DOUBLE EXPONENTIAL FAMILIES AND THEIR USE IN GENERALIZED LINEAR REGRESSION

BY

BRADLEY EFRON

TECHNICAL REPORT NO. 239

AUGUST 1985

PREPARED UNDER THE AUSPICES

OF

NATIONAL SCIENCE FOUNDATION GRANT MCS80-24649

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

DOUBLE EXPONENTIAL FAMILIES AND THEIR USE IN GENERALIZED LINEAR REGRESSION

By

Bradley Efron

TECHNICAL REPORT NO. 239

August 1985

PREPARED UNDER THE AUSPICES

OF

NATIONAL SCIENCE FOUNDATION GRANT MCS80-24649

Also prepared under Public Health Service Grant 5 R01 GM21215-11 and issued as Technical Report #107, Division of Biostatistics, Stanford University.

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

Double Exponential Families and Their Use in Generalized Linear Regression

Bradley Efron

Abstract

In one-parameter exponential families such as the binomial and the Poisson, the variance is a function of the mean. Double exponential families allow the introduction of a second parameter which controls variance independently of the mean. We use double families as constituent distributions in generalized linear regressions, in which both means and variances are allowed to depend on observed covariates. The theory is applied to two examples, a logistic regression and a large two-way contingency table. Close connections with previous ideas concerning generalized linear models are discussed.

DOUBLE EXPONENTIAL FAMILIES AND THEIR USE IN GENERALIZED LINEAR REGRESSION

Bradley Efron

1. Introduction.

Table 1 shows the proportions of subjects testing positive for the disease Toxoplasmosis in 34 cities of El Salvador. It was desired to assess the effect of rainfall on the proportions positive. Efron (1978) fit an ordinary logistic regression to this data and found that a cubic regression on rainfall was highly significant.

Cities with large sample sizes n_j have large influence when fitting an ordinary logistic regression. For example city 30 with $n_{30} = 75$ was roughly 7.5 times as influential as city 4, $n_4 = 10$. This happens because the model assumes binomial variation for the observed response y_j, in particular assuming that the variance of y_j is inversely proportional to sample size n_j.

The binomial model of variance may be untrustworthy. For instance given the realistic difficulties of obtaining the data in Table 1 the subjects may have been accrued in clumps, so that the statistician should really be using smaller values of n_j, not necessarily proportionately smaller in each city. This is the problem of over-dispersion, an habitual source of concern to users of binomial and Poisson models.

This paper concerns a class of regression families which allow the statistician to model overdispersion, while simultaneously carrying out the usual regression analyses for the mean as a function of the predictors. These are called double exponential families because they enjoy exponential family properties simultaneously for the mean and dispersion parameters.

Table 1 shows part of the outcome of such an analysis on the Toxoplasmosis data, presented in detail in Section 6. The analysis has adjusted all 34 sample sizes

City # j	Predictor (rainfall) x_j	Response (Proportion Pos.) y_j	Original Sample Size n_j	Effective Sample Size $n_j\theta_j$
1	1735	.500	4	2.4
2	1936	.300	10	6.9
3	2000	.200	5	3.1
4	1973	.300	10	6.9
5	1750	1.000	2	1.1
6	1800	.600	5	3.1
7	1750	.250	8	5.3
8	2077	.368	19	14.7
9	1920	.500	6	3.8
10	1800	.800	10	6.9
11	2050	.292	24	19.1
12	1830	.000	1	0.6
13	1650	.500	30	24.1
14	2200	.182	22	17.3
15	2000	.000	1	0.6
16	1770	.545	11	7.7
17	1920	.000	1	0.6
18	1770	.611	54	33.5
19	2240	.444	9	6.1
20	1620	.278	18	13.8
21	1756	.167	12	8.6
22	1650	.000	1	0.6
23	2250	.727	11	7.7
24	1796	.532	77	15.3
25	1890	.471	51	33.9
26	1871	.438	16	12.0
27	2063	.561	82	10.6
28	2100	.692	13	9.4
29	1918	.535	43	32.3
30	1834	.707	75	17.4
31	1780	.615	13	9.4
32	1900	.300	10	6.9
33	1976	.167	6	3.8
34	2292	.622	37	29.1

Table 1. Toxoplasmosis data , showing the proportions of subjects testing positive (y_j) and numbers tested (n_j) in 34 cities of El Salvador. The predictor (x_j), annual rainfall in mm., was used to fit an ordinary logistic regression, Efron (1978). The method of double exponential families fits an effective sample size (last column) as well as a logistic regression in rainfall. The effective sample size is based on the deviance of y_j from the fitted curve. This example is discussed in detail in Section 6.

downward, as shown in the last column. This is because the cubic model obtained by ordinary logistic regression was inadequate in the sense of failing a chi-square test for goodness of fit, see Table 2 of Efron (1978). In the absence of a better systematic model than cubic regression, the observed deviances of the y_j from the regression curve, which are too big according to ordinary binomial variation, are explained in terms of smaller effective sample sizes, i.e. bigger effective variances.

However the effective sample sizes are not equal ratios of the original sample sizes n_j. City 30 has been reduced from sample size 75 to 17.4, while city 4 has gone from 10 to 6.9. City 30 had only about three times the influence of city 4, rather than 7.5 times, in the double family fit of the y_j versus rainfall.

Double exponential families can be used to generalize any exponential family regression model, but we will be mainly interested in binomial and Poisson regressions. The toxoplasmosis example, Section 6, will be used to illustrate the binomial case. Poisson regression is discussed in terms of an intriguing example due to Mosteller and Parunak (1985): trying to identify which cells of a big two-way table most violate the hypothesis of independence, see Section 7.

Except for the emphasis on two-way exponential family behavior, the theory developed here is not original to this paper. Readers of the literature on Generalized Linear Models, in particular McCullagh and Nelder (1983), will find most of the ideas familiar, though stated from a different point of view. Double exponential families as models of overdispersion were introduced in Section 5 of Diaconis and Efron (1985) in the context of two-way contingency tables. As discussed in the Rejoinder of that paper, the idea is very similar to Nelder and Pregibon's extended quasi-likelihood, (1983). Bent Jorgensen (1985) gives a thorough mathematical treatment of a related construction, exponential dispersion models, though with differences that are important for applications to binomial and Poisson models. West's (1985)

scaled exponential families are particularly close in technical detail and interpretation to our development in Section 2. Pregibon's (1984) incisive review of McCullagh and Nelder's book strongly suggests the kind of applications we will be pursuing here.

Sections 2-5 present the definitions of double exponential families and their basic statistical properties. The paper finishes with the two examples, Sections 6 and 7. Without pretending to be a fully developed theory, the examples show the potential of double exponential families for robustifying standard logistic and Poisson regression analyses. Of course the same potential exists for the closely related ideas of the authors mentioned above.

2. Double Exponential Families.

Ordinary linear regression assumes that the observed responses y_j have normal distributions $N(\mu_j,V)$, with the expectations μ_j varying smoothly as a function of the predictors x_j. Logistic regression applies the same idea to the case where the response variables y_j are binomial, as in the Toxoplasmosis example of Section 1. Logistic regression is the most common example of general linear regression, in which the possible distributions of the responses y_j are members of a one-parameter exponential family.

The regression models of this paper are based on double exponential families, which allow us to conveniently model dispersion as well as mean response in a regression situation. This section describes the basic properties of double exponential families, the proofs appearing in Section 3.

We begin with an ordinary one-parameter exponential family of density functions,

$$g_{\mu,n}(y) = e^{n[\eta y-\psi(\mu)]} \quad [dG_n(y)] \ . \tag{2.1}$$

Here μ is the expectation parameter, $\mu = \int_{-\infty}^{\infty} y g_{\mu,n}(y)\, dG_n(y)$; y is the natural statistic; η is the natural, or canonical, parameter, a monotone function of μ ;

$\psi(\mu)$ is a normalizing function, chosen to make the density integrate to one; $G_n(y)$ is the carrier measure for the exponential family, so that $\text{Prob}_\mu\{A\} = \int_A g_{\mu,n}(y)dG_n(y)$ for measurable sets A ; and finally n is the sample size, in accordance with the familiar situation where y is actually the average of n independent quantities z_i , each of form (2.1) except with $n = 1$,

$$y = \sum_{i=1}^{n} z_i/n \qquad [z_i \overset{\text{ind}}{\sim} g_{\mu,1}] \ . \tag{2.2}$$

The terminology here follows Morris (1982), and Efron and Diaconis (1985).

Binomial family. As an example consider the rescaled binomial family with sample size n and probability parameter μ , say $y \sim \text{Bi}(n,\mu)/n$, $\mu \in [0,1]$:

$$g_{\mu,n}(y) = \binom{n}{ny} \mu^{ny}(1-\mu)^{n(1-y)} \quad \text{for} \quad y = 0,1/n,2/n,\ldots,1 \ . \tag{2.3}$$

The probability mass function $g_{\mu,n}(y)$ can be rewritten in form (2.1) by defining $\eta = \log \frac{\mu}{1-\mu}$, $\psi = -\log[2\cdot(1-\mu)]$, and G_n the discrete distribution putting mass $\binom{n}{ny}2^{-n}$ at $y = 0, 1/n, 2/n, \ldots, 1$. The extreme cases $\mu = 0$ and 1 technically do not belong to the exponential family, but it is convenient to include them here. The component variates z_i in (2.2) are Bernoulli's, $z_i = 1$ or 0 with probability μ and $1-\mu$ respectively. Note: The definitions and most of the results that follow are stated in ways that do not require the exponential family density to be expressed in canonical form (2.1).

The possible values of μ in (2.1), those for which $g_{\mu,n}(y)$ is a genuine density, lie in an interval of the real line. For easy discussion we will assume that y also lies in this interval. Then $\mu = y$ is the maximum likelihood estimate (MLE) of μ. That is, $g_{y,n}(y)$ maximizes $g_{\mu,n}(y)$ over the allowable choices of μ; see for example Efron (1978A).

We can now define double exponential families.

Definition. Given an exponential family (2.1), the family of density functions

$$\tilde{f}_{\mu,\theta,n}(y) = c(\mu,\theta,n)\ \theta^{\frac{1}{2}}\{g_{\mu,n}(y)\}^{\theta}\{g_{y,n}(y)\}^{1-\theta}\ \ [dG_n(y)] \tag{2.4}$$

is called a double exponential family with parameters μ, θ, and n. The constant $c(\mu,\theta,n)$ is defined to make $\int_{-\infty}^{\infty} \tilde{f}_{\mu,\theta,n}(y)\ dG_n(y) = 1$. Our intention is to use the densities (2.4) as constituents of a regression analysis, in which the unknown parameters μ and θ will both be estimated from the data.

