Supplementary Material for "A Flexible Regression Model for Count Data"

Kimberly F. Sellers

Department of Mathematics

Georgetown University, Washington, DC 20057

Galit Shmueli

Department of Decision, Operations & Information Technologies

Smith School of Business, University of Maryland, College Park, MD 20742

1 Iterative Reweighted Least Squares Estimation

Due to the GLM formulation, maximizing the likelihood function can be formulated as an iterative weighted least squares procedure. In the following we describe such a formulation.

Since the CMP distribution belongs to the exponential family, we can use the following formulation in order to obtain the MLE for $\boldsymbol{\beta}$ (Dobson, 2002, p.40):

$$\begin{aligned} \frac{\partial \log L_i}{\partial \beta_j} &= \frac{\partial \log L_i}{\partial \log \lambda_i} \cdot \frac{\partial \log \lambda_i}{\partial \beta_j} \\ &= \left(y_i - \frac{\partial \log Z(\lambda_i, \nu)}{\partial \log \lambda_i} \right) x_{ij} \\ &= (y_i - E(Y_i))\, x_{ij}, \end{aligned} \tag{1}$$

where $j = 0, \ldots, p$. To estimate ν (which is assumed unknown and constant across observa-

tions), we consider

$$\begin{aligned}\frac{\partial \log L_i}{\partial \nu} &= -\log y_i! - \frac{\partial \log Z(\lambda_i, \nu)}{\partial \nu} \\ &= -\log y_i! + E(\log Y_i!). \end{aligned} \tag{2}$$

By summing the results in Equations (1) and (2) over all n observations and equating each summation to zero, we obtain the normal equations for estimating $\boldsymbol{\beta}$ and ν:

$$\begin{aligned}\sum_{i=1}^{n} y_i x_{ij} &= \sum_{i=1}^{n} \left\{ x_{ij} \frac{\sum_{s=0}^{\infty} s e^{s\mathbf{X}_i\boldsymbol{\beta}}/(s!)^{\nu}}{\sum_{s=0}^{\infty} e^{s\mathbf{X}_i\boldsymbol{\beta}}/(s!)^{\nu}} \right\} \\ \sum_{i=1}^{n} \log y_i! &= \sum_{i=1}^{n} \left\{ \frac{\sum_{s=0}^{\infty} \log(s!) e^{s\mathbf{X}_i\boldsymbol{\beta}}/(s!)^{\nu}}{\sum_{s=0}^{\infty} e^{s\mathbf{X}_i\boldsymbol{\beta}}/(s!)^{\nu}} \right\}. \end{aligned}$$

These equations are non-linear in $\boldsymbol{\beta}$ and ν, and therefore require an iterative solution, starting with the Poisson estimates, $\boldsymbol{\beta}^{(0)}$ and $\nu^{(0)} = 1$, or any other initial values. Furthermore, due to the GLM formulation solving the $p+2$ normal equations above can be done via reweighted least squares of the form,[1]

$$\boldsymbol{\mathcal{X}'\mathcal{W}\mathcal{X}\theta^{(m)} = \mathcal{X}'\mathcal{W}T}, \tag{3}$$

where $\boldsymbol{X}$ is an $n \times (p+2)$ matrix that is the ordinary design matrix X with the additional right column

$$\begin{bmatrix} \frac{-\log Y_1! + E(\log Y_1!)}{Y_1 - E(Y_1)} \\ \vdots \\ \frac{-\log Y_n! + E(\log Y_n!)}{Y_n - E(Y_n)} \end{bmatrix}.$$

The weight matrix $\boldsymbol{W}$ is an $n \times n$ diagonal matrix with elements $\boldsymbol{W}_{ii} = \mathrm{Var}(Y_i)$; The vector $\boldsymbol{\theta}^{(m)}$ is the mth iteration of the estimated coefficient vector $[\hat{\boldsymbol{\beta}}', \hat{\nu}]'$; Finally, the vector $\boldsymbol{T}$ (of

[1] This is a generalization of the classic WLS formulation, as in Dobson (2002), to a two parameter case.

length n) has element i equal to

$$t_i = \sum_{j=0}^{p} x_{ij}\beta_j^{(m-1)} + \nu^{(m-1)}\frac{\log Y_i! + E(\log Y_i!)}{Y_i - E(Y_i)} - \frac{Y_i - E(Y_i)}{Var(Y_i)}.$$

This derivation is based on an initial Newton-Raphson iterative formulation of the form $\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)} + \boldsymbol{I^{-1}U}$, where $\boldsymbol{I}$ is the Fisher Information matrix defined in Section 2, and $\boldsymbol{U}$ is the score vector with $p+2$ elements equal to the right-hand side in Equations (1)-(2), summed over $i = 1, \ldots, n$. Left-multiplication of both sides by $\boldsymbol{I}$, and using $\boldsymbol{I} = \boldsymbol{\mathcal{X}'\mathcal{W}\mathcal{X}}$ leads to the formulation in Equation (3).

