A Bayesian regression approach to the poisson distribution

A Bayesian regression approach to the poisson distribution

Michiko Hiradate

I Count data

Regression analysis of the poisson distributed data has been discussed by several authors ([3],[4],[7]). But Nelder and Wedderburn formulated the generalized linear model in 1972 ([5]) and since then we can view this problem in a wider scope including the usual normal regression. The poisson is, theoretically and practically, the most important distribution, when we analyse the count data. But we often observe covariables, or factors, together with response variable in economic or sociological studies, or in other fields. In these cases, multiple regression analysis provides an interesting technique to interprete data and obtain information about an underlying mechanism. However, we somtimes obtain the poor fit in regression analysis, especially in analysing the count data. This may be because the count observations are aggregated over some factors, or important explanatory variables are not observed or unavailable. As a result, overdiversification is suspected.

Here we examine a method to accomodate such unexplained variations due to misspecification by introducing normally random part into the linear regression model for the poisson distribution.

II Model and a method of estimation

Nelder and Wedderburn propose the generalized linear model which relates the natural parameter or some function of it of exponential family distributions to explanatory variables ([5]). Likelihood function and linear regression part are given by

$$f(y \mid \theta, \phi) = \exp\{\phi[\theta y - a(\theta)]\} b(\phi, y) \quad \cdots\cdots (1)$$

and

$$\eta(\theta) = h' \beta \quad \cdots\cdots (2)$$

where θ is the natural parameter, ϕ is the scale parameter, h is a $p \times 1$ vector of explanatory variables, and β is a $p \times 1$ vector of regression coefficients including constant term. When y follows the standard poisson distribution, $\phi = 1$, $a(\theta) = e^{\theta}$. Continuous distributions of exponential family are as flexible as to account for the variation in the data by giving a variety of values to the scale parameter. However, discrete distributions have fixed value of ϕ and are not so flexible. West ([7]) discusses this problem in detail from the Bayesian point of view, and proposes scaled exponential family likelihood as an approximate sampling model which keeps the scale parameter free and uses the deviance function. The scaled exponential family likelihood has the same mean/variance relationship that of original formulation, and reduces to the original N-W model when $\phi = 1$. If we keep ϕ free and try to learn its behavior, some approximation is necessary, because factor $b(\phi, y)$ is usually difficult to be incorporated in the analysis of discrete data. Sweeting ([6]) discusses this problem in more general context with scale parameter having improper prior. Our formulation introduces normally random part into the regression relation (2), for the purpose of finding the sources of such variations and interpreting data in a more appropriate model.

Suppose we observe n samples of y, and let $y = (y_1 \ldots\ldots, y_n)$. We suppose the components of y follow independently the exponential family distributions of the same type. Then, likelihood function is given by

$$p(y \mid \boldsymbol{\theta}, \phi) = \prod_{i=1}^{n} \left\{ \exp \left[\phi(\theta_i y_i - a(\theta_i)) \right] b(y_i, \phi) \right\} \cdots\cdots\cdots (3)$$

where $\boldsymbol{\theta}' = (\theta_1, \cdots\cdots, \theta_n)$ is a vector of the natural parameters. We suppose the following linear relation instead of (2).

$$\eta(\theta_i) = h_i' \beta + u_i \cdots\cdots\cdots (4)$$

The added term u accomodates errors due to the misspecification or the failure of observing some important factors. In the case of the poisson of parameter λ_i , relation (4) may reduce to

$$\theta_i = \log_e \lambda_i = h_i' \beta + u_i \cdots\cdots\cdots (5)$$

Let **u** be a vector of u term. As for the prior distributions, we have no conjugate analysis here except the simplest case, and we assume β and **u** are independent and normally distributed, and also **u** have zero means. Then joint prior distribution is as follows

$$p(\beta, \mathbf{u} \mid b, B, V) = p(\beta \mid \mathbf{b}, B)\, p(\mathbf{u} \mid V)$$