The simplest example of (2.4) is the normal family, where we begin at (2.1) with $y \sim N(\mu,\sigma^2/n)$, μ unknown but σ^2 a fixed and known constant,

$$g_{\mu,n}(y) = \left(\frac{2\pi\sigma^2}{n}\right)^{-\frac{1}{2}} e^{-\frac{n}{2}\left(\frac{y-\mu}{\sigma}\right)^2}\ . \tag{2.5}$$

The variance of y is expressed as σ^2/n in accordance with convention (2.2) that $y = \Sigma_{i=1}^{n} z_i/n$, where $z_i \overset{ind}{\sim} g_{\mu,1} \sim N(\mu,\sigma^2)$. Then (2.4) gives

$$\tilde{f}_{\mu,\theta,n}(y) = c(\mu,\theta,n)\ \left(\frac{2\pi\sigma^2}{n\theta}\right)^{-\frac{1}{2}} e^{-\frac{n\theta}{2}\left(\frac{y-\mu}{\sigma}\right)^2}\ . \tag{2.6}$$

Both (2.5) and (2.6) are densities with respect to ordinary Lebesgue measure. In this case it is obvious that $c(\mu,\theta,n) = 1$, and that (2.6) is just the density for $y \sim N(\mu, \frac{\sigma^2}{n\theta})$.

In other words, if we start with the ordinary family $N(\mu,\sigma^2/n)$, with unknown mean μ but known variance σ^2/n, then the double family has unknown mean μ and also unknown variance $\sigma^2/n\theta$. Of course none of this is necessary in the normal case, where we could just take σ^2 unknown to begin with. Definition (2.4) allows us to add a dispersion parameter θ to families like the binomial, which originally have only one unknown parameter μ.

We will state several useful facts about double exponential families, which emphasize that the analogy with the normal family carries through surprisingly far. [For example Fact 4: that $\tilde{f}_{\mu,\theta,n}(y)$ represents approximately the same distribution as (2.1) except with the sample size changed from n to $n\theta$, as in (2.5), (2.6).]

The proofs of these facts are deferred until Section 3, as are the close connections with previous work by West, Nelder, McCullagh, Pregibon, and Jorgensen.

In addition to the definitions following (2.1), we need notation for the variance function and Kullback-Leibler distance of an ordinary exponential family $g_{\mu,1}$,

$$\text{variance function:}\quad V(\mu) = \mathrm{Var}_{\mu,1}\{z\} = \int_{-\infty}^{\infty}(z-\mu)^2\, g_{\mu,1}(z)\, dG_1(z) \qquad (2.7)$$

and

$$\text{Kullback-Leibler distance:}\quad I(\mu_1,\mu_2) = E_{\mu,1}\log\frac{g_{\mu_1,1}(z)}{g_{\mu_2,1}(z)} \ . \qquad (2.8)$$

These definitions apply to the case $n = 1$. For the general case $g_{\mu,n}$ considered in (2.1),

$$\mathrm{Var}_{\mu,n}(y) = V(\mu)/n \quad\text{and}\quad I_n(\mu_1,\mu_2) = E_{\mu,n}\log\frac{g_{\mu_1,n}(y)}{g_{\mu_2,n}(y)} = nI(\mu_1,\mu_2). \qquad (2.9)$$

Twice the Kullback-Leibler distance $I_n(y,\mu)$ is the Deviance $D(y,\mu)$, as discussed later

Fact 1: The constant $c(\mu,\theta,n)$ in (2.4) nearly equals 1 , so the family of densities (2.4) can be approximated by

$$f_{\mu,\theta,n}(y) = \theta^{\frac{1}{2}}\{g_{\mu,n}(y)\}^{\theta}\,\{g_{y,n}(y)\}^{1-\theta}\quad [dG_n(y)] \ . \qquad (2.10)$$

This approximation is usually quite accurate, as discussed in Section 3, and (2.10) is much more convenient to use than (2.4). In subsequent sections double exponential families will be applied in form (2.10). However it is easier to present some of the properties of double families in terms of definition (2.4).

Fact 2: The density $\tilde{f}_{\mu,n,\theta}(y)$ has mean value approximately μ and variance approximately $V(\mu)/n\theta$.

Fact 3: With θ and n fixed, (2.4) is an exponential family of densities indexed by μ ,

$$\tilde{f}_{\mu,\theta,n}(y) = a_{\theta,n}(\mu)\, b_{\theta,n}(y)\, e^{n\theta[\eta y - \psi(\mu)]} \quad [dG_n(y)] \ , \tag{2.11}$$

with natural statistic y , natural parameter η , and expectation parameter approximately equal to μ. The fact that the carrier distribution G_n is the same for $\tilde{f}_{\mu,\theta,n}(y)$ as for $g_{\mu,n}(y)$ is important for the binomial and Poisson applications of Sections 6 and 7.

Fact 4: The density $\tilde{f}_{\mu,\theta,n}(y)$ represents approximately the same probability distribution as (2.1) with n changed to $n\theta$,

$$\int_A \tilde{f}_{\mu,\theta,n}(y)\, dG_n(y) \doteq \int_A g_{\mu,n\theta}(y)\, dG_{n\theta}(y) \ , \tag{2.12}$$

for any interval A. The same statement holds with $f_{\mu,\theta,n}(y)$ replacing $\tilde{f}_{\mu,\theta,n}(y)$.

Fact 5: With μ and n fixed, (2.4) is an exponential family of densities indexed by θ

$$\tilde{f}_{\mu,\theta,n}(y) = c(\mu,\theta,n)\, \theta^{\frac{1}{2}}\, e^{-n\theta I(y,\mu)}\, g_{y,n}(y) \quad [dG_n(y)] \ , \tag{2.13}$$

with natural statistic $-nI(y,\mu)$, and natural parameter θ. The Deviance

$$D(y,\mu) \equiv 2n\, I(y,\mu) \tag{2.14}$$

has an approximate scaled chi-square distribution

$$D(y,\mu) \doteq \frac{1}{\theta}\, \chi^2_1 \tag{2.15}$$

for $y \sim \tilde{f}_{\mu,\theta,n}$.

Fact 6: The expected Fisher information matrix $\mathcal{J}_{\mu,\theta}$ for (μ,θ) in family (2.4) , n fixed, approximately equals

$$\mathcal{I}_{\mu,\theta} \doteq \begin{pmatrix} \frac{n\theta}{V(\mu)} & 0 \\ 0 & \frac{1}{2\theta^2} \end{pmatrix}. \tag{2.16}$$

Finally, here are some results concerning likelihood analyses of repeated independent observations from a double exponential family. Let

$$y_1, y_2, \ldots, y_J \overset{\text{ind}}{\sim} \tilde{f}_{\mu,\theta,n} \tag{2.17}$$

and define

$$\ell_{\mu,\theta,n} = \log \prod_{j=1}^{J} f_{\mu,\theta,n}(y_j) \ . \tag{2.18}$$

(It is now more convenient to work with the approximate density (2.18).)

Fact 7: The score functions based on the approximate likelihood (2.18) are

$$\frac{\partial \ell}{\partial \mu} = n\theta J \frac{\bar{y}-\mu}{V(\mu)} \quad \text{and} \quad \frac{\partial \ell}{\partial \theta} = \frac{J}{2\theta} - n \sum_{j=1}^{J} I(y_j,\mu) \tag{2.19}$$

where $\bar{y} = \Sigma_{j=1}^{J} y_j/J$. These give maximum likelihood estimates

$$\hat{\mu} = \bar{y} \quad \text{and} \quad \hat{\theta} = \frac{J}{2n \ \Sigma_{j=1}^{J} I(y_j,\bar{y})} \ . \tag{2.20}$$

Fact 8: The observed Fisher information matrix for (μ,θ) based on (2.18) is

$$\hat{i} \equiv -\begin{pmatrix} \partial^2\ell/\partial\mu^2 & \partial^2\ell/\partial\mu\partial\theta \\ \partial^2\ell/\partial\mu\partial\theta & \partial^2\ell/\partial\theta^2 \end{pmatrix}_{\hat{\mu},\hat{\theta}} = \begin{pmatrix} \frac{n\hat{\theta}J}{V(\hat{\mu})} & 0 \\ 0 & \frac{J}{2\hat{\theta}^2} \end{pmatrix}. \tag{2.21}$$

Facts 7 and 8 are discussed further in Section 3, emphasizing the close analogy with normal families (2.5)-(2.6). For example the fact that the off-diagonal elements in (2.19), and (2.14), are zero corresponds to the independence of $\hat{\mu}$ and $\hat{\sigma}^2$ in the normal situation.

To summarize this Section, we begin with an ordinary one-parameter exponential family $g_{\mu,n}$; definition (2.4) allows us to add a second parameter θ , which varies the dispersion of the family without changing the mean; the extended family behaves like $g_{\mu,n\theta}$, that is the original family with sample size changed from n to $n\theta$; the extended family is an exponential family in μ when θ and n are fixed, and an exponential family in θ when μ and n are fixed; and finally the extended family enjoys, at least approximately, some of the useful properties of the $N(\mu,\sigma^2)$ family.

3. Proofs.

We prove the facts stated in Section 2 and discuss them further, giving some idea of the errors involved in the various approximations. The material of this section is mainly technical and can be deferred until after Section 7 at the reader's preference.

Our theory depends on Hoeffding's representation of an exponential family density,

$$g_{\mu,n}(y) = g_{y,n}(y)\, e^{-nI(y,\mu)} \quad . \tag{3.1}$$

This follows easily from (2.1) and the expression

$$I(\mu_1,\mu_2) = (\eta_1-\eta_2)\mu_1 - (\psi(\mu_1)-\psi(\mu_2)) \tag{3.2}$$

for the Kullback-Leibler distance (2.8), see Efron (1978A).

Notice that the approximate version (2.10) of the double exponential family $\tilde{f}_{\mu,\theta,n}$ can be rewritten as

$$f_{\mu,\theta,n}(y) = \theta^{\frac{1}{2}}\, g_{y,n}(y)\, e^{-n\theta I(y,\mu)} \quad . \tag{3.3}$$

West (1985) uses (3.3) to begin his Bayesian investigation of scaled exponential families, also arriving at expressions quite similar to (2.10).

