example collapses the table for these three variables over the categories of perceived closeness to examine the *marginal table* for turnout and preference. I return below to the analysis of the full three-way table.

cell i, j is $\mu_{ij} \equiv E(Y_{ij}) = n\pi_{ij}$. If the variables are independent, then we have $\mu_{ij} = n\pi_{i+}\pi_{+j}$. Moreover, because $\mu_{i+} = \sum_{j=1}^{c} n\pi_{ij} = n\pi_{i+}$ and $\mu_{+j} = \sum_{i=1}^{r} n\pi_{ij} = n\pi_{+j}$, we may write $\mu_{ij} = \mu_{i+}\mu_{+j}/n$. Taking the log of both sides of this last equation produces

$$\eta_{ij} \equiv \log_e \mu_{ij} = \log_e \mu_{i+} + \log_e \mu_{+j} - \log_e n \tag{15.8}$$

That is, under independence, the log expected frequencies η_{ij} depend additively on the logs of the row marginal expected frequencies, the column marginal expected frequencies, and the sample size. As Fienberg (1980, pp. 13–14) points out, Equation 15.8 is reminiscent of a main-effects two-way ANOVA model, where $-\log_e n$ plays the role of the constant, $\log_e \mu_{i+}$ and $\log_e \mu_{+j}$ are analogous to "main-effect" parameters, and η_{ij} appears in place of the response-variable mean. If we impose ANOVA-like sigma constraints on the model, we may reparametrize Equation 15.8 as follows:

$$\eta_{ij} = \mu + \alpha_i + \beta_j \tag{15.9}$$

where $\alpha_+ \equiv \sum \alpha_i = 0$ and $\beta_+ \equiv \sum \beta_j = 0$. Equation 15.9 is the *loglinear model for independence* in the two-way table. Solving for the parameters of the model, we obtain

$$\begin{aligned} \mu &= \frac{\eta_{++}}{rc} \\ \alpha_i &= \frac{\eta_{i+}}{c} - \mu \\ \beta_j &= \frac{\eta_{+j}}{r} - \mu \end{aligned} \tag{15.10}$$

It is important to stress that although the loglinear model is *formally* similar to an ANOVA model, the *meaning* of the two models differs importantly: In analysis of variance, the α_i and β_j are main-effect parameters, specifying the partial relationship of the (quantitative) response variable to each explanatory variable. The loglinear model in Equation 15.9, in contrast, does not distinguish a response variable, and, because it is a model for independence, specifies that the row and column variables in the contingency table are *unrelated*; for this model, the α_i and β_j merely express the relationship of the log expected cell frequencies to the row and column marginals. The model for independence describes rc expected frequencies in terms of

$$1 + (r-1) + (c-1) = r + c - 1$$

independent parameters.

By analogy to the two-way ANOVA model, we can add parameters to extend the loglinear model to data for which the row and column classifications are not independent in the population but rather are related in an arbitrary manner:

$$\eta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \tag{15.11}$$

where $\alpha_+ = \beta_+ = \gamma_{i+} = \gamma_{+j} = 0$ for all i and j. As before, we may write the parameters of the model in terms of the log expected counts η_{ij}. Indeed, the solution for μ, α_i, and β_j are the same as in Equation 15.10, and

$$\gamma_{ij} = \eta_{ij} - \mu - \alpha_i - \beta_j$$

By analogy to the ANOVA model, the γ_{ij} in the loglinear model are often called "interactions," but this usage is potentially confusing. I will therefore instead refer to the γ_{ij} as *association parameters* because they represent deviations from independence.

Under the model in Equation 15.11, called the *saturated model* for the two-way table, the number of independent parameters is equal to the number of cells in the table,

$$1 + (r-1) + (c-1) + (r-1)(c-1) = rc$$

The model is therefore capable of capturing *any* pattern of association in a two-way table.

Remarkably, maximum-likelihood estimates for the parameters of a loglinear model (that is, in the present case, either the model for independence in Equation 15.9 or the saturated model in Equation 15.11) may be obtained by treating the observed cell counts Y_{ij} as the response variable in a Poisson GLM; the log expected counts η_{ij} are then just the linear predictor for the GLM, as the notation suggests.[29]

The constraint that all $\gamma_{ij} = 0$ imposed by the model of independence can be tested by a likelihood-ratio test, contrasting the model of independence (Equation 15.9) with the more general model (Equation 15.11). Because the latter is a saturated model, its residual deviance is necessarily 0, and the likelihood-ratio statistic for the hypothesis of independence H_0: $\gamma_{ij} = 0$ is simply the residual deviance for the independence model, which has $(r-1)(c-1)$ residual degrees of freedom. Applied to the illustrative two-way table for the *American Voter* data, we get $G_0^2 = 19.428$ with $(3-1)(2-1) = 2$ degrees of freedom, for which $p < .0001$, suggesting that there is strong evidence that intensity of preference and turnout are related.[30]

Maximum-likelihood estimates of the parameters of the saturated loglinear model are shown in Table 15.6. It is clear from the estimated association parameters $\widehat{\gamma}_{ij}$ that turning out to vote, $j = 1$, increases with partisan preference (and, of course, that *not* turning out to vote, $j = 2$, decreases with preference).

Three-Way Tables

The saturated loglinear model for a three-way $(a \times b \times c)$ table for variables A, B, and C is defined in analogy to the three-way ANOVA model, although, as in the case of two-way tables, the meaning of the parameters is different:

$$\eta_{ijk} = \mu + \alpha_{A(i)} + \alpha_{B(j)} + \alpha_{C(k)} + \alpha_{AB(ij)} + \alpha_{AC(ik)} + \alpha_{BC(jk)} + \alpha_{ABC(ijk)} \tag{15.12}$$

---

[29]* The reason that this result is remarkable is that a direct route to a likelihood function for the loglinear model leads to the multinomial distribution (discussed in Appendix D on probability and estimation), not to the Poisson distribution. That is, selecting n independent observations from a population characterized by cell probabilities π_{ij} results in cell counts following the multinomial distribution,

$$\begin{aligned} p(y_{11}, \ldots, y_{rc}) &= \frac{n!}{\prod_{i=1}^{r}\prod_{j=1}^{c} y_{ij}!} \prod_{i=1}^{r}\prod_{j=1}^{c} \pi_{ij}^{n_{ij}} \\ &= \frac{n!}{\prod_{i=1}^{r}\prod_{j=1}^{c} y_{ij}!} \prod_{i=1}^{r}\prod_{j=1}^{c} \left(\frac{\mu_{ij}}{n}\right)^{n_{ij}} \end{aligned}$$

Noting that the expected counts μ_{ij} are functions of the parameters of the loglinear model leads to the multinomial likelihood function for the model. It turns out that maximizing this multinomial likelihood is equivalent to maximizing the likelihood for the Poisson GLM described in the text (see, e.g., Fienberg, 1980, app. II).

[30]This test is very similar to the usual Pearson chi-square test for independence in a two-way table. See Exercise 15.3 for details, and for an alternative formula for calculating the likelihood-ratio test statistic G_0^2 directly from the observed frequencies, Y_{ij}, and estimated expected frequencies under independence, $\widehat{\mu}_{ij}$.

Table 15.6 Estimated Parameters for the Saturated Loglinear Model Fit in Table 15.4

	$\widehat{\gamma}_{ij}$		
i	$j=1$	$j=2$	$\widehat{\alpha}_i$
1	−0.183	0.183	0.135
2	−0.037	0.037	0.273
3	0.219	−0.219	−0.408
$\widehat{\beta}_j$	0.625	−0.625	$\widehat{\mu} = 5.143$

with sigma constraints specifying that each set of parameters sums to zero over each subscript; for example $\alpha_{1(+)} = \alpha_{12(i+)} = \alpha_{123(ij+)} = 0$. Given these constraints, we may solve for the parameters in terms of the log expected counts, with the solution following the usual ANOVA pattern; for example,

$$\begin{aligned}
\mu &= \frac{\eta_{+++}}{abc} \\
\alpha_{A(i)} &= \frac{\eta_{i++}}{bc} - \mu \\
\alpha_{AB(ij)} &= \frac{\eta_{ij+}}{c} - \mu - \alpha_{A(i)} - \alpha_{B(j)} \\
\alpha_{ABC(ijk)} &= \eta_{ijk} - \mu - \alpha_{A(i)} - \alpha_{B(j)} - \alpha_{C(k)} - \alpha_{AB(ij)} - \alpha_{AC(ik)} - \alpha_{BC(jk)}
\end{aligned}$$

The presence of the three-way term $\boldsymbol{\alpha}_{ABC}$ in the model implies that the relationship between any pair of variables (say, A and B) depends on the category of the third variable (say, C).[31]

Other loglinear models are defined by suppressing certain terms in the saturated model, that is, by setting parameters to zero. In specifying a restricted loglinear model, we will be guided by the principle of marginality:[32] Whenever a high-order term is included in the model, its lower-order relatives are included as well. Loglinear models of this type are often called *hierarchical*. Nonhierarchical loglinear models may be suitable for special applications, but they are not sensible in general (see Fienberg, 1980). According to the principle of marginality, for example, if $\boldsymbol{\alpha}_{AB}$ appears in the model, so do $\boldsymbol{\alpha}_A$ and $\boldsymbol{\alpha}_B$.

- If we set all of $\boldsymbol{\alpha}_{ABC}, \boldsymbol{\alpha}_{AB}, \boldsymbol{\alpha}_{AC}$, and $\boldsymbol{\alpha}_{BC}$ to zero, we produce the model of mutual independence, implying that the variables in the three-way table are completely unrelated:

$$\eta_{ijk} = \mu + \alpha_{A(i)} + \alpha_{B(j)} + \alpha_{C(k)}$$

- Setting $\boldsymbol{\alpha}_{ABC}, \boldsymbol{\alpha}_{AC}$, and $\boldsymbol{\alpha}_{BC}$ to zero yields the model

$$\eta_{ijk} = \mu + \alpha_{A(i)} + \alpha_{B(j)} + \alpha_{C(k)} + \alpha_{AB(ij)}$$

  which specifies (1) that variables A and B are related, controlling for (i.e., within categories of) variable C; (2) that this partial relationship is constant across the categories of variable C; and (3) that variable C is independent of variables A and B taken jointly—that is, if we form the two-way table with rows given by combinations of categories of A and B, and columns given by C, the two variables in this table are independent. Note that there are two other models of this sort: one in which $\boldsymbol{\alpha}_{AC}$ is nonzero and another in which $\boldsymbol{\alpha}_{BC}$ is nonzero.

[31] Here and below I use the shorthand notation $\boldsymbol{\alpha}_{ABC}$ to represent the whole set of $\alpha_{ABC(ijk)}$, and similarly for the other terms in the model.

[32] See Section 7.3.2.