2 Fisher Information Matrix Components Associated with CMP Coefficients

Due to the GLM formulation, we have a block Fisher Information matrix of the form

$$\mathbf{I} = \begin{pmatrix} \mathbf{I}^{\boldsymbol{\beta}} & \mathbf{I}^{\boldsymbol{\beta},\nu} \\ \mathbf{I}^{\boldsymbol{\beta},\nu} & I^{\nu} \end{pmatrix},$$

where, for $j, k \in \{0, \ldots, p\}$,

$$\begin{aligned}
I^{\beta}_{j,k} &= \sum_{i=1}^{n} x_{ij}x_{ik}\left\{\frac{\sum_{s=0}^{\infty} s^2 e^{sX_i\beta}/(s!)^{\nu}}{\sum_{s=0}^{\infty} e^{sX_i\beta}/(s!)^{\nu}} - \left[\frac{\sum_{s=0}^{\infty} s e^{sX_i\beta}/(s!)^{\nu}}{\sum_{s=0}^{\infty} e^{sX_i\beta}/(s!)^{\nu}}\right]^2\right\}, \\
I^{\nu} &= \sum_{i=1}^{n}\left\{\frac{\sum_{s=0}^{\infty}(\log s!)^2 e^{sX_i\beta}/(s!)^{\nu}}{\sum_{s=0}^{\infty} e^{sX_i\beta}/(s!)^{\nu}} - \left[\frac{\sum_{s=0}^{\infty}(\log s!) e^{sX_i\beta}/(s!)^{\nu}}{\sum_{s=0}^{\infty} e^{sX_i\beta}/(s!)^{\nu}}\right]^2\right\}, \text{ and} \\
I^{\beta_j,\nu} &= \sum_{i=1}^{n} x_{ij}\left\{\frac{\sum_{s=0}^{\infty} s\log s! e^{sX_i\beta}/(s!)^{\nu}}{\sum_{s=0}^{\infty} e^{sX_i\beta}/(s!)^{\nu}} - \left[\frac{\sum_{s=0}^{\infty} s e^{sX_i\beta}/(s!)^{\nu}}{\sum_{s=0}^{\infty} e^{sX_i\beta}/(s!)^{\nu}}\right]\left[\frac{\sum_{s=0}^{\infty}(\log s!) e^{sX_i\beta}/(s!)^{\nu}}{\sum_{s=0}^{\infty} e^{sX_i\beta}/(s!)^{\nu}}\right]\right\}.
\end{aligned}$$

3 Full Datasets and Diagnostics Under Various Regression Models

Tables 1 and 2 show a more complete comparison of regression models with regard to fitted values, MSE, AIC_C, and leverage for the airfreight example.

Table 1: Airfreight breakage example: data, and fitted values

Obs.	No. of Aircraft Transfers (X)	No. of Broken Ampules Upon Arrival (Y)	Poisson Fit	Linear Reg Fit	CMP Median Fit
1	1	16	13.69	13.56	14
2	0	9	10.52	10.25	10
3	2	17	17.83	17.95	18
4	0	12	10.52	10.25	10
5	3	22	23.21	23.75	23
6	1	13	13.69	13.56	14
7	0	8	10.52	10.25	10
8	1	15	13.69	13.56	14
9	2	19	17.83	17.95	18
10	0	11	10.52	10.25	10
AIC_C			52.11	49.37	**47.29**
MSE			2.210	2.363	**1.900**

Table 2: Airfreight breakage example: leverage results associated with various regression models

Obs.	No. of Aircraft Transfers (X)	No. of Broken Ampules Upon Arrival (Y)	Poisson Leverage	Linear Reg Leverage	CMP Leverage
1	1	16	0.103	0.1	0.154
2	0	9	0.183	0.2	0.194
3	2	17	0.183	0.2	0.273
4	0	12	0.183	0.2	0.226
5	**3**	**22**	**0.594**	**0.5**	**0.600**
6	1	13	0.103	0.1	0.379
7	0	8	0.183	0.2	0.238
8	1	15	0.103	0.1	0.112
9	2	19	0.183	0.2	0.365
10	0	11	0.183	0.2	0.458

Figure 1 shows the boxplots associated with the log-transformed leverage results for the CMP, Poisson, negative binomial, and linear regression models.

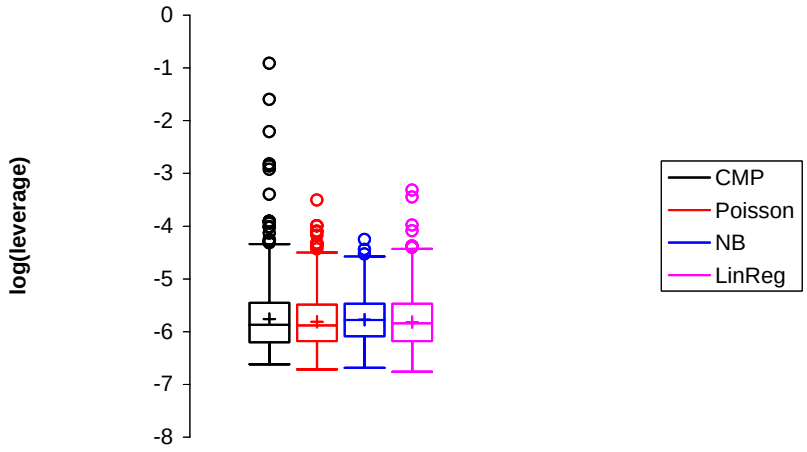

Figure 1: Boxplot of log-transformed leverage results associated with various regression models and the Toronto crash dataset.

References

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC, London.