In order to verify Fact 1 we need to evaluate

$$\frac{1}{c(\mu,\theta,n)} = \int_{-\infty}^{\infty} f_{\mu,\theta,n}(y)\,dG_n(y) = \theta^{\frac{1}{2}} \int_{-\infty}^{\infty} e^{-n\theta I(y,\mu)} g_{y,n\theta}(y) R_{n,\theta}(y)\,dG_{n\theta}(y) \ , \qquad (3.4)$$

where we have defined

$$R_{n,\theta}(y) = \frac{g_{y,n}(y)}{g_{y,n\theta}(y)} \frac{dG_n}{dG_{n\theta}}(y) \ . \qquad (3.5)$$

Using (3.1) again, for sample size $n\theta$, (3.4) becomes

$$\frac{1}{c(\mu,\theta,n)} = \theta^{\frac{1}{2}} \int_{-\infty}^{\infty} R_{n,\theta}(y)\ g_{\mu,n\theta}(y)\ dG_{n\theta}(y) \ . \qquad (3.6)$$

Applying the central limit theorem (CLT) to $y = \Sigma_{i=1}^{n} z_i/n$, $z_i \overset{iid}{\sim} g_{\mu,1}$, gives

$$g_{y,n}(y)\ G_n(dy) \doteq \left[\frac{n}{2\pi V(y)}\right]^{\frac{1}{2}} \qquad (3.7)$$

for dy a small interval containing y. Dividing by the corresponding approximation for sample size $n\theta$, $g_{y,n\theta}(y)G_{n\theta}(dy) \doteq [n\theta/2\pi V(y)]^{\frac{1}{2}}$, results in $R_{n,\theta}(y) \doteq \theta^{-\frac{1}{2}}$. Standard Edgeworth series arguments show that the error in (3.7) is a factor $[1+O_p(n^{-1})]$, so that actually

$$R_{n,\theta}(y) \doteq \theta^{-\frac{1}{2}}[1+O_p(n^{-1})] \ . \qquad (3.8)$$

Then (3.6) gives

$$c(\mu,\theta,n) = 1 + O(n^{-1}) \equiv 1 + \frac{C(\mu,\theta,n)}{n} \qquad (3.9)$$

say. This is a quantitative statement of Fact 1. Later we will calculate $C(\mu,\theta,n)$ explicitly.

The argument leading to (3.9) is presented in more detail in Section 5 of Diaconis and Efron (1985), applied to multidimensional exponential families. Notice

that it uses the central limit theorem in saddlepoint form, that is only at the center of the distribution where the CLT is most accurate. This is why the error terms in (3.8), (3.9) are $O(n^{-1})$ rather than $O(n^{-\frac{1}{2}})$. It seems as if $n\theta$ should be an integer for the argument to work, but that is not actually necessary for any of the common cases, normal, binomial, Poisson, etc. Saddlepoint arguments are also the basis of McCullagh and Nelder's (1983) generalized quasilikelihood. The rejoinder of Diaconis and Efron (1985) shows how close the two ideas are.

Gamma Family. Suppose that we begin at (2.1) with the Gamma family,

$$g_{\mu,n}(y) = \frac{y^{n-1}\, e^{-ny/\mu}}{(\mu/n)^n\, \Gamma(n)} \qquad (\mu, y > 0) \ . \tag{3.10}$$

Here we have taken the densities with respect to ordinary Lebesgue measure $dG_n(y) = dy$, and have not converted to canonical form (2.1), but that isn't necessary to calculate $R_{n,\theta}(y)$. Definition (3.5), and the fact that $dG_n/dG_{n\theta}(y) = 1$, gives

$$R_{n,\theta}(y) = \frac{g_{y,n}(y)}{g_{y,n\theta}(y)} = \frac{\Gamma(n\theta)\,(n/e)^n}{\Gamma(n)\,(n\theta/e)^n} \ . \tag{3.11}$$

In this case $R_{n,\theta}(y)$ doesn't depend on y. We obtain directly from (3.6) that

$$c(\mu,\theta,n) = 1/\{\theta^{\frac{1}{2}}\, R_{n,\theta}\} \doteq 1 - \frac{1}{12n}\,\frac{1-\theta}{\theta} \ , \tag{3.12}$$

the last approximation coming from Stirling's formula. Thus $C(\mu,\theta,n) \doteq -(1-\theta)/12\theta$ in (3.9). Note: $\tilde{f}_{\mu,\theta,n}$ exactly equals $g_{\mu,n\theta}$ for the Gamma family, so Fact 4 is exact. See the note following (3.16).

Edgeworth series methods provide good approximations for $C(\mu,n,\theta)$ in any double exponential family. A standard three-term Edgeworth expansion for $g_{y,n}(y)\, G_n(dy)$, Johnson and Kotz (1970), equation 46, Section 12.4, results in

$$R_{n,\theta}(y) \doteq \theta^{-\frac{1}{2}}\left[1 - \frac{1-\theta}{n\theta}\,\varepsilon(y)\right] \ , \quad \varepsilon(y) \equiv \frac{9\delta_y - 15\gamma_y^2}{72} \ , \tag{3.13}$$

where γ_μ is the skewness and δ_μ the kurtosis of $z \sim g_{\mu,1}$. Then by (3.6)

$$c(\mu,\theta,n) \doteq 1 + \frac{1}{n}\frac{1-\theta}{\theta}\,\varepsilon(\mu) \;; \tag{3.14}$$

so $C(\mu,\theta,n) \doteq [(1-\theta)/\theta]\,\varepsilon(\mu)$ in (3.9). The error in (3.14), which is $O(n^{-2})$, can be reduced by more careful integration of $R_{n,\theta}(y)$ in (3.6), as in the Poisson example of Section 4. For the binomial double exponential family, $g_{\mu,n}$ as in (2.3),

$$c(\mu,\theta,n) \doteq 1 + \frac{1}{12n}\frac{1-\theta}{\theta}\left[1 - \frac{1}{\mu(1-\mu)}\right] . \tag{3.15}$$

The calculations above verify Fact 4 as well as Fact 1. Following steps (3.4)-(3.6) and (3.13), we have for any interval A

$$\begin{aligned}\int_A f_{\mu,\theta,n}(y)\; dG_n(y) &= \int_A \theta^{\frac{1}{2}}\, R_{n,\theta}(y)\; g_{\mu,n\theta}(y)\; dG_{n\theta}(y) \\ &\doteq \int_A \left[1 - \frac{1-\theta}{n\theta}\,\varepsilon(y)\right]\; g_{\mu,n\theta}(y)\; dG_{n\theta}(y) \;,\end{aligned} \tag{3.16}$$

a similar result holding for $\tilde{f}_{\mu,\theta,n}$. The error in (2.12) is $O(n^{-1})$. Note: if $R_{n,\theta}(y)$ does not depend on y, as in the case of the Gamma family, then the representation $\int_A \tilde{f}_{\mu,\theta,n}(y)\; dG_n(y) = \int_A c(\mu,\theta,n)\theta^{\frac{1}{2}}\, R_{n,\theta}\; g_{\mu,n\theta}(y)\; dG_{n\theta}(y)$ shows that $c(\mu,n,\theta) = \{\theta^{\frac{1}{2}}\, R_{n,\theta}\}^{-1}$, and that Fact 4 holds exactly.

Since by definition, and by (3.3),

$$\tilde{f}_{\mu,\theta,n}(y) \equiv c(\mu,\theta,n)\; f_{\mu,\theta,n}(y) = c(\mu,\theta,n)\; \theta^{\frac{1}{2}}\; g_{y,n}(y)\; e^{-n\theta I(y,\mu)} \;, \tag{3.17}$$

(3.2) leads to

$$\tilde{f}_{\mu,\theta,n}(y) = [c(\mu,\theta,n)\; \theta^{\frac{1}{2}}]\,[e^{-n\theta\eta_y y - \psi(y)}\; g_{y,n}(y)]\, e^{n\theta[\eta y - \psi(\mu)]} \;, \tag{3.18}$$

which verifies Fact 3, (2.11). Here η_y is the value of η corresponding to $\mu = y$.

With θ and n fixed, (3.18) can be thought of as the density function for an exponential family with natural statistic y, natural parameter η, sample size n, and normalizing function

$$\tilde{\psi}(\mu) = \psi(\mu) - \frac{1}{n\theta} \log c(\mu,\theta,n) \ . \tag{3.19}$$

The mean and variance of y are obtained, as usual in exponential families, by differentiating $\tilde{\psi}$ with respect to η,

$$\begin{aligned} E_{\mu,\theta,n}\{y\} &= \frac{d}{d\eta}\tilde{\psi}(\mu) = \frac{d}{d\eta}\left[\psi(\mu) - \frac{1}{n\theta}\log c(\mu,\theta,n)\right] \\ \mathrm{Var}_{\mu,\theta,n}\{y\} &= \frac{1}{n\theta}\frac{d^2}{d\eta^2}\tilde{\psi}(\mu) = \frac{1}{n\theta}\frac{d^2}{d\eta^2}\left[\psi(\mu) - \frac{1}{n\theta}\log c(\mu,\theta,n)\right] \ . \end{aligned} \tag{3.20}$$

From the standard results $\frac{d}{d\eta}\psi(\mu) = \mu$, $\frac{d^2}{d\eta^2}\psi(\mu) = V(\mu)$, and (3.9),

$$\begin{aligned} E_{\mu,\theta,n}\{y\} &\doteq \mu - \frac{1}{n^2\theta}\frac{\partial}{\partial\eta}C(\mu,\theta,n) = \mu + O(n^{-2}) \\ \mathrm{Var}_{\mu,\theta,n}\{y\} &\doteq \frac{V(\mu)}{n\theta} - \frac{1}{n^3\theta^2}\frac{\partial^2}{\partial\eta^2}C(\mu,\theta,n) = \frac{V(\mu)}{n\theta}\left[1+O(n^{-2})\right] \ . \end{aligned} \tag{3.21}$$

Fact 2 is seen to be accurate to $O(n^{-2})$. The numerical results for the Poisson case presented in Section 4 confirm this high accuracy.

Fact 5, (2.13), is identical to (3.17). We can write (3.17) as

$$\tilde{f}_{\mu,\theta,n}(y) = e^{-\theta D(y,\mu)/2 + \frac{1}{2}\log\theta + \log c(\mu,\theta,n)} g_{y,n}(y) \ , \tag{3.22}$$

and then differentiate the normalizing function $-\frac{1}{2}\log\theta - \log c(\mu,\theta,n)$ to obtain the cumulants of the natural statistic $-D(y,\mu)/2$. We get for instance

$$E_{\mu,\theta,n}\{D\} = \frac{1}{\theta} + 2\,\frac{\partial}{\partial\theta}\log c(\mu,\theta,n) = \frac{1}{\theta} + O(n^{-1})$$

$$\mathrm{Var}_{\mu,\theta,n}\{D\} = \frac{2}{\theta^2} + 4\,\frac{\partial^2}{\partial\theta^2}\log c(\mu,\theta,n) = \frac{2}{\theta^2} + O(n^{-1})\ . \tag{3.23}$$

All of the cumulants of D approximately agree with those of a χ_1^2/θ variate, errors of order $O(n^{-1})$, verifying (2.15).