$$\propto \exp\left\{ -\frac{1}{2}(\beta - \mathbf{b})' B (\beta - \mathbf{b}) \right\} \cdot \exp\left\{ -\frac{1}{2} \mathbf{u}' V \mathbf{u} \right\} \cdots\cdots (6)$$

where **b** is a vector of prior means of β , B and V are precision matrices of β and **u** respectively. It is not practical to assume that precision matrix B and V are known completely. An unknown scale factor is assumed to exist in each matrix and it has some hyper prior distribution. Namely, we set $B = \gamma B_0$, $V = \nu V_0$, where B_0, V_0 are specified. Unknown factors γ, ν follow chi-square distributions for some f_0, d_0, k_0, g_0, that is, $f_0 \gamma \sim \chi^2_{d_0}$, $k_0 \nu \sim \chi^2_{g_0}$. Then integrating (6) with respect p (γ), p (ν) gives

$$p(\beta, \mathbf{u} \mid \mathbf{b}, f_0, d_0, k_0, g_0)$$

$$\propto [f_0 + (\beta - \mathbf{b})' B_0 (\beta - \mathbf{b})]^{-\frac{d_0+1}{2}} \cdot [k_0 + \mathbf{u}' V_0 \mathbf{u}]^{-\frac{g_0+1}{2}} \cdots\cdots\cdots (7)$$

The kernel of posterior distribution is given by making product of (3) and (7), and incorporating the relation (4). Posterior means of parameters are not obtaind analytically in this formulation, so we take the posterior modes as convenient estimates. Let $\Psi'=(\beta', \mathbf{u}')$, then the posterior score function and information function are defined respectively by

$$g(\Psi \mid y, \phi)=\frac{d}{d\Psi}\log p(\Psi \mid y, \phi) \qquad \cdots\cdots (8)$$

and

$$G(\Psi \mid y, \phi)=-\frac{d}{d\Psi'}g(\Psi \mid y, \phi) \qquad \cdots\cdots (9)$$

In the case of relation (5), they are

$$g(\Psi \mid y, \phi)=\begin{cases}\phi H\mathbf{z}-\bar{\gamma}(\Psi)B_0(\beta-\mathbf{b})\\ \mathbf{z}-\bar{\nu}(\Psi)V_0\mathbf{u}\end{cases}$$

where H is p×n matrix with column h_i,

$\mathbf{z}$ is n×1 vecter with elements $(y_i-a'(h_i'\beta+u_i))$, $i=1.....n$,

$$\bar{\gamma}(\Psi)=\frac{d_0+p}{f_0+(\beta-\mathbf{b})'B_0(\beta-\mathbf{b})}=E(\gamma \mid y, \Psi) \qquad \cdots\cdots (11)$$

$$\bar{\nu}(\Psi)=\frac{g_0+n}{k_0+\mathbf{u}'V_0\ \mathbf{u}}=E(\nu \mid y, \Psi) \qquad \cdots\cdots (12)$$

$G(\Psi \mid y, \phi)$

$$=\begin{pmatrix}\phi HA(\Psi)H'+\bar{\gamma}(\Psi)B_0-2D_1 & HA'(\Psi)\\ A(\Psi)H' & \phi A(\Psi)+\bar{\nu}(\Psi)V_0-2D_2\end{pmatrix} \qquad \cdots\cdots (13)$$

where $A(\Psi)$ is n×n matrix with diagonal elements $a''(h_i'\beta+u_i)$, and

$$D_1=\frac{\bar{\gamma}^2(\Psi)}{d_0+p}B_0(\mathbf{p}-\mathbf{b})(\mathbf{p}-\mathbf{b})'B_0'$$

$$D_2=\frac{\bar{\nu}^2(\Psi)}{g_0+n}V_0\ \mathbf{u}\,\mathbf{u}'\ V_0'.$$

The posterior modes are obtained iteratively by usual Newton- Raphson method. Let Ψ_i be the value of Ψ of i'th iteration, and the iteration equation is as follows

$$\Psi_{i+1}=\Psi_i+G(\Psi_i \mid y, \phi)^{-1}g(\Psi_i \mid y, \phi) \qquad \cdots\cdots (10)$$

