Ann. Inst. Statist. Math.
Vol. 44, No. 3, 417–434 (1992)

POSTERIOR MODE ESTIMATION FOR THE GENERALIZED LINEAR MODEL

D. M. EAVES[1]* AND T. CHANG[2]**

[1]*Department of Mathematics and Statistics, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6*
[2]*Department of Mathematics, University of Virginia, Charlottesville, VA 22903, U.S.A.*

(Received August 21, 1989; revised June 3, 1991)

Abstract. Posterior mode estimators are proposed, which arise from simply expressed prior opinion about expected outcomes, roughly as follows: a conjugate family of prior distributions is determined by a given variance function. Using a conjugate prior, a posterior mode estimator and its estimated (co-)variances are obtained through conventional maximum likelihood computations, by means of small alterations to the observed outcomes and/or to the modelled variance function. Within the conjugate family, for purposes of inference about the regression vector, a reference prior is proposed for a given choice of linear design of the canonical link. The resulting approximate reference inferences approximate the Bayesian inferences which arise from a "minimally informative" reference prior. A set of subjective prior upper and lower percentage points for the expected outcomes can be used to determine a conjugate family member. Alternatively, a set of subjective prior means and standard deviations determines a member. The subfamily of priors determinable by percentage points either includes or approximates the proposed reference prior.

Key words and phrases: Conjugate prior, contingency tables, exponential family, frequency counts, generalized linear model, Jeffreys prior, logistic regression, multinomial outcome, minimally informative prior, nonlinear regression, quasi-likelihood, reference prior, regression, variance function.

1. Introduction

This paper discusses the Bayesian analysis of the generalized linear model with special emphasis on (1) finding an analog to Jeffreys' prior, (2) techniques

* The research of the first-named author was funded in part by a Natural Sciences and Engineering Research Council grant.

** The second named author gratefully acknowledges the support of the National Science Foundation, grant #DMS-8901494 and of the Kansas Geological Survey where he visited during the term of the majority of this research.

for choosing subjective priors, and (3) an approximation to the Bayesian approach which can be easily computed.

Consider the simple experiment of observing y successes out of a small number k of independent Bernoulli trials for estimating an unknown success rate π or its logit $\theta = \log[\pi/(1-\pi)]$. When it can be calculated, the MLE is $\hat{\theta} = \log[y/(k-y)] = \log[\hat{\pi}/(1-\hat{\pi})]$. The estimated variance of $\hat{\theta}$ obtained through a standard information calculation is $\widehat{\text{var}}(\hat{\theta}) = 1/[k\hat{\pi}(1-\hat{\pi})]$. Given $\omega > 0$, if we think of $y + \omega/2$ as an altered number of successes from $k + \omega$ trials, the altered MLE is $\tilde{\theta} = \log[(y+\omega/2)/(k-y+\omega/2)]$, with similarly estimated variance $\widetilde{\text{var}}(\tilde{\theta}) = 1/[(k+\omega)\tilde{\pi}(1-\tilde{\pi})]$, where $\tilde{\pi} = 1/[1+\exp(-\tilde{\theta})] = (y+\omega/2)/(k+\omega)$. $\hat{\theta}$ is of course the case $\omega = 0$.

The function $MSE(\theta) = E_\theta[\tilde{\theta} - \theta]^2$ is identically infinite for the case $\omega = 0$, and plotting $MSE(\theta)$ vs θ for each of several values of $\omega > 0$ shows dramatic improvement as ω increases. From a sampling theory standpoint this militates against the use of very small ω, and against the MLE $\hat{\theta}$ in particular. $\tilde{\theta}$ with $\omega = 1$ is the Cox (1970) empirical logistic transform referred to in Dobson ((1983), Example 8.3), who reports that $\omega = 1$ minimizes the bias of $\tilde{\theta}$ and uniquely produces bias $= o(1/k^2)$ whereas other values produce $o(1/k)$.

$\tilde{\theta}$ is also the Posterior Modal Estimator (PME) from a $d\theta$-density proportional to $e^{\theta\omega/2}/(1+e^{\theta})^{\omega}$, which, in terms of π, is the symmetric Dirichlet $(\omega/2, \omega/2)$ distribution. As such, using the $\tilde{\theta}$ from a very small ω represents, in a sense, prior near-certainty that π is extremely close to 0 or 1, an unlikely attribute of almost any scientist conducting such an experiment. With $\omega = 1$, $\tilde{\theta}$ is the PME from Jeffreys' prior $d\theta$-density. If prior information is not strong, $\omega = 1$ (which attributes prior probability 0.025 to each of the situations $\pi < 0.006$ and $\pi > 0.994$) is proposed as a general-purpose "reference" value. The intent of the term "reference" is merely that $\omega = 1$ may conveniently be used as a value to whose consequent inferences other inferences may be referred for a standard comparison.

We further propose that for convenience inferences be based on "standard" maximum quasi-likelihood estimates (MQLE) and information calculations applied to the posterior $d\theta$-density, rather than based on the exact posterior distribution. This procedure, when applied to the simple binomial case, produces the asymmetric approximate 68% confidence or credibility interval,

$$\frac{1}{1+\exp(-\{\tilde{\theta} - [\widetilde{\text{var}}(\tilde{\theta})^{1/2}\})} < \pi < \frac{1}{1+\exp(-\{\tilde{\theta} + [\widetilde{\text{var}}(\tilde{\theta})]^{1/2}\})}.$$

In what follows we shall extend the proposal of a reference PME $\tilde{\theta}$ of the canonical parameter θ and of $\widetilde{\text{var}}(\tilde{\theta})$ to apply to a generalized linear model with any variance function $v(\mu)$ that is either multinomial, or univariate *quadratic*, i.e., $v(\mu) = r + s\mu + t\mu^2$, and to an independent set of samples with canonical links $\theta_i = \nu_i + x_i'\beta$. β will be the parameter of interest. The *canonical parameter* or link function $\theta = \Theta(\mu)$ is defined through $d\Theta/d\mu = 1/v(\mu)$. For example if $v(\mu)$ is the binomial(k) variance function $\mu(k-\mu)/k$, $0 < \mu < k$, then $\theta = \log[\mu/(k-\mu)]$. (In the case of a multivariate observation, $v(\mu)$ is a matrix and the *weight function* $1/v(\mu)$ becomes the inverse or a pseudoinverse $v(\mu)^-$.)

In Section 2 we propose a conjugate family of distributions based on the quasi-likelihood function for a generalized linear model with a quadratic variance function. For the given variance function, a Jeffreys-like prior is also proposed which is either a member of the conjugate family or else is approximated by the family. Section 3 discusses how informative conjugate priors might be specified in practice. The full versatility of generalized linear modelling is realized only when we consider sub-full linear designs for the canonical parameter vector $\theta = [\theta_1, \theta_2, \ldots]'$. The methods of Section 2 are extended to this situation in Section 4. In Section 5 several examples are given including $r \times c$ contingency tables.

2. Quasi-likelihood, conjugate families and Jeffreys priors

2.1 *Quasi-likelihood*

The quasi-likelihood concept may be thought of as an extension of the notion of the log-likelihood function of a distribution of exponential type. It may also be described as a framework for fitting to data models which specify only first and second moments, in which the variance of each observation y_i is a known function $v(\mu_i)$ (or perhaps $= \phi v(\mu_i)$ with $\phi > 0$ unknown) of expectation μ_i. The quasi-likelihood function $L(\theta, y)$ determined by a given variance function $v(\mu)$ is based on $dL(\Theta(\mu) \mid y)/d\mu = v(\mu)^{-}(y-\mu)$. The only multivariate model we shall consider here is that of a multinomially distributed observation, e.g., a trinomial(k) observation $[y_1, y_2]'$ where $y_h =$ observed frequency in category $\#h$. Detailed accounts of quasi-likelihood modelling and of properties of maximum quasi-likelihood estimates may be found in Wedderburn (1974), McCullagh (1983) and McCullagh and Nelder (1983).