Note: (2.15) extends the well-known result that $D(y,\mu)$ is approximately distributed as χ_1^2, error $O(n^{-1})$, when $y \sim g_{\mu,n}$. See for example McCullagh and Nelder (1983), Appendices A-D, and also Barndorff-Nielsen and Cox (1984).

Now let $\tilde{\ell}_{\mu,\theta,n} \equiv \log \tilde{f}_{\mu,\theta,n}(y)$. From (3.18), (3.17), and (3.9) the score functions are

$$\frac{\partial\tilde{\ell}}{\partial\mu} = n\theta\,\frac{y-\mu}{V(\mu)} + \frac{\partial}{\partial\mu}\log c(\mu,\theta,n) \doteq n\theta\,\frac{y-\mu}{V(\mu)} + \frac{1}{n}\,\frac{\partial C(\mu,\theta,n)}{\partial\mu}$$

and

$$\frac{\partial\tilde{\ell}}{\partial\theta} = -nI(y,\mu) + \frac{1}{2\theta} + \frac{\partial}{\partial\theta}\log c(\mu,\theta,n) \doteq -nI(y,\mu) + \frac{1}{2\theta} + \frac{1}{n}\,\frac{\partial C(\mu,\theta,n)}{\partial\theta}\ . \tag{3.24}$$

The second partial derivatives are

$$\frac{\partial^2\tilde{\ell}}{\partial\mu^2} \doteq -n\theta\left[\frac{1}{V(\mu)} + \frac{y-\mu}{V(\mu)}\,\frac{V'(\mu)}{V(\mu)}\right] + \frac{1}{n}\,\frac{\partial^2 C}{\partial\mu^2}\ , \qquad \frac{\partial^2\tilde{\ell}}{\partial\theta\partial\mu} \doteq \frac{y-\mu}{V(\mu)} + \frac{1}{n}\,\frac{\partial^2 C}{\partial\theta\partial\mu}$$

and

$$\frac{\partial^2\tilde{\ell}}{\partial\theta^2} \doteq -\frac{1}{2\theta^2} + \frac{1}{n}\,\frac{\partial^2 C}{\partial\theta^2}\ . \tag{3.25}$$

The expectations of minus the second derivatives are the entries of the expected Fisher information matrix $\mathcal{J}_{\mu,\theta}$. Fact 6 follows from (3.21). The errors in (2.16) are $O(n^{-1})$.

The approximate log likelihood $\ell_{\mu,\theta,n}$, (2.18), is

$$\ell_{\mu,\theta,n} = \frac{J}{2}\log\theta - n\theta J\, I(\bar{y},\mu) - n\theta \sum_{j=1}^{J} I(y_j,\bar{y}) - \sum_{j=1}^{J} \log g_{y_j,n}(y_j) \quad , \quad (3.26)$$

by (3.3) and the identity $\Sigma_{j=1}^{J} I(y_j,\mu) = JI(\bar{y},\mu) + \Sigma_{j=1}^{J} I(y_j,\mu)$, Efron (1978A). Facts 7 and 8 follow by differentiation of (3.26), remembering that $\frac{\partial}{\partial\mu} I(y,\mu) = -(y-\mu)/V(\mu)$. This last fact is emphasized in Wedderburn's important quasilikelihood paper of (1974). Double exponential families can be thought of as a way to remove the "quasi" from quasilikelihood; Nelder and Pregibon's extended quasilikelihood has similar intentions.

Remark A. The score function $\partial\ell/\partial\mu$ in (2.19) considered as a function of μ is identical to the score function $\frac{\partial}{\partial\mu}\log g_{\mu,n\theta J}(\bar{y})$ which would apply if we observed $y_1,y_2,\ldots,y_J \overset{\text{ind}}{\sim} g_{\mu,n\theta}$. In a likelihood sense, (2.19) is a more exact version of Fact 4: $f_{\mu,n,\theta} \doteq g_{\mu,n\theta}$.

Remark B. The score function $\partial\ell/\partial\theta$ in (2.19) considered as a function of θ is identical to the score function for θ based on observing $\Sigma_{j=1}^{J} D(y_j,\mu)$, if it were exactly true that

$$\sum_{j=1}^{J} D(y_j,\mu) \sim \frac{1}{\theta}\chi^2_J \quad . \quad (3.27)$$

In a likelihood sense, (2.19) is a more exact version of (2.15): $D(y_j,\mu) \doteq \chi^2_1/\theta$. Section 2 of West (1985) makes similar points.

Remark C. The fact that the off-diagonal elements in the information matrices are zero implies asymptotic independence for $\hat{\mu}$ and $\hat{\theta}$. This has important consequences: in estimating θ for instance, it doesn't matter whether or not μ is known, at least not in a first-order asymptotic sense. We will see this point arising in the regression situation of Sections 5-7.

Remark D. Double exponential families are also standard two-parameter exponential families. This can be seen from (3.18). The natural statistics are $(y,-\eta_y y)$, with corresponding natural parameters $(\theta\eta,\theta)$.

Remark E. In what follows we will define maximum likelihood estimates using approximate densities $f_{\mu,\theta,n}$ rather than exact densities $\tilde{f}_{\mu,\theta,n}$. The differences between approximate and exact MLEs are usually negligible. In the repeated sampling situation (2.17), the difference $d\hat{\mu}$ for the estimate of μ is of order $O_p(1/n^2J)$, compared to standard deviation $O(1/\sqrt{nJ})$ for $\hat{\mu}$. Likewise $d\hat{\theta} = O_p(1/nJ)$ compared to standard deviation $O(1/\sqrt{J})$ for $\hat{\theta}$.

Remark F. If (2.1) is the inverse Gaussian distribution, Johnson and Kotz (1970) p. 138, $g_{\mu,n}(y) = (n/2\pi y^3)^{\frac{1}{2}} \exp\{-n(y-\mu)^2/2\mu^2 y\}$ for $y > 0$, it turns out that $R_{n,\theta}(y) = \theta^{-\frac{1}{2}}$, (3.5). Thus $c(\mu,\theta,n) = 1$, (3.6). The argument following (3.22) shows that $D(y,\mu) \sim \chi^2_1/\theta$ exactly in this case. The inverse Gaussian family is discussed in Jorgensen (1985), West (1985), and McCullagh and Nelder (1983).

4. The Double Poisson Family.

This section describes the double exponential family based on the Poisson distribution. The contingency table analysis in Section 7 depends on the double Poisson family.

Suppose then that $g_{\mu,n}(y)$ at (2.1) is the probability mass function

$$g_\mu(y) = e^{-\mu}\mu^y/y! \qquad [y = 0,1,\ldots] \,. \tag{4.1}$$

The sample size n has been suppressed in (4.1) for a simple reason: the Poisson family is closed under convolutions, so $g_{\mu,n}(y)$ is the same family for all values of n. [This assumes that we have rescaled y, which was previously taken to be an averag of n quantities, so that its possible values are 0, 1, 2, ..., rather than 0, 1/n, 2/n, ... This is the usual way of using the Poisson in applications. For

comparison with previous results, the reader can take the suppressed sample size to be $n = 1$.]

Definition (2.10) gives

$$f_{\mu,\theta}(y) = (\theta^{\frac{1}{2}} e^{-\theta\mu})\left(\frac{e^{-y}y^{y}}{y!}\right)\left(\frac{e\mu}{y}\right)^{y} \qquad [y = 0,1,2,\ldots] \tag{4.2}$$

as the approximate probability mass function for the Double Poisson family. The carrier G_n for family (4.2) is counting measure, i.e. $\Sigma_{y\in A} f_{\mu,\theta}(y)$ is the probability assigned to any set A of non-negative integers. Expression (4.1) has been written in the usual Poisson form rather than canonical form (2.1), but that doesn't effect (2.10).

The exact Double Poisson density (2.4) is

$$\tilde{f}_{\mu,\theta}(y) = c(\mu,\theta)\, f_{\mu,\theta}(y) \tag{4.3}$$

where the factor $c(\mu,\theta)$ can be calculated as in (3.13)-(3.14),

$$\frac{1}{c(\mu,\theta)} = \sum_{y=0}^{\infty} f_{\mu,\theta}(y) \doteq 1 + \frac{1-\theta}{12\mu\theta}\left(1 + \frac{1}{\mu\theta}\right) . \tag{4.4}$$

Table 2 gives numerical results for the case $\mu = 10$. We see that formula (4.4) is reasonably good, and more importantly that the approximate density (4.2) nearly sums to 1, even for $\theta = .333$. Fact 2,

$$E_{\mu,\theta}\{y\} \doteq \mu , \quad SD_{\mu,\theta}\{y\} \doteq [\mu/\theta]^{\frac{1}{2}} \tag{4.5}$$

is highly accurate. (The expectation and standard deviation refer to the exact density $\tilde{f}_{\mu,\theta}(y)$.)

Let $Y_{\mu,\theta}$ be a random variable with probability mass function (4.3). Fact 4 says that

$$Y_{\mu,\theta} \doteq \frac{X}{\theta} \quad \text{where} \quad X \sim \text{Po}(\mu\theta) . \tag{4.6}$$

	θ			
	1	.75	.5	.333
1. Expected Value:	10.000	9.994	9.968	9.916
2. Standard Deviation:	3.162	3.653	4.482	5.480
3. (Theoretical SD, $[10/\theta]^{\frac{1}{2}}$):	(3.162)	(3.652)	(4.472)	(5.477)
4. $\Sigma_{y=0}^{\infty} f_{\mu,\theta}(y)$:	1.000	1.003	1.012	1.026
5. (Theoretical sum (4.4)):	(1.000)	(1.003)	(1.010)	(1.022)

Table 2. Double Poisson family obtained from Poisson distribution $g_\mu(y)$, $\mu = 10$. The parameter θ has almost no effect on the expectation 10, but changes the standard deviation to approximately $[10/\theta]^{\frac{1}{2}}$.

This looks different than (2.12), but only because of the rescaling of "y" mentioned previously. For example if $\mu = 10$ and $\theta = .5$, then $Y_{10,.5}$ should be approximately distributed as $2 \cdot Po(5)$. Table 3, which uses a smooth interpolation to correct for the fact that $2 \cdot Po(5)$ takes on only even values, shows that this approximation is excellent.