III Numerical examples

1. A modified formulation

We show two numerical examples in this section. For convenience, we modify the formulation. Assume we have m groups of observations on the poisson variable y and i'th group consists of r_i observations, and we have p explanatory variables with observations $h'_i = (h_{1i}, \cdots\cdots\cdots, h_{pi})$, where $h_{1i} = 1$ for all i, i=1, ..., m. Then we have m poisson parameters λ_i, i=1,....m and $\log_e \lambda_i = \theta_i$. θ_i is expressed by regression model as (5).

As for priors, we suppose in section 2 β and **u** have independent normal distributions, but between components of β and **u** , there may be correlations respectively. Here we simply assume that components of β and **u** distribute independently each other with different scale parameters γ_k, ν_i. Then, the prior distribution of β and **u** is

$$p(\beta, \mathbf{u} \mid \gamma, \nu, \mathbf{b}) \propto \prod_{k=1}^{p} \gamma_k^{\frac{1}{2}} \exp\left[-\frac{\gamma_k(\beta_k - b_k)^2}{2} \right] \cdot \prod_{i=1}^{m} \nu_i^{\frac{1}{2}} \exp\left[-\frac{\nu_i u_i^2}{2} \right].$$

As hyper priors of ν_i , γ_k , we assess that $f_0\gamma_k \sim \chi_{d_0}^2$, $k_0 \nu_i \sim \chi_{g_0}^2$. Then prior distribution of β and **u** is

$$p(\beta, \mathbf{u} \mid f_0, k_0, d_0, g_0) \propto \prod_{k=1}^{p} \left[(\beta_k - b_k)^2 + f_0 \right]^{-\frac{d_0+1}{2}} \cdot \prod_{i=1}^{m} \left[u_i^2 + k_0 \right]^{-\frac{g_0+1}{2}} .$$

Let $s_i = \sum_{j=1}^{r_i} y_{ij}$. Posterior is given by

$$p(\beta, \mathbf{u} \mid y, h) \quad \propto$$

$$\exp\left[\sum_{i=1}^{m} \left\{ s_i(h_i'\beta + u_i) - r_i e^{h_i'\beta + u_i} \right\} \right]$$
$$\times \prod_{k=1}^{p} \left[(\beta_k - b_k)^2 + f_0 \right]^{-\frac{d_0+1}{2}} \cdot \prod_{i=1}^{m} \left[u_i^2 + k_0 \right]^{-\frac{g_0+1}{2}} .$$

Let $\bar{\gamma}_k(\Psi) = \dfrac{d_0+1}{f_0 + (\beta_k - b_k)^2}$, $k=1, \cdots\cdots\cdots p$,

$\bar{\nu}_i(\Psi) = \dfrac{g_0+1}{k_0 + u_i^2}$, $i=1, \cdots\cdots\cdots, m$,

z be m×1 vector with elements $(s_i - r_i e^{h_i'\beta + u_i})$,

$\Gamma(\Psi)$ be p×p matrix with diagonal elements $\bar{\gamma}_k(\Psi)$,

and $N(\Psi)$ be $m \times m$ matrix with diagonal elements $\bar{\nu}_i(\Psi)$.
It follows similar relations like (11) and (12).