In terms of θ, L looks like

$$\text{(2.1a)} \qquad L(\theta \mid y) = y\theta - b(\theta) \qquad \text{for a model with } \operatorname{var}(y) = v(\mu),$$

or else

$$\text{(2.1b)} \qquad L(\theta, \phi \mid y) = [y\theta - b(\theta)]/\phi \qquad \text{for } \operatorname{var}(y) = \phi v(\mu),$$

with the unknown *dispersion parameter* ϕ. Here $b(\theta)$ is defined through $db(\Theta(\mu))/d\mu = v(\mu)^{-1}\mu$. (In the logistic trinomial case $y = [y_1, y_2]'$ in (2.1a), and $y\theta$ becomes $y'\theta$.) (2.1a) implies $\operatorname{var}_\theta(y) = d^2b(\theta)/d\theta^2$ while (2.1b) implies $\operatorname{var}_\theta(y) = \phi d^2b(\theta)/d\theta^2$. Table 1 lists the common quadratic variance functions and related entities, may be verified from the foregoing definitions, and appears to a large extent in Tables 2.1 and 8.1 of McCullagh and Nelder (1983).

2.2 *Conjugate families, and reference priors for one observation*

There is a vast literature concerning minimally informative reference priors for a given subparameter of interest. Recent expositions in Box and Tiao (1973), Bernardo (1979), Berger (1985) and Chang and Eaves (1990) are but a few references. When the quasi-likelihood (2.1a) is a log-likelihood Jeffreys' prior is a reference prior in the established sense. Thus for general quasi-likelihoods of the

Table 1. Common quadratic variance functions v, their canonical parameters θ, and $b(\theta)$.

Name	$v(\mu)$	$\theta = \Theta(\mu)$	$b(\theta)$
normal	1	μ	$\theta^2/2$
Poisson	μ	$\log(\mu)$	$\exp(\theta)$
binomial(k)	$\mu - \mu^2/k$	$\log[\mu/(k-\mu)]$	$k\log(1+e^\theta)$
gamma	μ^2	$-1/\mu$	$-\log(-\theta)$
neg. binom.(k)	$\mu + \mu^2/k$	$\log[\mu/(k+\mu)]$	$-k\log(1-e^\theta)$
trinomial(k)	$\mathrm{diag}(\mu) - \frac{1}{k}\mu\mu'$	$[\log(\mu_1/\mu_3), \log(\mu_2/\mu_3)]'$	$k\log[1+\exp(\theta_1)+\exp(\theta_2)]$

Table 2(a). v-conjugate distributions in terms of μ, expressed as functions proportional to $d\mu$-densities.

Name of model	$v(\mu)$	fnc.prop. to $d\mu$-density	μ-distribution family
normal	1	$\exp[-(\lambda/2)(\mu - \kappa/\lambda)^2]$	normal
Poisson	μ	$\mu^{\kappa-1}\exp(-\lambda\mu)$	gamma
binomial(k)	$\mu - \mu^2/k$	$\mu^{\kappa-1}(k-\mu)^{k\lambda-\kappa-1}$	Dirichlet
gamma	μ^2	$\mu^{-\lambda-2}\exp(-\kappa/\mu)$	inverse gamma
neg. binom.(k)	$\mu + \mu^2/k$	$\mu^\kappa(\mu+k)^{k\lambda-\kappa-1}$	—
quadratic	$r + s\mu + t\mu^2$	—	—
trinomial(k)	$\mathrm{diag}(\mu) - \frac{1}{k}\mu\mu'$	$\mu_1^{\kappa 1-1}\mu_2^{\kappa 2-1}$ $\cdot(k-\mu_1-\mu_2)^{\lambda k-\kappa 1-\kappa_2-1}$	Dirichlet

Table 2(b). Quadratic variance functions V and their conjugate reference parameters.

Name	$v(\mu)$	κ	λ
normal	1	0	0
Poisson	μ	1/2	0
binomial(k)	$\mu - \mu^2/k$	1/2	$1/k$
gamma	μ^2	0	-1
neg. binom.(k)	$\mu + \mu^2/k$	1/2	$-1/k$
quadratic	$r + s\mu + t\mu^2$	$s/2$	$-t$
trinomial(k)	$\mathrm{diag}(\mu) - \frac{1}{k}\mu\mu'$	$\begin{bmatrix}1/2\\1/2\end{bmatrix}$	$3/2k$

form (2.1a) we propose a Jeffreys-type prior calculated from the quasi-likelihood, as a reference prior. For (2.1b) we propose multiplying this prior by $d\phi/\phi$.

A $d\theta$-density that, for some κ and λ, is proportional to $\exp[\kappa\theta - \lambda b(\theta)]$, will be called a *$v$-conjugate $d\theta$-density*. Table 2(a) lists the corresponding distributions of μ under the change of variable $\theta \to \mu$. Note that $v(\mu) = r + s\mu + t\mu^2$ does not generally lead to closed expressions for $b(\theta)$ or the corresponding densities.

We verify below the following proposition:

PROPOSITION 2.1. *Jeffreys' $d\theta$-density is v-conjugate if and only if it arises*

from a quadratic (or multinomial) variance function $v(\mu)$. We shall term the κ and λ thus arising from a given $v(\mu)$ as the reference parameters for $v(\mu)$. The reference parameters of quadratic variance functions are listed in Table 2(b). *Since Jeffreys' prior is proportional to root-quasi-information $b^{(2)}(\theta)^{1/2}$, v-conjugacy means that for some κ and λ, $(1/2)\log[b^{(2)}(\theta)] = \text{const} + \kappa\theta - \lambda b(\theta)$. Substituting $\Theta(\mu)$ for θ and taking $d/d\mu$ immediately gives $v^{(1)}(\mu) = 2(\kappa - \lambda\mu)$, so that $v(\mu) = r + 2\kappa\mu - \lambda\mu^2$ is quadratic. Conversely if, for some s and t, $v^{(1)}(\mu) = s + 2t\mu$, then $(d/d\mu)\log(v(\mu)) = (s + 2t\mu)/v(\mu) = (d/d\mu)[s\Theta(\mu) + 2tb(\Theta(\mu))]$ shows that $\log[b^{(2)}(\theta)] = \text{const} + s\theta + 2tb(\theta)$, so that Jeffreys' prior is v-conjugate.*

Notice that when a v-conjugate prior $d\theta$-density is used, the posterior will also be a v-conjugate $d\theta$-density. The effect upon the likelihood of the κ in the conjugate prior is to replace y by $y + \kappa$, whereas the effect of λ is to alter the modelled variance $b^{(2)}(\theta)$ by the multiplicative factor $(1 + \lambda)$. Thus the use of v-conjugate priors is especially convenient and can be implemented with existing software.

Example. Consider a random sample $y_1, \ldots, y_n$ from a gamma family with known scale factor 1 and shape parameter α. Since $\mu = \alpha = v(\mu)$, this is a Poisson-type quasi-likelihood! The PME for θ satisfys $b'(\theta) = (\bar{y} + \kappa/n)/(1 + \lambda)$, so that $\tilde{\theta} = \log(\bar{y} + 1/2n)$. The estimated variance of $\tilde{\theta}$ is $\widetilde{\text{var}}(\tilde{\theta}) = [n(1 + \lambda)b''(\tilde{\theta})]^{-1} = 1/[n\exp(\tilde{\theta})]$. Therefore the proposed 68% confidence interval for $\theta = \log(\mu)$ is $\log(\bar{y} + 1/2n) \pm [n\bar{y} + 1/2]^{-1/2}$, while for μ its endpoints are $[\bar{y} + 1/2n]\exp\{\pm[n\bar{y} + 1/2]^{-1/2}\}$. Notice that in this example the exact Jeffreys prior density with respect to $d\mu$ is $[-d^2\log(\Gamma(\mu))/d\mu^2]^{-1/2}$, a considerably less tractable expression.

3. Subjective prior distributions

Other values of κ and λ than the reference values may be chosen, and used with equal convenience. Here we discuss the choosing of κs and λs as a means of expressing subjective prior opinion about the mean outcomes μ. Two schemes are proposed, one based on a prior mean $m = E(\mu)$ and standard deviation $s = SD(\mu)$. The second and more widely applicable scheme is based on prior upper and lower percentage points for μ.