Table 15.7 Voter Turnout by Perceived Closeness of the Election and Intensity of Partisan Preference, for the 1956 U.S. Presidential Election

		(C) Turnout	
(A) Perceived Closeness	*(B) Intensity of Preference*	*Voted*	*Did Not Vote*
One-sided	Weak	91	39
	Medium	121	49
	Strong	64	24
Close	Weak	214	87
	Medium	284	76
	Strong	201	25

- A third type of model has *two* nonzero two-way terms; for example, setting $\boldsymbol{\alpha}_{ABC}$ and $\boldsymbol{\alpha}_{BC}$ to zero, we obtain

$$\eta_{ijk} = \mu + \alpha_{A(i)} + \alpha_{B(j)} + \alpha_{C(k)} + \alpha_{AB(ij)} + \alpha_{AC(ik)}$$

  This model implies that (1) variables A and B have a constant partial relationship across the categories of variable C; (2) variables A and C have a constant partial relationship across the categories of variable B; and (3) variables B and C are independent within categories of variable A. Again, there are two other models of this type.
- Finally, consider the model that sets only the three-way term $\boldsymbol{\alpha}_{ABC}$ to zero:

$$\eta_{ijk} = \mu + \alpha_{A(i)} + \alpha_{B(j)} + \alpha_{C(k)} + \alpha_{AB(ij)} + \alpha_{AC(ik)} + \alpha_{BC(jk)}$$

  This model specifies that each pair of variables (e.g., A and B) has a constant partial association across the categories of the remaining variable (e.g., C).

These descriptions are relatively complicated because the loglinear models are models of association among variables. As we will see presently, however, if one of the variables in a table is taken as the response variable, then the loglinear model is equivalent to a logit model with a simpler interpretation.

Table 15.7 shows a three-way table cross-classifying voter turnout by perceived closeness of the election and intensity of partisan preference, elaborating the two-way table for the *American Voter* data presented earlier in Table 15.4.[33] I have fit all hierarchical loglinear models to this three-way table, displaying the results in Table 15.8. Here I employ a compact notation for the high-order terms in each fitted model: For example, AB represents the two-way term $\boldsymbol{\alpha}_{AB}$ and implies that the lower-order relatives of this term—μ, $\boldsymbol{\alpha}_A$, and $\boldsymbol{\alpha}_B$—are also in the model. As in the loglinear model for a two-way table, the saturated model has a residual deviance of 0, and consequently the likelihood-ratio statistic to test any model against the saturated model (within which all of the other models are nested, and which is the last model shown) is simply the residual deviance for the unsaturated model.

The first model in Table 15.8 is the model of complete independence, and it fits the data very poorly. At the other end, the model with high-order terms AB, AC, and BC, which may be used to test the hypothesis of no three-way association, H_0: all $\alpha_{ABC(ijk)} = 0$, also has a statistically significant likelihood-ratio test statistic (though not overwhelmingly so), suggesting that the association between any pair of variables in the contingency tables varies over the levels of the remaining variable.

[33] This table was also discussed in Chapter 14 (see Table 14.9 on page 371).

Table 15.8 Hierarchical Loglinear Models Fit to Table 15.7

High-Order Terms	*Residual Degrees of Freedom*		G_0^2	p
	General	*Table 15.7*		
A,B,C	$(a-1)(b-1)+(a-1)(c-1)(b-1)(c-1)$ $+(a-1)(b-1)(c-1)$	7	36.39	$\ll .0001$
AB,C	$(a-1)(c-1)+(b-1)(c-1)+(a-1)(b-1)(c-1)$	5	34.83	$\ll .0001$
AC,B	$(a-1)(b-1)+(b-1)(c-1)+(a-1)(b-1)(c-1)$	5	16.96	.0046
A,BC	$(a-1)(b-1)+(a-1)(c-1)+(a-1)(b-1)(c-1)$	6	27.78	.0001
AB,AC	$(b-1)(c-1)+(a-1)(b-1)(c-1)$	3	15.40	.0015
AB,BC	$(a-1)(c-1)+(a-1)(b-1)(c-1)$	4	26.22	$<.0001$
AC,BC	$(a-1)(b-1)+(a-1)(b-1)(c-1)$	4	8.35	.079
AB,AC,BC	$(a-1)(b-1)(c-1)$	2	7.12	.028
ABC	0	0	0.0	—

NOTE: The column labeled G_0^2 is the likelihood-ratio statistic for testing each model against the saturated model.

This approach generalizes to contingency tables of any dimension, although the interpretation of high-order association terms can become complicated.

Loglinear Models and Logit Models

As I explained, the loglinear model for a contingency table is a model for association among the variables in the table; the variables are treated symmetrically, and none is distinguished as the response variable. When one of the variables in a contingency table is regarded as the response, however, the loglinear model for the table implies a logit model (identical to the logit model for a contingency table developed in Chapter 14), the parameters of which bear a simple relationship to the parameters of the loglinear model for the table.

For example, it is natural to regard voter turnout in Table 15.7 as a dichotomous response variable, potentially affected by perceived closeness of the election and by intensity of partisan preference. Indeed, this is precisely what we did previously when we analyzed this table using a logit model.[34] With this example in mind, let us return to the saturated loglinear model for the three-way table (repeating Equation 15.12):

$$\eta_{ijk} = \mu + \alpha_{A(i)} + \alpha_{B(j)} + \alpha_{C(k)} + \alpha_{AB(ij)} + \alpha_{AC(ik)} + \alpha_{BC(jk)} + \alpha_{ABC(ijk)}$$

For convenience, I suppose that the response variable is variable C, as in the illustration. Let Ω_{ij} symbolize the response-variable logit within categories i, j of the two explanatory variables; that is,

$$\begin{aligned}\Omega_{ij} &= \log_e \frac{\pi_{ij1}}{\pi_{ij2}} = \log_e \frac{n\pi_{ij1}}{n\pi_{ij2}} = \log_e \frac{\mu_{ij1}}{\mu_{ij2}} \\ &= \eta_{ij1} - \eta_{ij2}\end{aligned}$$

Then, from the saturated loglinear model for η_{ijk},

$$\begin{aligned}\Omega_{ij} = &\left[\alpha_{C(1)} - \alpha_{C(2)}\right] + \left[\alpha_{AC(i1)} - \alpha_{AC(i2)}\right] \\ &+ \left[\alpha_{BC(j1)} - \alpha_{BC(j2)}\right] + \left[\alpha_{ABC(ij1)} - \alpha_{ABC(ij2)}\right]\end{aligned} \tag{15.13}$$

[34] See Section 14.3.

Noting that the first bracketed term in Equation 15.13 does not depend on the explanatory variables, that the second depends only upon variable A, and so forth, let us rewrite this equation in the following manner:

$$\Omega_{ij} = \omega + \omega_{A(i)} + \omega_{B(j)} + \omega_{AB(ij)} \tag{15.14}$$

where, because of the sigma constraints on the αs,

$$\begin{aligned}
\omega &\equiv \alpha_{C(1)} - \alpha_{C(2)} = 2\alpha_{C(1)} \\
\omega_{A(i)} &\equiv \alpha_{AC(i1)} - \alpha_{AC(i2)} = 2\alpha_{AC(i1)} \\
\omega_{B(j)} &\equiv \alpha_{BC(j1)} - \alpha_{BC(j2)} = 2\alpha_{BC(j1)} \\
\omega_{AB(ij)} &\equiv \alpha_{ABC(ij1)} - \alpha_{ABC(ij2)} = 2\alpha_{ABC(ij1)}
\end{aligned}$$

Furthermore, because they are defined as twice the αs, the ωs are also constrained to sum to zero over any subscript:

$$\omega_{A(+)} = \omega_{B(+)} = \omega_{AB(i+)} = \omega_{AB(+j)} = 0, \text{ for all } i \text{ and } j$$

Note that the loglinear-model parameters for the association of the *explanatory* variables A and B do not appear in Equation 15.13. This equation (or, equivalently, Equation 15.14), the saturated logit model for the table, therefore shows how the response-variable log-odds depend on the explanatory variables and their interactions. In light of the constraints that they satisfy, the ωs are interpretable as ANOVA-like effect parameters, and indeed we have returned to the binomial logit model for a contingency table introduced in the previous chapter: Note, for example, that the likelihood-ratio test for the three-way term in the loglinear model for the *American Voter* data (given in the penultimate line of Table 15.8) is identical to the likelihood-ratio test for the interaction between closeness and preference in the logit model fit to these data (see Table 14.11 on page 373).

A similar argument may also be pursued with respect to *any* unsaturated loglinear model for the three-way table: Each such model implies a model for the response-variable logits. Because, however, our purpose is to examine the effects of the explanatory variables on the response, and not to explore the association *between* the explanatory variables, we generally include $\boldsymbol{\alpha}_{AB}$ and its lower-order relatives in *any* model that we fit, thereby treating the association (if any) between variables A and B as given. Furthermore, a similar argument to the one developed here can be applied to a table of any dimension that has a response variable, and to a response variable with more than two categories. In the latter event, the loglinear model is equivalent to a *multinomial* logit model for the table, and in any event, we would generally include in the loglinear model a term of dimension one less than the table corresponding to all associations among the explanatory variables.

> Loglinear models for contingency tables bear a formal resemblance to analysis-of-variance models and can be fit to data as Poisson generalized linear models with a log link. The loglinear model for a contingency table, however, treats the variables in the table symmetrically—none of the variables is distinguished as a response variable—and consequently the parameters of the model represent the associations among the variables, not the effects of explanatory variables on a response. When one of the variables is construed as the response, the loglinear model reduces to a binomial or multinomial logit model.

15.3 Statistical Theory for Generalized Linear Models*

In this section, I revisit with greater rigor and more detail many of the points raised in the preceding sections.[35]

15.3.1 Exponential Families

As much else in modern statistics, the insight that many of the most important distributions in statistics could be expressed in the following common "linear-exponential" form was due to R. A. Fisher:

$$p(y; \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \tag{15.15}$$

where

- $p(y; \theta, \phi)$ is the probability function for the discrete random variable Y, or the probability-density function for continuous Y.
- $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions that vary from one exponential family to another (see below for examples).
- $\theta = g_c(\mu)$, the *canonical parameter* for the exponential family in question, is a function of the expectation $\mu \equiv E(Y)$ of Y; moreover, the *canonical link function* $g_c(\cdot)$ does not depend on ϕ.
- $\phi > 0$ is a *dispersion parameter*, which, in some families, takes on a fixed, known value, while in other families it is an unknown parameter to be estimated from the data along with θ.

Consider, for example, the normal or Gaussian distribution with mean μ and variance σ^2, the density function for which is given in Equation 15.1 (on page 382). To put the normal distribution in the form of Equation 15.15 requires some heroic algebraic manipulation, eventually producing[36]

$$p(y; \theta, \phi) = \exp\left\{\frac{y\theta - \theta^2/2}{\phi} - \frac{1}{2}\left[\frac{y^2}{\phi} + \log_e(2\pi\phi)\right]\right\}$$

with $\theta = g_c(\mu) = \mu$; $\phi = \sigma^2$; $a(\phi) = \phi$; $b(\theta) = \theta^2/2$; and $c(y, \phi) = -\frac{1}{2}\left[y^2/\phi + \log_e(2\pi\phi)\right]$.