$$g(\Psi \mid y)=\begin{pmatrix} Hz-\Gamma(\Psi)(\beta-b) \\ z-N(\Psi)u \end{pmatrix}$$

$$G(\Psi \mid y)=\begin{pmatrix} HA(\Psi)H'+\Gamma(\Psi)-\frac{2}{d_0+1}D_1 & HA(\Psi) \\ A(\Psi)H' & A(\Psi)+N(\Psi)-\frac{2}{g_0+1}D_1 \end{pmatrix}$$

where $A(\Psi)$ is $p \times p$ matrix with diagonal elements $r_i e^{h_i'\beta+u_i}$,

$$D_1=\Gamma(\Psi)^2(\beta-b)(\beta-b)'$$

$$D_2=N(\Psi)^2 u\ u'.$$

2 Electronic equipment

Jorgenson ([4]) discusses the number of failures of a complex piece of electronic equipment using regression analysis of the poisson distribution. Explanatory variables are times spent in two operating regimes at the cycle of operation. The observations are shown in Table 1. Dependent variable is the number of failures in the i'th cycle.
Jorgenson's formulation is

$$y_i=\beta_1 t_{1i}+\beta_2 t_{2i},\quad i=1,\ \cdots\cdots\cdots,n.$$

Alternatively, we set

$$\log\lambda_i=\beta_0+\beta_1 t_{1i}+\beta_2 t_{2i}+u_i.$$

Posterior modes and estimates of λ are also shown in Table 1, where we take f_0, k_0, d_0, g_0 as 5.

Table 1

T_1	T_2	y	$\hat{\lambda}$	$\hat{u}$
33.3	25.3	15	14.10	0.86
52.2	14.4	9	8.79	0.18
64.7	32.5	14	13.63	0.32
137.0	20.5	24	24.13	−0.11
125.9	97.6	27	27.35	−0.29
116.9	53.6	27	26.88	0.10
131.7	56.6	23	23.34	−0.29
85.0	87.3	18	18.05	−0.04
91.9	47.5	22	21.66	0.29

$\hat{\beta}_0 = 1.135$, $\hat{\beta}_1 = 0.015$, $\hat{\beta}_2 = 0.006$

3 Quine's sociological data

Aitken ([1]) analyses data of a sociological study about Australian school children. A response variable is the number of days absent from school during school year. Children were sampled by four factors, i.e. age, sex, cultural background and learning ability. Aitken analyses the data in terms of the analysis of variance, and seeks to find the minimal adequate model. He obtains six final minimal models and shows fitted values on each model. Fit seems to be not so good. We suppose the response variable is poisson variate, and the logarithm of parameter has a regression part and error part. Unfortunately, we cannot refer to Quine's original data but only Aitken's summarized data (sample means) with sample sizes. Data are shown in table 2. Under above modified formulation, λ_is are estimated and fitted to sample means. Explanatory variables are as follows,

C:1 = Aboriginal, 2 = white,

S:1 = Female, 2 = Male,

A:1 = Primary, 2 = First form, 3 = Second form, 4 = Third form,

L:1 = slow, 2 = average (learning ability).

We try also the case with explanatory variables including S×A, C×A without u-term, for comparison. Aitken obtains the best fit in this case under his formulation. The result is shown in table 2. Here we take f_o, k_o, d_o, g_o as 5. In table 2, $\hat{\lambda}_A$ is estimate for our formulation and $\hat{\lambda}_B$ is estimate for model {SA, CA} without u.