3.1 *Using a subjective prior mean and standard deviation of expected outcome*

For each of the variance functions named in Table 1, the θ-distribution proportional to $\exp[\kappa\theta - \lambda b(\theta)]d\theta$ corresponds to the distribution of μ indicated in Table 3 (see also Table 2(b)). Calculation of the corresponding prior expectation $E(\mu)$ and standard deviation $s = SD(\mu)$ is therefore straightforward. $s = SD(\mu)$ is shown in Table 3 and in all cases $E(\mu) = \kappa/\lambda$. The table may be used as follows: let m be a prior guess, however wild, at the value of μ, and let s be a rough guess at how far wrong m might be. (This approach to choosing κ and λ is appropriate only if the scientist is indifferent as to the algebraic sign of $\mu - m$). m and s may be used as prior mean and standard deviation of μ to determine adjustments to the observed outcomes and to the variance function, by solving $m = E(\mu)$ and

Table 3. Prior standard deviations s in terms of κ, λ and κ, λ in terms of m, s.

Name of var.fnc.	distribution of μ	$s = SD(\mu)$	κ	λ
normal	normal	$(m/\kappa)^{1/2}$	m/s^2	$1/s^2$
Poisson	gamma $(\kappa, 1/\lambda)$	$m/\kappa^{1/2}$	m^2/s^2	m/s^2
binomial(k)	μ/k is beta $(\kappa, \lambda k - \kappa)$	$[m(k-m)/(\lambda k+1)]^{1/2}$	λm	$[m(k-m) - s^2]/ks^2$
gamma	$1/\mu$ is gamma $(\lambda + 1, 1/\kappa)$	$m/(\lambda - 1)^{1/2}$	λm	$(m^2/s^2) + 1$
neg. binomial(k)	$\frac{\lambda k + 1}{k\kappa}\mu$ is $F[2\kappa, 2(\lambda k + 1)]$	$[m(k+m)/(\lambda k-1)]^{1/2}$	λm	$[m(k+m) + s^2]/ks^2$
trinomial(k)	μ/k is Dirichlet $(\kappa_1, \kappa_2, \lambda k - \kappa_1 - \kappa_2)$	$[m_h(k-m_h)/(\lambda k+1)]^{1/2}$	λm	$[m_1(k-m_1) - s_1^2]/ks_1^2$

$s = SD(\mu) = \text{var}(\mu)^{1/2}$ for κ and λ, for each row of W. The solutions are shown in columns κ and λ.

For the trinomial(k) parameter $\mu = [\mu_1, \mu_2, \mu_3]' = k\pi$, choices of $\text{var}(\mu_h)$ must satisfy $\text{var}(\pi_h) = \text{var}(\pi_1)E(\pi_h)[1 - E(\pi_h)]/[E(\pi_1)(1 - E(\pi_1))]$; thus it is that all three $[m_h(k - m_h) - s_h^2]/ks_h^2$ coincide. For the variance functions shown in Table 3, only the t-nomial(k) has reference values κ and λ which correspond to finite prior means and variances: for the t-nomial(k), $\kappa = [1/2, \ldots, 1/2]'$, $\lambda = t/(2k)$, $E(\mu_h) = k/t$ and $\text{var}(\mu_h) = (k^2/t^2)[(2t-2)/(t+2)]$. Berger (1985) and others have commented on the difficulty of obtaining reliable prior assessments of moments. Indeed, one may wish to consider a prior distribution whose mean and standard deviation are vaguely large. For this reason we now describe a more broadly applicable approach based on prior percentage points.

3.2 *Using subjective prior upper and lower percentage points of expected outcomes*

Assume the scientist can produce a subjective prior upper and lower $100\alpha\%$ percentage point $\boldsymbol{u}$ and $\boldsymbol{l}$, for the expected outcome μ. Sensitivity of inferences to these two *bounding parameters* should generally be explored as part of data analysis. We will now see how these two bounding parameters, together with the choice of the model variance function $v(\mu)$, will determine the κ and λ of a conjugate prior distribution of μ. We also note the manner in which the reference values of κ and λ can be approximated by considering external values of the pair $\boldsymbol{u}$, $\boldsymbol{l}$:

(i) *Normal variance function*: $\lambda = 4z_\alpha^2/(\boldsymbol{u} - \boldsymbol{l})^2$, where z_α is the upper $100\alpha\%$ point of the standard normal distribution. Furthermore $m = (\boldsymbol{u} + \boldsymbol{l})/2$ and $\kappa = \lambda m$. The reference values $\lambda = 0$ and $\kappa = 0$ are obtained by letting $\boldsymbol{u} - \boldsymbol{l}$ and $(\boldsymbol{u} - \boldsymbol{l})^2/|\boldsymbol{u} + \boldsymbol{l}|$ approach infinity.

(ii) *Poisson variance*: κ may be determined by solving $\chi^2[2\kappa, \alpha]/\chi^2[2\kappa, 1 - \alpha] = \boldsymbol{u}/\boldsymbol{l}$ for κ. This function of κ increases monotonically to infinity as κ decreases to 0; an approximate solution may be found by inspecting a χ^2 table. Then either $\lambda = \kappa/m$ or $\lambda = \chi^2[2\kappa, \alpha]/2\boldsymbol{u}$ or $\lambda = \chi^2[2\kappa, 1 - \alpha]/2\boldsymbol{l}$ may be used to find λ. The reference parameter pair $\kappa = 1/2$, $\lambda = 0$ corresponds to large values of $\boldsymbol{u}$ and $\boldsymbol{l}$ with $\boldsymbol{u}/\boldsymbol{l} = \{z[\alpha/2]/z[(1 - \alpha)/2]\}^2$.

(iii) Binomial(k) *variance*: With any κ and λ as prior parameters, Tables 2(b) and 3 imply that $[(\lambda k - \kappa)/\kappa]\mu/(k - \mu)$ has the $F[2\kappa, 2(\lambda k - \kappa)]$ distribution *a priori*. Therefore we may find κ, λ by setting $\nu_1 = 2\kappa$ and $\nu_2 = 2(\lambda k - \kappa)$ and solving for ν_1, ν_2 as follows: the condition $F[\nu_1, \nu_2, \alpha]F[\nu_2, \nu_1, \alpha] = [\boldsymbol{u}/(k - \boldsymbol{u})]/[\boldsymbol{l}/(k - \boldsymbol{l})]$ determines a contour in (ν_1, ν_2)-space. If m is also given, the solution is the intersection of this contour with the line $\nu_2 = (k - m)\nu_1/m$. If symmetric prior bounds are used, say $\boldsymbol{u} = \pi k$ and $\boldsymbol{l} = (1 - \pi)k$ given a π $(0.5 < \pi < 1)$, then an approximate solution may be found with a glance at the F-tables, since $\nu_1 = \nu_2$. Even with asymmetric bounds, it may be found by writing out a few values of $F[\nu_1, \nu_2, \alpha]F[\nu_2, \nu_1, \alpha]$; however an approximate contour map of this function (Fig. 1) is more convenient. Alternatively, starting with $\boldsymbol{u}$ and $\boldsymbol{l}$, the corresponding pair κ and λ may be found by solving simultaneously the two equations $(\nu_1/\nu_2)F[\nu_1, \nu_2, \alpha] = \boldsymbol{u}/(k - \boldsymbol{u})$, $(\nu_2/\nu_1)F[\nu_2, \nu_1, \alpha] = (k - \boldsymbol{l})/\boldsymbol{l}$ for ν_1, ν_2. Again, the solution for symmetric bounds is found with a glance at conventional F-

