Chapter 12

Maximum Likelihood Theory for Log-Linear Models

This chapter presents the basic theoretical results of fitting log-linear models by maximum likelihood. The level of mathematical sophistication is considerably higher than in the rest of the book. The presentation assumes knowledge of advanced calculus, mathematical statistics, and large sample theory. Although the results in this chapter are proven in a different manner than for regular linear models, the results themselves are quite similar in nature. The common linear structure of the two techniques leads to the well-known analogies between them. A familiarity with log-linear models at the level of, say Fienberg (1980), is assumed.

In order to simplify proofs, the $o(\cdot)$, $O(\cdot)$, $o_p(\cdot)$, $O_p(\cdot)$ notations have been used extensively. See Bishop, Fienberg, and Holland (1975) for a detailed discussion of these.

Section 1 introduces notation and recalls some analytic results. Section 2 discusses finite sample properties of maximum likelihood estimators. Section 3 treats asymptotic results. Section 4 examines how the theory applies to weighted least squares, obtaining variance estimates, and logit and multinomial response models. Section 5 contains two proofs that are more involved than the rest of the chapter.

12.1 Notation

All vectors are considered as column vectors unless otherwise stated.

We will frequently apply the same real valued function to each element of a vector. Let x be a vector in $\mathbf{R}^s$ and f a function from $\mathbf{R}$ to $\mathbf{R}$, then

the function from $\mathbf{R}^s$ to $\mathbf{R}^s$ that maps x elementwise is denoted

$$f(x) = \big(f(x_1), \ldots, f(x_s)\big)'.$$

The most common choice of f will be the log function; thus, $\log(x) = (\log x_1, \ldots, \log x_s)'$.

A diagonal matrix with the values of the vector x on the diagonal will be written $D(x)$. As usual, an $s \times 1$ vector of ones is written J_s with the subscript dropped when the dimension is clear.

Proofs using the $o_p(\cdot)$, $O_p(\cdot)$ notation are usually divided into an analytic argument and a stochastic argument. To reduce the length of the discussion, these arguments have frequently been run together. Therefore, if $f(x) = o(g(x))$ and X_N is a sequence of random variables, we write $f(X_N) = o(g(X_N))$. Two frequently used properties are (a) $o(O_p(N^{-\alpha})) = o_p(N^{-\alpha})$ for any $\alpha > 0$ and (b) $o(o_p(1)) = o_p(1)$. For example, if $g(X_N) = o_p(1)$, we have $f(X_N) = o(g(X_N)) = o(o_p(1)) = o_p(1)$.

If F is a function from $\mathbf{R}^s$ into $\mathbf{R}^t$ with $F(x) = (f_1(x), \ldots, f_t(x))'$, then the derivative of F at c is the $t \times s$ matrix of partial derivatives,

$$dF(c) = [\partial f_i / \partial x_j |_{x=c}].$$

If g maps $\mathbf{R}^s$ into $\mathbf{R}$, then $dg(c)$ is a $1 \times s$ *row* vector. The second derivative matrix of g at c is

$$d^2g(c) = d[(dg(x))'|_{x=c}] = \big[\partial^2 g / \partial x_i \partial x_j |_{x=c}\big],$$

which is an $s \times s$ matrix. Taylor's theorem can be written

$$g(x) = g(c) + dg(c)(x-c) + (x-c)'\big[d^2g(c)\big](x-c)/2 + o\big(\|x-c\|^2\big),$$

where $\|x-c\|^2 = (x-c)'(x-c)$. Critical points are points c, where $dg(c) = 0$. The chain rule can be written as a matrix product: If $f : \mathbf{R}^s \to \mathbf{R}^t$ and $g : \mathbf{R}^t \to \mathbf{R}^u$, then $d(g \circ f)(c) = dg(f(c))df(c)$.

12.2 Fixed Sample Size Properties

Consider a table of counts with q cells. The observations are denoted $n = (n_1, \ldots, n_q)'$. The n_i's are assumed to be the result of Poisson, multinomial, or product-multinomial sampling. Let $\mathrm{E}(n) = m$ and let X be a known $q \times p$ matrix. The log-linear model is $\log(m) \equiv \mu = Xb$ for some vector b. The log-linear model is simply the requirement that $\mu \in C(X)$. Unless otherwise indicated, X will be assumed to have rank p.

If the observations n_i in the cells are independent Poisson(m_i) random variables, the likelihood function is

$$\prod_{i=1}^{q} \big[e^{-m_i} m_i^{n_i} / n_i!\big]. \tag{1}$$

The log-likelihood is

$$\begin{aligned}\ell^{\mathbf{P}}(n,\mu) &= \sum_{i=1}^{q}[-m_i + n_i \log m_i - \log(n_i!)] \\ &= \sum_{i=1}^{q}[-e^{\mu_i} + n_i\mu_i - \log(n_i!)] \qquad (2) \\ &= \sum_{i=1}^{q} -e^{\mu_i} + n'\mu - \sum_{i=1}^{q}\log(n_i!).\end{aligned}$$

If the log-linear model holds, $\mu = Xb$, so

$$\ell^{\mathbf{P}}(n,\mu) = n'Xb - \sum_{i=1}^{q} e^{\mu_i} - \sum_{i=1}^{q}\log(n_i!).$$

Since the distribution of n is in the exponential family, $X'n$ is a complete sufficient statistic.

If the observations come from a product-multinomial sampling scheme, certain of the margins are fixed. Assume that there are r independent multinomials. (If $r = 1$, the sampling scheme is a simple multinomial.) Partition $\{1,\ldots,q\}$ into r sets $Q_1,\ldots,Q_r$, each set containing the indices for one of the multinomials. For $i = 1,\ldots,q$, $j = 1,\ldots,r$, let x_j be a vector with ith row, $x_{ij} = \delta_{Q_j}(i)$ where $\delta_{Q_j}(i)$ is one if $i \in Q_j$ and zero otherwise. Thus, x_j is a column of dummy variables indicating the jth population. By the sampling scheme, $n'x_j$ is fixed for $j = 1,\ldots,r$. In particular, $n'x_j = m'x_j = N_j$, the sample size for the jth multinomial. It will be convenient to combine the vectors x_j into a matrix, say $X_0 = [x_1,\ldots,x_r]$.

With product-multinomial sampling, there are two restrictions on the parameters: (a) $\mu \in C(X)$, and (b) $n'X_0 = m'X_0$. Estimates of the parameters also need to satisfy these conditions. If we assume that $C(X_0) \subset C(X)$, we will see that the MLE of m, based only on condition (a), will automatically satisfy condition (b).

We will assume throughout that $C(X_0) \subset C(X)$. For Poisson sampling, X_0 can be taken as a matrix of zeros. We also need the assumption that $J_q \in C(X)$. For product-multinomial sampling, $J_q \in C(X_0)$, so this is not a new restriction. For Poisson sampling, we are requiring that an overall mean (or its equivalent) be fitted.

Recall that the probability of an occurrence in the ith cell under product-multinomial sampling is m_i/N_j, where $i \in Q_j$, so the likelihood function is

$$\prod_{k=1}^{r}\left[\left(N_k!\Big/\prod_{i\in Q_k} n_i!\right)\prod_{i\in Q_k}\left(\frac{m_i}{N_k}\right)^{n_i}\right]. \qquad (3)$$

Let $\ell^{\mathbf{m}}(n,\mu)$ be the log of (3). For $\ell^{\mathbf{m}}(n,\mu)$ to be a log-likelihood, μ must have the property that $n'X_0 = m'X_0$, where $m = e^{\mu}$. In fact, $\ell^{\mathbf{m}}(n,\mu)$ is

only defined for such μ. In particular, the maximum likelihood estimate of μ, under product-multinomial sampling, must be a value of μ for which $\ell^{\mathbf{m}}(n,\mu)$ is defined. We will expand the domain of $\ell^{\mathbf{m}}(n,\mu)$ to include all real vectors μ. For a log-linear model $\mu \in C(X)$ with $C(X_0) \subset C(X)$, we will find the maximum of $\ell^{\mathbf{m}}(n,\mu)$ without reference to the condition $n'X_0 = m'X_0$. We will then observe that the value of μ that maximizes $\ell^{\mathbf{m}}(n,\mu)$ also satisfies the condition $n'X_0 = m'X_0$. This value of μ must be the maximum likelihood estimate.

We now proceed to expand the domain of $\ell^{\mathbf{m}}(n,\mu)$. Since $\sum_{i\in Q_k} n_i = \sum_{i\in Q_k} m_i = N_k$, (3) can be rewritten as

$$\prod_{k=1}^{r}\left[N_k! e^{N_k} N_k^{-N_k} \prod_{i\in Q_k} \left(m_i^{n_i} e^{-m_i}/n_i!\right)\right]$$

or

$$\left[\prod_{k=1}^{r} N_k! e^{N_k} N_k^{-N_k}\right]\left[\prod_{i=1}^{q} e^{-m_i} m_i^{n_i}/n_i!\right]. \tag{4}$$

The second term in (4) is exactly (1). The first term depends only on the N_k's. If we write $a(N_1,\ldots,N_r)$ as the log of the first term we can write the log-likelihood for product-multinomial sampling as

$$\ell^{\mathbf{m}}(n,\mu) = a(N_1,\ldots,N_r) + \ell^{\mathbf{P}}(n,\mu).$$

Since $\ell^{\mathbf{m}}(n,\mu)$ is defined only for values of μ satisfying $n'X_0 = m'X_0$, this relationship holds only for such values of μ. However, the relationship can be used to define the function $\ell^{\mathbf{m}}(n,\mu)$ for all values of μ, because $\ell^{\mathbf{P}}(n,\mu)$ is defined for all values of μ. Since the difference of $\ell^{\mathbf{P}}(n,\mu)$ and $\ell^{\mathbf{m}}(n,\mu)$ does not depend on μ, the maximums of the two functions, with respect to μ, will occur at the same place.

Rather than using either $\ell^{\mathbf{P}}(n,\mu)$ or $\ell^{\mathbf{m}}(n,\mu)$, remove the term in (2) that does not depend on μ to define

$$\ell(n,\mu) = n'\mu - \sum_{i=1}^{q} e^{\mu_i} = n'\mu - J'm. \tag{5}$$

For any of the sampling schemes considered, it will be enough to find MLEs by maximizing $\ell(n,\mu)$.

If the log-linear model $\mu \in C(X)$ holds, we need to maximize $\ell(n,\mu)$ subject to the condition that $\mu \in C(X)$. Since X is of full rank, μ can be written uniquely as $\mu = Xb$. We need to find the unconstrained maximum of

$$f_n(b) \equiv \ell(n,\mu).$$

Taking derivatives with respect to b (and μ)

$$df_n(b) = [d\ell(n,\mu)]\, d\mu(b),$$

$$d\ell(n,\mu) = d\left(n'\mu - \sum_{i=1}^{q} e^{\mu_i}\right) = n' - (e^{\mu_1},\ldots,e^{\mu_q}) = n' - m',$$

and

$$d\mu(b) = d(Xb) = X.$$

Substituting, we get

$$df_n(b) = (n-m)'X. \tag{6}$$

It should be recalled that m is a function of b ($m = m(b)$).

Critical points are found by setting the partial derivative matrix, $df_n(b)$, equal to zero. As will be seen below, $\hat{b}$, the MLE of b, must occur at a critical point, so $\hat{b}$ must satisfy

$$X'm(\hat{b}) = X'n. \tag{7}$$

In particular, since $C(X_0) \subset C(X)$, we have $X_0'm(\hat{b}) = X_0'n$. As indicated above, if $C(X_0) \subset C(X)$ the additional restriction on the MLEs from product-multinomial sampling is automatically satisfied.

By considering $d^2f_n(b)$, we can investigate the nature of the critical points.

$$d^2f_n(b) = d(df_n(b)') = d(X'n - X'm(b)) = -X'dm(b). \tag{8}$$

Write

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_q' \end{bmatrix} = [x_{ij}],$$

then

$$m(b) = (m_1,\ldots,m_q)' = \left(e^{x_1'b},\ldots,e^{x_q'b}\right)'.$$

Therefore,

$$\begin{aligned} dm(b) &= \begin{bmatrix} x_{11}e^{x_1'b} & \cdots & x_{1p}e^{x_1'b} \\ \vdots & & \vdots \\ x_{q1}e^{x_q'b} & \cdots & x_{qp}e^{x_q'b} \end{bmatrix} \\ &= \begin{bmatrix} x_1'e^{x_1'b} \\ \vdots \\ x_q'e^{x_q'b} \end{bmatrix} = \begin{bmatrix} x_1'm_1 \\ \vdots \\ x_q'm_q \end{bmatrix} \\ &= D(m)X, \end{aligned} \tag{9}$$

where $m = m(b)$. Substitution into (8) gives

$$d^2f_n(b) = -X'D(m(b))X. \tag{10}$$

In a log-linear model, m_i is always positive because $m_i = e^{\mu_i}$; therefore, $d^2 f_n(b)$ is negative definite and $f_n(b)$ is strictly convex. If $f_n(b)$ takes on its maximum, it must be at a critical point and the maximum is unique. The maximum is the MLE $\hat{b} = \hat{b}(n)$. $\hat{b}$ uniquely determines MLEs $\hat{\mu} = \hat{\mu}(n) = X\hat{b}$ and $\hat{m} = \hat{m}(n) = m(\hat{\mu}) = m(\hat{b})$.

A simple condition exists that ensures that $f_n(b)$ takes on its maximum.

Theorem 12.2.1. If there exists $\xi \perp C(X)$ such that $n_i + \xi_i > 0$ for $i = 1, \ldots, q$, then $\ell(n, \mu) = f_n(b)$ attains its maximum.

Proof.

$$\ell(n, \mu) = n'\mu - \sum_{i=1}^{q} e^{\mu_i}.$$

Let $\xi \perp C(X)$, then for $\mu \in C(X)$, $\ell(n, \mu) = g(\mu)$ where

$$\begin{aligned} g(\mu) &= (n + \xi)'\mu - \sum_{i=1}^{q} e^{\mu_i} \\ &= \sum_{i=1}^{q} (n_i + \xi_i)\mu_i - e^{\mu_i}. \end{aligned}$$

With all the $n_i + \xi_i$'s positive, as any $\mu_i \to -\infty$, $g(\mu) \to -\infty$. Similarly, if any $\mu_i \to \infty$, $g(\mu) \to -\infty$. This establishes that the convex function $g(\mu)$ must take on its maximum. Moreover, $g(\mu)$ must take on its maximum for $\mu \in C(X)$ because $C(X)$ is a closed set. □

In particular, if all the n_i's are positive, the MLE of μ exists. Henceforth, we assume that the (unique) MLE of μ exists.

In finding MLEs of μ and m, we relied on a particular parameterization $\mu = Xb$. Any alternative parameterization $\mu = Z\gamma$ where $C(Z) = C(X)$ is equally valid. Regardless of the parameterization, we are maximizing $\ell(n, \mu)$ subject to the constraint that $\mu \in C(X)$. Since we found a unique MLE $\hat{\mu}$, it must be valid for any parameterization. Similarly, $\hat{m}$ does not depend on the parameterization. In particular, Z need not be of full column rank. When Z is not of full rank, γ is not estimable. The problem of estimability can be handled as it is for standard linear models.

For testing hypotheses, the likelihood ratio test statistic (LRTS) is often used. To test

$$H_0 \colon \mu = \mu_0 \text{ versus } H_A \colon \mu \neq \mu_0 \qquad \text{where } \mu_0 \in C(X),$$

the LRTS is $-2[\ell(n, \mu_0) - \ell(n, \hat{\mu})]$. For testing

$$H_0 \colon \mu \in C(X_1) \text{ versus } H_A \colon \mu \notin C(X_1) \qquad \text{where } C(X_1) \subset C(X), \quad (11)$$

the LRTS is $-2[\ell(n,\hat{\mu}_1) - \ell(n,\hat{\mu})]$, where $\hat{\mu}_1$ is the MLE of μ for the log-linear model $\mu \in C(X_1)$. [The reduced model is also assumed to have $C(X_0) \subset C(X_1)$ and $J \in C(X_1)$.]