Now consider the binomial distribution in Equation 15.2 (page 382), where Y is the proportion of "successes" in n independent binary trials, and μ is the probability of success on an individual trial. Written after more algebraic gymnastics as an exponential family,[37]

$$p(y; \theta, \phi) = \exp\left[\frac{y\theta - \log_e(1 + e^\theta)}{1/n} + \log_e \binom{n}{ny}\right]$$

with $\theta = g_c(\mu) = \log_e[\mu/(1 - \mu)]$; $\phi = 1$; $a(\phi) = 1/n$; $b(\theta) = \log_e(1 + e^\theta)$; and $c(y, \phi) = \log_e \binom{n}{ny}$.

Similarly, the Poisson, gamma, and inverse-Gaussian families can all be put into the form of Equation 15.15, using the results given in Table 15.9.[38]

[35] The exposition here owes a debt to Chapter 2 of McCullagh and Nelder (1989), which has become the standard source on GLMs, and to the remarkly lucid and insightful briefer treatment of the topic by Firth (1991).

[36] See Exercise 15.4.

[37] See Exercise 15.5.

[38] See Exercise 15.6.

Table 15.9 Functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ for Constructing the Exponential Families

Family	$a(\phi)$	$b(\theta)$	$c(y, \phi)$
Gaussian	ϕ	$\theta^2/2$	$-\frac{1}{2}\left[y^2/\phi + \log_e(2\pi\phi)\right]$
Binomial	$1/n$	$\log_e(1+e^\theta)$	$\log_e\binom{n}{ny}$
Poisson	1	e^θ	$-\log_e y!$
Gamma	ϕ	$-\log_e(-\theta)$	$\phi^{-2}\log_e(y/\phi) - \log_e y - \log_e\Gamma(\phi^{-1})$
Inverse-Gaussian	ϕ	$-\sqrt{-2\theta}$	$-\frac{1}{2}\left[\log_e(\pi\phi y^3) + 1/(\phi y)\right]$

NOTE: In this table, n is the number of binomial observations, and $\Gamma(\cdot)$ is the gamma function.

The advantage of expressing diverse families of distributions in the common exponential form is that general properties of exponential families can then be applied to the individual cases. For example, it is true in general that

$$b'(\theta) \equiv \frac{db(\theta)}{d\theta} = \mu$$

and that

$$V(Y) = a(\phi)b''(\theta) = a(\phi)\frac{d^2b(\theta)}{d\theta^2} = a(\phi)v(\mu)$$

leading to the results in Table 15.2 (on page 382).[39] Note that $b'(\cdot)$ is the inverse of the canonical link function. For example, for the normal distribution,

$$\begin{aligned} b'(\theta) &= \frac{d(\theta^2/2)}{d\theta} = \theta = \mu \\ a(\phi)b''(\theta) &= \phi \times 1 = \sigma^2 \\ v(\mu) &= 1 \end{aligned}$$

and for the binomial distribution,

$$\begin{aligned} b'(\theta) &= \frac{d[\log_e(1+e^\theta)]}{d\theta} = \frac{e^\theta}{1+e^\theta} = \frac{1}{1+e^{-\theta}} = \mu \\ a(\phi)b''(\theta) &= \frac{1}{n} \times \left[\frac{e^\theta}{1+e^\theta} - \left(\frac{e^\theta}{1+e^\theta}\right)^2\right] = \frac{\mu(1-\mu)}{n} \\ v(\mu) &= \mu(1-\mu) \end{aligned}$$

The Gaussian, binomial, Poisson, gamma, and inverse-Gaussian distributions can all be written in the common linear-exponential form:

$$p(y; \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right]$$

(*Continued*)

[39] See Exercise 15.7.

(Continued)

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions that vary from one exponential family to another; $\theta = g_c(\mu)$ is the canonical parameter for the exponential family in question; $g_c(\cdot)$ is the canonical link function; and $\phi > 0$ is a dispersion parameter, which takes on a fixed, known value in some families. It is generally the case that $\mu = E(Y) = b'(\theta)$ and that $V(Y) = a(\phi)b''(\theta)$.

15.3.2 Maximum-Likelihood Estimation of Generalized Linear Models

The log likelihood for an individual observation Y_i follows directly from Equation 15.15 (page 402):

$$\log_e L(\theta_i, \phi; Y_i) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(Y_i, \phi)$$

For n independent observations, we have

$$\log_e L(\boldsymbol{\theta}, \phi; \mathbf{y}) = \sum_{i=1}^{n} \frac{Y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(Y_i, \phi) \tag{15.16}$$

where $\boldsymbol{\theta} \equiv \{\theta_i\}$ and $\mathbf{y} \equiv \{Y_i\}$.

Suppose that a GLM uses the link function $g(\cdot)$, so that[40]

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

The model therefore expresses the expected values of the n observations in terms of a much smaller number of regression parameters. To get estimating equations for the regression parameters, we have to differentiate the log likelihood with respect to each coefficient in turn. Let l_i represent the ith component of the log likelihood. Then, by the chain rule,

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \times \frac{d\theta_i}{d\mu_i} \times \frac{d\mu_i}{d\eta_i} \times \frac{\partial \eta_i}{\partial \beta_j} \text{ for } j = 0, 1, \ldots, k \tag{15.17}$$

After some work, we can rewrite Equation 15.17 as[41]

$$\frac{\partial l_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a_i(\phi)v(\mu_i)} \times \frac{d\mu_i}{d\eta_i} \times x_{ij}$$

Summing over observations, and setting the sum to zero, produces the maximum-likelihood estimating equations for the GLM,

$$\sum_{i=1}^{n} \frac{Y_i - \mu_i}{a_i v(\mu_i)} \times \frac{d\mu_i}{d\eta_i} \times x_{ij} = 0, \text{ for } j = 0, 1, \ldots, k \tag{15.18}$$

where $a_i \equiv a_i(\phi)/\phi$ does not depend upon the dispersion parameter, which is constant across observations. For example, in a Gaussian GLM, $a_i = 1$, while in a binomial GLM, $a_i = 1/n_i$.

[40] It is notationally convenient here to write β_0 for the regression constant α.

[41] See Exercise 15.8.

Further simplification can be achieved when $g(\cdot)$ is the canonical link. In this case, the maximum-likelihood estimating equations become

$$\sum_{i=1}^{n} \frac{Y_i x_{ij}}{a_i} = \sum_{i=1}^{n} \frac{\mu_i x_{ij}}{a_i}$$

setting the "observed sum" on the left of the equation to the "expected sum" on the right. We noted this pattern in the estimating equations for logistic-regression models in the previous chapter.[42] Nevertheless, even here the estimating equations are (except in the case of the Gaussian family paired with the identify link) nonlinear functions of the regression parameters and generally require iterative methods for their solution.

Iterative Weighted Least Squares

Let

$$\begin{aligned} Z_i &\equiv \eta_i + (Y_i - \mu_i)\frac{d\eta_i}{d\mu_i} \\ &= \eta_i + (Y_i - \mu_i)g'(\mu_i) \end{aligned}$$

Then

$$E(Z_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

and

$$V(Z_i) = \left[g'(\mu_i)\right]^2 a_i v(\mu_i)$$

If, therefore, we could compute the Z_i, we would be able to fit the model by weighted least-squares regression of Z on the Xs, using the inverses of the $V(Z_i)$ as weights.[43] Of course, this is not the case because we do not know the values of the μ_i and η_i, which, indeed, depend on the regression coefficients that we wish to estimate—that is, the argument is essentially circular. This observation suggested to Nelder and Wedderburn (1972) the possibility of estimating GLMs by *iterative weighted least-squares* (IWLS), cleverly turning the circularity into an iterative procedure:

1. Start with initial estimates of the $\widehat{\mu}_i$ and the $\widehat{\eta}_i = g(\widehat{\mu}_i)$, denoted $\widehat{\mu}_i^{(0)}$ and $\widehat{\eta}_i^{(0)}$. A simple choice is to set $\widehat{\mu}_i^{(0)} = Y_i$.[44]
2. At each iteration l, compute the *working response variable* Z using the values of $\widehat{\mu}$ and $\widehat{\eta}$ from the preceding iteration,

$$Z_i^{(l-1)} = \eta_i^{(l-1)} + \left(Y_i - \mu_i^{(l-1)}\right) g'\left(\mu_i^{(l-1)}\right)$$

[42] See Sections 14.1.5 and 14.2.1.

[43] See Section 12.2.2 for a general discussion of weighted least squares.

[44] In certain settings, starting with $\widehat{\mu}_i^{(0)} = Y_i$ can cause computational difficulties. For example, in a binomial GLM, some of the observed proportions may be 0 or 1—indeed, for binary data, this will be true for *all* the observations—requiring us to divide by 0 or to take the log of 0. The solution is to adjust the starting values, which are in any event not critical, to protect against this possibility. For a binomial GLM, where $Y_i = 0$, we can take $\widehat{\mu}_i^{(0)} = 0.5/n_i$, and where $Y_i = 1$, we can take $\widehat{\mu}_i^{(0)} = (n_i - 0.5)/n_i$. For binary data, then, all the $\widehat{\mu}_i^{(0)}$ are 0.5.

along with weights

$$W_i^{(l-1)} = \frac{1}{\left[g'\left(\mu_i^{(l-1)}\right)\right]^2 a_i v\left(\mu_i^{(l-1)}\right)}$$

3. Fit a weighted least-squares regression of $Z^{(l-1)}$ on the Xs, using the $W^{(l-1)}$ as weights. That is, compute

$$\mathbf{b}^{(l)} = \left(\mathbf{X}'\mathbf{W}^{(l-1)}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{(l-1)}\mathbf{z}^{(l-1)}$$

where $\underset{(k+1\times 1)}{\mathbf{b}^{(l)}}$ is the vector of regression coefficients at the current iteration; $\underset{(n\times k+1)}{\mathbf{X}}$ is (as usual) the model matrix; $\underset{(n\times n)}{\mathbf{W}^{(l-1)}} \equiv \text{diag}\left\{W_i^{(l-1)}\right\}$ is the diagonal weight matrix; and $\underset{(n\times 1)}{\mathbf{z}^{(l-1)}} \equiv \left\{Z_i^{(l-1)}\right\}$ is the working-response vector.

4. Repeat Steps 2 and 3 until the regression coefficients stabilize, at which point $\mathbf{b}$ converges to the maximum-likelihood estimates of the βs.

Applied to the canonical link, IWLS is equivalent to the Newton-Raphson method (as we discovered for a logit model in the previous chapter); more generally, IWLS implements Fisher's "method of scoring."