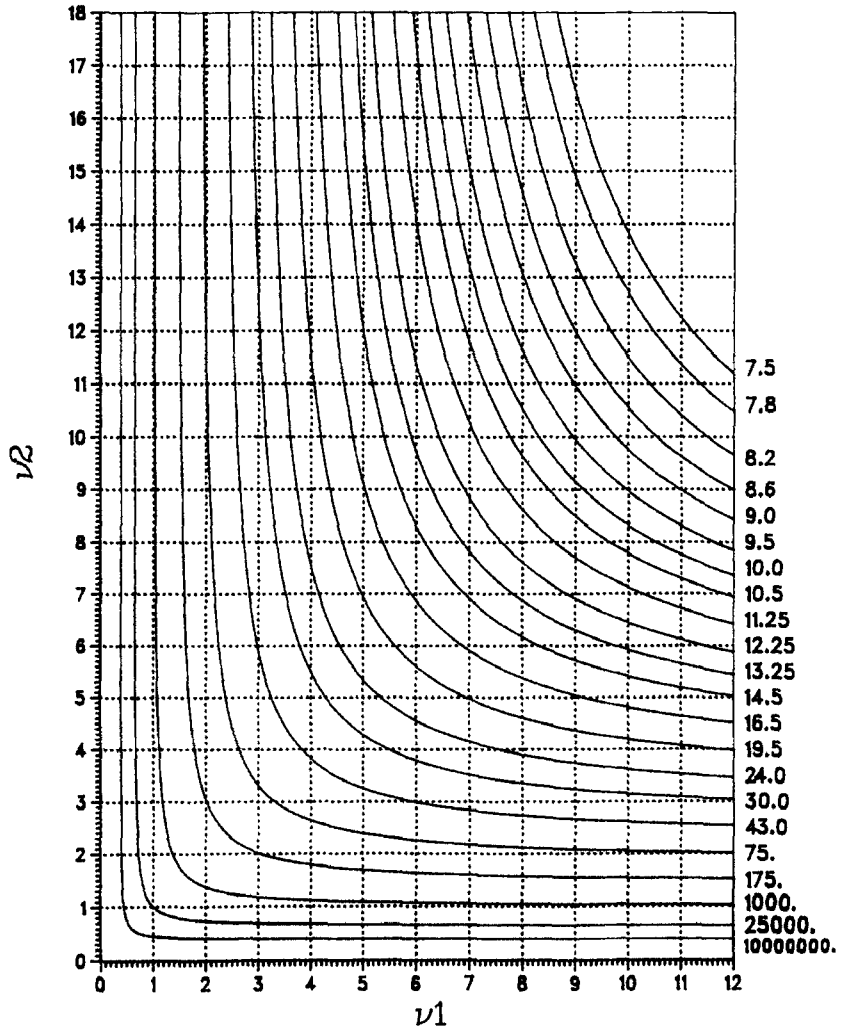


Fig. 1. Level contours of $F[\nu_1, \nu_2, .95]F[\nu_2, \nu_1, .95]$.

tables. With asymmetric bounds, the solution may be found by writing down a few values of $G[\nu_1, \nu_2, \alpha] = (\nu_1/\nu_2)F[\nu_1, \nu_2, \alpha]$, the upper $100\alpha\%$ points of chi-square ratios. Here again, an approximate contour map (Fig. 2) is more convenient. The curves with increasing slopes are $G[\nu_2, \nu_1, \alpha]$, needed below in (v). From $F[1, 1, 0.05] = 161.4$, it is seen for the binomial(k) that the reference parameters $\kappa = 1/2$ and $\lambda = 1/k$ correspond to upper and lower 5%-points $\boldsymbol{u} = 0.994k$ and $\boldsymbol{l} = 0.006k$.

(iv) *Gamma variance*: $2\kappa/\mu$ has the $\chi^2_{2(\lambda+1)}$ prior distribution, so that $\chi^2_{2(\lambda+1)}(\alpha)/\chi^2_{2(\lambda+1)}(1-\alpha) = \boldsymbol{u}/\boldsymbol{l}$ may be solved for λ. Then either $\kappa = \boldsymbol{u}\chi^2_{2(\lambda+1)} \cdot (1-\alpha)/2$ or $\kappa = \boldsymbol{l}\chi^2_{2(\lambda+1)}(\alpha)/2$ may be used, or $\kappa = \lambda m$ if a finite m is preferred. Convergence of (κ, λ) to the reference values $(0, -1)$ corresponds to the convergence condition that for arbitrarily large M, eventually $\boldsymbol{u}/\boldsymbol{l} > M$.

(v) *Negative binomial(k) variance*: From Table 3, $(\lambda k + 1)\mu/\kappa k$ has the

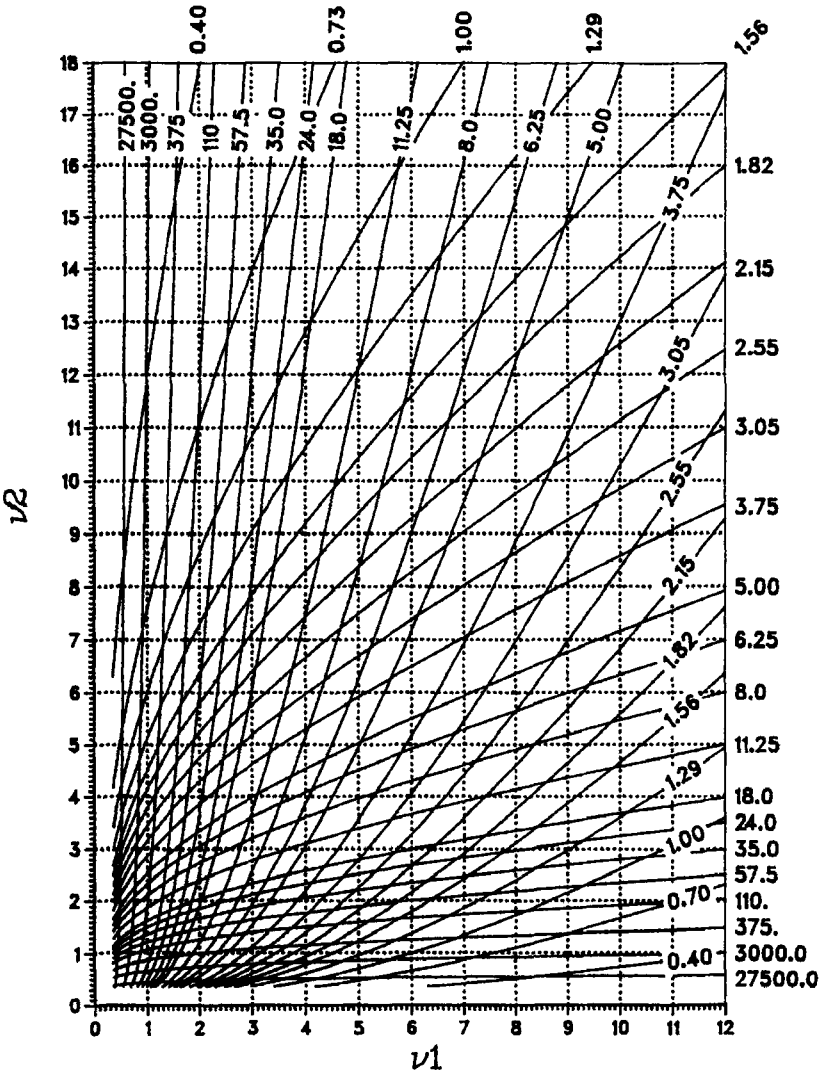


Fig. 2. Level contours of $G[\nu_1, \nu_2, .95] = (\nu_1/\nu_2)F[\nu_1, \nu_2, .95]$ (incr slopes) and of $G[\nu_2, \nu_1, .95]$ (decr slopes).

$F[2\kappa, 2(\lambda k + 1)]$ prior distribution. (κ, λ) will be found by setting $\nu_1 = 2\kappa$ and $\nu_2 = 2(\lambda k + 1)$ and solving for (ν_1, ν_2). Starting with $\boldsymbol{u}/\boldsymbol{l}$, the condition $F[\nu_1, \nu_2, \alpha]F[\nu_2, \nu_1, \alpha] = \boldsymbol{u}/\boldsymbol{l}$ determines a contour in (ν_1, ν_2)-space (Fig. 1). If, besides $\boldsymbol{u}/\boldsymbol{l}$, a finite m is available, then the solution is the intersection of this contour with the line $\nu_2 = (k/m)\nu_2 + 2$. Alternatively, starting with $\boldsymbol{u}$ and $\boldsymbol{l}$, the solution point is that of the system $G[\nu_1, \nu_2, \alpha] = \boldsymbol{u}/k$, $G[\nu_2, \nu_1, \alpha] = k/\boldsymbol{l}$, available from Fig. 2. Convergence of (κ, λ) to the reference values $(1/2, -1/k)$ (meaning that the resultant posterior distributions converge) is implied by $(\boldsymbol{u}, \boldsymbol{l}) = (k/\nu_2)(u, l)$ where u, l are the upper and lower points for $F[1, \nu_2]$ as $\nu_2 \to 0$.