In both cases, the LRTS simplifies considerably. We investigate the LRTS for hypothesis (11). From (5), $\ell(n,\hat{\mu}_1) - \ell(n,\hat{\mu}) = n'(\hat{\mu}_1 - \hat{\mu}) - J'(\hat{m}_1 - \hat{m})$. Since $J \in C(X_1) \subset C(X)$, from (7) we get $J'(\hat{m}_1 - \hat{m}) = J'n - J'n = 0$. Substitution gives

$$\ell(n,\hat{\mu}_1) - \ell(n,\hat{\mu}) = n'(\hat{\mu}_1 - \hat{\mu}).$$

Since $\hat{\mu}_1, \hat{\mu} \in C(X)$, (7) also gives

$$n'(\hat{\mu}_1 - \hat{\mu}) = n'M(\hat{\mu}_1 - \hat{\mu}) = \hat{m}'M(\hat{\mu}_1 - \hat{\mu}) = \hat{m}'(\hat{\mu}_1 - \hat{\mu}),$$

where $M = X(X'X)^{-1}X'$ is the perpendicular projection matrix onto $C(X)$. Thus, the LRTS is

$$-2[\hat{m}'(\hat{\mu}_1 - \hat{\mu})] = 2\sum_{i=1}^{q} \hat{m}_i \log(\hat{m}_i/\hat{m}_{1i}). \tag{12}$$

Contingency tables with structural zeros ($m_i = 0$) frequently occur. Under the sampling schemes considered here, $m_i = 0$ implies that $\Pr(n_i = 0) = 1$; therefore, without loss of generality, such cells can simply be dropped from the model and the theory does not change.

12.3 Asymptotic Properties

In establishing asymptotic properties, we let N go to infinity, where N is the total sample size in product-multinomial sampling and the sum of the expected values for the q cells in Poisson sampling. For product-multinomial sampling, we assume that the probabilities in each cell remain constant and that the individual sample sizes for each population remain in a fixed proportion; i.e., N_j/N remains constant. For Poisson sampling, we assume that the ratio of any pair of expected values remains constant.

Terminology appropriate for product-multinomial sampling will be used, but the results also hold for Poisson sampling. N will be referred to as the sample size. n_N is a sample of size N with $\mathrm{E}(n_N) = m_N$. By our assumptions, $m_N = Nm^*$ for some vector m^*. For multinomial sampling, m^* is the vector of probabilities. For product-multinomial sampling, m^* is the vector of normalized probabilities. The probabilities are normalized by the relative sizes of the populations so that $J'm^* = \sum_{i=1}^{q} m_i^* = 1$. In particular, if i is a cell in the jth multinomial population, m_i^* is $p_i(N_j/N)$ where p_i is the probability of getting an observation from the jth population in the ith cell. For Poisson sampling, m^* is the vector of expected values divided

by N, thus $m_N = Nm^*$. Note that for all sampling schemes, $J'm_N = N$, hence $J'm^* = 1$.

For a log-linear model to be valid, we need additional restrictions on m_N. In particular, writing $\mu_N = \log(m_N)$, we need $\mu_N \in C(X)$. Since $m_N = Nm^*$, we have $\mu_N = \log(m_N) = (\log N)J + \log(m^*)$. Since J is assumed to be in $C(X)$, it is sufficient to require $\log(m^*) \in C(X)$. Henceforth, we make the assumption that for some b^*, $\log(m^*) = Xb^*$. Define $\mu^* = Xb^*$; then, $\mu_N = \mu^* + (\log N)J$.

Some special matrices will be used frequently in the sequel. Let $D = D(m^*)$ and $A = X(X'DX)^{-1}X'D$. Let A_0 and A_1 be defined as A, with X_0 and X_1 replacing X. Note that A is a projection matrix onto $C(X)$. (In fact, A is the perpendicular projection matrix for the inner product determined by D.) A has the same form as a matrix giving best linear unbiased estimates in a regular linear model with covariance matrix D^{-1}. Taking $D^{1/2} = D(\sqrt{m^*})$ we see that $P = D^{1/2}AD^{-1/2}$ is the perpendicular projection matrix onto $C(D^{1/2}X)$ (with the usual inner product).

We first consider properties of n_N for large samples. These results are well known but necessary for the rest of the development.

Theorem 12.3.1.

(1) $N^{-1}n_N \to m^*$ a.s.,

(2) $N^{-1}n_N \xrightarrow{P} m^*$,

(3) $N^{-1/2}(n_N - m_N) \xrightarrow{L} N(0, D[I - A_0])$, and

(4) $N^{-1}(n_N - m_N) = O_p(N^{-1/2})$.

JUSTIFICATION. For product-multinomial sampling, (1) is a direct application of the strong law of large numbers applied to the elements of n_N. For Poisson sampling, (1) follows from Chebyshev's inequality and the Borel-Cantelli Lemma. Part (2) follows from (1). Part (4) follows from (3). It remains only to show (3).

(a) *Poisson sampling.* For Poisson sampling, $X_0 = 0$, so $A_0 = 0$. Since the n_{Ni}'s are independent, it suffices to show the $N^{-1/2}(n_{Ni} - m_{Ni})$ is asymptotically $N(0, m_i^*)$. We use a moment generating function argument. The moment generating function of a random variable w is $\varphi_w(u) = \mathrm{E}(e^{uw})$. These behave similarly to characteristic functions, but moment generating functions do not exist for some random variables. The moment generating function we need is

$$\varphi_{N^{-1/2}(n_{Ni}-m_{Ni})}(u) = \varphi_{n_{Ni}}(N^{-1/2}u)\exp\Big(-N^{-1/2}m_{Ni}u\Big),$$

where

$$\varphi_{n_{Ni}}(u) = \exp[-m_{Ni}(1 - e^u)].$$

Using a Taylor's expansion,

$$e^{au} = 1 + au + a^2u^2/2 + a^3\tilde{u}^3/6,$$

for some $\tilde{u} \in [0, u]$, we can write

$$\begin{aligned}
&\log \varphi_{N^{-1/2}(n_{Ni}-m_{Ni})}(u) \\
&\quad = -m_{Ni}\left[1 - e^{N^{-1/2}u}\right] - N^{-1/2}m_{Ni}u \\
&\quad = -m_{Ni}\left[1 - \left(1 + N^{-1/2}u + N^{-1}u^2/2 + N^{-3/2}\tilde{u}^3/6\right)\right] - N^{-1/2}m_{Ni}u \\
&\quad = N^{-1}m_{Ni}u^2/2 + N^{-3/2}m_{Ni}\tilde{u}^3/6 \\
&\quad = m_i^*u^2/2 + N^{-1/2}m_i^*\tilde{u}^3/6.
\end{aligned}$$

As $N \to \infty$,

$$\log \varphi_{N^{-1/2}(n_{Ni}-m_{Ni})}(u) \to m_i^*u^2/2,$$

so

$$N^{-1/2}(n_{Ni} - m_{Ni}) \xrightarrow{L} N(0, m_i^*).$$

(b) *Multinomial Sampling.* For multinomial sampling, $X_0 = J$, and $A_0 = JJ'D$, so $D[I - A_0] = D - DJJ'D$. Part (3) is then the standard large sample result for multinomials, which is an immediate consequence of the multivariate central limit theorem.

(c) *Product-Multinomial Sampling.* The different multinomial populations are asymptotically normal and they are independent by assumption. It remains only to establish that $D[I - A_0]$ is the correct block diagonal covariance matrix. Write n_N so that all observations in the first multinomial population are listed first, the second population is listed second, etc. Considering the implied structure of X_0, it is easily seen that $(X_0'DX_0)^{-1} = ND^*$, where $D^* = \text{Diag}\left(N_1^{-1}, \ldots, N_r^{-1}\right)$, and that $D[I - A_0] = D - NDX_0D^*X_0'D$. It is easily seen that $D - NDX_0D^*X_0'D$ is precisely the block diagonal matrix needed. □

The main result used in finding asymptotic properties of maximum likelihood estimates is a relationship between the MLE $\hat{\mu}_N$ and the observations n_N.

Lemma 12.3.2. $N^{1/2}(\hat{\mu}_N - \mu_N) - \left(AD^{-1}\right)N^{-1/2}(n_N - m_N) \xrightarrow{P} 0.$

Proof. See Section 5. □

Since the asymptotic distribution of $N^{-1/2}(n_N - m_N)$ is known, the lemma gives the asymptotic distribution of $N^{1/2}(\hat{\mu}_N - \mu_N)$ and, by a Taylor's expansion, the asymptotic distribution of $N^{-1/2}(\hat{m}_N - m_N)$.

Theorem 12.3.3.

(1) $N^{1/2}(\hat{\mu}_N - \mu_N) \xrightarrow{L} N\big(0, [A - A_0]\, D^{-1}\big)$.

(2) $\hat{\mu}_N - \mu_N \xrightarrow{P} 0$.

(3) $N^{-1/2}(\hat{m}_N - m_N) \xrightarrow{L} N(0, D\,[A - A_0])$.

(4) $N^{-1}\hat{m}_N \xrightarrow{P} m^*$.

Proof.

(1) Theorem 12.3.1 and Lemma 12.3.2 imply that

$$N^{1/2}(\hat{\mu}_N - \mu_N) \xrightarrow{L} N\big(0, AD^{-1}[D(I - A_0)]D^{-1}A'\big).$$

It is easily seen that

$$\begin{aligned} AD^{-1}[D(I - A_0)]D^{-1}A' &= AD^{-1}A' - AA_0D^{-1}A' \\ &= AD^{-1} - A_0D^{-1} \\ &= (A - A_0)D^{-1}. \end{aligned}$$

(2) From (1), $N^{1/2}(\hat{\mu}_N - \mu_N) = O_p(1)$, so $\hat{\mu}_N - \mu_N = O_p\big(N^{-1/2}\big) = o_p(1)$.

(3) Recall that $\exp(y) = (e^{y_1}, \ldots, e^{y_q})'$. Taylor's theorem gives

$$\begin{aligned} \exp(y) &= \exp(x) + [d\exp(x)]\,(y - x) + o(\|y - x\|) \\ &= \exp(x) + D\big(\exp(x)\big)\,(y - x) + o(\|y - x\|). \end{aligned} \tag{1}$$

Let $y = \hat{\mu}_N - \big(\frac{1}{2}\log N\big)\, J$ and $x = \mu_N - \big(\frac{1}{2}\log N\big)\, J$. Since $\hat{\mu}_N - \mu_N = o_p(1)$, rearranging equation (1) and evaluating $\exp(y)$ and $\exp(x)$ gives

$$N^{-1/2}\hat{m}_N - N^{-1/2}m_N - \Big[D\Big(N^{-1/2}m_N\Big)\Big](\hat{\mu}_N - \mu_N) = o\big(o_p(1)\big) = o_p(1).$$

Since $m_N = Nm^*$, we have

$$N^{-1/2}(\hat{m}_N - m_N) - [D]\, N^{1/2}(\hat{\mu}_N - \mu_N) = o_p(1).$$

Applying part (1) of the theorem, we see that

$$N^{-1/2}(\hat{m}_N - m_N) \xrightarrow{L} N\big(0, D(A - A_0)D^{-1}D\big)\,,$$

but $D(A - A_0)D^{-1}D = D(A - A_0)$.

(4) $N^{-1/2}(\hat{m}_N - m_N) = O_p(1)$ by part (3), so $N^{-1}(\hat{m}_N - m_N) = O_p\big(N^{-1/2}\big) = o_p(1)$. Now part (4) follows by observing that $N^{-1}m_N = m^*$. □

It is of interest to note that Theorem 12.3.3 implies $\hat{\mu}_N - (\log N)J \xrightarrow{P} \mu^*$, so $\hat{\mu}_N$ is not consistent for μ^*.

The following corollary will be useful in examining the likelihood ratio test.

Corollary 12.3.4. $\hat{b}_N - b_N = O_p\big(N^{-1/2}\big)$, therefore $\hat{b}_N - b_N \xrightarrow{P} 0$.

Proof. From Theorem 12.3.3, $X\hat{b}_N - Xb_N = \hat{\mu}_N - \mu_N = O_p\big(N^{-1/2}\big)$. It follows that $\hat{b}_N - b_N = (X'X)^{-1}X'\Big[X\hat{b}_N - Xb_N\Big] = (X'X)^{-1}X'O_p\big(N^{-1/2}\big) = O_p\big(N^{-1/2}\big)$. □

We now consider the problem of testing

$$H_0 : \mu_N = \mu_{0N} \quad \text{versus} \quad H_A : \mu_N \neq \mu_{0N}. \tag{2}$$

As the sample size N gets larger, m_N gets larger and so does μ_N. Using a fixed value for the hypothesis, say $H_0 : \mu_N = \mu_0$, is not appropriate.

μ_{0N} must be in $C(X)$ and compatible with having a sample size of N, i.e., $J'm_{0N} = N$. In accordance with our other assumptions, we consider only the case where $\mu_{0N} = (\log N)J + \mu_0^*$ with $\mu_0^* \in C(X)$, and $J'm_0^* = 1$. m_{0N}, b_{0N}, m_0^*, and b_0^* are defined in the usual way. Note that (2) is equivalent to

$$H_0 : m^* = m_0^* \quad \text{versus} \quad H_A : m^* \neq m_0^*,$$

so one can think of the hypothesis as being on the (normalized) vector of probabilities.

The asymptotic distribution theory for the more interesting hypothesis

$$H_0 : \mu_N \in C(X_1) \quad \text{versus} \quad H_A : \mu_N \notin C(X_1), \tag{3}$$

where $C(X_0) \subset C(X_1) \subset C(X)$, can be handled quite simply after dealing with (2).

We want to show that the likelihood ratio test statistic (LRTS), $-2[\ell(n_N, \mu_{0N}) - \ell(n_N, \hat{\mu}_N)]$, has an asymptotic χ^2 distribution.

Theorem 12.3.5. If $\mu_N = \mu_{0N}$, then $-2[\ell(n_N, \mu_{0N}) - \ell(n_N, \hat{\mu}_N)] \xrightarrow{L} \chi^2(p - r)$.

Proof. The proof is in four parts. The first three find statistics

asymptotically equivalent to the LRTS. The last one establishes the distribution of the LRTS.

(a) Let $f_n(b) = \ell(n, \mu(b))$. A Taylor's expansion of $f_n(b)$ about $\tilde{b}$ gives

$$\begin{aligned} f_n(b) &= f_n(\tilde{b}) + [df_n(\tilde{b})](b - \tilde{b}) + \frac{1}{2}(b - \tilde{b})' \left[d^2 f_n(\tilde{b})\right](b - \tilde{b}) \\ &\quad + o\left(\|b - \tilde{b}\|^2\right). \end{aligned}$$

Substituting for $df_n(\tilde{b})$ and $d^2 f_n(\tilde{b})$ as found in (12.2.6) and (12.2.10), we get

$$\begin{aligned} f_n(b) - f_n(\tilde{b}) - \left[n - m(\tilde{b})\right]' X(b - \tilde{b}) \\ + \frac{1}{2}(b - \tilde{b})' \left[X' D\big(m(\tilde{b})\big) X\right](b - \tilde{b}) = o\left(\|b - \tilde{b}\|^2\right). \end{aligned} \quad (4)$$

Apply equation (4) with $n = n_N$, $\tilde{b} = \hat{b}_N$, and $b = b_N$. From Corollary 12.3.4, $\hat{b}_N - b_N = o_p(1)$, so

$$\begin{aligned} \ell(n_N, \mu_N) - \ell(n_N, \hat{\mu}_N) - (n - \hat{m}_N)' X(b_N - \hat{b}_N) \\ + \frac{1}{2}(b_N - \hat{b}_N)' [X' D(\hat{m}_N) X](b_N - \hat{b}_N) = o\big(o_p(1)\big). \end{aligned} \quad (5)$$

By (12.2.7), $(n_N - \hat{m}_N)' X = 0$. After multiplying by -2, (5) becomes

$$-2[\ell(n_N, \mu_N) - \ell(n_N, \hat{\mu}_N)] - (\mu_N - \hat{\mu}_N)' D(\hat{m}_N)(\mu_N - \hat{\mu}_N) = o_p(1). \quad (6)$$

The quadratic form in (6) can be rewritten as

$$N^{1/2}(\mu_N - \hat{\mu}_N)' D\big(N^{-1}\hat{m}_N\big)\, N^{1/2}(\mu_N - \hat{\mu}_N). \quad (7)$$

(b) For random variables Y_N and Z_N, it is well known that if $Y_N \xrightarrow{L} Y$ and $Z_N \xrightarrow{P} 0$, then $Y_N Z_N \xrightarrow{P} 0$. Repeated application of this gives the result: if $Y_N \xrightarrow{L} Y$ and $Z_N \xrightarrow{P} Z$, then $Y_N' Z_N Y_N - Y_N' Z Y_N \xrightarrow{P} 0$ where Y_N is a q vector and Z_N is a $q \times q$ matrix. Let $Z_N = D\big(N^{-1}\hat{m}_N\big)$ in (7). Since $N^{-1}\hat{m}_N \xrightarrow{P} m^*$,

$$\begin{aligned} N^{1/2}(\mu_N - \hat{\mu}_N)' D\big(N^{-1}\hat{m}_N\big)\, N^{1/2}(\mu_N - \hat{\mu}_N) \\ - N^{1/2}(\mu_N - \hat{\mu}_N)' D N^{1/2}(\mu_N - \hat{\mu}_N) \xrightarrow{P} 0. \end{aligned}$$

(c) Applying Lemma 12.3.2 gives

$$\begin{aligned} N^{1/2}(\mu_N - \hat{\mu}_N)' D N^{1/2}(\mu_N - \hat{\mu}_N) \\ - N^{-1/2}(n_N - m_N)' D^{-1} A' D A D^{-1} N^{-1/2}(n_N - m_N) \xrightarrow{P} 0. \end{aligned} \quad (8)$$

It is easily seen that

$$D^{-1}A'DAD^{-1} = AD^{-1} = D^{-1/2}PD^{-1/2}.$$

Recall that, $D^{-1/2} = D\left(1/\sqrt{m^*}\right)$ and P is the perpendicular projection matrix onto $C(D^{-1/2}X)$. Rewrite the second quadratic form in (8) as

$$\left[D^{-1/2}N^{-1/2}(n_N - m_N)\right]' P \left[D^{-1/2}N^{-1/2}(n_N - m_N)\right]. \tag{9}$$

(d) From Theorem 12.3.1, $D^{-1/2}N^{-1/2}(n_N - m_N) \xrightarrow{L} Y$, where $Y \sim N\left(0, D^{1/2}[I - A_0]D^{-1/2}\right)$. As in part (c), it is easy to see that $D^{1/2}[I - A_0]D^{-1/2} = I - P_0$, where P_0 is the perpendicular projection matrix onto $C(D^{-1/2}X_0)$. The quadratic form (9) converges in distribution to $Y'PY$. By Theorem 1.3.6 in Christensen (1996b), $Y'PY$ will have a χ^2 distribution with $tr[P(I - P_0)]$ degrees of freedom if

$$(I - P_0)P(I - P_0)P(I - P_0) = (I - P_0)P(I - P_0). \tag{10}$$

Since $C(P_0) \subset C(P)$, we have $PP_0 = P_0$. Simplifying gives both sides of (10) as $(P - P_0)$.