Note: Stirling's formula $y! = (2\pi)^{\frac{1}{2}} y^{y+.5} e^{-y}$, substituted into (4.2) gives

$$f_{\mu,\theta}(y) \doteq \theta e^{-\mu\theta} (\mu\theta)^{\theta y} / (\theta y)! \,, \tag{4.7}$$

$(\theta y)! \equiv \Gamma(\theta y + 1)$, which is a nice restatement of (4.6).

Notice that X/θ in (4.6) has mean μ and standard deviation $[\mu/\theta]^{\frac{1}{2}}$, as in (4.5). Why not simply use the family $Po(\mu\theta)/\theta$ to do generalized Poisson regression? Because this family has a different sample space for each choice of θ, namely $\{\theta, \theta^{-1}, 2\theta^{-1}, 3\theta^{-1}, \ldots\}$, and so can't be used for a regression analysis of data taking integral values. Jorgensen (1985) comments on this fact.

The double Poisson family (4.2) is defined on $\{0,1,2,\ldots,\}$; allows us to individually adjust the mean and standard deviation of y using the parameters μ and θ; and only involves rescaled Poisson distributions, in the approximate sense of (4.6).

y:	0	1	2	3	4	5	6	7	8
[1]	0.0047	0.0090	0.0173	0.0288	0.0425	0.0570	0.0703	0.0810	0.0876
[2]	0.0034	0.0085	0.0169	0.0284	0.0421	0.0567	0.0702	0.0810	0.0878

y:	9	10	11	12	13	14	15	16	17
[1]	0.0897	0.0875	0.0815	0.0728	0.0626	0.0519	0.0417	0.0324	0.0245
[2]	0.0900	0.0878	0.0818	0.0731	0.0629	0.0522	0.0420	0.0327	0.0247

y:	18	19	20	21	22	23	24
[1]	0.0180	0.0129	0.0090	0.0061	0.0041	0.0027	0.0017
[2]	0.0182	0.0130	0.0091	0.0062	0.0041	0.0027	0.0017

Table 3. Row [1] is the Double Poisson density $\tilde{f}_{10,5}(y)$; Row [2] is the density for 2*Po(5). In order to facilitate comparison, density [2] has been divided by 2, and interpolated for odd values of y.

5. Regression Models.

We now consider the regression situation, where we observe independent double exponential family variates $y_1, y_2, \ldots, y_J$, with approximate density functions (2.10),

$$y_j \overset{ind}{\sim} f_{\mu_j,\theta_j,n_j} \qquad j = 1,2,\ldots,J \ . \tag{5.1}$$

Each y_j has its own values of the unknown mean and dispersion parameters μ_j and θ_j, which will be related by regression models. The sample sizes n_j may be different as in the Toxoplasmosis example, but are known.

The regression model for the means μ_j will be the usual general linear model, where the natural parameters η_j are assumed to equal $t_j'\alpha$ for some unknown p-dimensional parameter vector α. The p-dimensional vector t_j is an observed covariate, for example $t_j' = (1, x_j, x_j^2, x_j^3)$ for the cubic logistic regression on rainfall referred to in Section 1.

Let $T = (t_1, t_2, \ldots, t_J)$ be the $p \times J$ matrix with jth column t_j. Then the regression equation determining the J means is

$$[\eta] = T'\alpha \ , \tag{5.2}$$

where $[\eta] = (\eta_1,\eta_2,\ldots,\eta_J)'$. In McCullagh and Nelder's terminology (1983), we are using the canonical link function in (5.2), i.e. modelling $\eta(\mu)$ rather than μ itself as a linear function of the covariates.

Notation: if $v_1, v_2, \ldots, v_J$ are J quantities, then $[q(v)]$ denotes the vector $(q(v_1),\ldots,q(v_J))'$, and $\text{diag}[q(v)]$ the $J\times J$ diagonal matrix with $q(v_j)$ as the jth diagonal element; thus $[e^v] = (e^{v_1},e^{v_2},\ldots,e^{v_J})'$, etc.

We also need a regression model for the dispersion parameters θ_j. Let s_j be an H-dimensional covariate vector, which might or might not be the same as t_j. (In the examples of Sections 6 and 7 it is not the same.). We will assume a logistic-like model for the θ_j in terms of an unknown H-dimensional parameter vector β,

$$\theta_j = \frac{M}{1+e^{-\lambda_j}} \quad \text{where} \quad \lambda_j = s_j'\beta \qquad j = 1,2,\ldots,J \ . \tag{5.3}$$

The fixed constant M in (5.3), chosen by the statistician, is the maximum allowable value for any θ_j. (M is taken equal to 1.25 in our examples.) Letting S be the $H\times J$ matrix $(s_1,s_2,\ldots,s_J)$, the regression equation determining the j dispersions is -

$$[\lambda] = S'\beta \ , \tag{5.4}$$

analogous to (5.2).

There is nothing "natural" about specification (5.3). It is justified by certain practical considerations, discussed later.

Let

$$\ell_{\alpha,\beta} = \log \prod_{j=1}^{J} f_{\mu_j,\theta_j,n_j}(y_j) \tag{5.5}$$

be the approximate likelihood of the data $[y]$ as a function of the parameter vectors α and β, which determine $[\mu]$ and $[\theta]$ via (5.2) and (5.4). Also let

$$D_j = 2n_j\, I(y_j,\mu_j) \tag{5.6}$$

denote the jth deviance.

The maximum likelihood equations for α and β based on (5.5) are quite simple.

Fact 9: The score vectors $\frac{\partial \ell}{\partial \alpha} = (\frac{\partial \ell}{\partial \alpha_1}, \ldots, \frac{\partial \ell}{\partial \alpha_p})'$ and $\frac{\partial \ell}{\partial \beta} = (\frac{\partial \ell}{\partial \beta_1}, \ldots, \frac{\partial \ell}{\partial \beta_H})'$ are

$$\frac{\partial \ell}{\partial \alpha} = T \cdot \mathrm{diag}[n\theta] \cdot [y-\mu]$$

and (5.7)

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{2} S \cdot \mathrm{diag}[1-\theta/M] \cdot [1-\theta D] \quad .$$

Here η_j and λ_j are obtained from α and β by (5.2), (5.4), and then determine the values of μ_j, θ_j, and D_j appearing in (5.7). The corresponding MLE's $\hat{\alpha}$ and $\hat{\beta}$ are obtained by solving (5.7) for zero,

$$\frac{\partial \ell}{\partial \alpha} = 0 \ , \quad \frac{\partial \ell}{\partial \beta} = 0 \ . \tag{5.8}$$

Let ∇_α indicate the gradient operator $(\frac{\partial}{\partial \alpha_1}, \ldots, \frac{\partial}{\partial \alpha_p})'$, so $\nabla_\alpha \ell = \frac{\partial \ell}{\partial \alpha}$, and likewise $\nabla_\mu = (\frac{\partial}{\partial \mu_1}, \ldots, \frac{\partial}{\partial \mu_J})'$, $\nabla_\beta = (\frac{\partial}{\partial \beta_1}, \ldots, \frac{\partial}{\partial \beta_H})'$, and $\nabla_\theta = (\frac{\partial}{\partial \theta_1}, \ldots, \frac{\partial}{\partial \theta_J})'$. Remembering that $d\mu_j/d\eta_j = V(\mu_j)$, it is straightforward to show that the operators are related by

$$\nabla_\alpha = T \cdot \mathrm{diag}[V] \cdot \nabla_\mu$$

and (5.9)

$$\nabla_\beta = S \cdot \mathrm{diag}[\theta(1 - \frac{\theta}{M})] \cdot \nabla_\theta \ .$$

Without going into details, this gives (5.7) and also

Fact 10: The expected Fisher information matrix for (α,β) approximately equals

$$\mathcal{J}_{\alpha,\beta} \doteq \begin{pmatrix} T \cdot \mathrm{diag}[n\theta V] \cdot T' & 0 \\ 0 & \frac{1}{2} S \cdot \mathrm{diag}[1 - \frac{\theta}{M}]^2 \cdot S' \end{pmatrix} . \tag{5.10}$$

Formula (5.10) agrees with Fact 6 in the repeated sampling situation (2.17).

The method of scoring can be used to find the MLE's $\hat{\alpha}$ and $\hat{\beta}$. Suppose (α^o,β^o) are initial estimates of (α,β). The improved estimate $(\alpha^o+d\alpha,\ \beta^o+d\beta)$ has

$$d\alpha = \{\mathcal{J}_{\alpha^o}\}^{-1} \frac{\partial \ell}{\partial \alpha^o}$$

and (5.11)

$$d\beta = \{\mathcal{J}_{\beta^o}\}^{-1} \frac{\partial \ell}{\partial \beta^o} .$$

Here $\frac{\partial \ell}{\partial \alpha^o}$ and $\frac{\partial \ell}{\partial \beta^o}$ are the score functions (5.7) evaluated at (α^o,β^o) ; $\mathcal{J}_{\alpha^o}$ is the upper block $T\cdot\mathrm{diag}[n\theta V]\cdot T'$ in (5.10) evaluated at (α^o,β^o) , and likewise $\mathcal{J}_{\beta^o}$

The motivation behind the awkward-looking parameterization (5.3) is simple: we want the θ_j to be positive, and we do not want them to get too large. Values of $\theta_j > 1$ correspond to underdispersion, for instance in logistic regression to proportions y_j with variance less than $\mu_j(1-\mu_j)/n_j$. Technically there is nothing to stop us from using such values, and in fact (2.1) defines a genuine density for $\theta \in (0,\infty)$ in all of common cases, normal, binomial, Poisson, etc. However there are often good physical reasons for not believing in underdispersion, especially in binomial and Poisson situations.

Formula (5.3) allows us to set a maximum value M for θ_j. A value slightly greater than one, $M = 1.25$, was used in the examples of Sections 6 and 7 to avoid having $\theta_j = 1$ be on the boundary of the allowable parameter space. In neither example did the choice of M much effect the fitted regression for the mean, (5.2); in the second example it had a mild effect on the dispersion regression (5.4), as will be mentioned.

6. The Toxoplasmosis Data.

The Toxoplasmosis data of Table 1 was analyzed using binomial double exponential families. The response variables y_j , observed proportion of subjects testing

positive for Toxoplasmosis in city j, were assumed to have independent binomial double exponential distributions as in (5.1). That is, $g_{\mu,n}(y)$ in (2.10) was the rescaled binomial density (2.3).