Estimating the Dispersion Parameter

Note that we do not require an estimate of the dispersion parameter to estimate the regression coefficients in a GLM. Although it is in principle possible to estimate ϕ by maximum likelihood as well, this is rarely done. Instead, recall that $V(Y_i) = \phi a_i v(\mu_i)$. Solving for the dispersion parameter, we get $\phi = V(Y_i)/a_i v(\mu_i)$, suggesting the *method of moments* estimator

$$\widetilde{\phi} = \frac{1}{n-k-1}\sum \frac{(Y_i - \widehat{\mu}_i)^2}{a_i v(\widehat{\mu}_i)} \tag{15.19}$$

The estimated asymptotic covariance matrix of the coefficients is then obtained from the last IWLS iteration as

$$\widehat{\mathcal{V}}(\mathbf{b}) = \widetilde{\phi}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}$$

Because the maximum-likelihood estimator $\mathbf{b}$ is asymptotically normally distributed, $\widehat{\mathcal{V}}(\mathbf{b})$ may be used as the basis for Wald tests of the regression parameters.

The maximum-likelihood estimating equations for generalized linear models take the common form

$$\sum_{i=1}^{n} \frac{Y_i - \mu_i}{a_i v(\mu_i)} \times \frac{d\mu_i}{d\eta_i} \times x_{ij} = 0, \text{ for } j = 0, 1, \ldots, k$$

These equations are generally nonlinear and therefore have no general closed-form solution, but they can be solved by iterated weighted least squares (IWLS). The estimating equations for the coefficients do not involve the dispersion parameter, which (for models in which the dispersion is not fixed) then can be estimated as

$$\widetilde{\phi} = \frac{1}{n-k-1}\sum \frac{(Y_i - \widehat{\mu}_i)^2}{a_i v(\widehat{\mu}_i)}$$

The estimated asymptotic covariance matrix of the coefficients is

$$\widehat{\mathcal{V}}(\mathbf{b}) = \widetilde{\phi}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}$$

where $\mathbf{b}$ is the vector of estimated coefficients and $\mathbf{W}$ is a diagonal matrix of weights from the last IWLS iteration.

Quasi-Likelihood Estimation

The argument leading to IWLS estimation rests only on the linearity of the relationship between $\eta = g(\mu)$ and the Xs, and on the assumption that $V(Y)$ depends in a particular manner on a dispersion parameter and μ. As long as we can express the transformed mean of Y as a linear function of the Xs, and can write down a variance function for Y (expressing the conditional variance of Y as a function of its mean and a dispersion parameter), we can apply the "maximum-likelihood" estimating equations (Equation 15.18 on page 404) and obtain estimates by IWLS—even without committing ourselves to a particular conditional distribution for Y.

This is the method of *quasi-likelihood estimation*, introduced by Wedderburn (1974), and it has been shown to retain many of the properties of maximum-likelihood estimation: Although the quasi-likelihood estimator may not be maximally asymptotically efficient, it is consistent and has the same asymptotic distribution as the maximum-likelihood estimator of a GLM in an exponential family.[45] We can think of quasi-likelihood estimation of GLMs as analogous to least-squares estimation of linear regression models with potentially non-normal errors: Recall that as long as the relationship between Y and the Xs is linear, the error variance is constant, and the observations are independently sampled, the theory underlying OLS estimation applies—although the OLS estimator may no longer be maximally efficient.[46]

The maximum-likelihood estimating equations, and IWLS estimation, can be applied whenever we can express the transformed mean of Y as a linear function of the Xs, and can write the conditional variance of Y as a function of its mean and (possibly) a dispersion parameter—even when we do not specify a particular conditional distribution for Y. The resulting quasi-likelihood estimator shares many of the properties of maximum-likelihood estimators.

[45] See, for example, McCullagh and Nelder (1989, chap. 9) and McCullagh (1991).
[46] See Chapter 9.

15.3.3 Hypothesis Tests

Analysis of Deviance

Originally (in Equation 15.16 on page 404), I wrote the log likelihood for a GLM as a function $\log_e L(\boldsymbol{\theta}, \phi; \mathbf{y})$ of the canonical parameters $\boldsymbol{\theta}$ for the observations. Because $\mu_i = g_c^{-1}(\theta_i)$, for the canonical link $g_c(\cdot)$, we can equally well think of the log likelihood as a function of the expected response, and therefore can write the maximized log likelihood as $\log_e L(\widehat{\boldsymbol{\mu}}, \phi; \mathbf{y})$. If we then dedicate a parameter to each observation, so that $\widehat{\mu}_i = Y_i$ (e.g., by removing the constant from the regression model and defining a dummy regressor for each observation), the log likelihood becomes $\log_e L(\mathbf{y}, \phi; \mathbf{y})$. The *residual deviance* under the initial model is twice the difference in these log likelihoods:

$$\begin{aligned} D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) &\equiv 2[\log_e L(\mathbf{y}, \phi; \mathbf{y}) - \log_e L(\widehat{\boldsymbol{\mu}}, \phi; \mathbf{y})] \\ &= 2\sum_{i=1}^{n}[\log_e L(Y_i, \phi; Y_i) - \log_e L(\widehat{\mu}_i, \phi; Y_i)] \\ &= 2\sum_{i=1}^{n}\frac{Y_i\,[g(Y_i) - g(\widehat{\mu}_i)] - b\,[g(Y_i)] + b\,[g(\widehat{\mu}_i)]}{a_i} \end{aligned} \tag{15.20}$$

Dividing the residual deviance by the estimated dispersion parameter produces the *scaled deviance*, $D^*(\mathbf{y}; \widehat{\boldsymbol{\mu}}) \equiv D(\mathbf{y}; \widehat{\boldsymbol{\mu}})/\widetilde{\phi}$. As explained in Section 15.1.1, deviances are the building blocks of likelihood-ratio and F-tests for GLMs.

Applying Equation 15.20 to the Gaussian distribution, where $g_c(\cdot)$ is the identity link, $a_i = 1$, and $b(\theta) = \theta^2/2$, produces (after some simplification)

$$D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = \sum(Y_i - \widehat{\mu})^2$$

that is, the residual sum of squares for the model. Similarly, applying Equation 15.20 to the binomial distribution, where $g_c(\cdot)$ is the logit link, $a_i = n_i$, and $b(\theta) = \log_e(1 + e^\theta)$, we get (after quite a bit of simplification)[47]

$$D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = 2\sum n_i\left[Y_i \log_e \frac{Y_i}{\widehat{\mu}_i} + (1 - Y_i)\log_e \frac{1 - Y_i}{1 - \widehat{\mu}_i}\right]$$

> The residual deviance for a model is twice the difference in the log likelihoods for the saturated model, which dedicates one parameter to each observation, and the model in question:
>
> $$\begin{aligned} D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) &\equiv 2[\log_e L(\mathbf{y}, \phi; \mathbf{y}) - \log_e L(\widehat{\boldsymbol{\mu}}, \phi; \mathbf{y})] \\ &= 2\sum_{i=1}^{n}\frac{Y_i\,[g(Y_i) - g(\widehat{\mu}_i)] - b\,[g(Y_i)] + b\,[g(\widehat{\mu}_i)]}{a_i} \end{aligned}$$
>
> Dividing the residual deviance by the estimated dispersion parameter produces the scaled deviance, $D^*(\mathbf{y}; \widehat{\boldsymbol{\mu}}) \equiv D(\mathbf{y}; \widehat{\boldsymbol{\mu}})/\widetilde{\phi}$.

[47] See Exercise 15.9, which also develops formulas for the deviance in Poisson, gamma, and inverse-Gaussian models.

Testing General Linear Hypotheses

As was the case for linear models,[48] we can formulate a test for the general linear hypothesis

$$H_0\colon \underset{(q\times k+1)}{\mathbf{L}}\ \underset{(k+1\times 1)}{\boldsymbol{\beta}} = \underset{(q\times 1)}{\mathbf{c}}$$

where the hypothesis matrix $\mathbf{L}$ and right-hand-side vector $\mathbf{c}$ contain pre-specified constants; usually, $\mathbf{c} = \mathbf{0}$. For a GLM, the Wald statistic

$$Z_0^2 = (\mathbf{Lb} - \mathbf{c})' \left[\mathbf{L}\widehat{\mathcal{V}}(\mathbf{b})\,\mathbf{L}'\right]^{-1} (\mathbf{Lb} - \mathbf{c})$$

follows an asymptotic chi-square distribution with q degrees of freedom under the hypothesis. The simplest application of this result is to the Wald statistic $Z_0 = B_j/\mathrm{SE}(B_j)$, testing that an individual regression coefficient is zero. Here, Z_0 follows a standard-normal distribution under H_0: $\beta_j = 0$ (or, equivalently, Z_0^2 follows a chi-square distribution with one degree of freedom).

Alternatively, when the dispersion parameter is estimated from the data, we can calculate the test statistic

$$F_0 = \frac{(\mathbf{Lb} - \mathbf{c})' \left[\mathbf{L}\widehat{\mathcal{V}}(\mathbf{b})\,\mathbf{L}'\right]^{-1} (\mathbf{Lb} - \mathbf{c})}{q}$$

which is distributed as $F_{q,n-k-1}$ under H_0. Applied to an individual coefficient, $t_0 = \pm\sqrt{F_0} = B_j/\mathrm{SE}(B_j)$ produces a t-test on $n - k - 1$ degrees of freedom.

> To test the general linear hypothesis H_0: $\mathbf{L}\boldsymbol{\beta} = \mathbf{c}$, where the hypothesis matrix $\mathbf{L}$ has q rows, we can compute the Wald chi-square test statistic $Z_0^2 = (\mathbf{Lb} - \mathbf{c})' \left[\mathbf{L}\widehat{\mathcal{V}}(\mathbf{b})\mathbf{L}'\right]^{-1} (\mathbf{Lb} - \mathbf{c})$, with q degrees of freedom. Alternatively, if the dispersion parameter is estimated from the data, we can compute the F-test statistic $F_0 = (\mathbf{Lb} - \mathbf{c})' \left[\mathbf{L}\widehat{\mathcal{V}}(\mathbf{b})\,\mathbf{L}'\right]^{-1} (\mathbf{Lb} - \mathbf{c})\,/q$ on q and $n - k - 1$ degrees of freedom.

Testing Nonlinear Hypotheses

It is occasionally of interest to test a hypothesis or construct a confidence interval for a *nonlinear* function of the parameters of a linear or generalized linear model. If the nonlinear function in question is a differentiable function of the regression coefficients, then an approximate asymptotic standard error may be obtained by the *delta method*.[49]

Suppose that we are interested in the function

$$\gamma \equiv f(\boldsymbol{\beta}) = f(\beta_0, \beta_1, \ldots, \beta_k)$$

where, for notational convenience, I have used β_0 to denote the regression constant. The function $f(\boldsymbol{\beta})$ need not use *all* the regression coefficients (see the example below). The

[48] See Section 9.4.4.