The theorem follows by observing that under H_0, $\mu_N = \mu_{0N}$, so the asymptotic distribution of (9) is the same as that of the LRTS, and that $tr[P(I - P_0)] = tr[P - P_0] = tr(P) - tr(P_0) = p - r$. □

We would like to show that the likelihood ratio test is consistent. That is, if $\mu_{0N} \neq \mu_N$, then $-2[\ell(n_N, \mu_{0N}) - \ell(n_N, \hat{\mu}_N)] \xrightarrow{P} \infty$. Consistency is an immediate result of the following theorem.

Theorem 12.3.6.

$$-2N^{-1}[\ell(n_N, \mu_{0N}) - \ell(n_N, \hat{\mu}_N)] \xrightarrow{P} -2[\ell(m^*, \mu_0^*) - \ell(m^*, \mu^*)].$$

If $\mu_{0N} - \mu_N \equiv \mu_0^* - \mu^* \neq 0$, the right-hand side is strictly positive.

Proof.

$$\begin{aligned} &-2N^{-1}[\ell(n_N, \mu_{0N}) - \ell(n_N, \hat{\mu}_N)] \qquad (11)\\ &= \quad -2N^{-1}[\ell(n_N, \mu_{0N}) - \ell(n_N, \mu_N)] - 2N^{-1}[\ell(n_N, \mu_N) - \ell(n_N, \hat{\mu}_N)]. \end{aligned}$$

Consider the second term of (11). From (6),

$$\begin{aligned} &-2N^{-1}[\ell(n_N, \mu_N) - \ell(n_N, \hat{\mu}_N)]\\ &\qquad - N^{-1}(\mu_N - \hat{\mu}_N)'D(\hat{m}_N)(\mu_N - \hat{\mu}_N) = o_p\left(N^{-1}\right) = o_p(1). \end{aligned}$$

Since $(\mu_N - \hat{\mu}_N) \xrightarrow{P} 0$ and $N^{-1}\hat{m}_N \xrightarrow{P} m^*$, we have $N^{-1}(\mu_N - \hat{\mu}_N)'D(\hat{m}_N)(\mu_N - \hat{\mu}_N) \xrightarrow{P} 0$. Therefore, $-2N^{-1}[\ell(n_N, \mu_N) - \ell(n_N, \hat{\mu}_N)] \xrightarrow{P} 0$.

Now consider the first term on the right of (11). By definition,

$$\begin{aligned} & -2N^{-1}[\ell(n_N, \mu_{0N}) - \ell(n_N, \mu_N)] \\ &= -2N^{-1}[n_N'\mu_{0N} - J'm_{0N}] + 2N^{-1}[n_N'\mu_N - J'm_N] \\ &= -2N^{-1}[n_N'(\mu_{0N} - \mu_N) - J'(m_{0N} - m_N)]\,. \end{aligned}$$

Since $\mu_{0N} - \mu_N = \mu_0^* - \mu^*$, $N^{-1}m_{0N} = m_0^*$, $N^{-1}m_N = m^*$, and $N^{-1}n_N \xrightarrow{P} m^*$, we have

$$\begin{aligned} -2N^{-1}[\ell(n_N, \mu_{0N}) - \ell(n_N, \mu_N)] &\xrightarrow{P} -2[m^*(\mu_0^* - \mu^*) - J'(m_0^* - m^*)] \\ &= -2[\ell(m^*, \mu_0^*) - \ell(m^*, \mu^*)]\,. \end{aligned}$$

As discussed in Lemma 12.5.3, μ^* is the unique MLE of μ when the data are m^* [i.e. $\mu^* = \hat{\mu}(m^*)$], so $\ell(m^*, \mu^*)$ is the unique maximum of $\ell(m^*, \mu)$. If $\mu_0^* \neq \mu^*$, $-2[\ell(m^*, \mu_0^*) - \ell(m^*, \mu^*)] > 0$. □

We now consider the problem of testing (3).

Theorem 12.3.7. If $\mu_N \in C(X_1)$, then $-2[\ell(n_N, \hat{\mu}_{1N}) - \ell(n_N, \hat{\mu}_N)] \xrightarrow{L} \chi^2(p - p_1)$, where $\hat{\mu}_{1N}$ is the MLE of μ_N under the model $\mu_N \in C(X_1)$ and $r(X_1) = p_1$.

Proof.

$$\begin{aligned} & -2[\ell(n_N, \hat{\mu}_{1N}) - \ell(n_N, \hat{\mu}_N)] \\ &\qquad = -2[\ell(n_N, \mu_N) - \ell(n_N, \hat{\mu}_N)] + 2[\ell(n_N, \mu_N) - \ell(n_N, \hat{\mu}_{1N})]\,. \end{aligned}$$

Since $C(X_1) \subset C(X)$, the proof of Theorem 12.3.5 applies to both terms on the right-hand side. In particular, (9) can be applied as is and also with X_1 substituted for X. If P_1 is the perpendicular projection matrix onto $C(D^{-1/2}X_1)$, then $P_1 = PP_1 = P_1P$, so the LRTS is asymptotically equivalent to $\left[D^{-1/2}N^{-1/2}(n_N - m_N)\right]'(P - P_1)\left[D^{-1/2}N^{-1/2}(n_N - m_N)\right]$. Since $(P - P_1)(I - P_0)(P - P_1) = (P - P_1)$, verification of the conditions of Theorem 1.3.6 is trivial. The LRTS has a χ^2 distribution with $r(P - P_1) = p - p_1$ degrees of freedom. □

Theorem 12.3.8 establishes the consistency of the likelihood ratio test for hypothesis (3).

Theorem 12.3.8.

$$-2N^{-1}[\ell(n_N, \hat{\mu}_{1N}) - \ell(n_N, \hat{\mu}_N)] \xrightarrow{P} -2[\ell(m^*, \hat{\mu}_1(m^*)) - \ell(m^*, \mu^*)]\,.$$

$\mu_N \notin C(X_1)$ if and only if the right-hand side is positive.

Proof. The proof involves several arguments from the proof of Lemma 12.3.2, so it is deferred until Section 5. □

Finally, we establish the asymptotic equivalence under H_0 of the LRTS and the Pearson test statistic (PTS). The Pearson test statistic is

$$(\hat{m}_N - \hat{m}_{1N})' D^{-1}(\hat{m}_{1N})(\hat{m}_N - \hat{m}_{1N}) \tag{12}$$

for the hypothesis (3).

Theorem 12.3.9. If $\mu_N \in C(X_1)$, then

$$-2[\ell(n_N, \hat{\mu}_{1N}) - \ell(n_N, \hat{\mu}_N)] - (\hat{m}_N - \hat{m}_{1N})' D^{-1}(\hat{m}_{1N})(\hat{m}_N - \hat{m}_{1N}) \xrightarrow{P} 0.$$

Proof. The PTS can be written elementwise as

$$\sum_{i=1}^{q} (\hat{m}_{Ni} - \hat{m}_{1Ni})^2 / \hat{m}_{1Ni}. \tag{13}$$

The elementwise form for the LRTS is found in (12.2.12). Note that (13) is equivalent to

$$N \sum_{i=1}^{q} \left[N^{-1}(\hat{m}_{Ni} - \hat{m}_{1Ni})\right]^2 / N^{-1}\hat{m}_{1Ni}$$

and the LRTS is

$$2N \sum_{i=1}^{q} \left(N^{-1}\hat{m}_{Ni}\right) \left[\log\left(N^{-1}\hat{m}_{Ni}\right) - \log\left(N^{-1}\hat{m}_{1Ni}\right)\right]. \tag{14}$$

To simplify notation, let $(x, y) = (N^{-1}\hat{m}_N, N^{-1}\hat{m}_{1N})$. Taking a second-order expansion of (14) about (m^*, m^*) gives

$$\begin{aligned} & 2N \sum_{i=1}^{q} x_i [\log x_i - \log y_i] \\ = \; & 2N \sum_{i=1}^{q} m_i^* [\log m_i^* - \log m_i^*] \\ & + 2N \sum_{i=1}^{q} [\log x_i - \log y_i + x_i(1/x_i)]\,|_{(x,y)=(m^*,m^*)}(x_i - m_i^*) \end{aligned}$$

$$
\begin{aligned}
&+ 2N \sum_{i=1}^{q} \left[-x_i(1/y_i)\right] |_{(x,y)=(m^*,m^*)} (y_i - m_i^*) \\
&+ N \sum_{i=1}^{q} \left[x_i^{-1}\right] |_{x=m^*} (x_i - m_i^*)^2 \\
&+ 2N \sum_{i=1}^{q} \left[-y_i^{-1}\right] |_{y=m^*} (x_i - m_i^*)(y_i - m_i^*) \\
&+ N \sum_{i=1}^{q} \left[x_i y_i^{-2}\right] |_{(x,y)=(m^*,m^*)} (y_i - m_i^*)^2 \\
&+ No\big(\|x - m^*\|^2 + \|y - m^*\|^2\big) .
\end{aligned}
$$

This easily simplifies to

$$
\begin{aligned}
& 2N \sum_{i=1}^{q} x_i [\log x_i - \log y_i] \\
&= \quad 2N \sum_{i=1}^{q} (x_i - y_i) \\
&\qquad + N \sum_{i=1}^{q} (1/m_i^*) \left[(x_i - m_i^*)^2 - 2(x_i - m_i^*)(y_i - m_i^*) + (y_i - m_i^*)^2\right] \\
&\qquad + No\big(\|x - m^*\|^2 + \|y - m^*\|^2\big) \\
&= \quad N \sum_{i=1}^{q} (x_i - y_i)^2 / m_i^* + No\big(\|x - m^*\|^2 + \|y - m^*\|^2\big) .
\end{aligned}
$$

The last equality is because $J \in C(X_1) \subset C(X)$, so that by (12.2.7), $NJ'x = J'n_N = NJ'y$ and $J'(x - y) = 0$.

In our regular notation,

$$
\begin{aligned}
2 \sum_{i=1}^{q} \hat{m}_{Ni} \left[\log\big(\hat{m}_{Ni}/\hat{m}_{1Ni}\big)\right] &= \sum_{i=1}^{q} (\hat{m}_{Ni} - \hat{m}_{1Ni})^2 / m_{Ni} \\
&\quad + No\big(\|N^{-1}(\hat{m}_N - m_N)\|^2 + \|N^{-1}(\hat{m}_{1N} - m_N)\|^2\big) .
\end{aligned}
$$

Investigating the argument of $o(\cdot)$,

$$
\begin{aligned}
&\|N^{-1}(\hat{m}_N - m_N)\|^2 + \|N^{-1}(\hat{m}_{1N} - m_N)\|^2 \\
&\qquad = \left[O_p\left(N^{-1/2}\right)\right]^2 + \left[O_p\left(N^{-1/2}\right)\right]^2 = O_p\big(N^{-1}\big) .
\end{aligned}
$$

Since $o(O_p(N^{-1})) = o_p(N^{-1})$ and $No_p(N^{-1}) = o_p(1)$, we have the LRTS asymptotically equivalent to

$$
\sum_{i=1}^{q} (\hat{m}_{Ni} - \hat{m}_{1Ni})^2 / m_{Ni}
$$

$$= (\hat{m}_N - \hat{m}_{1N})' D^{-1}(m_N)(\hat{m}_N - \hat{m}_{1N}) \quad (15)$$
$$= N^{-1/2}(\hat{m}_N - \hat{m}_{1N})' D^{-1}(m^*) N^{-1/2}(\hat{m}_N - \hat{m}_{1N}).$$

Under H_0, $D^{-1}\big(N^{-1}\hat{m}_{1N}\big) \xrightarrow{P} D^{-1}(m^*)$, so (15) is asymptotically equivalent to (12) the PTS. □

A similar argument holds to show that the LRTS and the PTS are asymptotically equivalent under H_0 for testing the hypothesis (2). Results similar to Theorems 12.3.6 and 12.3.8 also exist for the PTS. For a more extensive discussion of the asymptotic properties of log-linear models, see Haberman (1974a).

12.4 Applications

a) *Weighted Least Squares.* We now consider two methods of obtaining estimates for log-linear models. The first is the Newton-Raphson technique, which is an iterative method for obtaining the maximum likelihood estimates. The Newton-Raphson method can be performed by doing a series of weighted least squares regression analyses. The second method is an approximate technique based on the asymptotic results that we have derived. It is a noniterative weighted least squares regression approach.

The Newton-Raphson technique is an iterative procedure for finding where a function equals the zero vector. Let g be a function mapping $\mathbf{R}^p$ into $\mathbf{R}^p$. We wish to find b_* such that $g(b_*) = 0$. Let b_0 be an initial guess for b_*. Newton-Raphson defines (recursively) a sequence b_t that converges to b_*. By Taylor's theorem, if b_{t+1} is near b_t, we have the approximate equality

$$g(b_{t+1}) = g(b_t) + [dg(b_t)]\,\delta_t\,,$$

where $\delta_t = b_{t+1} - b_t$. The Newton-Raphson technique assumes that b_t is known, sets $g(b_{t+1}) = 0$, and seeks to find b_{t+1}, i.e.,

$$0 = g(b_t) + [dg(b_t)]\,\delta_t\,,$$

so

$$\delta_t = -\,[dg(b_t)]^{-1} g(b_t) \quad (1)$$

and

$$b_{t+1} = b_t + \delta_t.$$

For finding maximum likelihood estimates, we wish to find where the derivative of $f_n(b) \equiv \ell(n, Xb)$ is zero. With $g(b) \equiv [df_n(b)]'$ and substituting (12.2.6) and (12.2.10) into (1), we get

$$\delta_t = \big[X' D\big(m(b_t)\big) X\big]^{-1} X'\big(n - m(b_t)\big).$$

The sequence b_t converges to a critical point of $\ell(n, Xb)$ which, as we have seen, must be the MLE of b under fairly weak conditions.

A weighted least squares computer program can be used to execute the Newton-Raphson procedure. Fit the model $Y = Xb+e$, $\mathrm{E}(e) = 0$, $\mathrm{Cov}(e) = D\big(m(b_t)\big)^{-1}$ where Y is taken as

$$\begin{aligned} Y &= Xb_t + D\big(m(b_t)\big)^{-1}\big(n - m(b_t)\big) \\ &= \log(m_t) + D(m_t)^{-1}(n - m_t). \end{aligned}$$

Let b_{t+1} denote the estimate of b obtained from this procedure. Clearly,

$$\begin{aligned} b_{t+1} &= \big[X'D\big(m(b_t)\big)X\big]^{-1}X'D(m_t)Y \\ &= b_t + \big[X'D\big(m(b_t)\big)X\big]^{-1}X'\big(n - m(b_t)\big), \end{aligned}$$

which is the Newton-Raphson value for b_{t+1}.