(vi) *t-nomial(k) variance*: t parameters of the Dirichlet prior distribution of $\pi = \mu/k$ must be specified, e.g. the vector $m = [m_1, \ldots, k - \sum_{h=1}^{t-1} m_h]'$ of prior

expected values of $\mu = [\mu_1, \ldots, \mu_{t-1}]'$ and any one of the prior $s_h^2 = \text{var}(\mu_h)$ determine λ as in Table 3. (The values m_h and just one s_h^2 determine the other s_h^2 through $[s_1^2, \ldots, s_t^2] = [m_1(k - m_1), \ldots, m_h(k - m_t)]/(\lambda k + 1)$; then, $\kappa_h = \lambda m_h$.) Alternatively, given the pair of values m_h and its corresponding upper/lower odds ratio $r_h = [\boldsymbol{u}_h/(k - \boldsymbol{u}_h)]/\boldsymbol{l}_h(k - \boldsymbol{l}_h)]$ for just one h, λ may be found by finding the intersection of the (ν_1, ν_2)-contour $F[\nu_1, \nu_2, \alpha]F[\nu_2\nu_1, \alpha] = r_h$ with the line $\nu_h = (k - m_h)\nu_1/m_h$ (Fig. 1), from which $\lambda = (\nu_1 + \nu_2)/2k$. Then if all the m_h are known, the κ_h may be determined from $\kappa_h = \lambda m_h$. For $t = 2, 3$ and 4 the reference values $\kappa_h = 1/2$, $\lambda = t/2k$ produce upper 5% points $0.994k$, $0.90k$ and $0.77k$ respectively.

4. Prior distributions for the vector β of regression coefficients

4.1 *Univariate data*

We now consider a vector $\boldsymbol{y} = [y_1, \ldots, y_n']$ of univariate observations with mean $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]'$ and diagonal $n \times n$ variance-covariance matrix $V = V(\mu) = \text{diag}[v(\mu_1), \ldots, v(\mu_n)]$. Suppose v-conjugate priors, either reference or subjective, have been chosen for each θ_i with conjugate parameters κ_i and λ_i. These could differ, for example, if different subjective priors are used for $i = 1, 2, \ldots$. We shall assume the affine structure $\theta = \nu + X\beta$, where X is $n \times d$ and of full rank. Most applications in the generalized linear model literature satisfy this affine canonical link condition. Vector ν is a known *offset*, and is null in many applications. For example $\nu = 0$ with $x_i' = (\text{row } \#i \text{ of } X) = [1, t_i]$ gives a generalized simple regression model.

A choice of prior for each θ_i corresponding to d linearly independent rows x_i' of course determines a distribution for β. This latter distribution in turn determines a distribution for every θ_i, some of which might be incompatible with priors already chosen. In this section we propose a means of choosing a prior for β in such a way that the originally chosen priors for the θ_i are adjusted towards compatibility.

Let W have the same set of distinct rows as X but with no rows repeated. Write n_j $j = 1, \ldots, m$ for the number of times row w_j of W appears in X, and (by re-indexing) θ_j for the corresponding θ. If W is invertible then the above incompatibility does not arise. W is invertible for some common designs X, e.g., if X represents a one-way anova, or a higher-way anova with all interactions. In this situation $\theta = W\beta$, where θ is redefined to be $[\theta_1, \ldots, \theta_m]'$. By a change of variable the $d\beta$-prior $p(\beta \mid X)$ is v-conjugate and is proportional to $f(\beta \mid W) = |\det(W)| \exp\{\sum_{j=1}^m [\kappa_j\theta_j - \lambda_j b(\theta_j)]\}$. Note that the quasi-likelihood of $\boldsymbol{y}$ is $\text{const} + \sum_j [(\sum_{i:j(i)=j} y_i)\theta_j - n_j b(\theta_j)]$, where $j(i) = (\text{row index number in } W \text{ of } X\text{-row } \#i)$. It follows that the PME $\tilde{\beta}$ and $\widetilde{\text{var}}(\tilde{\beta})$ may be found exactly in the manner of finding the MQLE $\hat{\beta}$ and $\widehat{\text{var}}(\hat{\beta})$, only replacing $\boldsymbol{y}$ with $\tilde{\boldsymbol{y}} = [\tilde{y}_1, \ldots, \tilde{y}_n]'$ where $\tilde{y}_i = y_i + \kappa/n_{j(i)}$, and replacing the original variance function $v(\mu_{j(i)})$ in MQLE program with $v(\mu_{j(i)})(1 + \lambda/n_{j(i)})$. Thus, identical adjustments are applied to all cases of a given distinct covariate pattern. $\text{var}(y_i)$ and deviance increments can be specified with GLIM macros, while in BMDP3R $1/v(\mu_i)$ can be specified with a WEIGHT formula.

In the general case where W is not necessarily invertible and reference priors are used, Jeffreys' prior distribution $p(\beta \mid X) \propto \sqrt{\det\{X'VX\}}$ where $v(\mu_i) = v(\Theta^{-1}(\nu_i + x_i'\beta))$. Thus

$$p(\beta \mid X) \propto \sqrt{\det\{W'NUW\}}, \tag{4.1}$$

where

$$N = \operatorname{diag}[n_1, \ldots, n_m] \quad \text{and} \quad U = \operatorname{diag}[v(\mu_1), \ldots, v(\mu_m)].$$

The Cauchy-Binet theorem (Marcus and Minc (1966)) or Binet-Cauchy theorem (Pearl (1973)) states that $\sqrt{\det\{W'NUW\}} = [\sum_S \det\{W(S)'N(S)U(S) \cdot W(S)\}]^{1/2}$, where the variable summation index S is any d-element subset of the j-index set $\{1, \ldots, m\}$, $W(S)$ is the submatrix of W formed by those rows w_j of W with index j in S, and $N(S)$ and $U(S)$ are the corresponding diagonal $d \times d$ submatrices of N and U. $p(\beta \mid X)$ can be written in the form $p(\beta \mid X) \propto [\sum_S \omega(S) f(\beta \mid W(S))^2]^{1/2}$, where $\omega(S) = (\pi_{j:j\in S} n_j) / \sum_{S'} (\pi_{j:j\in S'} n_j)$ and $f(\beta \mid W(S)) = |\det\{W(S)\}| [\det\{U(S)\}]^{1/2} = |\det\{W(S)\}| \exp\{\sum_{j:j\in S} [\kappa\theta_j - \lambda_j b(\theta_j)]\}$ as above. Here $b(\theta_j)$ is as described in Table 2(b), and values of κ and λ_j in Table 2(b) are used. (We note that for the multinomial and negative binomial the reference λ might depend on j, but κ will not.)

Thus $p(\beta \mid X)$ is a weighted root-mean-square "average" of exponentials and is intractable when W is not invertible. In this case we propose a convenient product-form approximation: we replace this root-weighted-mean-square average $p(\beta \mid X)$ of the densities $f(\beta \mid W(S))$ with their similarly weighted geometric average,

$$q(\beta \mid X) \propto \prod_S f(\beta \mid W(S))^{\omega(S)} \propto \exp\left\{\sum_{j=1}^m \omega_j [\kappa\theta_j - \lambda_j b(\theta_j)]\right\}, \tag{4.2}$$

$$\theta_j = \nu_j + w_j'\beta$$

where for each j, $\omega_j = \sum_{S:j\in S} \omega(S)$. $q(\beta \mid X)$ will be called the *conjugate approximate reference prior* for the given model. If subjective κ_j and λ_j are used we propose the obvious modification of (4.2). To obtain the resulting PMEs one need only replace each element y_i of $\boldsymbol{y}$ with $y_i + \omega_{j(i)}\kappa/n_{j(i)}$ and replace its variance function $v(\mu_{j(i)})$ with $(1 + \omega_{j(i)}\lambda/n_{j(i)})v(\mu_i)$.