In the original analysis of this data, Efron (1978) fit an ordinary logistic regression predicting y_j in terms of a cubic function of the annual rainfall x_j for city j. Let X_j be the standardized value of the rainfall, $X_j = (x_j - \bar{x})/\{\Sigma_{j=1}^{34}(x_j-\bar{x})^2/33\}^{\frac{1}{2}}$. Then the natural parameter $\eta_j = \log \mu_j/(1-\mu_j)$ for city j was modeled as

$$\eta_j = \alpha_0 + \alpha_1 X_j + \alpha_2 X_j^2 + \alpha_3 X_j^3 \ . \tag{6.1}$$

This same specification was used for the double logistic analysis. In the notation of Section 5, $t_j' = (1, X_j, X_j^2, X_j^3)$, $p = 4$, and (6.1) was the jth row of (5.2).

The dispersion parameters θ_j were modeled as in (5.3), (5.4) with $M = 1.25$. Let N_j be the standardized value of the sample size n_j for city j, $N_j = (n_j - \bar{n})/\{\Sigma(n_j-\bar{n})^2/33\}^{\frac{1}{2}}$. Then $s_j' = (1, N_j, N_j^2)$, $H = 3$, and the jth row of (5.4) was

$$\lambda_j = \beta_0 + \beta_1 N_j + \beta_2 N_j^2 \ , \tag{6.2}$$

a quadratic regression on sample size.

The results of the analysis are shown in Figure 1. The regression of mean response μ_j on rainfall differs moderately from the ordinary fit, especially near $x_j = 2100$ mm. The regression of the dispersion parameter θ_j on n_j is interesting. Cities with n_j near 30 were assigned $\hat{\theta}_j \doteq .80$, so that their effective sample size $n_j\hat{\theta}_j$ was not much less than the actual size n_j. Cities with n_j much smaller than 30 or, in particular, much larger than 30 were strongly downweighted. The full results appear in Table 4.

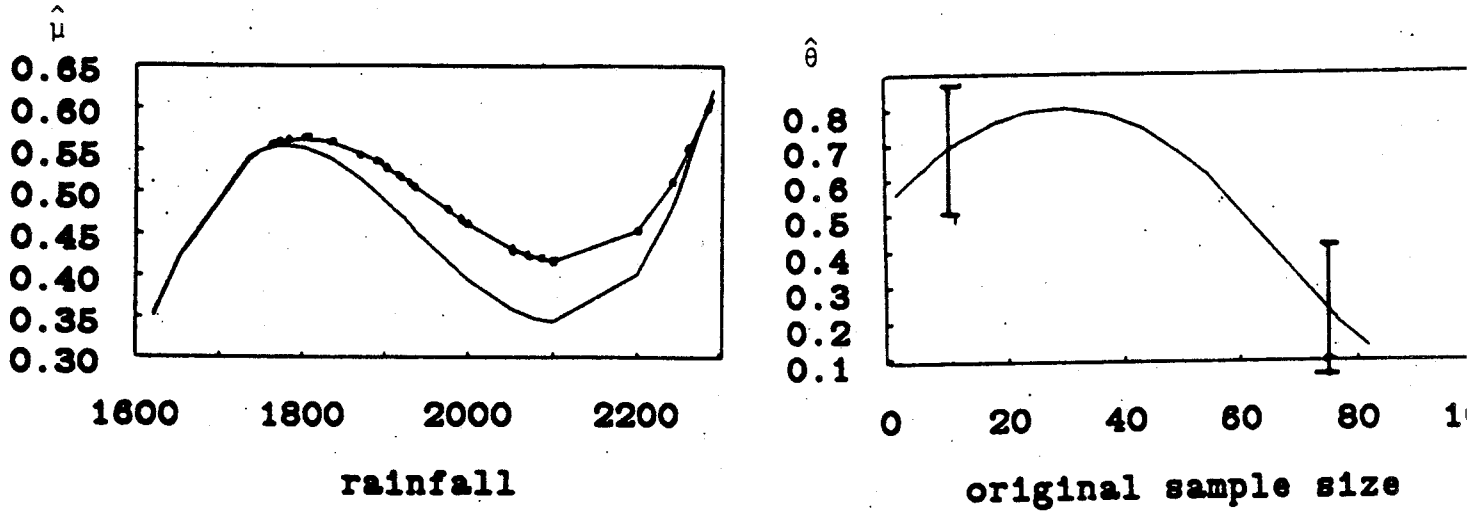

Figure 1. Results of the double logistic fit (6.1), (6.2) to the Toxoplasmosis data. The left panel compares the fitted regression of mean response μ as a function of rainfall with the corresponding regression for the ordinary logistic analysis (dotted curve). The right panel shows the fitted dispersion parameter $\hat{\theta}_j$ as a function of sample size n_j. Cities with large n_j were assigned small values of $\hat{\theta}_j$. Estimated standard errors of $\hat{\theta}_j$ are indicated for cities 4 and 30.

The absolute value of the deviance residual from an ordinary logistic regression (or any general linear model) is defined to be

$$\hat{R}_j \equiv \{D(y_j,\hat{\mu}_j)\}^{\frac{1}{2}} = \{2n_j\, I(y_j,\hat{\mu}_j)\}^{\frac{1}{2}}\,, \qquad (6.3)$$

as on page 30 of McCullagh and Nelder (1983). If the regression model is correct then the $\hat{R}_j$'s should have an approximately half-normal distribution. Figure 2 compares the absolute deviance residuals for $\hat{\mu}_j$ as obtained from the double fit to those obtained from the ordinary logistic regression. The residuals are shown individually in the last two columns of Table 4. (In order to compare the magnitude of the residuals from the two fits, $\hat{R}_j$ for the double fit is not defined here as $\{2n_j\hat{\theta}_j\, I(y_j,\hat{\mu}_j)\}^{\frac{1}{2}} = \hat{\theta}_j^{\frac{1}{2}}\, D(y_j,\hat{\mu}_j)^{\frac{1}{2}}$; we will use this definition in Section 7.)

In 24 of the 34 cities, $\hat{R}_j(\text{double}) \leq \hat{R}_j(\text{ordinary})$, but in a couple of cities, particularly city 27 which has the largest sample size, $\hat{R}_j(\text{double})$ is much bigger

j	n_j	$\hat{\theta}_j$	y_j	FITTED MEANS $\hat{\mu}_j$		RESIDUALS $\hat{R}_j$	
				double	ordinary	double	ordinary
[1]	4	0.607	0.500	0.541	0.539	0.16	0.16
[2]	10	0.691	0.300	0.455	0.506	1.00	1.32
[3]	5	0.623	0.200	0.394	0.461	0.93	1.22
[4]	10	0.691	0.300	0.419	0.480	0.77	1.16
[5]	2	0.574	1.000	0.549	0.549	1.55	1.55
[6]	5	0.623	0.600	0.553	0.563	0.21	0.17
[7]	8	0.666	0.250	0.549	0.549	1.72	1.72
[8]	19	0.772	0.368	0.348	0.422	0.19	0.48
[9]	6	0.638	0.500	0.471	0.517	0.14	0.08
[10]	10	0.691	0.800	0.553	0.563	1.64	1.58
[11]	24	0.795	0.292	0.359	0.432	0.69	1.42
[12]	1	0.557	0.0	0.542	0.560	1.25	1.28
[13]	30	0.804	0.500	0.422	0.421	0.86	0.87
[14]	22	0.788	0.182	0.401	0.454	2.20	2.69
[15]	1	0.557	0.0	0.394	0.461	1.00	1.11
[16]	11	0.703	0.545	0.555	0.558	0.06	0.09
[17]	1	0.557	0.0	0.471	0.517	1.13	1.21
[18]	54	0.621	0.611	0.555	0.558	0.84	0.79
[19]	9	0.679	0.444	0.474	0.506	0.17	0.37
[20]	18	0.766	0.278	0.349	0.353	0.64	0.68
[21]	12	0.714	0.167	0.551	0.552	2.75	2.76
[22]	1	0.557	0.0	0.422	0.421	1.05	1.05
[23]	11	0.703	0.727	0.498	0.522	1.55	1.39
[24]	77	0.199	0.532	0.554	0.563	0.39	0.55
[25]	51	0.664	0.471	0.499	0.536	0.40	0.92
[26]	16	0.751	0.438	0.515	0.546	0.62	0.86
[27]	82	0.129	0.561	0.353	0.427	3.84	2.44
[28]	13	0.724	0.692	0.343	0.417	2.56	2.00
[29]	43	0.750	0.535	0.473	0.518	0.82	0.22
[30]	75	0.232	0.707	0.540	0.559	2.96	2.63
[31]	13	0.724	0.615	0.555	0.561	0.44	0.40
[32]	10	0.691	0.300	0.490	0.530	1.22	1.47
[33]	6	0.638	0.167	0.416	0.477	1.31	1.60
[34]	37	0.788	0.622	0.622	0.611	0.01	0.14

Table 4. Double logistic fit to Toxoplasmosis data, compared to ordinary logistic fit.

than $\hat{R}_j$(ordinary). The double fitting process essentially gave up trying to fit those cities in order to reduce the majority of the residuals. Whether or not the double fit is better is certainly debatable, but it is clear that the double family method can potentially robustify a standard analysis. Double family regression is not automatically preferrable to the ordinary method; it can be a useful supplement to reveal possible weaknesses and limitations of ordinary analyses.

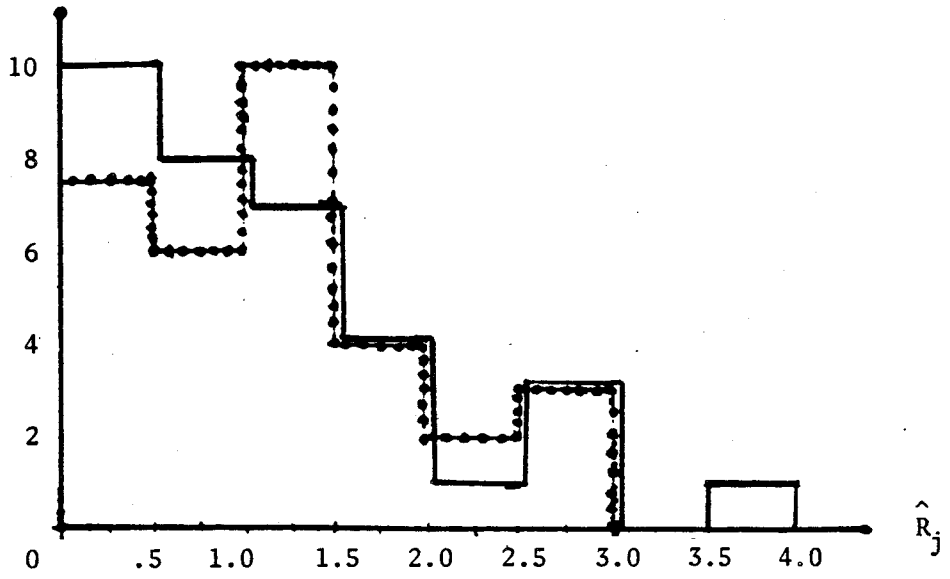

Figure 2. Histogram of deviance residuals (6.3) for double fit (solid line) compared to ordinary fit (dotted line).