The second method is based on the asymptotic results of Theorem 12.3.3 and the fact, shown in Theorem 10.1.3 in Christensen (1996b), that best linear unbiased estimates (BLUEs) for the linear model $Y = X\beta+e$, $\mathrm{E}(e) = 0$, $\mathrm{Cov}(e) = V$, are the same as those for the model $Y = X\beta + e$, $\mathrm{E}(e) = 0$, $\mathrm{Cov}(e) = V + XUX'$, where U is any non-negative definite matrix.

The saturated log-linear model, $\mu \in \mathbf{R}^q$ always fits the data. For the saturated model $\hat{\mu} = \log(n)$ and $A = I$. If N is large, Theorem 12.3.3 gives the asymptotic relation

$$\log(n) - \mu \sim N\big(0, (I - A_0)D^{-1}(m)\big)\,. \tag{2}$$

It will be convenient to rewrite the term $A_0D^{-1}(m)$. It is easily seen that

$$A_0D^{-1}(m) = X_0\left(X_0'D(m)X_0\right)^{-1}X_0'.$$

Now consider the term $X_0'D(m)X_0$. The matrix X_0 has the same structure as the design matrix for a one-way ANOVA model, say

$$y_{ij} = \mu_i + e_{ij}.$$

Exploiting the simple form of X_0 and the fact that $X_0'n = X_0'm = (N_1, \ldots, N_r)'$, it is easily seen that

$$X_0'D(m)X_0 = D(X_0'n).$$

We now have

$$A_0D^{-1}(m) = X_0D^{-1}(X_0'n)X_0',$$

and the asymptotic distribution of $\log(n)$ is

$$\log(n) - \mu \sim N\big(0, D^{-1}(m) - X_0D^{-1}(X_0'n)X_0'\big)\,. \tag{3}$$

Imposing the linear constraint $\mu = Xb$, the asymptotic distribution (3) leads to fitting the linear model

$$\log(n) = Xb + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = D^{-1}(m) - X_0 D^{-1}(X_0' n) X_0'. \quad (4)$$

By Theorem 10.1.3, the BLUEs in model (4) are the same as those in

$$\log(n) = Xb + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = D^{-1}(m). \quad (5)$$

Of course, m is unknown, so (5) cannot be used directly. Estimating m with n gives the model

$$\log(n) = Xb + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = D^{-1}(n). \quad (6)$$

Note that n is the MLE of m under the saturated model. One virtue of model (6) is that it can be fit with any regression program that does weighted regression.

Besides the rationale just given, there are two other justifications for using this approximate procedure. First, if we take $m_0 = n$ in the Newton-Raphson algorithm, then the first step of the algorithm is precisely fitting model (6). Second, for product-multinomial data, fitting model (6) is the same procedure as that proposed by Grizzle, Starmer, and Koch (1969). In their paper, they consider modeling the vector of probabilities $p = (p_1, \ldots, p_q)'$. Their method specifies a generalized linear model $F(p) = Xb$, where F is a quite general function from $\mathbf{R}^q$ into $\mathbf{R}^q$. In particular, one can choose $F(p) = \log(m)$. With this choice of F, their estimation procedure amounts to a weighted least squares analysis with the covariance matrix

$$\mathrm{Cov}(e) = D^{-1}(n) - X_0 D^{-1}(X_0' n) X_0'.$$

Because $C(X_0) \subset C(X)$, the best linear unbiased estimates under this covariance matrix are the same as those using model (6).

b) *Asymptotic Variances Under Saturated Models.* The relation (2) is the basis for a number of asymptotic variance formulas commonly used with saturated models. For a saturated model, $\mathrm{Cov}(\hat{\mu}) = D^{-1}(m) - A_0 D^{-1}(m)$.

Suppose that we write a saturated model $\mu = Xb$, where $X = [X_0, X_1]$ and $b' = [b_0', b_1']$. The parameter b_0 is forced into the model to deal with the product-multinomial sampling. [Recall that to deal with product-multinomial sampling, we always assume that $C(X_0) \subset C(X)$.] For a linear function $\rho'\mu$ where $\rho \perp X_0$, one gets $\rho'\mu = \rho' X_0 b_0 + \rho' X_1 b_1 = \rho' X_1 b_1$ and $\mathrm{Var}(\rho'\hat{\mu}) = \rho' D^{-1}(m)\rho$ because $\rho' A_0 D^{-1}(m)\rho = 0$. The maximum likelihood estimate of $\rho' D^{-1}(m)\rho$ is $\rho' D^{-1}(n)\rho$. In other words, for estimable functions of the parameters that are not forced into the model (i.e., functions involving only b_1), the estimate of the variance of $\rho'\hat{\mu} = \rho' X_1 \hat{b}_1$ is $\rho' D^{-1}(n)\rho$.

EXAMPLE 12.4.1. Consider now a 2×3 table. One log-odds ratio is $\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}$, with estimate $\log(n_{11}n_{22}/n_{12}n_{21})$. The estimated variance is then $n_{11}^{-1} + n_{12}^{-1} + n_{21}^{-1} + n_{22}^{-1}$.

c) *Logit and Multinomial Response Models.* Suppose that the sampling scheme is product-multinomial where each multinomial has exactly two categories. Without loss of generality, we can write $n = (n_{11}, \ldots, n_{r1}, n_{12}, \ldots, n_{r2})'$ where the pairs (n_{i1}, n_{i2}) are the multinomial outcomes. (This two-subscript notation will be used for all vectors discussed.)

Logits are defined by $\log(m_{i1}/m_{i2}) = \mu_{i1} - \mu_{i2}$. Let $\eta = (\mu_{11} - \mu_{12}, \ldots, \mu_{r1} - \mu_{r2})'$ be the vector of logits. A logit model is a model $\eta = Z\beta_*$ for some β_*, where Z is an $r \times p$ matrix.

We wish to show that the logit model defines a log-linear model for μ. Let I_r be an $r \times r$ identity matrix, and let $L' = [I_r, -I_r]$, so that $\eta = L'\mu$. Then the restriction placed on μ is that $\mu \in \mathcal{M}$, where $\mathcal{M} = \{\mu | L'\mu = Z\beta_* \text{ for some } \beta_*\}$. Because of the product-multinomial sampling, we have $X_0 = [I_r, I_r]'$. Now let

$$X_* = \begin{bmatrix} Z \\ 0_{rp} \end{bmatrix},$$

where 0_{rp} is an $r \times p$ matrix of zeros. It is easily seen that $\mathcal{M} = \{\mu | L'\mu = L'X_*\beta_* \text{ for some } \beta_*\}$. Arguing as in Section 3.3 of Christensen (1996b), $\mathcal{M}$ is a vector space, so the logit model has defined a log-linear model.

EXAMPLE 12.4.2. For a 3×2 table, a linear logit model is defined by $\mu_{i1} - \mu_{i2} = \gamma_0 + \gamma_1 t_i$, $i = 1, 2, 3$. The equation $L'\mu = L'X_*\beta_*$ becomes

$$\begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{21} \\ \mu_{31} \\ \mu_{12} \\ \mu_{22} \\ \mu_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix}.$$

Continuing as in Section 3.3, $\mathcal{M}$ can be rewritten as $\mathcal{M} = \{\mu | \mu = \mu_0 + \mu_1, \text{ where } \mu_0 \perp C(L) \text{ and } \mu_1 \in C(X_*)\}$. Thus, $\mathcal{M}$ is the space spanned by the columns of X_* and any spanning set for the orthogonal complement of $C(L)$. In particular, X_0 is a matrix with $C(X_0) = C(L)^\perp$. We can write the log-linear model as $\mu = X_0\beta_0 + X_*\beta_*$.

It was assumed at the beginning of the discussion that the sampling was product-multinomial with two categories in each multinomial. Normally, we consider only log-linear models $\mu = Xb$ such that $C(X_0) \subset C(X)$. In the current case, we have defined the logit model, i.e., $\mu \in \mathcal{M}$, $\mathcal{M} = \{\mu | L'\mu = L'X_*\beta_* \text{ for some } \beta_*\}$. We have then shown that $\mathcal{M} = C(X_0, X_*)$, so that a logit model must satisfy the condition $C(X_0) \subset \mathcal{M}$. It is interesting to note that even if the two-category product-multinomial sampling scheme

had not been assumed, the logit model would still correspond to a log-linear model consistent with that sampling scheme.

EXAMPLE 12.4.3. Write the log-linear version of the logit model as $\mu = X\beta$, where $X = [X_0, X_*]$ and $\beta' = [\beta_0', \beta_*']$. Because the MLEs satisfy $X'n = X'\hat{m}$, we have $X_0'n = X_0'\hat{m}$, i.e., $n_{i\cdot} = \hat{m}_{i\cdot}$ for $i = 1, \ldots, r$. The log-linear model must be of the form

$$\log(m_{ij}) = u_{1(i)} + \cdots.$$

The notation we have used is quite general, but it lends itself best to two-dimensional tables. Consider a four-dimensional log-linear model with a logit model in the last variable. If the observations are n_{hijk}, we have done nothing but substitute the three subscripts hij for the one subscript k. The argument presented here implies that the MLEs must satisfy $n_{hij\cdot} = \hat{m}_{hij\cdot}$ and the log-linear model must be of the form

$$\log(m_{hijk}) = u_{123(hij)} + \cdots.$$

Consider now the problem of estimating $\rho_1'\eta$; we can write $\rho = L\rho_1$ so that $\rho_1'\eta = \rho_1'L'\mu = \rho'\mu$. Because of the particular structures of L and X_* and the fact that $C(L) \perp C(X_0)$, the estimate of $\rho_1'\eta$ is

$$\rho_1'\hat{\eta} = \rho'\hat{\mu} = \rho_1'L'(X_0\hat{\beta}_0 + X_*\hat{\beta}_*) = \rho_1'L'X_*\hat{\beta}_* = \rho_1'Z\hat{\beta}_*.$$

The estimates in the logit model come directly from the log-linear model and all the asymptotic distribution results continue to apply. In particular, the estimate of $\eta = Z\beta_*$ is $\hat{\eta} = Z\hat{\beta}_*$, where $\hat{\beta}_*$ is estimated from the log-linear model.

Consider a logit model $\eta = Z\beta_*$ and a corresponding log-linear model $\mu = X\beta$, where $X = [X_0, X_*]$ and $\beta' = [\beta_0', \beta_*']$. We wish to be able to test the adequacy of a reduced logit model, say $\eta = Z_1\gamma_*$, where $C(Z_1) \subset C(Z)$. If the log-linear model corresponding to $\eta = Z_1\gamma_*$, say $\mu = X_1\gamma$, has $C(X_1) \subset C(X)$, then the test can proceed immediately from log-linear model theory. If Z_1 is a $r \times p_1$ matrix, we can write $X_{1*}' = [Z_1', 0_{rp_1}']$ and $X_1 = [X_0, X_{1*}]$. Clearly, if $C(Z_1) \subset C(Z)$, we have $C(X_{1*}) \subset C(X_*)$ and $C(X_1) \subset C(X)$, so the test can proceed.

The hypothesis that a logit model $\eta = Z_1\gamma_*$ fits the data relative to a general log-linear model $\mu = X\beta$ is equivalent to hypothesizing, for X_{1*} with $C(X_{1*}) \subset C(X)$, that $\mu \in \mathcal{M}$, where $\mathcal{M} = \{\mu | \mu \in C(X)$ and $L'\mu = L'X_{1*}\gamma$ for some $\gamma\}$. We can rewrite $\mathcal{M}$ as $\mathcal{M} = \{\mu | \mu = \mu_0 + \mu_1,$ where $\mu_1 \in C(X_{1*}),\ \mu_0 \in C(X)$ and $\mu_0 \perp C(L)\}$. Thus, $\mathcal{M}$ is the space spanned by the columns of X_{1*} and any spanning set for the subspace of $C(X)$ orthogonal to $C(L)$. The usual test for lack of fit of a logit model is $H_0 : \mu \in \mathcal{M}$ versus $H_A : \mu \in \mathbf{R}^q$, i.e., $C(X) = \mathbf{R}^q$.

Many types of multinomial response models can be written as log-linear models using the method outlined here. An exception are continuation ratio models. They do not correspond to a single log-linear model.

d) *Estimation of Parameters.* Estimation of parameters in log-linear models is very similar to that in standard linear models. A standard linear model

$$Y = X\beta + e, \qquad \mathrm{E}(e) = 0$$

implies that

$$\mathrm{E}(Y) = X\beta.$$

The least squares estimate of $X\beta$ is $\hat{Y} = MY$. The least squares estimate of $\rho' X\beta$ is $\rho' M\hat{Y} = \rho' MY$.

Similarly, in a log-linear model we have

$$\log(m) \equiv \mu = Xb.$$

Computer programs often give the MLE of m, i.e., $\hat{m}$. From this, one can obtain $\hat{\mu} = \log(\hat{m})$. Because $\hat{\mu} \in C(X)$, the MLE of $\rho' Xb$ is $\rho'\hat{\mu} = \rho' M\hat{\mu}$.

The key to finding the estimate of an estimable function $\lambda'\beta$ or $\lambda' b$ is in obtaining $M\rho$ so that $\lambda' = \rho' X = \rho' MX$. Given $M\rho$, estimates in the standard linear model can be obtained from Y and estimates in a log-linear model can be obtained from $\hat{\mu}$. Finding such a vector $M\rho$ depends only on λ and X. It does not depend on whether a linear or a log-linear model is being fitted. Christensen (1996b) discusses, in great detail, how to find estimates of estimable functions for standard linear models. The procedure amounts to finding $M\rho$. Precisely the same vectors $M\rho$ work for log-linear models. In other words, if one knows how to use Y to estimate something in a standard linear model, exactly the same technique applied to $\hat{\mu}$ will give the estimate in a log-linear model.

EXAMPLE 12.4.4. Consider a two-dimensional table with parameterization

$$\mu_{ij} = \gamma + \alpha_i + \beta_j + (\alpha\beta)_{ij}\,.$$

In discussions of log-linear models, this model would commonly be written as

$$\mu_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}\,,$$

but it is the same model with either parameterization. Estimates follow just as in a two-way ANOVA. To simplify this as much as possible, let $z_{ij} = \hat{\mu}_{ij}$ and assume the "usual" side conditions, then

$$\begin{aligned}
\hat{\gamma} &= \bar{z}_{\cdot\cdot}\,,\\
\hat{\alpha}_i &= \bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot}\,,\\
\hat{\beta}_i &= \bar{z}_{\cdot j} - \bar{z}_{\cdot\cdot}\,,\\
\widehat{(\alpha\beta)}_{ij} &= z_{ij} - \bar{z}_{i\cdot} - \bar{z}_{\cdot j} + \bar{z}_{\cdot\cdot}\,.
\end{aligned}$$

It seems very reasonable (to me at any rate) to restrict estimation to estimable functions of b. In that case, the choice of side conditions is of no importance.

Tests and confidence intervals for $\rho' Xb$ can be based on Theorem 12.3.3. A large sample approximation is

$$\frac{\rho'\hat{\mu} - \rho' Xb}{\sqrt{\rho'(A - A_0)D^{-1}(m)\rho}} \sim N(0,1).$$

Of course, $AD^{-1}(m)$ has to be estimated in order to get a standard error. [$A_0 D^{-1}(m)$ does not depend on unknown parameters.] As indicated in application (b), variances are easy to find in the saturated model; unfortunately, the estimable functions of b are generally quite complicated in the saturated model. If one is willing to use side conditions, the side conditions can sometimes give the illusion that the estimable functions are not complicated.

12.5 Proofs of Lemma 12.3.2 and Theorem 12.3.8

Two results from advanced calculus are needed. Recall that if $F : \mathbf{R}^q \times \mathbf{R}^p \to \mathbf{R}^p$, then $dF(x,y)$ is a p by $q+p$ matrix. Partition $dF(x,y)$ into a $p \times q$ matrix, say $d_x F = [\partial F_i/\partial x_j]$, and a $p \times p$ matrix, $d_y F = [\partial F_i/\partial y_j]$.

The Implicit Function Theorem. If $F : \mathbf{R}^q \times \mathbf{R}^p \to \mathbf{R}^p$, $F(a,c) = 0$, $F(a,y)$ is differentiable, and $d_y F$ is nonsingular at $y = c$, then $F(x,y) = 0$ determines y uniquely as a function of x in a neighborhood of (a,c). This unique function, say $\xi(x)$, is differentiable and satisfies $\xi(a) = c$ and $F(x, \xi(x)) = 0$ for x in a neighborhood of a.