Note that v, d and $\boldsymbol{n} = [n_1, \ldots, n_m]' / \sum_j n_j$ completely determine $q(\beta \mid X)$ in the sense that a linear recording $\beta = M\gamma$ leads to $q(M\gamma \mid XM)$.

If $\boldsymbol{y}$ has variance-covariance $\phi V(\mu)$ with unknown $\phi > 0$, then $q(\beta \mid X) d\beta d\phi/\phi$ is proposed as the reference prior distribution for inferences about β. In estimating linear functions of β this leads to approximate confidence/credibility regions based on t-distributions rather than the normal.

It can be shown that $\sum_j \omega_j = d$, and if the design is balanced then every $\omega_j = d/m$. As for unbalanced designs, although calculation of the model adjustment weights ω_j is easy in specific examples with small m (Section 5, alternative formulation of Cox data), general formulas may be somewhat complicated even in simple general cases:

LEMMA 4.1. *Suppose distinct covariate patterns labeled* $1, 2, \ldots, A$ *each receive a observations, while the remaining patterns,* $A+1, A+2, \ldots, A+B$ *each receive* b. *Then* $\omega = [\omega_1, \ldots, \omega_m]'$ *is proportional to* $[c, \ldots, c, e, \ldots, e]'$, *where* $c = E[Ha^H b^{(d-H)}]/A$ *and* $e = E[(d-H)a^H b^{(d-H)}]/B$, *where* H *has the hypergeometric distribution, i.e., that of the number of class "A" items obtained in* d *selections without replacement from* $A+B$ *items.*

When W is invertible, exact posterior probabilities can sometimes be calculated in terms of a tabled distribution; for example, the posterior distribution in a binomial model is then a product of Dirichlet distributions. Thus for purposes of comparing the recommended approximation with an exact calculation we may calculate, for example, the exact posterior probability of the recommended approximate 68.26% posterior credibility interval $\tilde{\theta} \pm [\widetilde{\text{var}}(\tilde{\theta})]^{1/2}$ in the case of binomial data with $k = 2$ dichotomous trials, using the reference prior distribution: when this is done, the approximate credibility interval for estimating success probability $\pi = \mu/2$, arising from $y = 1$ observed success, is $0.240 < \pi < 0.760$ with exact posterior credibility 0.632. For $y = 2$ these figures become $0.515 < \pi < 0.959$ and 0.582. (This can be verified from F-tables since the exact posterior distribution of $\log(\theta)$ is a scaled F.)

Conditional distributions of θ of the form $\exp\{\sum_{j=1}^m [\kappa\theta_j - \lambda b(\theta_j)]\} d\theta_1 \cdots d\theta_m$ were used by Albert (1988) in a proposed two-stage hierarchical scheme.

4.2 *Multinomial models*

Suppose μ is a t-vector of t-nomial mean frequency counts. Then $v(\mu)$ is no longer just a number but is the $t \times t$ matrix $\text{diag}(\mu) - \mu\mu'/k$, which, although non-invertible, has a pseudo-inverse $\text{diag}\{1/\mu_i, \ldots, 1/\mu_i\}$. It is this diagonal form of the reweighting matrix which allows the t-variate analysis of t-nomial(k) data to be done with univariate MQLE programs, by arranging the entire string of t observed frequency counts as though it were a univariate outcome vector and by imposing the constraint $\hat{\mu}_1 + \cdots + \hat{\mu}_t = k$. This constraint removes one parameter, say θ_t, leaving $\theta_h = \log(\pi_h/\pi_t)$ for only $h = 1, \ldots, (t-1)$. Aitkin *et al.* ((1989), Section 5.6) describe how this is done with GLIM (including the important situation where covariates are involved). With more conventional iterated reweighted least squares (IRLS) programs such as BMDP3R the constraint may be imposed algebraically with the control language, using columns of 0s and 1s to indicate the t categories.

In order to convert from MQL Estimation to reference PM Estimation, the required adjustment to the data is simply to add $\omega_{j(i)}/2n_{j(i)}$ to each $y_{ih}, h = 1, \ldots, t$. Unlike the binomial analysis already discussed, no adjustment to the variance function is required. This is because the "conditioning" approach discussed above to modelling the multinomial is really a form of modelling independent Poisson variates. To implement a conjugate adjustment parameter κ in general (not necessarily reference values), the adjustment is: add $\kappa_{j(i)}\omega_{j(i)}/n_{j(i)}$ to each y_{ih}. The conditioning constraint becomes $\sum_{h=1}^t \hat{\mu}_{ih} = [k_i + t\kappa_{j(i)}\omega_{j(i)}]/n_{j(i)}$.

4.3 *A trinomial example*

The pneumoconiosis (miners' black lung disease) frequency count data appears in Aitkin *et al.* ((1989), p. 235). The three categories of severity are normal, mild

and severe. The explanatory variable is log (number of years of the miner's exposure through working at the mineface), with numbers of years $x = 5.8$, 15, 21.5, 27.5, 33.5, 39.5, 46 and 51.5. The corresponding 8 triples of observed frequency counts were 98 0 0; 51 2 1; 34 6 3; 35 5 8; 32 10 9; 23 7 8; 12 6 10; 4 2 5. The model as coded was $\theta_{xh} = \log(\pi_{xh}/\pi_{x1}) = \beta_h + \gamma_h \log(x)$, $h = 2, 3$, where π_{xh} = true proportion of the x-exposed population in severity category $\#h$. Suppose we guess *a priori* that the mild and severe categories are similar in that they share a common intercept around $\beta_h = -10.5$ and slope $\gamma_h = 2.5$ for both $h = 2, 3$. Thus our prior means m_{xh} for μ_{xh} are such that $[m_{x1}, m_{x2}, m_{x3}] \propto [1, e^{-10.5}x^{2.5}, e^{-10.5}x^{2.5}]$. Suppose we further decide that the prior standard deviation of each π_{xh} is about 50% of the maximum possible, i.e., that $1/(\lambda_x k_x + 1) = 1/4$. This gives $\lambda_x = 3/k_x$ and $\kappa_{xh} = 3m_{xh}/k_x$, where $k_x = \sum_{h=1}^{3} y_{xh}$. Note that $\omega_j = 4/8$ and $n_j = 1$ for all $j = 1, 2, \ldots, 8$. Consequently the subjective adjustment is to add $1.5m_{xh}/k_x$ to each y_{xh}. Table 4 shows the resulting estimates using the program GLIM both for standard MQLE and for the above subjective adjustments. The consequent changes are negligible, in part because the data happens to agree closely with the prior opinion.

Table 4. Pneumoconiosis: logistic intercepts β and slopes γ, with standard errors.