Table 2

C	S	A	L	$\bar{y}$	r	$\hat{\lambda}_A$	$\hat{u}$	$\hat{\lambda}_B$
1	1	1	1	9.00	3	8.971	0.0730	14.34
1	1	1	2	13.00	5	12.907	0.3998	10.96
1	1	2	1	9.00	3	9.059	−0.1493	21.02
1	1	2	2	10.50	2	10.521	−0.0366	16.06
1	1	3	1	37.00	4	36.747	1.0188	30.81
1	1	3	2	27.29	7	27.321	0.6855	23.54
1	1	4	2	27.14	7	27.069	0.4441	34.50
1	2	1	1	3.00	1	3.878	−0.8330	21.35
1	2	1	2	21.25	4	21.033	0.8209	16.32
1	2	2	1	22.60	10	22.523	0.6942	21.66
1	2	2	2	11.40	5	11.405	−0.0232	16.55
1	2	3	1	36.38	8	35.008	0.9031	21.97
1	2	3	2	2.00	1	3.245	−1.5122	16.79
1	2	4	2	14.56	9	14.587	−0.2414	17.03
2	1	1	1	30.00	3	29.613	1.2863	8.55
2	1	1	2	5.33	6	5.420	−0.4488	6.53
2	1	2	1	6.14	7	6.225	−0.5055	12.43
2	1	2	2	3.50	2	3.995	−0.9859	9.50
2	1	3	1	29.33	3	29.049	0.8027	18.07
2	1	3	2	9.14	7	9.207	−0.3831	13.81
2	1	4	2	27.28	7	27.209	0.4683	20.07
2	2	1	1	25.00	1	23.994	1.0086	12.73
2	2	1	2	18.50	4	18.308	0.7013	9.73
2	2	2	1	6.00	11	6.061	−0.5994	12.81
2	2	2	2	11.00	6	11.008	−0.0396	9.79
2	2	3	1	6.22	9	6.316	−0.7904	12.89
2	2	3	2	1.00	1	2.315	−1.8308	9.85
2	2	4	2	13.50	10	13.535	−0.2972	9.91

$\hat{\beta}_0 = 1.804$, $\hat{\beta}_C = -0.0191$, $\hat{\beta}_S = 0.0672$

$\hat{\beta}_A = 0.2321$, $\hat{\beta}_L = 0.0369$

IV Discussion

It is not a easy task to interprete the estimated u terms appropriately, but we may examine in a usual analysis of residuals manner. As mentioned above, we cannot refer to the original Quine's data, so we estimate λ s and examine the fit to sample means only. Then we do not have the problem of over-diversification. However, if we try to fit to the original individual observations, extravariation may arise. This problem may be approached either by the coumpound distribution which is a mixture with respect to some distribution for the parameter, or by the method that West proposes, which keeps scale parameter unristricted and uses his approximate likelihood. But, if we can consider extravariations as consequences of the misspecification, u-term may convey some information about the kind of misspecification, or, missing explanatory variables. That some important explanatory variables are overlooked may mean, in this example, observations of dependent variable which belong to different distributions are aggregated into the same distribution, and then give rise to the extravariations. An ad hoc method to deal with this problem may be to group observations according to u values. But some formal procedure for exploiting the information that u conveys is necessary.

Finally, as for the prior distribution of u-term, some smooth prior might be used, which is related to values of the explanatry variables in the sense of expressing the judgement that the changes of values of u are not so large among y having similar values of explanatory variables, as in Blight and Ott ([2]).

References

1. Aitken, M.A. 1978. The analysis of unbalanced cross-classifications (with discussion). J. R. S. S. A 141, 195--223.

2. Blight, B. J. N., Ott, L. 1975. A Bayesian approach to model inadequacy for polynomial regression. Biometrika 62, 79--88.

3. Frome, E. L., Kutner, M. H., Beauchamp, J. 1973. Regression analysis of poisson-

distributed data. J. A. S. A. 68, no. 334, 935-940.

4. Jorgenson, D.W. 1961. Multiple regression analysis of a poisson process. J. A. S. A. 235-245.

5. Nelder, J. A.˙ Wedderburn, R. W. M. 1972. Generalized Linear Models. J. R. S. S. A. No.135, 370-384.

6. Sweeting, T. 1981. Scale Parameters; a Bayesian Treatment. J. R. S. S. B. 43 No. 3, 333-338

7. West, M. 1985. Generalized linear models. Scale parameters, Outlier Accomodation and Prior distributions. Bayesian statistics 2, J. M. Bernardo, M. H. A. DeGroot, D. V. Lindley, A. F. M. Smith (Eds.) North-Holland, 531 - - 558.