Proof. See Bartle (1964). □

Corollary 12.5.1. $d\xi(x) = -[d_y F]^{-1}[d_x F]$ where $y = \xi(x)$.

Proof. See Bartle (1964). □

Lemma 12.5.2. If a is a scalar and n is a q vector of counts, then

(1) $\hat{m}(an) = a\hat{m}(n)$

(2) $\hat{\mu}(an) = [\log(a)]J + \hat{\mu}(n)$.

Proof. $\hat{m}(an)$ is the unique solution of $[an - m]'X = 0$ with

$\log(\hat{m}(an)) \in C(X)$. We will show that $a\hat{m}(n)$ is also a solution with $\log(a\hat{m}(n)) \in C(X)$, so $\hat{m}(an) = a\hat{m}(n)$. $\hat{m}(n)$ is the unique solution of $[n - m]'X = 0$ with $\log(\hat{m}(n)) \in C(X)$. Clearly, if $[n - \hat{m}(n)]'X = 0$, then $[an - a\hat{m}(n)]'X = 0$, but $\log(a\hat{m}(n)) = [\log(a)]J + \log(\hat{m}(n)) \in C(X)$ because both J and $\log(\hat{m}(n))$ are in $C(X)$.

Taking logs gives $\hat{\mu}(an) = [\log(a)]J + \hat{\mu}(n)$. □

Lemma 12.5.3. $\hat{\mu}(m^*) = \mu^*$ and $\hat{m}(m^*) = m^*$.

Proof. By definition, $m^* = m(b^*)$, so b^* is a solution of $[m^* - m(b)]'X = 0$. Since $\hat{\mu}(m^*)$ is unique, we must have $\hat{\mu}(m^*) = Xb^* = \mu^*$.

$$\hat{m}(m^*) = \exp[\hat{\mu}(m^*)] = \exp[\mu^*] = m^*. \qquad \square$$

Lemma 12.3.2 $N^{1/2}(\hat{\mu}_N - \mu_N) - \left(AD^{-1}\right) N^{-1/2}(n_N - m_N) \xrightarrow{P} 0.$

Proof. The MLE $\hat{\mu}_N$ is defined by $\hat{\mu}_N = X\hat{b}_N$, where $\hat{b}_N$ is a function of n_N which is defined implicitly as the solution to $df_{n_N}(b) = [n_N - m(b)]'X = 0$.

The proof follows from investigating the properties of the Taylor's expansion

$$\hat{\mu}(n) = \hat{\mu}(n_0) + d\hat{\mu}(n_0)(n - n_0) + o(\|n - n_0\|). \tag{1}$$

The expansion is applied with $n = N^{-1}n_N$ and $n_0 = N^{-1}m_N = m^*$. Rewriting (1) gives

$$\hat{\mu}(N^{-1}n_N) - \hat{\mu}(m^*) - d\hat{\mu}(m^*)(N^{-1}n_N - m^*) = o(\|N^{-1}n_N - m^*\|). \tag{2}$$

We examine the terms $\hat{\mu}(N^{-1}n_N) - \hat{\mu}(m^*)$ and $d\hat{\mu}(m^*)$ separately.

(a) We show that for any observations vector n_N,

$$\hat{\mu}(N^{-1}n_N) - \hat{\mu}(m^*) = \hat{\mu}(n_N) - \mu_N.$$

By Lemmas 12.5.2 and 12.5.3,

$$\hat{\mu}(N^{-1}n_N) - \hat{\mu}(m^*) = [\log N^{-1}]J + \hat{\mu}(n_N) - \mu^*.$$

Since $\mu_N = [\log N]J + \mu^*$, we have the result.

(b) We characterize the $q \times q$ matrix $d\hat{\mu}(m^*)$. $\hat{\mu}(n) = X\hat{b}(n)$, so $d\hat{\mu}(n) = X[d\hat{b}(n)]$, with $\hat{b}(n)$ defined implicitly as a zero of $F(n, b) = X'[n - m(b)]$. For any fixed vector b_0, let $n_0 = m(b_0)$. Then $F(n_0, b_0) = 0$, so by the Implicit Function Theorem, there exists $\hat{b}(n)$ such that if n is close to n_0, $F(n, \hat{b}(n)) = 0$ and (from Corollary 12.5.1) $d\hat{b}(n) = -[d_{\hat{b}}F]^{-1}[d_nF]$. To

find $d\hat{b}(n)$, we need $dF(n,b) = [X', -X'dm(b)]$. From (12.2.9), $dm(b) = D(m(b))X$, so $dF(n,b) = [X', -X'D(m(b))X]$,

$$d\hat{b}(n) = [X'D(\hat{m})X]^{-1}X',$$

and $d\hat{\mu}(n) = X[X'D(m(b))X]^{-1}X'$. In particular, $d\hat{\mu}(n_0)$ is always defined.

We need $d\hat{\mu}(m^*)$. From Lemma 12.5.3, we have that $F(m^*, \hat{b}(m^*)) = 0$, so $d\hat{\mu}(m^*)$ is defined and $d\hat{\mu}(m^*) = X[X'D(\hat{m}(m^*))X]^{-1}X'$. Again, from Lemma 12.5.3, $\hat{m}(m^*) = m^*$, so $D(\hat{m}(m^*)) = D(m^*) = D$ and $d\hat{\mu}(m^*) = X[X'DX]^{-1}X' = AD^{-1}$.

(c) Using $\|N^{-1}n_N - m^*\| = O_p\left(N^{-1/2}\right)$ and the results of (a) and (b) in (2) gives

$$\hat{\mu}(n_N) - \mu_N - \left(AD^{-1}\right) N^{-1}(n_N - m_N) = o\Big(O_p\Big(N^{-1/2}\Big)\Big) = o_p\Big(N^{-1/2}\Big).$$

Multiplying through by $N^{1/2}$ gives

$$N^{1/2}(\hat{\mu}_N - \mu_N) - \left(AD^{-1}\right) N^{-1/2}(n_N - m_N) = o_p(1). \qquad \square$$

Theorem 12.3.8.

$$-2N^{-1}[\ell(n_N, \hat{\mu}_{1N}) - \ell(n_N, \hat{\mu}_N)] \xrightarrow{P} -2[\ell(m^*, \hat{\mu}_1(m^*)) - \ell(m^*, \mu^*)].$$

$\mu_N \notin C(X_1)$ if and only if the right-hand side is positive.

Proof.

$$\begin{aligned} &-2N^{-1}[\ell(n_N, \hat{\mu}_{1N}) - \ell(n_N, \hat{\mu}_N)] \\ &\quad = -2N^{-1}[\ell(n_N, \hat{\mu}_{1N}) - \ell(n_N, \mu_N)] + 2N^{-1}[\ell(n_N, \hat{\mu}_N) - \ell(n_N, \mu_N)]. \end{aligned}$$

As in Theorem 12.3.6,

$$2N^{-1}[\ell(n_N, \hat{\mu}_N) - \ell(n_N, \mu_N)] \xrightarrow{P} 0,$$

so we need only investigate the behavior of

$$\begin{aligned} &-2N^{-1}[\ell(n_N, \hat{\mu}_{1N}) - \ell(n_N, \mu_N)] \\ &\qquad\qquad = -2N^{-1}[n_N'(\hat{\mu}_{1N} - \mu_N) - J'(\hat{m}_{1N} - m_N)]. \end{aligned}$$

From Theorem 12.3.1, $N^{-1}n_N \xrightarrow{P} m^*$. As in the proof of Lemma 12.3.2, $\hat{\mu}_{1N} - \mu_N = \hat{\mu}_1\left(N^{-1}n_N\right) - \mu^*$ and $N^{-1}(\hat{m}_{1N} - m_N) = \hat{m}_1\left(N^{-1}n_N\right) - m^*$. By the continuity of $\hat{m}_1(\cdot)$ and $\hat{\mu}_1(\cdot)$ (ensured by the Implicit Function Theorem), $\hat{m}_1\left(N^{-1}n_N\right) \xrightarrow{P} \hat{m}_1(m^*)$ and $\hat{\mu}_1\left(N^{-1}n_N\right) \xrightarrow{P} \hat{\mu}_1(m^*)$, so

$$\begin{aligned} &-2N^{-1}[\ell(n_N, \hat{\mu}_{1N}) - \ell(n_N, \mu_N)] \\ &\xrightarrow{P} \quad -2[m^{*\prime}(\hat{\mu}_1(m^*) - \mu^*) - J'(\hat{m}_1(m^*) - m^*)] \\ &= \quad -2[\ell(m^*, \hat{\mu}_1(m^*)) - \ell(m^*, \mu^*)]. \end{aligned}$$

Since $\hat{\mu}(m^*) = \mu^*$, $\ell(m^*, \mu^*)$ is the unique maximum of $\ell(m^*, \mu)$ for $\mu \in C(X)$. Since $\hat{\mu}_1(m^*)$ is in $C(X)$, if $\hat{\mu}_1(m^*) \neq \mu^*$,

$$-2[\ell(m^*, \hat{\mu}_1(m^*)) - \ell(m^*, \mu^*)] > 0.$$

This occurs whenever $\mu^* \notin C(X_1)$ because $\hat{\mu}_1(m^*) \in C(X_1)$. Finally, $\mu^* \notin C(X_1)$ if and only if $\mu_N \notin C(X_1)$. □

Chapter 13

Bayesian Binomial Regression

Standard methods for analyzing binomial regression data rely on asymptotic inferences. Bayesian methods performed using simple computations apply for any sample size. We discuss Bayesian inferences for binomial regression with an emphasis on inferences for the probability of "success." Furthermore, we illustrate diagnostic tools, perform model selection among non-nested models, and examine the sensitivity of the Bayesian methods. This chapter is closely related to Bedrick, Christensen, and Johnson (1997) and to earlier drafts of that article.

Section 1 introduces Bayesian binomial regression. Section 2 discusses standard Bayesian inference procedures with an emphasis on the predictive distribution. Section 3 presents Bayesian diagnostics including influence measures, global model checking methods, and a procedure for selection of the appropriate link function. Section 4 discusses computations.

13.1 Introduction

The purpose of this chapter is to illustrate the simplicity of a fully Bayesian approach to binomial regression models. Historically, it has been difficult to specify realistic prior information on regression coefficients in nonlinear models, and computations for inference and diagnostics were difficult due to intractable integrations. These difficulties no longer exist. We illustrate a fairly complete analysis for two data sets using methods that are simple and easy to apply. In particular, we discuss a method for specifying the prior distribution that focuses on binomial probabilities, rather than es-

oteric regression coefficients. For computations, we focus on Monte Carlo methods because of their flexibility and their ease of implementation. We show how Monte Carlo sampling is used for prediction, making inferences on regression coefficients and probabilities, diagnostics, model checking, link selection, and sensitivity analysis of the prior.

Leonard (1972) first discussed Bayesian hierarchical models for binomial data. Zellner and Rossi (1984) gave a overview of Bayesian methods for binomial regression models. Johnson and Geisser (1985) and Johnson (1985) introduced general Bayesian predictive and estimative case deletion diagnostics that apply to binomial regression. We integrate these ideas along with Box's (1980) work on model checking to provide a variety of tools appropriate for analyzing binomial response data.

Consider regression data (y_i, x_i'), $i = 1, \ldots, n$, where the x_i's are known k vectors of covariates and the y_i's are independent binomial random variables with N_i trials. The probability of success p for any *single* trial y with covariate x is $F(x'\beta)$, i.e., $F(x'\beta) \equiv p \equiv \Pr(y = 1|x, \beta)$. Here, the vector β is an unknown k vector of regression coefficients. Although the function $F(\cdot)$ could be an arbitrary cdf, we consider logistic, probit, and complementary log-log regression models in which $F(x'\beta)$ is modeled as one of

$$F(x'\beta) = \begin{cases} e^{x'\beta} \Big/ \left[1 + e^{x'\beta}\right] & \text{Logistic} \\ \Phi(x'\beta) & \text{Probit} \\ 1 - \exp\left[-e^{x'\beta}\right] & \text{Complementary log-log .} \end{cases}$$

Here, $\Phi(u)$ is the cdf of a standard normal distribution. The success probability p is related to β through $F^{-1}(p) = x'\beta$, which is the link function from Chapter 9. For the logistic, probit, and complementary log-log models, $F^{-1}(p) = \log\{p/(1-p)\}$, $\Phi^{-1}(p)$, and $\log\{-\log(1-p)\}$, respectively. The likelihood function for the complete data $Y = (y_1, \ldots, y_n)'$ is

$$L(\beta|Y) \equiv \prod_{i=1}^{n} L(\beta|y_i) \equiv \prod_{i=1}^{n} \binom{N_i}{y_i} [F(x_i'\beta)]^{y_i} [1 - F(x_i'\beta)]^{N_i - y_i}. \quad (1)$$

For a prior distribution on β, say $\pi(\beta)$, obtaining posterior and predictive distributions requires computing the posterior of β,

$$\pi(\beta|Y) = \frac{L(\beta|Y)\pi(\beta)}{\int L(\beta|Y)\pi(\beta)d\beta}.$$

Most interesting aspects of a Bayesian analysis can be obtained from various integrals involving this posterior density. Integrals involving $\pi(\beta|Y)$ are intractable, so we must use approximations.

Monte Carlo methods yield a discrete approximation to the posterior distribution that takes values β^r with probability $\tilde{q}_r$, $r = 1, \ldots, t$. Methods

for obtaining a discrete approximation are discussed in Section 4. Given a function $h(\beta)$, the posterior expectation $\mathrm{E}\{h(\beta) \mid Y\}$ is approximated by

$$\int h(\beta)\pi(\beta|Y)d\beta \doteq \sum_{r=1}^{t} h(\beta^r)\tilde{q}_r. \tag{2}$$

Typically, the Strong Law of Large Numbers implies that the error in the approximation converges almost surely to zero as the simulation sample size t increases.

13.2 Bayesian Inference

13.2.1 Specifying the Prior and Approximating the Posterior

Bayesian inference requires the specification of a prior distribution $\pi(\beta)$. In the past, several methods of specifying priors for binomial regression problems have been used. The standard approach has been to assume either a normal distribution for β or the "noninformative" diffuse prior $\pi(\beta) = 1$. These are convenient in large sample situations where the posterior on β is approximately normal. See Zellner and Rossi (1984) for relevant discussion. Another type of prior focuses on the assessment of "success" probabilities for various choices of covariate values, rather than on the assessment of regression coefficients.

Example 13.2.1. Consider a simple situation with $k = 2$. Imagine that we are recruiting statistics students into a graduate program. We will attempt to recruit from two populations: domestic students ($i = 1$) and international students ($i = 2$). If N_1 domestic students apply and N_2 international students apply, assuming independence of students we successfully recruit $y_1 \sim \mathrm{Bin}(N_1, p_1)$ domestic students and $y_2 \sim \mathrm{Bin}(N_2, p_2)$ international students. We can write a one-way ANOVA logit model

$$\log\{p_i/(1-p_i)\} = \mu + \alpha_i,$$

$i = 1, 2$. This model is overparameterized, so we impose the side condition $\alpha_1 = 0$ to make the model a logistic regression. We now have

$$\log\{p_1/(1-p_1)\} = \mu, \qquad \log\{p_2/(1-p_2)\} = \mu + \alpha_2.$$

The graduate advisor has specified prior distributions $p_1 \sim \mathrm{Beta}(4, 4)$ and $p_2 \sim \mathrm{Beta}(4, 1)$, reflecting (in part) the beliefs that about $80\% = \mathrm{E}(p_2) = 4/(4+1)$ of the international students and half, $[4/(4+4)]$, of the domestic students will be successfully recruited. The prior specification includes the

assumption that p_1 and p_2 are independent. Having placed a joint distribution on p_1 and p_2, it is a calculus problem to determine the corresponding distribution on the "regression" parameters μ and α_2. We discuss the exact procedure later. While we assumed that the distributions of p_1 and p_2 were independent, the approach can, in theory, be carried out with any joint distribution for p_1 and p_2. The problem is not in doing the calculus, but in specifying a realistic joint distribution when the independence assumption is not appropriate.

The independence assumption is a key part of the procedure. With p_1 and p_2 independent, if we were told the value of p_1, we should not be inclined to revise our thinking about p_2. That certainly seems reasonable if we are told that p_1 is something near its expected value .5. It seems less reasonable if we are told, say, that $p_1 \geq .95$. Knowing that $p_1 \geq .95$ would probably make us want to revise our distribution of p_2 to make larger values more probable. However, .95 is 2.7 prior standard deviations above the prior mean for p_1, so this event is extremely unlikely under the prior specification. If $p_1 \geq .95$ is more likely than the original prior specification allows, the entire prior should be recalibrated, at which point the independence assumption may be called in question. However, if, after reflection, those situations that might cause concern about the independence assumption are thought unlikely, then we believe the independence assumption is reasonable.