Coefficient	Standard MQLE		PME with the above subjective prior	
	Estimate	Est'd SE	Estimate	Est'd SE
β_2	−8.936	1.052	−8.972	1.031
β_3	−11.97	1.322	−11.88	1.287
γ_2	2.165	0.3045	2.174	0.2985
γ_3	3.067	0.3736	3.035	0.3640

5. Some special cases and examples

5.1 *Designs with* $\text{var}(y) = \phi\,\text{diag}[v(\mu_{11}, \ldots, v(\mu_{1k}); \ldots; v(\mu_{m1}), \ldots, v(\mu_{mk})]$, $\mu_{ji} = \mu_j$

Here we consider $n_j = k$ for all j, so that $\omega_j = d/m$, and we will suppose a common variance adjustment parameter λ is to be used for all m distinct rows of X (i.e., of W). The consequent replacement in L of each $b(\theta_{ji})$ with $(1+d\lambda/mk)b(\theta_{ji})$ amounts to modelling each $\text{var}(y_{ji})$ as $\phi(1+d\lambda/mk)v(\mu_{ji})$ rather than as $\phi v(\mu_{ji})$. Changing λ has no effect on $\tilde{\beta}$. If ϕ is to be set $= 1$ the estimate of $\text{var}(\tilde{\beta})$ is now $\widetilde{\text{var}}(\tilde{\beta}) = [X'Y(\tilde{\mu})X]^{-1}/(1 + d\lambda/mk)$. Otherwise if $\phi > 0$ is unknown, $\tilde{\phi} = \chi^2/(km - d) = (y - \tilde{\mu})'V^{-1}(y - \tilde{\mu})/(km - d)$ is a convenient estimate of ϕ and $\widetilde{\text{var}}(\tilde{\beta}) = \tilde{\phi}[X'V(\tilde{\mu})X]^{-1}/(1 + d\lambda/mk)$. Thus the reference adjustment of changing from $\widehat{\text{var}}(\hat{\beta})$ to $\widetilde{\text{var}}(\tilde{\beta})$ is one of scaling upward if the correctly fitted variance function is $v(\mu) = r + s\mu + t\mu^2$ with $t > 0$, since $\lambda = -t$ (Table 2(b)). Data exhibiting such a variance function with $r > 0$ and $s = 0$ could arise as follows: the outcomes $y_{i1}, y_{i2}, \ldots$ could be random observations with each y_{ik} from a population with mean and variance m_i and tm_i^2 (e.g., as with gamma

populations with a shape parameter $1/t$ common to all covariate patterns), where the mean m_i for covariate pattern $\#i$ is itself a random variable with mean μ_i and variance τ: it follows that $\text{var}(y_{ij}) = v(\mu_i) = (t+1)\tau + t\mu_i^2$.

5.2 *Log-linear and logistic models for $r \times c$ frequency counts*

5.2.1 *General*

Let y_{jh} = observed number of cross-classified occurrences in row $j = 1, \ldots, r$, column $h = 1, \ldots, c$. Table 5 exhibits the reference adjustment weights $\omega_{jh} = \omega$ for four models which might be fitted. It is understood that when a logistic t-nomial(k) model is fitted, the t-variate coding described in Subsection 4.2 is used. Thus no adjustment to the variance function μ_{jh} is made, and the term "logistic" implies, as does "log-linear", that the variance function for each observed count is μ_{jh}. In the case of log-linear modelling this is often rationalized as Poisson variance; in logistic t-nomial modelling it is the reciprocal of a diagonal element of the $t \times t$ diagonal matrix $V(\mu)^-$. Here $n_j = 1$ so the reference-adjusted observed outcomes are $y_{jh} + \omega_j/2$ for both log-linear and logistic t-nomial. Here $y_{..} = \sum_{jh} y_{jh}$ and $y_{\cdot h} = \sum_j y_{jh}$. Goodman ((1970), p. 229) has suggested in effect the general use of $\omega = 1$ and $\lambda = 0$ (sometimes in close agreement with Table 5) because "this adjustment of the y's reduces both the asymptotic bias and mean-squared-error of $\hat{\theta}_{jh}$".

Table 5. Reference weights for fitting two-way frequency counts.

Model	$d/m = \omega$
rc independent Poisson variates, additive effects	$(r + c - 1)/rc$
rc indep. Poisson variates, with interactions (full model)	$rc/rc = 1$
rc-nomial outcome (no covariates; conditioning on $y_{..}$)	$(rc - 1)/rc$
r-nomial outcome (one c-level covariate factor; conditioning on $y_{\cdot h}$)	$r(c-1)/rc$

5.2.2 *A 2×2 example*

Consider the 2×2 table $y_{11} = 0$, $y_{12} = 2$, $y_{21} = 2$, $y_{22} = 0$. The null hypothesis is that the two binomial(2) samples $[y_{1h}, y_{2h}]'$ $(h = 1, 2)$ are from a common population, as opposed to the full model of two populations with possibly different success rates $\pi = \mu/2$. The reference adjustment is to add $1/4$ to each y_{jh}. The results of three common tests are shown in Table 6. Here the deviances and Pearson χ^2 for the null and alternative models, and the Wald statistic for the larger model, are calculated in the usual way as in Aitkin *et al.* (1989) or McCullagh and Nelder (1983), but with the adjusted data and adjusted fitted model. Thus, e.g., the Wald statistic here is $\tilde{\beta}'[\widehat{\text{var}}(\tilde{\beta})]^{-1}\tilde{\beta}$. Each test statistic follows approximately a χ_1^2 distribution, central under the null hypothesis. (A Bayesian might prefer to point out that, conditionally given $\tilde{\beta}$, $\widehat{\text{var}}(\tilde{\beta})^{-1/2}(\beta - \tilde{\beta})$ has approximately a standard normal distribution.) These results may be compared with the "exact" p-level 0.125 based on permutational distributions (CYTEL (1990)), wherein for each h, a population of $y_{\cdot h}$ individuals is allocated randomly among the r outcome categories. Using a convenient coding of the

full model, the two estimated coefficients with SEs are $\tilde{\theta}_{12} = \log(\tilde{\pi}_{12}/\tilde{\pi}_{11}) = 2.197 \pm 2.357$ and $\tilde{\theta}_{22} - \tilde{\theta}_{12} = \log(\tilde{\pi}_{22}/\tilde{\pi}_{12}) - \log(\tilde{\pi}_{21}/\tilde{\pi}_{11}) = 2.556 \pm 3.333$. No such normal approximation or calculation can be made with the unadjusted likelihood.

Table 6.

	Δdeviance	ΔPearson χ^2	Wald
unadjusted	$5.545 = \chi^2_1(0.018)$	$4.000 = \chi^2_1(0.045)$	incalculable
adjusted	$2.945 = \chi^2_1(0.086)$	$2.556 = \chi^2_1(0.110)$	$1.738 = \chi^2_1(0.187)$

5.2.3 *A* 3×9 *example*

070000011
The table 111111100 appeared in CYTEL (1990), with the results in Ta-
080000011
ble 7, the two permutational p-levels being those observed for Δdeviance and ΔPearson χ^2 respectively. Treating this as 9 trinomial samples, the reference adjustment is to add 4/9 to each y_{jh}. The results are Δdeviance = 8.792, ΔPearsonχ^2 = 9.164, and $\tilde{\beta}'[\widetilde{\text{var}}(\tilde{\beta})]^{-1}\tilde{\beta} = 7.316$, for all of which the χ^2_{16} distribution carries no suggestion that not all 9 trinomial observations come from a common population. Does this contradict the small exact $p-$levels $\cong 0.04$? These are difficult to interpret in the absence of a specified alternative model. Looking at two columns at a time, there is little or no evidence that any one of the sparsely sampled trinomial populations $[\pi_{1h}, \pi_{2h}, \pi_{3h}]'$ (i.e., for $h \neq 2$) differs from $[\pi_{12}, \pi_{22}, \pi_{32}]'$. Also there is no evidence that the $[\pi_{1h}, \pi_{2h}, \pi_{3h}]'$ $(h \neq 2)$ are not all the same. The decision to treat the sparse columns unconsolidatedly expresses a prior belief that these 8 populations differ and rejects any "pooling" of the data columns $h \neq 2$. On the other hand if we knew *a priori* that they were the same, we would collapse them and analyse the 3×2 table with columns $[7, 1, 8]'$ and $[2, 6, 2]'$, with success probabilities γ_{jh} $(\gamma_{\cdot h} = 1)$. The results of the tests of $H_0 : [\gamma_{11}, \gamma_{21}, \gamma_{31}]' = [\gamma_{12}, \gamma_{22}, \gamma_{32}]'$ are then shown in Table 8 in good mutual agreement, and in close agreement between the $\chi^2_2(0.040)$ and the exact permutational p-level of the 3×9 table. It appears that the choice of alternative hypothesis matters.

Table 7.

	Δdeviance	ΔPearson χ^2	Exact Permutational p-levels
unadjusted	$\chi^2_{16}(0.119)$	$\chi^2_{16}(0.169)$	0.045, 0.041

Table 8.