[49] The delta method (Rao, 1973) is described in Appendix D on probability and estimation. The method employs a first-order (i.e., linear) Taylor-series approximation to the nonlinear function. The delta method is appropriate here because the maximum-likelihood (or quasi-likelihood) estimates of the coefficients of a GLM are asymptotically normally distributed. Indeed, the procedure described in this section is applicable *whenever* the parameters of a regression model are normally distributed and can therefore be applied in a wide variety of contexts—such as to the nonlinear regression models described in Chapter 17. In small samples, however, the delta-method approximation to the standard error may not be adequate, and the bootstrapping procedures described in Chapter 21 will usually provide more reliable results.

maximum-likelihood estimator of γ is simply $\widehat{\gamma} = f(\widehat{\boldsymbol{\beta}})$ (which, as an MLE, is also asymptotically normal), and the approximate sampling variance of $\widehat{\gamma}$ is then

$$\widehat{V}(\widehat{\gamma}) \approx \sum_{j=0}^{k}\sum_{j'=0}^{k} v_{jj'} \times \frac{\partial \widehat{\gamma}}{\partial \widehat{\beta}_j} \times \frac{\partial \widehat{\gamma}}{\partial \widehat{\beta}_{j'}}$$

where $v_{jj'}$ is the j, j'th element of the estimated asymptotic covariance matrix of the coefficients, $\widehat{V}(\widehat{\boldsymbol{\beta}})$.

To illustrate the application of this result, imagine that we are interested in determining the maximum or minimum value of a quadratic partial regression.[50] Focusing on the partial relationship between the response variable and a particular X, we have an equation of the form

$$E(Y) = \cdots + \beta_1 X + \beta_2 X^2 + \cdots$$

Differentiating this equation with respect to X, we get

$$\frac{dE(Y)}{dX} = \beta_1 + 2\beta_2 X$$

Setting the derivative to 0 and solving for X produces the value at which the function reseaches a minimum (if β_2 is positive) or a maximum (if β_2 is negative),

$$X = -\frac{\beta_1}{2\beta_2}$$

which is a nonlinear function of the regression coefficients β_1 and β_2.

For example, in Section 12.3.1, using data from the Canadian Survey of Labour and Income Dynamics (the "SLID"), I fit a least-squares regression of log wage rate on a quadratic in age, a dummy regressor for sex, and the square of education, obtaining (repeating, and slightly rearranging, Equation 12.7 on page 280):

$$\begin{aligned}\widehat{\log_2 \text{Wages}} = {} & 0.5725 + 0.1198 \times \text{Age} - 0.001230 \times \text{Age}^2 \\ & (0.0834) \quad (0.0046) \quad (0.000059) \\ & + 0.3195 \times \text{Male} + 0.002605 \times \text{Education}^2 \\ & (0.0180) \quad (0.000113) \\ R^2 = {} & .3892\end{aligned}$$

Imagine that we are interested in the age $\gamma \equiv -\beta_1/(2\beta_2)$ at which wages are at a maximum, holding sex and education constant. The necessary derivatives are

$$\begin{aligned}\frac{\partial \widehat{\gamma}}{\partial B_1} &= -\frac{1}{2B_2} = -\frac{1}{2(-0.001230)} = 406.5 \\ \frac{\partial \widehat{\gamma}}{\partial B_2} &= \frac{B_1}{2B_2^2} = \frac{0.1198}{2(-0.001230)^2} = 39{,}593\end{aligned}$$

Our point estimate of γ is

$$\widehat{\gamma} = -\frac{B_1}{2B_2} = -\frac{0.1198}{2 \times 0.001230} = 48.70 \text{ years}$$

[50] See Section 17.1 for a discussion of polynomial regression. The application of the delta method to finding the minimum or maximum of a quadratic curve is suggested by Weisberg (2005, sect. 6.1.2).

The estimated sampling variance of the age coefficient is $\widehat{V}(B_1) = 2.115 \times 10^{-5}$, and of the coefficient of age-squared, $\widehat{V}(B_2) = 3.502 \times 10^{-9}$; the estimated sampling covariance for the two coefficients is $\widehat{C}(B_1, B_2) = -2.685 \times 10^{-7}$. The approximate estimated variance of $\widehat{\gamma}$ is then

$$\begin{aligned}\widehat{\mathcal{V}}(\widehat{\gamma}) \approx & \left(2.115 \times 10^{-5}\right) \times 406.5^2 - \left(2.685 \times 10^{-7}\right) \times 406.5 \times 39,593 \\ & - \left(2.685 \times 10^{-7}\right) \times 406.5 \times 39,593 + \left(3.502 \times 10^{-9}\right) \times 39,593^2 \\ = & \ 0.3419\end{aligned}$$

Consequently, the approximate standard error of $\widehat{\gamma}$ is $\text{SE}(\widehat{\gamma}) \approx \sqrt{0.3419} = 0.5847$, and an approximate 95% confidence interval for the age at which income is highest on average is $\gamma = 48.70 \pm 1.96(0.5847) = (47.55, 49.85)$.

The delta method may be used to approximate the standard error of a nonlinear function of regression coefficients in a GLM. If $\gamma \equiv f(\beta_0, \beta_1, \ldots, \beta_k)$, then

$$\widehat{\mathcal{V}}(\widehat{\gamma}) \approx \sum_{j=0}^{k} \sum_{j'=0}^{k} v_{jj'} \frac{\partial \widehat{\gamma}}{\partial \widehat{\beta}_j} \frac{\partial \widehat{\gamma}}{\partial \widehat{\beta}_{j'}}$$

15.3.4 Effect Displays

Let us write the GLM in matrix form, with linear predictor

$$\underset{(n\times 1)}{\boldsymbol{\eta}} = \underset{(n\times k+1)}{\mathbf{X}} \underset{(k+1\times 1)}{\boldsymbol{\beta}}$$

and link function $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$, where $\boldsymbol{\mu}$ is the expectation of the response vector $\mathbf{y}$. As described in Section 15.3.2, we compute the maximum-likelihood estimate $\mathbf{b}$ of $\boldsymbol{\beta}$, along with the estimated asymptotic covariance matrix $\widehat{\mathcal{V}}(\mathbf{b})$ of $\mathbf{b}$.

Let the rows of $\mathbf{X}^*$ include regressors corresponding to all combinations of values of explanatory variables appearing in a high-order term of the model (or, for a continuous explanatory variable, values spanning the range of the variable), along with typical values of the remaining regressors. The structure of $\mathbf{X}^*$ with respect to interactions, for example, is the same as that of the model matrix $\mathbf{X}$. Then the fitted values $\widehat{\boldsymbol{\eta}}^* = \mathbf{X}^*\mathbf{b}$ represent the high-order term in question, and a table or graph of these values—or, alternatively, of the fitted values transformed to the scale of the response variable, $g^{-1}(\widehat{\boldsymbol{\eta}}^*)$—is an effect display. The standard errors of $\widehat{\boldsymbol{\eta}}^*$, available as the square-root diagonal entries of $\mathbf{X}^*\widehat{\mathcal{V}}(\mathbf{b})\mathbf{X}^{*\prime}$, may be used to compute pointwise confidence intervals for the effects, the end-points of which may then also be transformed to the scale of the response.

For example, for the Poisson regression model fit to Ornstein interlocking-directorate data, the effect display for assets in Figure 15.6(a) (page 390) is constructed by letting assets range between its minimum value of 0.062 and maximum of 147.670 billion dollars, fixing the dummy variables for nation of control and sector to their sample means—that is, to the observed proportions of the data in each of the corresponding categories of nation and sector. As noted previously, this is an especially simple example, because the model includes no interactions. The model was fit with the log link, and so the estimated effects, which in general are on the scale of the linear predictor, are on the log-count scale; the right-hand axis of the graph shows the corresponding count scale, which is the scale of the response variable.

Effect displays for GLMs are based on the fitted values $\widehat{\boldsymbol{\eta}}^* = \mathbf{X}^*\mathbf{b}$, representing a high-order term in the model; that is, $\mathbf{X}^*$ has the same general structure as the model matrix $\mathbf{X}$, with the explanatory variables in the high-term order ranging over their values in the data while other explanatory variables are set to typical values. The standard errors of $\widehat{\boldsymbol{\eta}}^*$, given by the square-root diagonal entries of $\mathbf{X}^*\widehat{\mathcal{V}}(\mathbf{b})\mathbf{X}^{*\prime}$, may be used to compute pointwise confidence intervals for the effects.

15.4 Diagnostics for Generalized Linear Models

Most of the diagnostics for linear models presented in Chapters 11 and 12 extend relatively straightforwardly to GLMs. These extensions typically take advantage of the computation of maximum-likelihood and quasi-likelihood estimates for GLMs by iterated weighted least squares, as described in Section 15.3.2. The final weighted-least-squares fit linearizes the model and provides a quadratic approximation to the log likelihood. Approximate diagnostics are then either based directly on the WLS solution or are derived from statistics easily calculated from this solution. Seminal work on the extension of linear least-squares diagnostics to GLMs was done by Pregibon (1981), Landwehr, Pregibon, and Shoemaker (1984), Wang (1985, 1987), and Williams (1987). In my experience, and with the possible exception of added-variable plots for non-Gaussian GLMs, these extended diagnostics typically work reasonably well.

15.4.1 Outlier, Leverage, and Influence Diagnostics

Hat-Values

Hat-values, h_i, for a GLM can be taken directly from the final iteration of the IWLS procedure for fitting the model,[51] and have the usual interpretation—except that, unlike in a linear model, the hat-values in a GLM depend on the response variable Y as well as on the configuration of the Xs.

Residuals

Several kinds of residuals can be defined for GLMs:

- Most straightforwardly (but least usefully), *response residuals* are simply the differences between the observed response and its estimated expected value: $Y_i - \widehat{\mu}_i$, where

$$\widehat{\mu}_i = g^{-1}(\widehat{\eta}_i) = g^{-1}(A + B_1X_{i1} + B_2X_{i2} + \cdots + B_kX_{ik})$$

- *Working residuals* are the residuals from the final WLS fit. These may be used to define partial residuals for component-plus-residual plots (see below).

[51]* The hat-matrix is

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$$

where $\mathbf{W}$ is the weight matrix from the final IWLS iteration.

- *Pearson residuals* are casewise components of the *Pearson goodness-of-fit statistic* for the model:[52]

$$\frac{\widetilde{\phi}^{1/2}(Y_i - \widehat{\mu}_i)}{\sqrt{\widehat{V}(Y_i|\eta_i)}}$$

  where $\widetilde{\phi}$ is the estimated dispersion parameter for the model (Equation 15.19 on page 406) and $V(y_i|\eta_i)$ is the conditional variance of the response (given in Table 15.2 on page 382).
- *Standardized Pearson residuals* correct for the conditional response variation and for the differential leverage of the observations:

$$R_{Pi} \equiv \frac{Y_i - \widehat{\mu}_i}{\sqrt{\widehat{V}(Y_i|\eta_i)(1 - h_i)}}$$

- *Deviance residuals*, G_i, are the square-roots of the casewise components of the residual deviance (Equation 15.20 on page 408), attaching the sign of the corresponding response residual.
- *Standardized deviance residuals* are

$$R_{Gi} \equiv \frac{G_i}{\sqrt{\widetilde{\phi}(1 - h_i)}}$$

- Several different approximations to studentized residuals have been proposed. To calculate exact studentized residuals would require literally refitting the model deleting each observation in turn and noting the decline in the deviance; this procedure, of course, is computationally unattractive. Williams suggests the approximation

$$E_i^* \equiv \sqrt{(1 - h_i)R_{Gi}^2 + h_i R_{Pi}^2}$$

  where, once again, the sign is taken from the response residual. A Bonferroni outlier test using the standard normal distribution may be based on the largest absolute studentized residual.