Lack of independence can also occur if the international students were thought to be very similar to the domestic students regardless of the behavior of the domestic students. In this case, knowing $\tilde{p}_1$ is highly informative about $\tilde{p}_2$ and our prior is not appropriate.

The main idea in Example 13.2.1 was to specify prior distributions for p_1 and p_2 rather than on the regression parameters μ and α_2. We do this because p_1 and p_2 have natural interpretations. In a simple logistic regression,

$$\log\{p/(1-p)\} = \beta_0 + \beta_1 \tau\,,$$

there are again only two regression parameters (β_0 and β_1), but there is no obvious choice for probabilities p_1 and p_2 at which to specify the prior. In such cases, we must pick two values, say $\tilde{\tau}_1$ and $\tilde{\tau}_2$, and specify prior distributions for $\tilde{p}_1$, the probability of success when $\tau = \tilde{\tau}_1$, and $\tilde{p}_2$, the probability of success when $\tau = \tilde{\tau}_2$.

EXAMPLE 13.2.2. *O-Ring Data.*
Consider fitting a simple regression model on temperature to the data in Table 2.1. Let p_i be the probability that any O-ring fails in case i and model this as $F^{-1}(p_i) = \beta_0 + \beta_1 \tau_i = x_i'\beta$, where τ_i is the temperature. Our prior is defined by giving independent distributions to the probabilities of O-ring failure at temperatures $\tilde{\tau}_1 = 55$ and $\tilde{\tau}_2 = 75$ degrees Fahrenheit. Write

$$\beta_0 + \beta_1 \tilde{\tau}_i = [1, \tilde{\tau}_i] \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \tilde{x}_i'\beta$$

and define $\tilde{p}_1$ and $\tilde{p}_2$ by $\tilde{p}_i = F(\tilde{x}_i'\beta)$. The $\tilde{\tau}_i$'s should be chosen in the expected range of the observed temperatures but far enough apart so that information about the corresponding probabilities can be reasonably assumed independent. The selected temperatures should also be amenable to expert opinion. Our priors on $\tilde{p}_1$ and $\tilde{p}_2$ are Beta$(1, .577)$ and Beta$(.577, 1)$ respectively. The prior on $\tilde{p}_1$ was chosen because it has a "J" shape and gives $\Pr[\tilde{p}_1 > 1/2] = 2/3$. The prior on $\tilde{p}_2$ has a "J" shape and gives $\Pr[\tilde{p}_2 < 1/2] = 2/3$.

The prior on $\beta = (\beta_0, \beta_1)'$ is determined using the change-of-variables method. Under the logistic model, the prior on β is a data augmentation prior (DAP) in the sense that it has the same functional form as the likelihood function, i.e.,

$$\pi(\beta) \propto \prod_{i=1}^{2} [F(\tilde{x}_i'\beta)]^{\tilde{y}_i} [1 - F(\tilde{x}_i'\beta)]^{\tilde{N}_i - \tilde{y}_i},$$

where $\tilde{N}_1 = \tilde{N}_2 = 1.577$, $\tilde{y}_1 = 1$, and $\tilde{y}_2 = .577$. With this DAP, the prior on $\tilde{p}_1$ can be thought of as one prior O-ring failure out of 1.577 trials at $\tilde{\tau}_1 = 55$, and for $\tilde{p}_2$, it can be thought of as .577 prior O-ring failures out of 1.577 trials at $\tilde{\tau}_2 = 75$. The weight attached to the prior is equivalent to $\tilde{N}_1 + \tilde{N}_2$ "prior" observations, about 3. The posterior density for β also has the same functional form as the likelihood, i.e.,

$$\pi(\beta|Y) \propto \prod_{i=1}^{n} [F(x_i'\beta)]^{y_i} [1 - F(x_i'\beta)]^{N_i - y_i} \prod_{i=1}^{2} [F(\tilde{x}_i'\beta)]^{\tilde{y}_i} [1 - F(\tilde{x}_i'\beta)]^{\tilde{N}_i - \tilde{y}_i}.$$

Many standard computer programs, e.g., GLIM and SPLUS, can be used to find the posterior mode β_M and an asymptotic dispersion matrix $\Sigma(\beta_M)$ for the posterior. To compute the mode, simply augment the observed data with a prior "binomial" observation at 55 degrees consisting of 1.577 trials and 1 observed O-ring failure and include a prior observation at 75 degrees with 1.577 trials and .577 O-ring failures. The posterior mode of β is the maximum likelihood estimate (MLE) from the augmented data. The asymptotic covariance matrix computed from the augmented data is the asymptotic dispersion matrix for the posterior. These quantities are of interest in themselves and can also be used to create a good discrete approximation to the posterior.

Figures 13.1 and 13.2 give contour plots of the prior and posterior distributions on β, respectively. Note the high correlation between β_0 and β_1 in both the prior and the posterior. The posterior exhibits appreciable skewness, with longer tails in the direction of small slopes and large intercepts. The high correlation between β_0 and β_1 is largely eliminated if we standardize the temperature to have mean zero, i.e., if we change the model to $\text{logit}(p_i) = \beta_0 + \beta_1(\tau_i - \bar{\tau}.)$. For some problems, this may be preferable to ease the computational burden. As there were no computational difficulties

with these data, and as prediction and model validation are independent of the regression parameterization, we consider only the original version of the model.

In general, we derive the prior on β for a model with k regression parameters from a prior elicited on success probabilities $\tilde{p}_i$ at k suitably selected predictor vectors $\tilde{x}_i$. We place independent Beta$(\tilde{y}_i, \tilde{N}_i - \tilde{y}_i)$ priors on the $\tilde{p}_i$, regardless of the choice of the link function. For an arbitrary link function, the induced prior on β has the form

$$\pi(\beta) \propto \prod_{i=1}^{k} [F(\tilde{x}_i'\beta)]^{\tilde{y}_i - 1} [1 - F(\tilde{x}_i'\beta)]^{\tilde{N}_i - \tilde{y}_i - 1} f(\tilde{x}_i'\beta),$$

where $f(\cdot)$ is the first derivative of the function $F(\cdot)$. In the case of logistic regression,

$$\pi(\beta) \propto \prod_{i=1}^{k} [F(\tilde{x}_i'\beta)]^{\tilde{y}_i} [1 - F(\tilde{x}_i'\beta)]^{\tilde{N}_i - \tilde{y}_i}, \tag{1}$$

which has the same form as the likelihood function. Therefore, (1) is a data augmentation prior (DAP), so named because the likelihood times the prior has the form of a likelihood with additional "prior" data $(\tilde{y}_i, \tilde{N}_i)$, $i = 1, \ldots, k$. In other words, for the logistic model we can think of the parameters of the prior distribution as a prior sample size $\tilde{N}_i$ and a prior number of successes $\tilde{y}_i$ corresponding to the vector of predictors $\tilde{x}_i$.

Incidentally, this procedure can also be executed with priors for the $\tilde{p}_i$'s other than betas. In fact, with different link functions, different distributions on the $\tilde{p}_i$'s lead to different DAPs. (Note that the likelihood depends on the link function, so DAPs depend on the link function.)

We now consider our primary example.

EXAMPLE 13.2.3. *Trauma Data.*
We analyze data on a randomly selected subset of 300 patients admitted to the University of New Mexico Trauma Center between the years 1991 and 1994. For each patient, we have their injury severity score (ISS), their revised trauma score (RTS), their AGE, the type of injuries (TI), that is, whether they were blunt ($TI = 0$), e.g., the result of a car crash, or penetrating ($TI = 1$), e.g., gunshot wounds, and the dependent variable, whether the patient eventually survived the injuries. The ISS is an overall index of a patient's injuries based on the approximately 1300 injuries catalogued in the Abbreviated Injury Scale. The ISS can take on values from 0 for a patient with no injuries to 75 for a patient with severe injuries in three or more body areas. The RTS is an index of physiologic injury and is constructed as a weighted average of an incoming patient's systolic blood pressure, respiratory rate, and Glasgow Coma Scale. The RTS takes on values from 0 for a patient with no vital signs to 7.84 for a patient with

FIGURE 13.1. O-Ring Data: Prior on β

FIGURE 13.2. O-Ring Data: Posterior on β

normal vital signs. The data are available electronically from STATLIB as well as from my web homepage:

`http://stat.unm.edu/~fletcher`

Additional information is given in the Preface.

Figure 13.3 gives side-by-side boxplots comparing the 278 survivors and 22 fatalities on RTS, ISS, and AGE. Seventeen of the 225 patients with blunt injuries died. Five of the 75 patients with penetrating injuries died.

The data were provided by Dr. Turner Osler, a trauma surgeon at the University of Vermont and former head of the Burn Unit at the University of New Mexico Trauma Center. Dr. Osler proposed a logistic regression model to estimate the probability of a patient's death using an intercept and predictors ISS, RTS, patient's AGE (used as a surrogate for physiologic reserve), TI, and an interaction between AGE and TI. Similar logistic models are used by trauma centers throughout the United States. Dr. Osler's expert opinions formed the basis for our prior distribution.

To induce a proper prior distribution on the $k = 6$ dimensional vector β, we require a joint distribution on death probabilities for 6 sets of conditions $\tilde{x}_i' = (1, ISS_i, RTS_i, AGE_i, TI_i, AGE_i \times TI_i)$. Based on discussions with our expert and two-dimensional plots of the data, we defined a 2^4 factorial design having ISS at levels 25 and 41, RTS at levels 3.34 and 7.84, AGE at levels 10 and 60, and TI at levels 0 and 1. The idea was to pick values of the variables that were relatively extreme within the data but still had substantial probabilities for both success and failure. The prior conditions were chosen as a 1/4 replicate of this 2^4 with two center points. However, the center points were taken to be values that could actually exist — none of ISS, RTS, and TI are truly continuous variables. In fact, TI is a binary variable, so one "center point" was taken with $TI = 0$ and the other with $TI = 1$. Bedrick, Christensen, and Johnson (1996) (henceforth referred to as BCJ) recommend calculating the condition number of the matrix $\tilde{X} = (\tilde{x}_1, \ldots, \tilde{x}_k)'$ to ascertain that the chosen $\tilde{x}_i$'s are not too close or too far apart. See Belsley (1991) for discussion of condition numbers. Beta priors were found to be suitable for the $\tilde{p}_i$'s with parameters given in Table 13.1. Figure 13.4 gives plots of the priors on the $\tilde{p}_i$'s as well as the posteriors. The priors are generally consistent with the posteriors. Relative to the amount of data, the priors are not overwhelming, being the equivalent of $57.5 = \sum_{i=1}^{6} \tilde{N}_i$ observations compared to 300 data points. (The posterior densities were obtained by sampling from the discrete approximate posterior and smoothing the samples.)

Our initial discussion with Dr. Osler involved eliciting 1'st, 50'th, and 99'th percentiles for each $\tilde{p}_i$. These actually overspecify a beta distribution. We wrote a computer program to find the beta distributions that most nearly satisfied the specifications, plotted these distributions, and validated them with our expert.

The first probability $\tilde{p}_1$ corresponds to an individual that "has good

FIGURE 13.3. Trauma Data: Box Plots

FIGURE 13.4. Trauma Data: Priors and Posteriors on $\tilde{p}$'s

TABLE 13.1. Trauma Data: Prior Specification

	Design for Prior						Beta $(\tilde{y}_i, \tilde{N}_i - \tilde{y}_i)$	
i	$\tilde{x}_i'$						$\tilde{y}_i$	$\tilde{N}_i - \tilde{y}_i$
1	1	25	7.84	60	0	0	1.1	8.5
2	1	25	3.34	10	0	0	3.0	11.0
3	1	41	3.34	60	1	60	5.9	1.7
4	1	41	7.84	10	1	10	1.3	12.0
5	1	33	5.74	35	0	0	1.1	4.9
6	1	33	5.74	35	1	35	1.5	5.5

physiology, is 'not bad hurt,' does not have a lot of reserve," and for whom there is "added uncertainty due to age." The Beta(1.1, 8.5) suitably reflects Dr. Osler's uncertainty about $\tilde{p}_1$. The median of his prior is around .09. The second type of individual "has bad physiology, is very ill, but is young and resilient and is not so bad hurt." The prior for $\tilde{p}_2$ is Beta(3, 11) with median around .20. Incidentally, "bad physiology" and "very ill" apparently refer to bad RTS scores, while how badly hurt one is relates to ISS. The third individual has "bad physiology, a pretty bad injury, and there is much more uncertainty here due to the age factor." The prior is Beta(5.9, 1.7) with median around .8. Prior individual four "is young, resilient, and has a big injury." The prior is Beta(1.3, 12) with a median of around .07.

Dr. Osler had more difficulty with the 5'th and 6'th types of individuals because their conditions were both less extreme and more related than those already considered. The priors for $\tilde{p}_5$ and $\tilde{p}_6$ are Beta(1.1,4.9) with approximate median .15, and Beta(1.5, 5.5) with approximate median .19, respectively.

The assumption of independence seems reasonable with the possible exception of $\tilde{p}_5$ and $\tilde{p}_6$. If our expert were told that $\tilde{p}_5 = .3$, he would definitely want to revise his probability for $\tilde{p}_6$ upward. This is because he is fairly confident that the difference between these two probabilities, $\tilde{p}_6 - \tilde{p}_5$, is positive but reasonably small, while he is less certain about the magnitude of the probabilities themselves. Having $\tilde{p}_6 - \tilde{p}_5$ small but positive is reflecting his perception that penetrating injuries are worse than blunt ones but not a lot worse.

Because of our concern about possible lack of independence for the two values $\tilde{p}_6$ and $\tilde{p}_5$ only, we also considered a prior in which the information about $\tilde{p}_6$ was left out of the specification. This results in a partially informative prior (see BCJ, Sec. 4.1, for a full discussion) which is an improper DAP using five prior observations instead of the six required for a proper DAP. We found that all statistical inferences were essentially the same for the two priors, so we have presented results only for the full prior.

Finally, it should be pointed out that the process of coming up with a prior is very much a collaboration between the expert and the statisticians. The judgement and expertise of both are needed. It is also quite a bit of

work for everyone involved.

Bedrick, Christensen, and Johnson (1996, 1997) give further details on this approach to specifying priors for regression problems, including discussions of priors with order restrictions on the $\tilde{p}_i$'s and partial prior information. As mentioned above, a particularly useful form of partial prior information is specifying $k' < k$ values $\tilde{p}_i$.

The genesis of this approach lies with Tsutakawa (1975), Tsutakawa and Lin (1986), and Grieve (1988) who considered independent prior distributions on two probabilities of "success" in simple linear binomial regression problems. Tsutakawa and Lin (1986) argued that eliciting information about success probabilities should be much easier than eliciting information about regression coefficients, a position with which we heartily agree. This is clearly true if one entertains the possibility of two or more models, such as logistic regression versus probit regression. The regression coefficients for these two models require separate elicitations, whereas if one has elicited a prior for probabilities, it is straightforward to induce the requisite prior on β for either model.

BCJ extended the Tsutakawa approach to generalized linear models (GLMs) with multiple covariates. For a hypothetical observation $\tilde{y}_i$ with covariate vector $\tilde{x}_i$, BCJ specify a prior on the mean value $\mathrm{E}(\tilde{y}_i|\tilde{x}_i)$. This is done for k locations $\tilde{x}_i$, $i = 1, \ldots, k$, where k is the common dimension of the $\tilde{x}_i$'s. The prior on the regression coefficient vector β is induced by transforming the distribution on the $\mathrm{E}(\tilde{y}_i|\tilde{x}_i)$'s into a distribution on β. BCJ call such priors conditional means priors (CMPs) and elaborate on the approach in considerable detail. The conditional means provide parameters that are more intuitive than regression coefficients and thus easier to specify prior information for. BCJ also make connections between CMPs and DAPs. (Note that to make the GLM approach apply to binomial regression, just as in Chapter 9, the y_i's have to be defined as binomial proportions rather than our usual binomial counts.)

A key feature in this approach is assuming prior independence of the $\mathrm{E}(\tilde{y}_i|\tilde{x}_i)$'s. This assumption might be unreasonable if the $\tilde{x}_i$'s are "too close" together (cf. Grieve, 1988). There are also technical difficulties if they are "too far apart." BCJ (Sec. 5) examined these issues in detail.