	Δdeviance	ΔPearson χ^2	Wald statistic
unadjusted	$9.362 = \chi^2_2(0.009)$	$8.989 = \chi^2_2(0.011)$	$6.840 = \chi^2_2(0.033)$
adjusted	$8.422 = \chi^2_2(0.015)$	$8.186 = \chi^2_2(0.017)$	$6.433 = \chi^2_2(0.040)$

5.3 *Logistic binomial simple regression: the Cox data*

At each of four ages t_i, k_i items were randomly selected from the population of t_i-hour-old items and tested; of these y_{i1} failed and y_{i2} did not fail (Table 9). This data has appeared widely (Cox (1970), Jennrich and Moore (1975), BMDP ((1983), Chapter 14)). The model fitted was $\log[\mu_{i1}/(k_i - \mu_{i1})] = \alpha + \beta t_i$, $\mu_{i2} = k_i - \mu_{i1}$, $V(\mu_i)^- = \text{diag}[1/\mu_{i1}, 1/\mu_{i2}]$ which, together with $\widehat{\text{var}}(\hat{\alpha}, \hat{\beta}) = [X'Y(\hat{\mu})X]^{-1}$ where $V(\mu) = \text{diag}[v(\mu_i)] = \text{diag}[\mu_{i1}(k_i - \mu_{i1})/k_i]$, leads to MQLEs $\hat{\alpha} = -5.172 \pm 0.692$ and $\hat{\beta} = 0.0753 \pm 0.0212$. (An alternative "unconsolidated" analysis of Table 9,

$$\text{with } X = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_{387} \end{bmatrix}, \text{ each } k_i = 1, \text{ and } \boldsymbol{n} \propto [55, 157, 159, 16],$$

produced $\hat{\alpha} = -5.115$ and $\hat{\beta} = 0.0720$.) For posterior modal estimation, the distinct rows of

$$X = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ 1 & t_4 \end{bmatrix}$$

imply all $\omega_j = 2/4$, leading to the reference adjustments whereby each y_{ih} is replaced by $y_{ih} + 1/4$ and $[1 + 1/2k_i]v(\mu_i)$ replaces $v(\mu_i) = \text{diag}(\mu_i) - (1/k_i)\mu_i\mu_i'$ (2×2). The results $\tilde{\alpha} = -5.099 \pm 0.701$ and $\tilde{\beta} = 0.0744 \pm 0.0219$ testify to a dataset of sufficient strength that the omission of $q(\alpha, \beta \mid X)$ from $q(\alpha, \beta \mid X)d\alpha d\beta$ results in only negligible distortion. (The unconsolidated reference PME analysis requires unbalanced adjustment weights $\omega = [0.378, 0.748, 0.751, 0.123]$ and produces similar results.)

Table 9. The Cox ingor data.

age in hours	t_i	=	7,	14,	27,	51
number observed to fail	y_{i1}	=	0,	2,	7,	3
number observed to not fail	y_{i2}	=	55,	155,	152,	13

5.4 *MQLE vs. reference PME for a smaller ingot experiment*

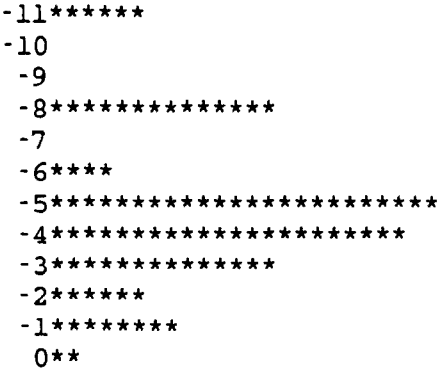
Although MQLE and reference PME produce similar inferences on moderately informative datasets, a small-sample MQLE vs. reference PME comparison highlights the difficulties of MQLE with small samples: 100 random "ingot" samples were generated with four binomial(k_i) observations each, with $\boldsymbol{k} = [5, 16, 16, 2]$, about one-tenth the sample size of the Cox data. $\alpha = -5$ and $\beta = 0.1$ were simulated. Summaries of the 100 resulting MQLEs and PMEs are shown in Table 10. Standard calculations of credibility intervals for functions of $[\alpha, \beta]$ are based on the assumption of approximate posterior normality. These are the same as the standard calculations of confidence intervals, which, from the sampling distribution standpoint, are based on the assumption of approximate normality of the

distribution of $[\hat{\alpha}, \hat{\beta}]$, or of $[\tilde{\alpha}, \tilde{\beta}]$. The qualities of these two normality assumptions (about sampling and posterior distributions) are closely related, but it is more convenient to compare the quality of the normality assumption, as it applys to $[\tilde{\alpha}, \tilde{\beta}]$ vs. $[\hat{\alpha}, \hat{\beta}]$, from the latter standpoint, that of sampling distributions: the grouped histograms in Fig. 3 are typical of comparisons in small-sample logistic regression, between the qualities of the normality assumption for MQLE and reference PME respectively. This illustrates an example of superiority of a reference PME, at least from a sampling distribution standpoint.

Table 10. Means and standard deviations of 100 estimates from $\alpha = -5$ and $\beta = 0.1$.

estimate	$\hat{\alpha}$	$\hat{\beta}$	$\tilde{\alpha}$	$\tilde{\beta}$
means	−12.44	0.240	−4.619	0.078
standard deviations	19.09	0.506	2.563	0.089

(a) 100 values of $\hat{\alpha}$



(b) 100 values of $\tilde{\alpha}$

Fig. 3.

6. Conclusions

The purposes of conjugate PMEs are to deal conveniently with small samples and with prior information. The availability of PMEs from an infinite array of conjugate priors allows the scientist either to adopt some standard reference prior or else to choose a prior by specifying some form of prior knowledge. The MQLE may be intractible and/or misleading for small or sparse datasets. (This needs further investigation, especially with models for continuous data.) The usual Jeffreys' prior is in some cases analytically intractible; hence we have proposed a convenient conjugate reference PME related to Jeffreys' and minimally informative priors. The specification of prior information can be conveniently carried out by specifying the general ranges of expected outcomes μ_i in terms of prior upper and lower percentage points of prior distributions.

Acknowledgements

The authors are grateful to Nicolas Hengartner, an NSERC Summer Scholarship student at Simon Fraser University, for cheerfully and efficiently creating graphs.

References

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989). *Statistical Modelling in GLIM*, Oxford, New York.

Albert, J. H. (1988). Computational methods using a Bayesian hierarchical generalized linear model, *J. Amer. Statist. Assoc.*, **83**, 1037–1044.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer, New York.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion), *J. R. Statist. Soc. Ser. B*, **41**, 113–147.

BMDP Statistical Software (1983). University of California Press, Berkeley.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Massachusetts.

Chang, T. and Eaves, D. M. (1990). Reference priors for the orbit in a group model, *Ann. Math. Stud.*, **18**, 1595–1614.

Cox, D. R. (1970). *The Analysis of Binary Data*, Chapman and Hall, London.

CYTEL Software Corp. (1990). Publicity for the package *StatXact*, based on work of Mehta, C. and Patel, N., Cambridge, Massachusetts.

Dobson, A. J. (1983). *An Introduction to Statistical Modelling*, Chapman and Hall, New York.

Goodman, L. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications, *J. Amer. Statist. Assoc.*, **65**, 226–256.

Jennrich, R. I. and Moore, R. H. (1975). Maximum likelihood estimation by means of nonlinear least squares, BMDP Technical Report, #9, BMDP Statistical Software, Inc., Los Angeles, California.

Marcus, M. and Minc, H. (1966). *Modern University Algebra*, Macmillan, New York.

McCullagh, P. (1983). Quasi-likelihood functions, *Ann. Statist.*, **11**, 59–67.

McCullagh, P. and Nelder, J. (1983). *Generalized Linear Models*, Chapman and Hall, London.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models, *J. R. Statist. Soc. Ser. A*, **135**, 370–384.

Pearl, M. (1973). *Matrix Theory and Finite Mathematics*, McGraw-Hill, New York.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalised linear models, and the Gauss-Newton method, *Biometrika*, **61**, 439–447.