Influence Measures

An approximation to Cook's distance influence measure is

$$D_i \equiv \frac{R_{Pi}^2}{\widetilde{\phi}(k + 1)} \times \frac{h_i}{1 - h_i}$$

This is essentially Williams's definition, except that I divide by the estimated dispersion $\widetilde{\phi}$ to scale D_i as an F-statistic rather than as a chi-square statistic.

Approximate values of influence measures for individual coefficients, DFBETA_{ij} and DFBETAS_{ij}, may be obtained directly from the final iteration of the IWLS procedure.

Wang (1985) suggests an extension of added-variable plots to GLMs that works as follows: Suppose that the focal regressor is X_j. Refit the model with X_j removed, extracting the working residuals from this fit. Then regress X_j on the other Xs by WLS, using the weights from the last IWLS step, obtaining residuals. Finally, plot the working residuals from the first regression against the residuals for X_j from the second regression.

[52] The Pearson statistic, an alternative to the deviance for measuring the fit of the model to the data, is the sum of squared Pearson residuals.

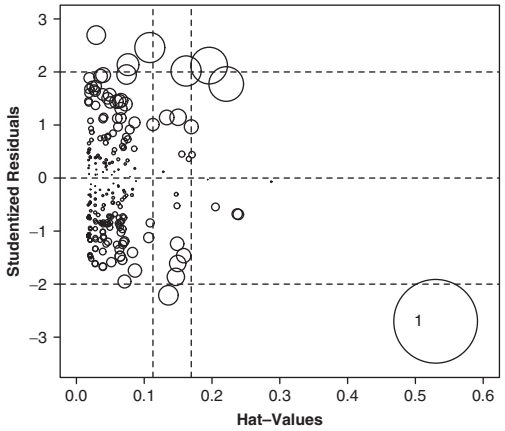

Figure 15.7 Hat-values, studentized residuals, and Cook's distances from the quasi-Poisson regression for Ornstein's interlocking-directorate data. The areas of the circles are proportional to the Cook's distances for the observations. Horizontal lines are drawn at −2, 0, and 2 on the studentized-residual scale, vertical lines at twice and three times the average hat-value.

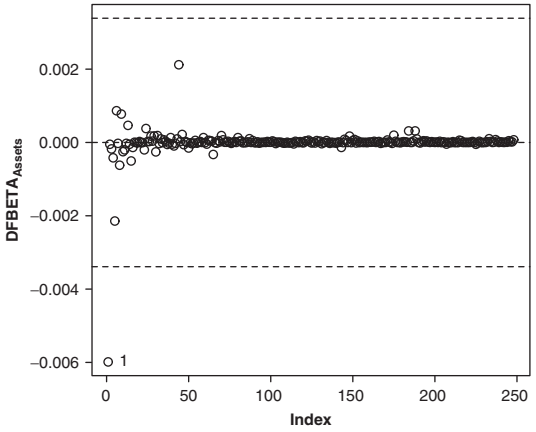

Figure 15.8 Index plot of DFBETA for the assets coefficient. The horizontal lines are drawn at 0 and $\pm\text{SE}(B_{\text{Assets}})$.

Figure 15.7 shows hat-values, studentized residuals, and Cook's distances for the quasi-Poisson model fit to Ornstein's interlocking directorate data. One observation—Number 1, the corporation with the largest assets—stands out by combining a very large hat-value with the biggest absolute studentized residual.[53] This point is not a statistically significant outlier, however (indeed, the Bonferroni p-value for the largest studentized residual exceeds 1). As shown in the DFBETA plot in Figure 15.8, Observation 1 makes the coefficient of assets substantially smaller than it would otherwise be (recall that the coefficient for assets is 0.02085). [54] In this case, the approximate DFBETA is quite accurate: If Observation 1 is deleted, the assets coefficient increases to 0.02602.

[53] Unfortunately, the data source does not include the names of the firms, but Observation 1 is the largest of the Canadian banks, which, in the 1970s, was (I believe) the Royal Bank of Canada.

[54] I invite the reader to plot the DFBETA values for the other coefficients in the model.

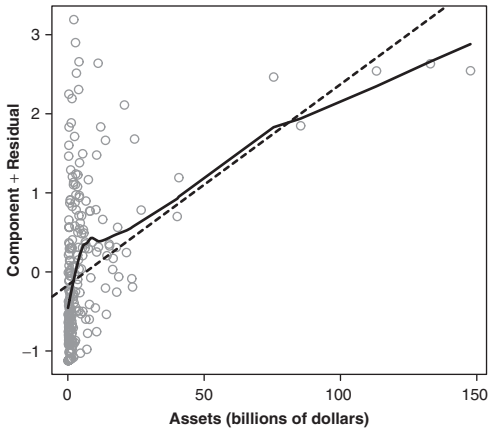

Figure 15.9 Component-plus-residual plot for assets in the interlocking-directorate quasi-Poisson regression. The broken line shows the least-squares fit to the partial residuals; the solid line is for a nonrobust lowess smooth with a span of 0.9.

Before concluding that Observation 1 requires special treatment, however, consider the check for nonlinearity in the next section.

15.4.2 Nonlinearity Diagnostics

Component-plus-residual and CERES plots also extend straightforwardly to GLMs. Nonparametric smoothing of the resulting scatterplots can be important to interpretation, especially in models for binary response variables, where the discreteness of the response makes the plots difficult to examine. Similar (if typically less extreme) effects can occur for binomial and count data.

Component-plus-residual and CERES plots use the linearized model from the last step of the IWLS fit. For example, the partial residual for X_j adds the working residual to $B_j X_{ij}$; the component-plus-residual plot then graphs the partial residual against X_j. In smoothing a component-plus-residual plot for a non-Gaussian GLM, it is generally preferable to use a nonrobust smoother.

A component-plus-residual plot for assets in the quasi-Poisson regression for the interlocking-directorate data is shown in Figure 15.9. Assets is so highly positively skewed that the plot is different to examine, but it is nevertheless apparent that the partial relationship between number of interlocks and assets is nonlinear, with a much steeper slope at the left than at the right. Because the bulge points to the left, we can try to straighten this relationship by transforming assets down the ladder of power and roots. Trial and error suggests the log transformation of assets, after which a component-plus-residual plot for the modified model (Figure 15.10) is unremarkable.

Box-Tidwell constructed-variable plots[55] also extend straightforwardly to GLMs: When considering the transformation of X_j, simply add the constructed variable $X_j \log_e X_j$ to the model and examine the added-variable plot for the constructed variable. Applied to assets in Ornstein's quasi-Poisson regression, this procedure produces the constructed-variable plot in Figure 15.11, which suggests that evidence for the transformation is spread throughout the data. The coefficient for assets $\times \log_e$assets in the constructed-variable regression is -0.02177 with a standard error of 0.00371; the Wald-test statistic $Z_0 = -0.02177/0.00371 = -5.874$ therefore indicates strong evidence for the transformation of assets. By comparing the coefficient of assets in the *original*

[55] See Section 12.5.2.

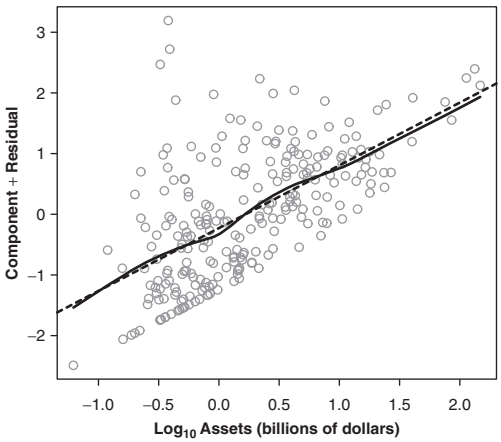

Figure 15.10 Component-plus-residual plot following the log-transformation of assets. The lowess fit is for a span of 0.6.

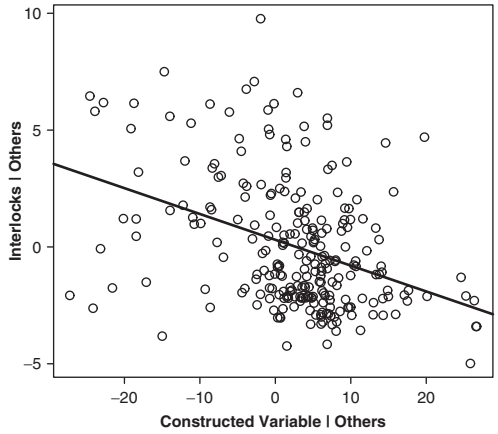

Figure 15.11 Constructed variable plot for the transformation of assets in the interlocking-directorate quasi-Poisson regression.

quasi-Poisson regression (0.02085) with the coefficient of the constructed variable, we get the suggested power transformation

$$\widetilde{\lambda} = 1 + \frac{-0.02177}{0.02085} = -0.044$$

that is, essentially the log-transformation, $\lambda = 0$.

Finally, it is worth noting the relationship between the problems of influence and nonlinearity in this example: Observation 1 was influential in the original regression because its very large assets gave it high leverage and because unmodelled nonlinearity put the observation below the erroneously linear fit for assets, pulling the regression surface towards it. Log-transforming assets fixes both these problems.

Alternative effect displays for assets in the transformed model are shown in Figure 15.12. Panel (a) in this figure graphs assets on its "natural" scale; on this scale, of course, the fitted partial relationship between log-interlocks and assets is nonlinear. Panel (b) uses a log scale for assets, rendering the partial relationship linear.

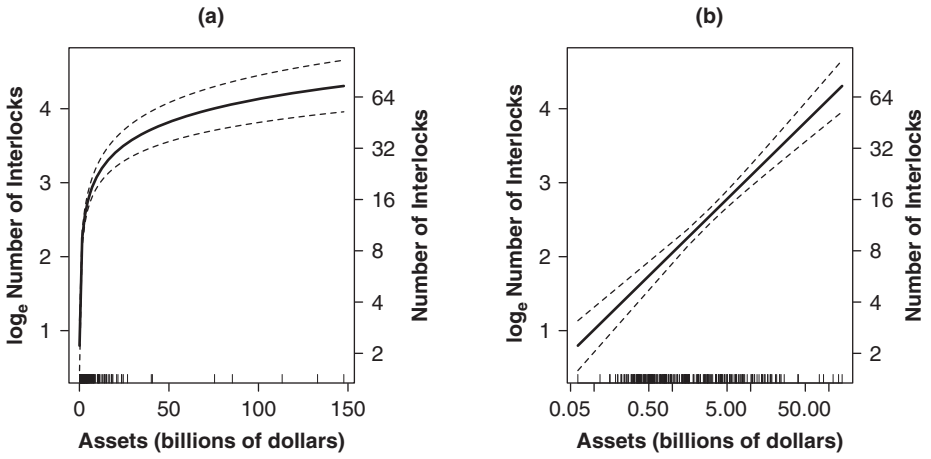

Figure 15.12 Effect displays for assets in the quasi-Poisson regression model in which assets has been log-transformed. Panel (a) plots assets on its "natural" scale, while panel (b) uses a log scale for assets. Rug plots for assets appear at the bottom of the graphs. The broken lines give pointwise 95% confidence intervals around the estimated effect.