13.2.2 Predictive Probabilities

The predictive probability of success in one new trial y with covariate x is

$$\Pr(y = 1|Y, x) = \mathrm{E}[F(x'\beta)|Y, x] = \int F(x'\beta)\pi(\beta|Y)\, d\beta. \tag{2}$$

Using the discrete approximation to the posterior gives

$$\Pr(y = 1|Y, x) \doteq \sum_{r=1}^{t} F(x'\beta^r)\tilde{q}_r.$$

FIGURE 13.5. O-Ring Data: Predictive Probabilities and MLEs

Unless specifically stated otherwise, the examples henceforth use logistic models.

EXAMPLE 13.2.2 CONTINUED. Figure 13.5 presents the Bayesian predictive probability of O-ring failure, $\Pr(y = 1|Y, x)$, and the MLE of the probability of an O-ring failure, $F(x'\hat{\beta}_{ml})$, as temperature varies from 30 degrees to 80 degrees. Predictive probabilities are less than the MLEs for temperatures below about 67 and are greater for larger temperatures.

The predictive probability of "success" can be interpreted in two ways. It is the subjective probability of at least one O-ring failure on the next flight at the given temperature. It is also the Bayes estimate of the proportion of flights at the given temperature in which there would be at least one O-ring failure. With the second interpretation, one may be interested in interval estimates. Geisser (1982) established that posterior interval estimates for probabilities can be viewed as (asymptotic) prediction intervals for the proportion of successes from a large number of future trials.

Figure 13.6 contains a plot of the predictive probabilities of O-ring failure as x varies from 30 to 80 degrees along with 90% interval estimates. In other words, it gives $\mathrm{E}[F(x'\beta)|Y, x]$ and a Bayesian posterior interval estimate based on our subjective prior. The Bayesian interval is obtained by determining the 5'th and 95'th percentiles of the approximate posterior distribution for $F(x'\beta)$, i.e., the distribution that takes values $F(x'\beta^r)$ with probability $\tilde{q}_r$. For low temperatures, the posterior distribution of $F(x'\beta)$ is highly skewed to the left; thus, the mean $\mathrm{E}[F(x'\beta)|Y, x]$ is lower than the median.

FIGURE 13.6. O-Ring Data: Predictive Probabilities and 90% Intervals

EXAMPLE 13.2.3 CONTINUED. Figure 13.7 presents predictive probabilities of death as a function of ISS for blunt and penetrating traumas. These are given for various values of RTS and AGE. Note that for 60-year-olds, there is essentially no difference in the probability of death due to blunt or penetrating injury. However for 10-year-olds, the probability of death is higher for a penetrating injury.

13.2.3 Inference for Regression Coefficients

The posterior mean $\mathrm{E}(\beta|Y)$ and covariance matrix

$$\mathrm{Cov}(\beta|Y) = \mathrm{E}(\beta\beta'|Y) - \mathrm{E}(\beta|Y)\mathrm{E}(\beta|Y)'$$

of the regression coefficients are approximated by $\hat{\beta} = \sum_{r=1}^{t} \beta^r \tilde{q}_r$ and

$$\widehat{\mathrm{Cov}}(\beta|Y) = \left[\sum_{r=1}^{t} \beta^r \beta^{k\prime} \tilde{q}_r\right] - \hat{\beta}\hat{\beta}',$$

respectively. The cdf of β_j $(j = 1, \ldots, k)$ and histograms for approximating the marginal posterior density can be obtained from probabilities of the form

$$\Pr(a < \beta_j \leq b|Y) = \int I_{(a,b]}(\beta_j)\pi(\beta|Y)d\beta \doteq \sum_{r=1}^{t} I_{(a,b]}(\beta_j^r)\tilde{q}_r$$

where $I_{(a,b]}(\beta_j)$ is 1 if $a < \beta_j \leq b$ and 0 otherwise.

FIGURE 13.7. Trauma Data: Predictive Probabilities

EXAMPLE 13.2.2 CONTINUED. Table 13.2 presents posterior means, standard deviations, and percentiles of β_0 and β_1 for the O-ring data. Using our prior, $\Pr(\beta_1 < 0|Y) > .99$, which suggests the slope is not zero. Figure 13.8 gives the Bayesian marginal posterior density for β_1 in the O-ring data. As before, this was actually generated by smoothing 5000 samples from the approximate posterior distribution, i.e., using Rubin's (1987) SIR algorithm.

TABLE 13.2. Posterior Marginal Distribution: O-Rings

	Full Data		Case 18 Deleted	
	β_0	β_1	β_0	β_1
$\hat{\beta}_i = \mathrm{E}(\beta_i\|Y)$	12.97	−.2018	16.92	−.2648
Std. Dev.$(\beta_i\|Y)$	5.75	.0847	7.09	.1056
5%	4.56	−.355	6.85	−.459
25%	9.04	−.251	11.98	−.324
50%	12.44	−.194	16.13	−.252
75%	16.20	−.144	20.86	−.191
95%	23.38	−.077	29.96	−.114

EXAMPLE 13.2.3 CONTINUED. Table 13.3 presents posterior means, standard deviations, and percentiles for the β_j's from the trauma data along with the maximum likelihood estimates, and asymptotic standard errors as well as posterior summaries obtained from the diffuse prior $\pi(\beta) = 1$. In addition, the informative prior gives $\Pr(\beta_1 > 0|Y) > .99$, which suggests that the coefficient of ISS is not zero. Recall that low values of RTS are bad for the patient, so the tendency of the RTS coefficients to be negative is reasonable. Central 90% posterior intervals for the β_j's are about 3/4's as wide using the informative prior as with the diffuse prior.

13.2.4 INFERENCE FOR LD_α

With the O-ring data, it is of interest to estimate the temperature at which the chance of O-ring failure is, say 50%, or some other prespecified amount α. This percentile is often called the LD_α in bioassay problems (LD for "lethal dose"), and satisfies $LD_\alpha = \{F^{-1}(\alpha) - \beta_0\}/\beta_1$. The LD_α is a function of the vector β, so its approximate posterior distribution is easily obtained. The approximate posterior takes on the value $\{F^{-1}(\alpha) - \beta_0^r\}/\beta_1^r$ with probability $\tilde{q}_r$.

Table 13.4 presents the posterior median and central 90% intervals for LD_α using five values of α for the O-ring data. In particular, the Bayesian analysis gives 69.8 degrees as the posterior median temperature at which the chance of O-ring failure is .25. The tails of the LD_α's are very heavy due to a non-negligible probability of getting β_1 values near zero.

FIGURE 13.8. O-Ring Data: Marginal Density for β_1

TABLE 13.3. Fitted Trauma Model

	Informative Posterior Summaries Based on informative prior				Maximum Likelihood	
Variable	Estimate	Std. Error	.05%	.95%	Estimate	Std. Error
Intercept	−1.79	1.10	−3.54	.02	−2.73	1.62
ISS	.07	.02	.03	.10	.08	.03
RTS	− .60	.14	− .82	−.37	− .55	.17
AGE	.05	.01	.03	.07	.05	.01
TI	1.10	1.06	− .66	2.87	1.34	1.33
AGE × TI	− .02	.03	− .06	.03	− .01	.03
	Posterior Summaries Based on diffuse prior					
Variable	Estimate	Std. Error	.05%	.95%		
Intercept	−2.81	1.60	−5.34	−.18		
ISS	.09	.03	.05	.13		
RTS	− .59	.17	− .86	−.32		
AGE	.06	.02	.03	.09		
TI	1.46	1.36	− .79	3.69		
AGE × TI	− .01	.03	− .07	.05		

TABLE 13.4. Posterior Summaries for LD_α's

	Full Data Percentiles				Case 18 Deleted Percentiles		
α	5%	50%	95%	α	5%	50%	95%
.90	30.2	52.9	60.4	.90	39.8	55.1	61.2
.75	43.4	58.5	64.0	.75	48.9	59.4	64.0
.50	55.9	64.2	68.5	.50	57.5	63.8	67.5
.25	65.1	69.8	76.4	.25	64.1	68.1	73.0
.10	70.3	75.4	88.3	.10	68.3	72.4	80.9

13.3 Diagnostics

In this section, we examine a variety of influence diagnostics based on deleting cases. We also explore Box's (1980) method of model checking. Finally, we consider the choice of an appropriate link function and an associated case deletion diagnostic.

13.3.1 Case Deletion Influence Measures

Case deletion diagnostics were pioneered by Cook (1977), Belsley, Kuh and Welsch (1980), and Pregibon (1981). Johnson and Geisser (1982, 1983, 1985) introduced Bayesian predictive and estimative case deletion diagnostics for the linear model and Johnson (1985) introduced diagnostics for the estimation of probabilities in logistic regression. Here we present the Johnson-Geisser influence measures for this nonlinear Bayesian setting. Our purpose is to detect those cases that, upon deletion from the data, noticeably affect inferences. For example, if the predictive probability of O-ring failure were to change radically upon deletion of a single case, it is incumbent upon us to report and quantify that fact. It may or may not be appropriate to delete such cases in a final analysis.

The effect of case deletion on the posterior of β is easily formulated. Recalling (13.1.1), the likelihood for β based on all the data except y_i is

$$L(\beta|Y_{(i)}) = \frac{L(\beta|Y)}{L(\beta|y_i)}$$

where $Y_{(i)}$ denotes the data Y with y_i deleted. It follows that

$$\pi(\beta|Y_{(i)}) = \frac{L(\beta|Y_{(i)})\pi(\beta)}{\int L(\beta|Y_{(i)})\pi(\beta)d\beta} = \frac{\pi(\beta|Y)/L(\beta|y_i)}{\int \pi(\beta|Y)/L(\beta|y_i)d\beta}. \tag{1}$$

If we renormalize the probability weights in our discrete approximation,

$$\tilde{q}_{r(i)} = \frac{\tilde{q}_r/L(\beta^r|y_i)}{\sum_{k=1}^{t} \tilde{q}_k/L(\beta^k|y_i)},$$

then the distribution taking values β^r with probability $\tilde{q}_{r(i)}$ gives a discrete approximation to the posterior (1). Expectations with respect to $\pi(\beta|Y_{(i)})$ are evaluated using this approximate distribution.

Estimative Influence

Kullback-Leibler (KL) divergences can be used as in Johnson and Geisser (1985) and Pettit and Smith (1985) to measure the discrepancy between full and reduced data posteriors. The KL divergence with respect to the

posterior density with the i'th case deleted is defined as

$$D_{1i}^{\beta} \equiv \int \log\left[\frac{\pi(\beta|Y_{(i)})}{\pi(\beta|Y)}\right] \pi(\beta|Y_{(i)})d\beta \geq 0.$$

A large value of D_{1i}^{β} indicates that deletion of case i results in a different posterior for β than if it were retained, possibly resulting in different inferences for β.

We now present a computational formula for D_{1i}^{β}. The predictive probability that a future binomial observation y with covariate vector x_i equals the observed y_i value, given $Y_{(i)}$, can be expressed in two equivalent ways:

$$\Pr(y = y_i|Y_{(i)}, x_i) = \int L(\beta|y_i)\pi(\beta|Y_{(i)})d\beta = \frac{L(\beta|y_i)\pi(\beta|Y_{(i)})}{\pi(\beta|Y)}. \quad (2)$$

To see this, note that from (1),

$$\frac{L(\beta|y_i)\pi(\beta|Y_{(i)})}{\pi(\beta|Y)} = \frac{1}{\int \pi(\beta|Y)/L(\beta|y_i)d\beta}.$$

Also, from (1),

$$\int L(\beta|y_i)\pi(\beta|Y_{(i)})d\beta = \frac{\int L(\beta|y_i)\pi(\beta|Y)/L(\beta|y_i)d\beta}{\int \pi(\beta|Y)/L(\beta|y_i)d\beta} = \frac{1}{\int \pi(\beta|Y)/L(\beta|y_i)d\beta}.$$

Equation (2) gives

$$\begin{aligned} D_{1i}^{\beta} &= \int \log\left[\frac{\Pr(y = y_i|Y_{(i)}, x_i)}{L(\beta|y_i)}\right] \pi(\beta|Y_{(i)})d\beta \\ &= \log\Pr(y = y_i|Y_{(i)}, x_i) - \int \log L(\beta|y_i)\pi(\beta|Y_{(i)})d\beta \\ &\doteq \log\left\{\sum_{r=1}^{t} L(\beta^r|y_i)\tilde{q}_{r(i)}\right\} - \sum_{r=1}^{t} \log L(\beta^r|y_i)\tilde{q}_{r(i)}. \end{aligned}$$

The KL divergence with respect to the posterior based on all observations is defined as

$$D_{2i}^{\beta} \equiv \int \log\left[\frac{\pi(\beta|Y)}{\pi(\beta|Y_{(i)})}\right] \pi(\beta|Y)d\beta.$$

Using equation (2),

$$D_{2i}^{\beta} \doteq \sum_{r=1}^{t} \log L(\beta^r|y_i)\tilde{q}_r - \log\left\{\sum_{r=1}^{t} L(\beta^r|y_i)\tilde{q}_{r(i)}\right\}.$$

The symmetric divergence is defined to be the sum of the divergences for the deleted and full posteriors, $D_i^{\beta} \equiv D_{1i}^{\beta} + D_{2i}^{\beta}$.

Predictive Influence

The predictive distribution for a single trial is Bernoulli, i.e., takes on the values 0 and 1. The symmetric KL divergence is used to measure the discrepancy between full and reduced data predictive distributions. The symmetric KL divergence between two Bernoulli distributions with probabilities p and q reduces to

$$J(p,q) \equiv (p-q)\log\left(\frac{p(1-q)}{(1-p)q}\right).$$

As in Johnson (1985), we define a symmetric predictive divergence diagnostic for predicting new observations at the original data locations when case i is deleted:

$$D_i^p \equiv \sum_{j=1}^{n} J\left(\Pr(y=1|Y,x_j), \Pr(y=1|Y_{(i)},x_j)\right).$$

Here, $\Pr(y=1|Y,x)$ is the predictive probability of success from all the data as defined in (13.2.2), and $\Pr(y=1|Y_{(i)},x)$ is the predictive probability of success based on all the data except case i:

$$\Pr(y=1|Y_{(i)},x) = \int F(x'\beta)\pi(\beta|Y_{(i)})d\beta \doteq \sum_{r=1}^{t} F(x'\beta^r)\tilde{q}_{r(i)}.$$

The symmetric predictive divergence diagnostic for predicting observations at an arbitrary set of locations, say x_j^f, $j=1,\ldots,r$, is

$$D_i^f \equiv \sum_{j=1}^{r} J\left(\Pr(y=1|Y,x_j^f), \Pr(y=1|Y_{(i)},x_j^f)\right).$$

A large value of D_i^p or D_i^f indicates that deletion of case i results in different predictive probabilities than if it were retained, possibly resulting in different inferences or decisions.

Example 13.3.1. *O-Ring Data.*
Figure 13.9 gives index plots of D_i^p and D_i^f for the O-ring data. The new locations used in defining D_i^f were $x_j^f = 31, 33, 35, \ldots, 51$. The plots for D_{1i}^β and D_{2i}^β were similar to the plot of D_i^p, so they are not included. Case 18, which corresponds to the flight where O-rings failed at the highest launch temperature, consistently stands out. Note that the values of D_i^f are larger for cases with low temperatures. This occurs because the predictions being made are also at low temperatures. The estimative measures and the predictive measure D_i^p are qualitatively similar for these data, although they need not be in general.

FIGURE 13.9. O-Ring Data: Index Plots of D_i^p and D_i^f

The influence of case 18 was evaluated by repeating the analysis with case 18 deleted. Summary statistics for the Bayesian analysis are provided in Tables 13.2 and 13.4. Omitting case 18, the probability of O-ring failure increased at low temperatures and decreased at high temperatures. The difference in the predictive probabilities for the analyses with and without case 18 are not dramatic. The actual influence of case 18 on the posterior summaries is minor.

EXAMPLE 13.3.2. *Trauma Data.*
Computing the D_i^p's we found cases 52 and 232 to be most influential. Case 52 is a 66-year-old person who had very little wrong with him ($ISS = 9$, $RTS = 7.84$) with a penetration injury who died. Case 232 is a very sick and damaged ($RTS = 2.19$, $ISS = 50$) 50-year-old person with a blunt injury who managed to survive. An ISS score of 50 is characteristic of a person who has very severe injuries to two different parts of the body. The actual statistics are $D_{52}^p = .46$ and $D_{232}^p = .41$, with the next highest value being $D_{173}^p = .25$. Figure 13.10 contains an index plot of the difference in the predictive probabilities of death, $p(y = 1|Y, x_j) - p(y = 1|Y_{(52)}, x_j)$. These probabilities depend on the specified prior. Note that having deleted a case in which a relatively healthy person died, most of the probability differences are very near 0 but slightly positive; e.g., most peoples probabilities of death have decreased. Moreover, all of the changes in probabilities are relatively small. Deletion of case 232 seems to change the regression coefficients even less and would seem to have even less effect on the fitted probabilities.