Remark G. We must have $\Sigma\hat{R}_j^2(\text{ordinary}) \leq \Sigma\hat{R}_j^2(\text{double})$, because

$$\Sigma\hat{R}_j^2 = \Sigma 2n_j\, I(y_j,\hat{\mu}_j) = 2 \log \prod_i \frac{g_{y_j,n_j}(y_j)}{g_{\hat{\mu}_j,n_j}(y_j)}, \tag{6.4}$$

twice the log likelihood ratio, and the ordinary fitting method amounts to minimizing this ratio. In this sense Figure 2 is generic; it may be possible to reduce a majority of the residuals by another fitting technique, but some cases must then be fitted considerably less well.

Remark H. The Toxoplasmosis data compares $n_j = 697$ cases. An ordinary logistic regression would produce the same answers working directly from the 697 Bernoulli variates, and their associated rainfalls, as with the data grouped by cities. This isn't true of the double logistic analysis, even aside from the fact that in this example the n_j were used as covariates. The problem here is partly technical, having to do with the breakdown of Section 2's various approximations in the $n = 1$

binomial situation: for example this situation results in $\hat{\mu} = \bar{y}$ and $\hat{\theta} = .5 \cdot [\hat{\mu} \log \hat{\mu} + (1-\hat{\mu}) \log(1-\hat{\mu})]^{-1}$ at (2.20), so that $\hat{\theta}$ is a function of $\hat{\mu}$ rather than being statistically independent.

Remark I. Here are the MLE's for the vector α from the two fits, with estimated standard errors in brackets,

$$\begin{aligned} \hat{\alpha}(\text{Double}) &= (-.071[.14],\ -.620[.23],\ -.170[.11],\ .272[.09]) \\ \hat{\alpha}(\text{Ordinary}) &= (.099[.10],\ -.448[.16],\ -.187[.09],\ .213[.06]) \ . \end{aligned} \tag{6.5}$$

The standard errors for $\hat{\alpha}$(Double) are larger not because we have estimated more parameters (the independence between the α and β blocks of $\dot{\ell}_{\alpha,\beta}$ in (5.10) eliminates this kind of increase) but rather because the $\hat{\theta}_j$ are less than one. The double model says that the y_j have approximate variance about $\mu_j(1-\mu_j)/n_j\theta_j$ rather than the ordinary variance $\mu_j(1-\mu_j)/n_j$, and so $\hat{\alpha}$ has a larger estimated dispersion matrix.

Which standard errors are correct? This depends on whether the double family is thought of as a convenience for robustly fitting situations like the Toxoplasmosis example where there are obvious inadequacies in the original regression model; or whether the reduced sample size $n_j\theta_j$ is thought to have a genuine physical basis, for example due to imperfect random sampling of the subjects. In the former case, the ordinary standard errors are preferrably even if we prefer $\hat{\alpha}$(Double) as a point estimate.

Remark J. The MLE and standard errors for β in the double fit were

$$\hat{\beta} = (.509[.83],\ .405[1.01],\ -.515[.52]) \ . \tag{6.6}$$

The standard errors, computed from (5.10), are enormous, but result in less enormous standard errors for the individual $\hat{\theta}_j$, as shown on the right panel of Figure 1. The shape of the regression curve for $\hat{\theta}_j$ shown on the right panel of Figure 1 is

highly uncertain. In general it seems more difficult to estimate the θ_j than the μ_j, so regression model (5.4) should be kept simple. Notice that the accuracy of the estimate $\hat{\theta}_j$ depends on the magnitude of J, not of the n_j, as seen for instance in (2.21).

7. Analysis of a Two-Way Table.

The 19×5 two-way contingency table, Table 5, is taken from Mosteller and Parunak (1985). They consider the question of which entries are extreme, in the sense of deviating from the hypothesis of independence. We will consider the same question using an analysis based on the double Poisson family.

		Distance in Miles				
	(0)	(0,.25)	(.25,.5)	(.5,1)	(>1)	
j:	1	2	3	4	5	total
i:						
[1,]	20	102	54	38	32	246
[2,]	33	136	86	58	63	376
[3,]	27	122	68	51	53	321
[4,]	2	10	8	5	4	29
[5,]	11	82	34	35	32	194
[6,]	10	53	25	17	20	125
[7,]	39	185	88	100	71	483
[8,]	34	179	70	78	71	432
[9,]	26	78	24	26	20	174
[10,]	24	88	32	41	29	214
[11,]	8	44	16	28	42	138
[12,]	15	75	30	35	35	190
[13,]	11	32	5	11	23	82
[14,]	12	28	5	18	7	70
[15,]	2	10	4	2	6	24
[16,]	3	8	4	6	8	29
[17,]	1	2	0	3	3	9
[18,]	13	5	3	9	7	37
[19,]	20	36	19	20	29	124
Total	311	1275	575	581	555	3297

Table 5. Archealogical contingency table, 19 types of artifacts, 5 distances from permanent water site. Which cells are extreme, in the sense of deviating from the hypothesis of independence? (Condensed from a 19×6 table in Mosteller and Parunak (1985, pg. 190), based on data collected by Laurel Casjens.)

We suppose that y_{ij}, the number of counts observed in row i and column j of Table 5, follows a double Poisson distribution (4.2),

$$y_{ij} \overset{\text{ind}}{\sim} \text{Po}(\mu_{ij},\theta_{ij}) \ , \tag{7.1}$$

where

$$\log \mu_{ij} \equiv \eta_{ij} = \rho_i + \gamma_j \ , \tag{7.2}$$

the sum of a row and a column effect.

If we replace (7.1) by $y_{ij} \overset{\text{ind}}{\sim} \text{Po}(\mu_{ij})$ (that is, take $\theta_{ij} \equiv 1$), then (7.2) is the usual hypothesis of independence for a two-way table. Let r_i equal the proportion of the observations in row i, and likewise c_j for the proportion in column j. Then the MLE of μ_{ij} under the usual hypothesis of independence is

$$\hat{\mu}^0_{ij} = n_+ \cdot r_i c_j \qquad (n_+ = \Sigma_i \Sigma_j \, y_{ij} = 3297) \ . \tag{7.3}$$

The θ_{ij} were modeled in form (5.3), with $M = 1.25$. Regression (5.4) was taken quadratic in $\hat{\mu}^o_{ij}$,

$$\lambda_{ij} = \beta_2 + \beta_1 \hat{\mu}^o_{ij} + \beta_2 (\hat{\mu}^o_{ij})^2 \ . \tag{7.4}$$

The ijth column of the covariate matrix S is $(1, \hat{\mu}^o_{ij}, (\hat{\mu}^o_{ij})^2)'$. The fact that this depends on the data is briefly discussed below.

The results of the double-family fit are easy to summarize:

- The fitted value of $\hat{\mu}_{ij}$ hardly deviated from $\hat{\mu}^o_{ij}$, (7.3).
- The fitted values of $\hat{\theta}_{ij}$ ranged from .51 to 1.25, with many more cases near the low end. The fitted values are tabulated in Table 6, and graphed versus $\hat{\mu}^o_{ij}$ in Figure 3. According to this graph, cells with low fitted values of $\hat{\mu}^o_{ij}$ are overdispersed compared to standard Poisson variability, while those with high $\hat{\mu}^o_{ij}$ enjoy underdispersion.

Now to answer the question of identifying extreme cells, posed by Mosteller and Parunak. The scaled deviance residual

$$r_{ij} \equiv \{\hat{\theta}_{ij}\, D(y_{ij},\hat{\mu}^{o}_{ij})\}^{\frac{1}{2}} \tag{7.5}$$

should have an approximate half-normal distribution according to (2.15). The right panel of Figure 3, which is the histogram of the 95 r_{ij} values, clearly indicates two outliers. Table 7 shows that these are cells (11,5) and (18,1). No other cell is obviously deviant, the nearest being cell (2,3) with the insignificant value $r_{ij} = 2.28$.

In summary, Table 5 shows a general background level of deviations from independence, concentrated in cells with low values of $\hat{\mu}^{o}_{ij}$, as seen in Figure 3 and Table 6; two of the cells are more extremely deviant, even after this background effect is taken into account. These conclusions agree with those of Mosteller and Parunak, reached by different arguments.

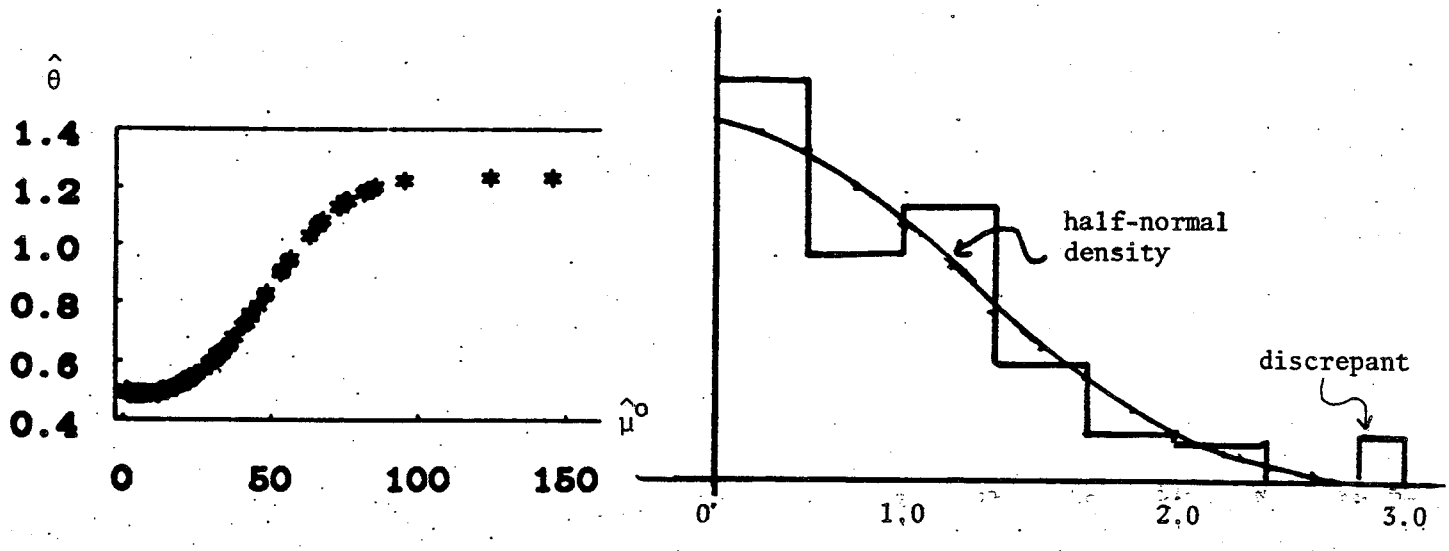

Figure 3. The left panel plots dispersion parameter $\hat{\theta}_{ij}$ versus $\hat{\mu}^{o}_{ij}$, the usual estimate of cell mean assuming independence. Cells with small $\hat{\mu}^{o}_{ij}$ have small predicted $\hat{\theta}_{ij}$, indicating overdispersion compared to standard Poisson variability. The right panel is a histogram of the scaled residuals $\{\hat{\theta}_{ij}\, D(y_{ij},\hat{\mu}_{ij})\}^{\frac{1}{2}}$, compared to a standard half-normal density. Two of the scaled residuals are obviously discrepant.