Most of the standard diagnostics for linear models extend relatively straightforwardly to GLMs. These extensions typically take advantage of the computation of maximum-likelihood and quasi-likelihood estimates for GLMs by iterated weighted least squares. Such diagnostics include studentized residuals, hat-values, Cook's distances, DFBETA and DFBETAS, added-variable plots, component-plus-residual plots, and the constructed-variable plot for transforming an explanatory variable.

Exercises

Exercise 15.1. Testing overdisperison: Let $\delta \equiv 1/\omega$ represent the inverse of the scale parameter for the negative-binomial regression model (see Equation 15.4 on page 392). When $\delta = 0$, the negative-binomial model reduces to the Poisson regression model (why?), and consequently a test of H_0: $\delta = 0$ against the one-sided alternative hypothesis H_a: $\delta > 0$ is a test of overdispersion. A Wald test of this hypothesis is straightforward, simply dividing $\widehat{\delta}$ by its standard error. We can also compute a likelihood-ratio test contrasting the deviance under the more specific Poisson regression model with that under the more general negative-binomial model. Because the negative-binomial model has one additional parameter, we refer the likelihood-ratio test statistic to a chi-square distribution with one degree of freedom; as Cameron and Trivedi (1998, p. 78) explain, however, the usual right-tailed p-value obtained from the chi-square distribution must be halved. Apply this likelihood-ratio test for overdispersion to Ornstein's interlocking-directorate regression.

Exercise 15.2. *Zero-inflated count regression models:

(a) Show that the mean and variance of the response variable Y_i in the zero-inflated Poisson (ZIP) regression model, given in Equations 15.5 and 15.6 on page 393, are

$$E(Y_i) = (1 - \pi_i)\mu_i$$
$$V(Y_i) = (1 - \pi_i)\mu_i(1 + \pi_i\mu_i)$$

(*Hint*: Recall that there are two sources of zeroes: observations in the first latent class, whose value of Y_i is *necessarily* 0, and observations in the second latent class, whose value *may be* zero. Probability of membership in the first class is π_i, and in the second $1 - \pi_i$.) Show that $V(Y_i) > E(Y_i)$ when $\pi_i > 0$.

(b) Derive the log likelihood for the ZIP model, given in Equation 15.7 (page 394).

(c) The *zero-inflated negative-binomial* (ZINB) *regression model* substitutes a negative-binomial GLM for the Poisson-regression submodel of Equation 15.6 on page 393:

$$\log_e \mu_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$
$$p(y_i|x_1, \ldots, x_k) = \frac{\Gamma(y_i + \omega)}{y!\Gamma(\omega)} \times \frac{\mu_i^{y_i}\omega^{\omega}}{(\mu_i + \omega)^{\mu_i+\omega}}$$

Show that $E(Y_i) = (1 - \pi_i)\mu_i$ (as in the ZIP model) and that

$$V(Y_i) = (1 - \pi_i)\mu_i[1 + \mu_i(\pi_i + 1/\omega)]$$

When $\pi_i > 0$, the conditional variance is greater in the ZINB model than in the standard negative-binomial GLM, $V(Y_i) = \mu_i + \mu_i^2/\omega$; why? Derive the log likelihood for the ZINB model. [*Hint*: Simply substitute the negative-binomial GLM for the Poisson-regression submodel in Equation 15.7 (page 394)].

Exercise 15.3. The usual Pearson chi-square statistic for testing for independence in a two-way contingency table is

$$X_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(Y_{ij} - \widehat{\mu}_{ij}\right)^2}{\widehat{\mu}_{ij}}$$

where the Y_{ij} are the observed frequencies in the table, and the $\widehat{\mu}_{ij}$ are the estimated expected frequencies under independence. The estimated expected frequencies can be computed from the maximum-likelihood estimates for the loglinear model of independence, or they can be computed directly as $\widehat{\mu}_{ij} = Y_{i+}Y_{+j}/n$. The likelihood-ratio statistic for testing for independence can also be computed from the estimated expected counts as

$$G_0^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} Y_{ij}\log_e\frac{Y_{ij}}{\widehat{\mu}_{ij}}$$

Both test statistics have $(r - 1)(c - 1)$ degrees of freedom. The two tests are asymptotically equivalent, and usually produce similar results. Applying these formulas to the two-way table for voter turnout and intensity of partisan preference in Table 15.4 (page 395), compute both test statistics, verifying that the direct formula for G_0^2 produces the same result as given in the text.

Exercise 15.4. *Show that the normal distribution can be written in exponential form as

$$p(y; \theta, \phi) = \exp\left\{\frac{y\theta - \theta^2/2}{\phi} - \tfrac{1}{2}\left[\frac{y^2}{\phi} + \log_e(2\pi\phi)\right]\right\}$$

where $\theta = g_c(\mu) = \mu$; $\phi = \sigma^2$; $a(\phi) = \phi$; $b(\theta) = \theta^2/2$; and $c(y, \phi) = -\frac{1}{2}\left[y^2/\phi + \log_e(2\pi\phi)\right]$.

Exercise 15.5. *Show that the binomial distribution can be written in exponential form as

$$p(y;\theta,\phi) = \exp\left[\frac{y\theta - \log_e(1+e^\theta)}{1/n} + \log_e\binom{n}{ny}\right]$$

where $\theta = g_c(\mu) = \log_e[\mu/(1-\mu)]$; $\phi = 1$; $a(\phi) = 1/n$; $b(\theta) = \log_e(1+e^\theta)$; and $c(y,\phi) = \log_e\binom{n}{ny}$.

Exercise 15.6. *Using the results given in Table 15.9 (on page 403), verify that the Poisson, gamma, and inverse-Gaussian families can all be written in the common exponential form

$$p(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

Exercise 15.7. *Using the general result that the conditional variance of a distribution in an exponential family is

$$V(Y) = a(\phi)\frac{d^2b(\theta)}{d\theta^2}$$

and the values of $a(\cdot)$ and $b(\cdot)$ given in Table 15.9 (on page 403), verify that the variances of the Gaussian, binomial, Poisson, gamma, and inverse-Gaussian families are, consecutively, ϕ, $\mu(1-\mu)/n$, μ, $\phi\mu^2$, and $\phi\mu^3$.

Exercise 15.8. *Show that the derivative of the log likelihood for an individual observation with respect to the regression coefficients in a GLM can be written as

$$\frac{\partial l_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a_i(\phi)v(\mu_i)} \times \frac{d\mu_i}{d\eta_i} \times x_{ij}, \text{ for } j = 0, 1, \ldots, k$$

(See Equation 15.17 on page 404.)

Exercise 15.9. *Using the general expression for the residual deviance,

$$D(\mathbf{y};\widehat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}\frac{Y_i\left[g(Y_i) - g(\widehat{\mu}_i)\right] - b\left[g(Y_i)\right] + b\left[g(\widehat{\mu}_i)\right]}{a_i}$$

show that the deviances for the several exponential families can be written in the following forms:

Family	*Residual Deviance*
Gaussian	$\sum(Y_i - \widehat{\mu}_i)^2$
Binomial	$2\sum\left[n_iY_i\log_e\frac{Y_i}{\widehat{\mu}_i} + n_i(1-Y_i)\log_e\frac{1-Y_i}{1-\widehat{\mu}_i}\right]$
Poisson	$2\sum\left[Y_i\log_e\frac{Y_i}{\widehat{\mu}_i} - (Y_i - \widehat{\mu}_i)\right]$
Gamma	$2\sum\left[-\log_e\frac{Y_i}{\widehat{\mu}_i} + \frac{Y_i - \widehat{\mu}_i}{\widehat{\mu}_i}\right]$
Inverse-Gaussian	$\sum\frac{(Y_i - \widehat{\mu}_i)^2}{Y_i\widehat{\mu}_i^2}$

Exercise 15.10. *Using the SLID data, Table 12.1 in Section 12.3.2 (on page 283) reports the results of a regression of log wages on sex, the square of education, a quadratic in age, and interactions between sex and education-squared, and between sex and the quadratic for age.

(a) Estimate the age γ_1 at which women attain on average their highest level of wages, controlling for education. Use the delta method to estimate the standard error of $\widehat{\gamma}_1$. *Note:* You will need to refit the model to obtain the covariance matrix for the estimated regression coefficients.
(b) Estimate the age γ_2 at which men attain on average their highest level of wages, controlling for education. Use the delta method to estimate the standard error of $\widehat{\gamma}_2$.
(c) Let $\gamma_3 \equiv \gamma_1 - \gamma_2$, the difference between the ages at which men and women attain their highest wage levels. Compute $\widehat{\gamma}_3$. Use the delta method to find the standard error of $\widehat{\gamma}_3$ and then test the null hypothesis H_0: $\gamma_3 = 0$.