FIGURE 13.10. Trauma Data: Index Plot of $p(y = 1|Y, x_j) - p(y = 1|Y_{(52)}, x_j)$

13.3.2 Model Checking

We consider two methods for model checking. The first is a global model check due to Box (1980). This involves finding the probability that a new vector Y_* has a marginal probability smaller than that of the vector Y that we actually observed, i.e.,

$$\Pr[p(Y_*) \leq p(Y)],$$

where

$$p(Y) = \int L(\beta|Y)\pi(\beta)d\beta.$$

This is essentially a *P value*, so small values are of significance. For the O-ring data, this value is approximated as .58. The probability is large, so there is no indication of a substantial problem with the model. If the improper diffuse prior $\pi(\beta) = 1$ is used, the required marginal distribution of the data may not exist.

Another model check considers the criterion for one element of the Y vector at a time, i.e.,

$$\Pr[p(y_{i*}) \leq p(y_i)].$$

This can be viewed as a Bayesian outlier check because we are assessing whether each observation is unusual relative to the model. For the O-ring data, all of these values are 1 except the two identical cases 13 and 14 that give .43 and case 18 that gives .37. This diagnostic gives no indication of substantial problems with the model.

The model checking computations were performed by sampling from the prior distribution. We sample pairs $(\tilde{p}_1, \tilde{p}_2)$ and solve the equations $F^{-1}(\tilde{p}_i) = \beta_0 + \beta_1 \tilde{x}_i$, $i = 1, 2$, to obtain samples of β_0 and β_1. Sampling the pairs $(\tilde{p}_1, \tilde{p}_2)$ is easy with our prior because the $\tilde{p}_i$'s have independent beta distributions. Given a sample $\beta^r_\#$, $r = 1, \ldots, v$, from the prior,

$$p(Y) \doteq \frac{1}{v} \sum_{r=1}^{v} L(\beta^r_\#|Y).$$

For an individual component,

$$p(y_i) \doteq \frac{1}{v} \sum_{r=1}^{v} L(\beta^r_\#|y_i).$$

Computing $\Pr[p(Y_*) \leq p(Y)]$ for a new vector Y_* requires an additional round of sampling. For each $\beta^r_\#$, $r = 1, \ldots, v$, generate new independent random variables y_{ir*}, $i = 1, \ldots, n$, that are $\text{Bin}(N_i, F(x_i'\beta^r_\#))$, respectively. The y_{ir*}'s form vectors Y_{r*} for which we can compute $p(Y_{r*})$ as above. $\Pr[p(Y_*) \leq p(Y)]$ is approximated by the proportion of $p(Y_{r*})$'s that are no greater than $p(Y)$. Computation of $\Pr[p(y_{i*}) \leq p(y_i)]$ is similar.

Rubin (1988) advocated Bayesian model checks using predictive rather than marginal distributions. On the O-ring data, Rubin's analogues of the global and local model checks lead to identical conclusions. Chaloner and Brant (1988) check for outliers using the posterior of β. Similar methods also apply to the trauma data.

13.3.3 Link Selection

We now allow the Bayesian paradigm to indicate which of the three link function models is most appropriate for the data: logistic (M_1), probit (M_2), or complementary log-log (M_3). Bayes factors for comparing models M_j and M_k are numbers BF_{jk} such that

$$\frac{P(M_j|Y)}{P(M_k|Y)} = \Big[BF_{jk}\Big]\frac{P(M_j)}{P(M_k)}.$$

The Bayes factor is the multiplier that changes the prior odds for the models into the posterior odds. It is a simple application of Bayes's theorem to show that

$$BF_{jk} = \frac{p(Y|M_j)}{p(Y|M_k)}.$$

$p(Y|M_1)$ was computed previously as $p(Y)$; it is the marginal probability of obtaining Y from the logistic model. Computing $p(Y|M)$ for an alternative model M involves integrating the corresponding likelihood function with respect to the induced prior on β for that model. As in the logistic case, $p(Y|M)$ is estimated using samples generated from the prior on the $\tilde{p}_j$'s.

Example 13.3.1 continued. *O-Ring Data.*
For the O-ring data, the Bayes factors under our prior are $BF_{21} = 1.086$, $BF_{31} = 1.403$, and, thus, $BF_{32} = BF_{31}/BF_{21} = 1.403/1.086 = 1.292$. None of these values is large enough to suggest a serious preference for one of the three models. In particular, if the prior odds for the probit versus logit models are 1, the posterior odds are merely 1.086.

Example 13.3.2 continued. *Trauma Data.*
For the trauma data, the Bayes factors under our prior are $BF_{21} = 1.05$, $BF_{13} = 20.72$, and, thus,

$$BF_{23} = BF_{21}/BF_{31} = BF_{21}BF_{13} = 1.05(20.72) = 21.83\,.$$

There is a suggestion against the complementary log-log model, but there is little to choose from between the logistic and probit models. If the prior odds for the probit versus logit models are 1, the posterior odds are merely 1.05. (These numbers were based on an importance sample of 10,000 observations. Based on only 2000 observations, we got $BF_{21} = 1.97$, BF_{13} was about the same. Similarly, the values D^p_{52}, D^p_{232}, and D^p_{173} were .525,

FIGURE 13.11. O-Ring Data: Bayes Factors with Case Deletion

.519, and .234 based on 2000 samples rather than those reported earlier. The numbers changed, but the message did not!)

We also formed a link selection case deletion diagnostic by computing

$$BF_{jk(i)} = \frac{p(Y_{(i)}|M_j)}{p(Y_{(i)}|M_k)},$$

where

$$p(Y_{(i)}|M) = \int L(\beta|Y_{(i)}, M)\pi(\beta|M)\, d\beta.$$

Figure 13.11 contains a simultaneous plot of $BF_{21(i)}$ versus i and $BF_{31(i)}$ versus i with $i = 1, \ldots, 23$ for the O-ring data. The full data Bayes factors are given by the intercept at "case index 0." Case 18 has the largest effect on the Bayes factors. The deletion of case 18 decreases the posterior odds for the complementary log-log relative to the logistic link, and increases the posterior odds for the probit over the logistic. The actual effect of this case on the Bayes factors is small, and so our decision to use the logistic link is not altered by case deletion.

13.3.4 Sensitivity Analysis

The sensitivity of posterior inferences to the choice of the prior can be evaluated by recalculating posterior summaries based on alternative priors. In situations where the prior changes radically, Monte Carlo samples

from the new posteriors might be needed. When changes in the prior are not dramatic, renormalization of the original Monte Carlo weights might be sufficient. For example, the posterior based on a prior $\pi^*(\beta)$ is approximated by the discrete distribution taking values β^r with probabilities $\tilde{q}_r^*$, where

$$\tilde{q}_r^* = \frac{\pi^*(\beta^r)\tilde{q}_r/\pi(\beta^r)}{\sum_{k=1}^t \pi^*(\beta^k)\tilde{q}_k/\pi(\beta^k)}.$$

EXAMPLE 13.3.1 CONTINUED. *O-Ring Data.*
We used two additional priors to evaluate the sensitivity of our analysis. Each of the priors is a product of independent beta distributions placed at $\tilde{\tau}_1 = 55$ and $\tilde{\tau}_2 = 75$ degrees. Prior II [$\tilde{p}_1 \sim$ Beta(.9,.1) and $\tilde{p}_2 \sim$ Beta(.1,.9)] places a prior (mean) probability of .9 for O-ring failure at 55 degrees and prior probability of .1 for O-ring failure at 75 degrees, while making the beliefs equivalent to one prior observation. Prior III placed Jeffrey's "noninformative" Beta(.5,.5) priors on the $\tilde{p}$'s. The posteriors using the original prior and Prior III were similar, whereas the posterior using Prior II was similar to the posterior obtained from the original prior after omitting case 18. Given the small effect of case 18 on our original analysis, we felt that our posterior analysis was not overly sensitive to these changes in the prior.

EXAMPLE 13.3.2 CONTINUED. *Trauma Data.*
To examine sensitivity to the prior specifications, we considered case deletions of the "prior observations." In Figure 13.12 we present plots of $p(y = 1|Y, x_j, \tilde{Y}) - p(y = 1|Y, x_j, \tilde{Y}_{(i)})$, where each is a predictive probability of success but based on different prior information. Here, the data are the same and the priors involve case deletion. In Figure 13.10, the data involve case deletion but the priors are the same. Note that $\tilde{Y}_{(i)}$ represents partial prior information in the sense of BCJ.

13.4 Posterior Computations and Sample Size Calculation

In recent years, Bayesian analysis has been performed by using numerical integrations (Naylor and Smith, 1982; Smith et al., 1985), by using the analytic Laplace approximation (Leonard,1982; Tierney and Kadane, 1986; Kass et al., 1988), and by using Monte Carlo methods (Zellner and Rossi, 1984; Gelfand and Smith, 1990; Dellaportas and Smith, 1993). See Gelman et al. (1995, Chaps. 9-11) for a nice summary of these methods. We prefer Monte Carlo methods to Laplace approximations in regression problems because when performing many predictions, only a single Monte Carlo sample is necessary to perform all predictions, while the Laplace method

FIGURE 13.12. Trauma Data: $p(y = 1|Y, x_j, \tilde{Y}) - p(y = 1|Y, x_j, \tilde{Y}_{(i)})$

requires a separate analytic approximation for each prediction. We prefer Monte Carlo methods to numerical integration because of their potential to deal with high-dimensional problems. Monte Carlo methods provide a discrete approximation to the posterior distribution. We discuss a variant of importance sampling that is especially simple when used with a DAP.

In importance sampling, one chooses a density function $g(\beta)$ that is similar in shape to the known kernel of the posterior $L(\beta|Y)\pi(\beta)$ with tails that do not decay more rapidly than the tails of the posterior. Then sample $\beta^1, \ldots, \beta^t$ from the distribution with density $g(\beta)$. For $r = 1, \ldots, t$, compute the weights

$$q_r = q(\beta^r) = \frac{L(\beta^r|Y)\pi(\beta^r)}{g(\beta^r)} \tag{1}$$

and

$$\tilde{q}_r = q_r \Big/ \sum_{k=1}^{t} q_k.$$

The discrete approximation to the posterior distribution takes values β^r with probability $\tilde{q}_r$.

Under fairly weak assumptions (cf. Geweke, 1989), the approximation in (13.1.2) has a large sample normal distribution with estimated variance

$$\hat{\sigma}_h^2 = \sum_{r=1}^{t} \{h(\beta^r) - \bar{\theta}_h\}^2 \tilde{q}_r \tag{2}$$

where $\bar{\theta}_h$ is the approximation from (13.1.2). The variance of $\bar{\theta}_h$ depends critically on the tails of $g(\beta)$ through the weight function $q(\beta)$ of (1). Geweke (1989) concluded that to attain high efficiency across a variety of functions, $q(\beta)$ should be reasonably constant with small tails. If the tails of the importance function were allowed to decrease much more rapidly than the tails of the posterior density, the normalized weights $\tilde{q}_r$ could be dominated by individual importance samples in the tail of the approximate posterior. This needlessly inflates the variance of $\bar{\theta}_h$. Similar difficulties can arise with any renormalization of the weights for dealing with case deletions or different priors.

A natural choice for the importance density $g(\beta)$ is a multivariate Student's t density with v degrees of freedom, with location equal to the posterior mode β_M, and dispersion proportional to $\Sigma(\beta_M)$, the asymptotic posterior covariance matrix evaluated at the mode. The approximate $N(\beta_M, \Sigma(\beta_M))$ posterior density is an alternative possibility, but the thin tails of the normal often cause problems; see Zellner and Rossi (1984). Johnson (1987) gives simple algorithms for generating the multivariate normal and $t(v)$ distributions.

Prior to selecting the importance sampling density $g(\beta)$, plot the kernel of $\pi(\beta|Y)$ along the asymptotic principal component directions and choose

the degrees of freedom v so that the tails of $g(\beta)$ are at least as heavy as those of $\pi(\beta|Y)$. Specifically, with $\Sigma(\beta_M) = TT'$, where the columns of T are orthogonal, plot $g(\beta_M + \delta T e_i)$ as a function of δ in each of the unit directions e_i, $i = 1, ..., k$, and similarly for the kernel of the posterior. (e_i is a vector of 0s except for a 1 in the i'th place.) In cases of extreme asymmetry along these directions, we recommend sampling from split-t distributions. The split-t distributions allow for different tail heights in each direction, in addition to asymmetry about the mode; see Geweke (1989) for details.

Figure 13.13 gives a plot of the posterior kernel and the normal, $t(6)$, and split-$t(6)$ densities in the direction of the first principal component for the O-ring data. Each function was normalized to have a maximum value of 1. The plot of the posterior reflects the skewness seen in Figure 13.2. The normal density is inferior to the $t(6)$ density as an importance function because the normal underestimates the posterior upper tail in this direction. The weights $q(\beta)$ in this direction at 3, 4, and 4.5 standard deviations above zero are 5, 40, and 150 times greater than the weight at zero for the normal density. For the $t(6)$ density, this ratio is below 3. The corresponding plot along the second principal component was similar, with the exception that the posterior is skewed to the left. The split-$t(6)$ density has heavier tails than the posterior in each direction and reproduces the shape in the center of the posterior. We concluded that the split-$t(6)$ is best among the three importance functions, with the $t(6)$ a close second. Heavier tails on the t distribution could have been obtained by reducing the degrees of freedom, but this was unnecessary. The posterior summaries based on both $t(6)$ and split-$t(6)$ sampling were obtained; they were similar.

For the O-ring data, we decided on the importance sample size by first generating a pilot study of 500 samples. Prediction was a primary goal. We decided that the estimates for the probability of O-ring failure $F(x'\beta)$ and success $1 - F(x'\beta)$ at the 23 observed lift-off temperatures must be accurate. The maximum coefficient of variation across estimates under our prior was 4.4%. To reduce this to a target value 2%, the sample size needed to be increased by a factor of $(2.2)^2 = 4.84$, to approximately 2500. We decided to sample 5000 observations. The estimated maximum coefficient of variation for the parameters of interest based on 5000 samples was 1.4%. Similar methods were used for the trauma data, with a pilot study of 2500 samples and a total sample of 10,000 from a split-$t(6)$.

We noted earlier that β_M and $\Sigma(\beta_M)$ are easily computed using standard software when the prior is a DAP. An interesting special case is the improper prior $\pi(\beta) = 1$, where β_M is the MLE $\hat{\beta}_{ml}$ based on the original data and $\Sigma(\beta_M)^{-1}$ is the observed Fisher information evaluated at $\hat{\beta}_{ml}$. For non-DAP priors, the posterior mode β_M must be computed using specialized software for numerical maximization. Typically, β_M is the solution to $S(\beta) = 0$, where $S(\beta)$ is the vector of partial derivatives of the log of the posterior kernel, i.e., $\log\{L(\beta|Y)\pi(\beta)\}$. The inverse of minus one times the matrix of second partial derivatives of the log kernel evaluated at β_M

FIGURE 13.13. O-Ring Data: Importance Function Diagnostic Plots

serves as $\Sigma(\beta_M)$. Alternatively, the maximum likelihood estimate and the inverse of either the observed or expected Fisher information can be used in place of β_M and $\Sigma(\beta_M)$, cf. Berger (1985, p. 224).

Except for this added computational component, there is no intrinsic problem with using importance sampling with arbitrary priors. Importance sampling can be inefficient when the shape of the posterior density is hard to match. This difficulty might arise when the posterior is highly non-concave or restricted to a subset of the natural parameter space. However, this problem usually does not occur when normal, diffuse, and DAP priors are used with logistic or probit models because the posterior densities are concave. Dellaportas and Smith's (1993) rejection method is also attractive in these cases because the posterior mode is not needed to sample the posterior. Unfortunately, no single method works well, regardless of the prior and link function.

In related work, Smith and Gelfand (1992) presented an introduction to the use of importance sampling and the rejection method. They use the prior distribution as an importance function, while we use the posterior mode β_M and asymptotic posterior dispersion matrix $\Sigma(\beta_M)$ to determine an importance function. Casella and George (1992) explained the Gibbs sampler. When applicable, this provides a random sample of size t from the posterior, so $\tilde{q}_r = 1/t$ for all r. Dellaportas and Smith (1993) combine the Gibbs sampler with the rejection method to obtain samples from the posterior in generalized linear models.

The recent books by Carlin and Louis (1996), Gelman et al. (1995), and

Gilks, Richardson and Spiegelhalter (1996) examine a variety of complex modeling problems that can be handled easily using Bayesian methods. Standard references for Bayesian prediction are Aitchison and Dunsmore (1975) and Geisser (1993).