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.56	1.24	0.77	0.77	0.75
[2,]	0.67	1.25	1.08	1.09	1.05
[3,]	0.62	1.25	0.96	0.97	0.93
[4,]	0.51	0.51	0.51	0.51	0.51
[5,]	0.54	1.17	0.66	0.66	0.64
[6,]	0.51	0.85	0.55	0.56	0.55
[7,]	0.81	1.25	1.21	1.22	1.20
[8,]	0.74	1.25	1.17	1.17	1.15
[9,]	0.53	1.10	0.62	0.62	0.61
[10,]	0.54	1.21	0.70	0.70	0.68
[11,]	0.52	0.92	0.57	0.57	0.56
[12,]	0.53	1.15	0.65	0.65	0.64
[13,]	0.51	0.63	0.52	0.52	0.52
[14,]	0.51	0.59	0.52	0.52	0.51
[15,]	0.51	0.51	0.51	0.51	0.51
[16,]	0.51	0.51	0.51	0.51	0.51
[17,]	0.52	0.51	0.51	0.51	0.51
[18,]	0.51	0.52	0.51	0.51	0.51
[19,]	0.51	0.84	0.55	0.55	0.55

Table 6. Fitted values of $\hat{\theta}_{ij}$ from the double Poisson analysis.

Remark K. Regression (7.4) uses the observed values $\hat{\mu}^o_{ij}$ as if they were fixed covariates. Technically we should have modified the fitting equations (5.7), but this wasn't done. In general, the block diagonal nature of (5.10) makes data-dependent specifications like (7.4) not overly worrisome. In specific, we might think of this analysis as conditional on the marginal totals of Table 5, in which case (7.4) is a legitimate regression model.

Remark L. A more serious concern about our analysis involves the large values of $\hat{\theta}_{ij}$, near the maximum M, connected with large values of $\hat{\mu}^o_{ij}$. Isn't this a sign of overfitting, reflecting the influence of a large value of y_{ij} in Table 5 on its own fitted value $\hat{\mu}^o_{ij}$?

It turns out that we can correct for this effect rather easily. Without going into the motivation, which is based on analogy with ordinary least squares residual analysis, here is the corrected procedure: (a) for each (i,j), construct a new

	[,1]	[,2]	[,3]	[,4]	[,5]
1,]	0.47	0.55	1.16	0.83	1.26
2,]	0.20	0.91	2.28	1.09	0.22
3,]	0.36	0.27	1.30	0.76	0.07
4,]	0.30	0.27	0.81	0.04	0.25
5,]	1.35	0.57	0.24	0.02	0.08
6,]	0.36	0.47	0.34	0.90	0.13
7,]	0.77	0.28	0.12	1.67	1.04
8,]	0.83	0.86	1.00	0.14	0.02
9,]	1.58	1.04	1.18	0.82	1.43
10,]	0.66	0.47	0.94	0.38	0.92
11,]	0.94	1.06	1.31	0.67	2.87
12,]	0.45	0.10	0.60	0.17	0.53
13,]	0.84	0.02	2.12	0.69	1.72
14,]	1.40	0.12	1.75	1.07	1.01
15,]	0.10	0.16	0.11	0.86	0.69
16,]	0.14	0.73	0.39	0.27	0.97
17,]	0.13	0.62	1.29	0.71	0.79
18,]	2.84	2.06	1.12	0.66	0.28
19,]	1.77	1.39	0.36	0.14	1.52

Table 7. The scaled deviance residuals r_{ij}, (7.5). The two outliers seen in the right panel of Figure 3 are cells (11,5) and (18,1).

table by replacing the actual entry y_{ij} with $\hat{\mu}^o_{ij}$, keeping the other entries fixed. (b) Recompute (7.3) for this new table, obtaining say $\hat{\mu}^o_{ij(ij)}$. (c) Compute $D(y_{ij},\hat{\mu}_{ij(ij)})$, the Poisson deviance of the observation y_{ij} from $\hat{\mu}^o_{ij(ij)}$, and then

$$D^\dagger_{ij} \equiv \{D(y_{ij},\hat{\mu}^o_{ij})\, D(y_{ij},\hat{\mu}^o_{ij(ij)})\}^{\frac{1}{2}} . \tag{7.6}$$

(e) Substitute $1-\theta D^\dagger$ for $1-\theta D$ in (5.7), and solve $\dot{\ell}/\partial\beta = 0$ obtaining $\hat{\theta}^\dagger$. (f) Finally, compute $r^\dagger_{ij} = \{\hat{\theta}^\dagger_{ij} D^\dagger_{ij}\}^{\frac{1}{2}}$.

The dagger notation is taken from Stone's (1974) paper on cross-validation. Table 8 shows the cross-validated dispersion estimate $\hat{\theta}^\dagger_{ij}$, and the corresponding scaled residuals $r^\dagger_{ij}$. Comparison with Table 6 shows that the $\hat{\theta}^\dagger_{ij}$ are 10-20% smaller than the $\hat{\theta}_{ij}$, and in particular there are only five $\hat{\theta}^\dagger_{ij} \geq 1$, compared

	[,1]	[,2]	[,3]	[,4]	[,5]	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.47	1.05	0.57	0.57	0.56	0.50	0.86	1.38	0.69	1.26
[2,]	0.53	1.24	0.76	0.77	0.74	0.33	1.06	2.39	1.02	0.03
[3,]	0.50	1.20	0.67	0.68	0.66	0.47	0.25	1.43	0.69	0.13
[4,]	0.43	0.44	0.43	0.43	0.43	0.32	0.29	0.87	0.03	0.29
[5,]	0.46	0.86	0.52	0.52	0.51	1.33	0.88	0.02	0.11	0.09
[6,]	0.44	0.61	0.47	0.47	0.46	0.38	0.62	0.51	0.84	0.17
[7,]	0.59	1.25	0.95	0.96	0.92	0.84	0.18	0.45	1.76	1.27
[8,]	0.56	1.25	0.86	0.87	0.83	0.89	1.24	0.65	0.22	0.21
[9,]	0.45	0.78	0.50	0.50	0.50	1.58	1.35	0.93	0.67	1.40
[10,]	0.46	0.93	0.54	0.54	0.53	0.60	0.66	0.72	0.43	0.97
[11,]	0.44	0.65	0.47	0.48	0.47	1.05	1.25	1.30	0.56	2.69
[12,]	0.45	0.84	0.51	0.52	0.51	0.51	0.19	0.44	0.21	0.41
[13,]	0.43	0.51	0.45	0.45	0.44	0.77	0.04	2.04	0.69	1.66
[14,]	0.43	0.49	0.44	0.44	0.44	1.32	0.15	1.66	1.11	1.07
[15,]	0.43	0.44	0.43	0.43	0.43	0.12	0.19	0.07	0.85	0.66
[16,]	0.43	0.44	0.43	0.43	0.43	0.11	0.78	0.35	0.28	0.93
[17,]	0.43	0.43	0.43	0.43	0.43	0.11	0.65	1.21	0.73	0.77
[18,]	0.43	0.45	0.43	0.43	0.43	2.79	2.15	1.07	0.66	0.22
[19,]	0.44	0.61	0.47	0.47	0.46	1.57	1.65	0.43	0.30	1.26

Table 8. Cross-validated values $\hat{\theta}_{ij}$ (left side) and scaled residuals $r^{\dagger}_{ij}$ (right side).

with 18 such $\hat{\theta}_{ij}$. Cells (11,5) and (18,1) are still noticeably most deviant, though less so than in Table 7. Finally cell (2,3) is now more deviant, $r^{\dagger}_{2,3} = 2.39$, though still not significantly so. In short, the results confirm our less careful previous analysis.

References

Barndorff-Nielsen, O. and Cox, D. R. (1984). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. JRSS B 46, 483-495.

Diaconis, P. and Efron, B. (1985). Testing the independence of a two-way table: new interpretations of the chi-square statistic. To appear in Ann. Stat.

Efron, B. (1978). Regression and ANOVA with zero-one data: measures of residual variation. JASA 73, 113-121.

Efron, B. (1978A). The geometry of exponential families. Ann. Stat. 6, 362-376.

Johnson, N. and Kotz, S. (1970). Continuous Univariate Distributions-2. Houghton-Mifflin, Boston.

Jorgensen, B. (1985). Exponential dispersion models. To appear, JRSS-B.

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

Morris, C. (1982). Natural exponential families with quadratic variance functions. Ann. Stat. 10, 65-80.

Mosteller, F. and Parunak, A. (1985). Identifying Extreme Cells In a Sizeable Contingency Table: Probabilities and Exploratory Approaches", pp. 189-224, Exploring Data Tables, Trends, and Shapes. Wiley, New York.

Nelder, J. and Pregibon, D. (1983). Quasilikelihood Models and Data Analysis. Bell Laboratories Technical Memorandum.

Pregibon, D. (1984). Review of Generalized Linear Models. Ann. Stat. 12, 1589-1596.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. JRSS B 36, 111-147.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika 61, 439-447.

West, M. (1985). Generalized linear models: scale parameters, outlier accommodation and prior distributions. Bayesian Statistics 2. North-Holland, Amsterdam.