Exercise 15.11. Coefficient quasi-variances: Coefficient quasi-variances for dummy-variable regressors were introduced in Section 7.2.1. Recall that the object is to approximate the standard errors for pairwise *differences* between categories,

$$\text{SE}(C_j - C_{j'}) = \sqrt{\widehat{V}(C_j) + \widehat{V}(C_{j'}) - 2 \times \widehat{C}(C_j, C_{j'})}$$

where C_j and $C_{j'}$ are two dummy-variable coefficients for an m-category polytomous explanatory variable; $\widehat{V}(C_j)$ is the estimated sampling variance of C_j; and $\widehat{C}(C_j, C_{j'})$ is the estimated sampling covariance of C_j and $C_{j'}$. By convention, we take C_m (the coefficient of the baseline category) and its standard error, $\text{SE}(C_m)$, to be 0. We seek coefficient quasi-variances $\widetilde{V}(C_j)$, so that

$$\text{SE}(C_j - C_{j'}) \approx \sqrt{\widetilde{V}(C_j) + \widetilde{V}(C_{j'})}$$

for all pairs of coefficients C_j and $C_{j'}$, by minimizing the total log relative error of approximation, $\sum_{j<j'} \left[\log(\text{RE}_{jj'})\right]^2$, where

$$\text{RE}_{jj'} \equiv \frac{\widetilde{V}(C_j - C_{j'})}{\widehat{V}(C_j - C_{j'})} = \frac{\widetilde{V}(C_j) + \widetilde{V}(C_{j'})}{\widehat{V}(C_j) + \widehat{V}(C_j) - 2 \times \widehat{C}(C_j, C_{j'})}$$

Firth (2003) cleverly suggests implementing this criterion by fitting a GLM in which the response variable is $Y_{jj'} \equiv \log_e[\widehat{V}(C_j - C_{j'})]$ for all unique pairs of categories j and j'; the linear predictor is $\eta_{jj'} \equiv \beta_j + \beta_{j'}$; the link function is the exponential link, $g(\mu) = \exp(\mu)$ (which is, note, *not* one of the common links in Table 15.1); and the variance function is constant, $V(Y|\eta) = \phi$. The quasi-likelihood estimates of the coefficients β_j are the quasi-variances $\widetilde{V}(C_j)$. For example, for the Canadian occupational prestige regression described in Section 7.2.1, where the dummy variables pertain to type of occupation (professional and managerial, white collar, or blue collar), we have

Pair (j,j′)	$Y_{jj'} = \log_e[\widehat{V}(C_j - C_{j'})]$
Professional, White Collar	$\log_e(2.771^2)$=2.038
Professional, Blue Collar	$\log_e(3.867^2)$=2.705
White Collar, Blue Collar	$\log_e(2.514^2)$=1.844

and model matrix

$$\mathbf{X} = \begin{bmatrix} (\beta_1) & (\beta_2) & (\beta_3) \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

With three unique pairs and three coefficients, we should get a perfect fit: As I mentioned in Section 7.2.1, when there are only three categories, the quasi-variances perfectly recover the estimated variances for pairwise differences in coefficients. Demonstrate that this is the case by fitting the GLM. Some additional comments:

- The computation outlined here is the basis of Firth's `qvcalc` package (described in Firth, 2003) for the R statistical programming environment.
- The computation of quasi-variances applies not only to dummy regressors in linear models but to all models with a linear predictor for which coefficients and their estimated covariance matrix are available—for example, the GLMs described in this chapter.
- Quasi-variances may be used to approximate the standard error for any linear combination of dummy-variable coefficients, not just for pairwise differences.
- Having found the quasi-variance approximations to a set of standard errors, we can then compute and report the (typically small) maximum relative error of these approximations. Firth and De Menezes (2004) give more general results for the maximum relative error for *any* contrast of coefficients.

Summary

- A generalized linear model (or GLM) consists of three components:
  1. A random component, specifying the conditional distribution of the response variable, Y_i (for the ith of n independently sampled observations), given the values of the explanatory variables in the model. In the initial formulation of GLMs, the distribution of Y_i was a member of an exponential family, such as the Gaussian (normal), binomial, Poisson, gamma, or inverse-Gaussian families of distributions.
  2. A linear predictor—that is a linear function of regressors,
  $$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{ik} X_k$$
  3. A smooth and invertible linearizing link function $g(\cdot)$, which transforms the expectation of the response variable, $\mu_i \equiv E(Y_i)$, to the linear predictor:
  $$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{ik} X_k$$
- A convenient property of distributions in the exponential families is that the conditional variance of Y_i is a function of its mean μ_i and, possibly, a dispersion parameter ϕ. In addition to the familiar Gaussian and binomial families (the latter for proportions), the Poisson family is useful for modeling count data, and the gamma and inverse-Gaussian families for modeling positive continuous data, where the conditional variance of Y increases with its expectation.
- GLMs are fit to data by the method of maximum likelihood, providing not only estimates of the regression coefficients but also estimated asymptotic standard errors of the coefficients.

- The ANOVA for linear models has an analog in the analysis of deviance for GLMs. The residual deviance for a GLM is $D_m \equiv 2(\log_e L_s - \log_e L_m)$, where L_m is the maximized likelihood under the model in question, and L_s is the maximized likelihood under a saturated model. The residual deviance is analogous to the residual sum of squares for a linear model.
- In GLMs for which the dispersion parameter is fixed to 1 (binomial and Poisson GLMs), the likelihood-ratio test statistic is the difference in the residual deviances for nested models. For GLMs in which there is a dispersion parameter to estimate (Gaussian, gamma, and inverse-Gaussian GLMs), we can instead compare nested models by an incremental F-test.
- The basic GLM for count data is the Poisson model with log link. Frequently, however, when the response variable is a count, its conditional variance increases more rapidly than its mean, producing a condition termed overdispersion and invalidating the use of the Poisson distribution. The quasi-Poisson GLM adds a dispersion parameter to handle overdispersed count data; this model can be estimated by the method of quasi-likelihood. A similar model is based on the negative-binomial distribution, which is not an exponential family. Negative-binomial GLMs can nevertheless be estimated by maximum likelihood. The zero-inflated Poisson regression model may be appropriate when there are more zeroes in the data than is consistent with a Poisson distribution.
- Loglinear models for contingency tables bear a formal resemblance to ANOVA models and can be fit to data as Poisson GLMs with a log link. The loglinear model for a contingency table, however, treats the variables in the table symmetrically—none of the variables is distinguished as a response variable—and consequently the parameters of the model represent the associations among the variables, not the effects of explanatory variables on a response. When one of the variables is construed as the response, the loglinear model reduces to a binomial or multinomial logit model.
- The Gaussian, binomial, Poisson, gamma, and inverse-Gaussian distributions can all be written in the common linear-exponential form:

$$p(y; \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right]$$

  where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions that vary from one exponential family to another; $\theta = g_c(\mu)$ is the canonical parameter for the exponential family in question; $g_c(\cdot)$ is the canonical link function; and $\phi > 0$ is a dispersion parameter, which takes on a fixed, known value in some families. It is generally the case that $\mu = E(Y) = b'(\theta)$ and that $V(Y) = a(\phi)b''(\theta)$.
- The maximum-likelihood estimating equations for generalized linear models take the common form

$$\sum_{i=1}^{n} \frac{Y_i - \mu_i}{a_i v(\mu_i)} \times \frac{d\mu_i}{d\eta_i} \times x_{ij} = 0, \text{ for } j = 0, 1, \ldots, k$$

  These equations are generally nonlinear and therefore have no general closed-form solution, but they can be solved by iterated weighted least squares (IWLS). The estimating equations for the coefficients do not involve the dispersion parameter, which (for models in which the dispersion is not fixed) then can be estimated as

$$\widetilde{\phi} = \frac{1}{n - k - 1} \sum \frac{(Y_i - \widehat{\mu}_i)^2}{a_i v(\widehat{\mu}_i)}$$

  The estimated asymptotic covariance matrix of the coefficients is

$$\widehat{\mathcal{V}}(\mathbf{b}) = \widetilde{\phi}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}$$

where $\mathbf{b}$ is the vector of estimated coefficients and $\mathbf{W}$ is a diagonal matrix of weights from the last IWLS iteration.

- The maximum-likelihood estimating equations, and IWLS estimation, can be applied whenever we can express the transformed mean of Y as a linear function of the Xs and can write the conditional variance of Y as a function of its mean and (possibly) a dispersion parameter—even when we do not specify a particular conditional distribution for Y. The resulting quasi-likelihood estimator shares many of the properties of maximum-likelihood estimators.
- The residual deviance for a model is twice the difference in the log likelihoods for the saturated model, which dedicates one parameter to each observation, and the model in question:

$$\begin{aligned} D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) &\equiv 2[\log_e L(\mathbf{y}, \phi; \mathbf{y}) - \log_e L(\widehat{\boldsymbol{\mu}}, \phi; \mathbf{y})] \\ &= 2 \sum_{i=1}^{n} \frac{Y_i \left[g(Y_i) - g(\widehat{\mu}_i)\right] - b\left[g(Y_i)\right] + b\left[g(\widehat{\mu}_i)\right]}{a_i} \end{aligned}$$

  Dividing the residual deviance by the estimated dispersion parameter produces the scaled deviance, $D^*(\mathbf{y}; \widehat{\boldsymbol{\mu}}) \equiv D(\mathbf{y}; \widehat{\boldsymbol{\mu}})/\widetilde{\phi}$.
- To test the general linear hypothesis H_0: $\mathbf{L}\boldsymbol{\beta} = \mathbf{c}$, where the hypothesis matrix $\mathbf{L}$ has q rows, we can compute the Wald chi-square test statistic

$$Z_0^2 = (\mathbf{Lb} - \mathbf{c})' \left[\mathbf{L}\widehat{\mathcal{V}}(\mathbf{b})\,\mathbf{L}'\right]^{-1} (\mathbf{Lb} - \mathbf{c})$$

  with q degrees of freedom. Alternatively, if the dispersion parameter is estimated from the data, we can compute the F-test statistic

$$F_0 = \frac{(\mathbf{Lb} - \mathbf{c})' \left[\mathbf{L}\widehat{\mathcal{V}}(\mathbf{b})\,\mathbf{L}'\right]^{-1} (\mathbf{Lb} - \mathbf{c})}{q}$$

  on q and $n - k - 1$ degrees of freedom.
- The delta method may be used to approximate the standard error of a nonlinear function of regression coefficients in a GLM. If $\gamma \equiv f(\beta_0, \beta_1, \ldots, \beta_k)$, then

$$\widehat{\mathcal{V}}(\widehat{\gamma}) \approx \sum_{j=0}^{k} \sum_{j'=0}^{k} v_{jj'} \times \frac{\partial \widehat{\gamma}}{\partial \widehat{\beta}_j} \times \frac{\partial \widehat{\gamma}}{\partial \widehat{\beta}_{j'}}$$

- Effect displays for GLMs are based on the fitted values $\widehat{\boldsymbol{\eta}}^* = \mathbf{X}^*\mathbf{b}$, representing a high-order term in the model; that is, $\mathbf{X}^*$ has the same general structure as the model matrix $\mathbf{X}$, with the explanatory variables in the high-term order ranging over their values in the data, while other explanatory variables are set to typical values. The standard errors of $\widehat{\boldsymbol{\eta}}^*$, given by the square-root diagonal entries of $\mathbf{X}^*\widehat{\mathcal{V}}(\mathbf{b})\mathbf{X}^{*\prime}$, may be used to compute pointwise confidence intervals for the effects.
- Most of the standard diagnostics for linear models extend relatively straightforwardly to GLMs. These extensions typically take advantage of the computation of maximum-likelihood and quasi-likelihood estimates for GLMs by iterated weighted least squares. Such diagnostics include studentized residuals, hat-values, Cook's distances, DFBETA and DFBETAS, added-variable plots, component-plus-residual plots, and the constructed-variable plot for transforming an explanatory variable.

Recommended Reading

- McCullagh and Nelder (1989), the "bible" of GLMs, is a rich and interesting—if generally difficult—text.
- Dobson (2001) presents a much briefer overview of generalized linear models at a more moderate level of statistical sophistication.
- Aitkin, Francis, and Hinde's (2005) text, geared to the statistical computer package GLIM for fitting GLMs, is still more accessible.
- A chapter by Firth (1991) is the best brief treatment of generalized linear models that I have read.
- Long (1997) includes an excellent presentation of regression models for count data (though not from the point of view of GLMs); an even more extensive treatment may be found in Cameron and Trivedi (1998).