A COM-Poisson Mixed Model with Normal Random Effects for Clustered Count Data

Darcy Steeg Morris*[1]
U.S. Census Bureau, Washington, DC, USA - darcy.steeg.morris@census.gov

Kimberly F. Sellers[1]
Georgetown University and U.S. Census Bureau, Washington, DC, USA - kfs7@georgetown.edu

Abstract

Clustered count data are commonly modeled using Poisson regression with random effects to account for the correlation induced by clustering. The Poisson mixed model allows for over-dispersion via the nature of the within-cluster correlation, however, departures from equi-dispersion may also exist due to the underlying count process mechanism. We extend the cross-sectional Conway-Maxwell-Poisson (COM-Poisson) regression model – a generalized regression model for count data in light of inherent data dispersion – to incorporate normal-distributed random effects for the analysis of clustered count data. We demonstrate model flexibility of the COM-Poisson mixed model via simulated examples.

Keywords: correlated count data; Poisson regression; over-dispersion; under-dispersion.

1. Introduction

The most widely used regression for count response data is the Poisson model. The Poisson regression model considers the relationship between count response and explanatory data under the strict assumption of equi-dispersion (i.e. the variance equals the mean). The Poisson random effects regression model – an extension of the standard Poisson regression model – is commonly used for correlated count data where multiple outcome measurements are available for each cluster.[2] To account for within-cluster correlation, the model combines the standard Poisson count model with a cluster-specific term that reflects cluster-level heterogeneity (Cameron and Trivedi 1998, Winkelmann 2008). Models of this type assume that data are independent between cluster and that within-cluster correlation is adequately controlled for through the cluster-specific random effect. The introduction of the additional randomness due to the random intercept allows the cluster-specific rates to vary in a way that cannot be accounted for by observables. Allowing this additional variability naturally relaxes the strict equi-dispersion assumption of the Poisson distribution.

The underlying count outcome, however, may exhibit under-dispersion or additional over-dispersion that is not adequately modeled by the correlation structure alone. Such data requires a more flexible count model to account for underlying count dispersion as well as clustering. A popular alternative to handle over-dispersion is the negative binomial (NB) model – with or without random effects. The negative binomial model can address additional over-dispersion, but it does not address data under-dispersion. Alternatively, the Conway-Maxwell-Poisson (COM-Poisson) model allows for both over- and under-dispersed count data. Marginal COM-Poisson models have been proposed to jointly account for within-cluster association and dispersion (Khan and Jowaheer 2013, Choo-Wosoba et. al. 2016). Using the mixed model framework (i.e. conditional models), we extend the COM-Poisson generalized linear model (GLM) (Sellers and Shmueli 2010) to study maximum likelihood estimation of a COM-Poisson mixed model with normal-distributed random

[1]This paper is intended to inform interested parties of research and to encourage discussion. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

[2]We define clustered data in the general sense, as the studied models can be applied to various specific types of clustered/correlated count data such as longitudinal, repeated measures, spatial, and family data.

effects.

2. COM-Poisson Distribution and Regression Model

The COM-Poisson distribution is a flexible distribution for count data that allows for over- or under-dispersion (Conway and Maxwell 1962, Shmueli et. al. 2005). The COM-Poisson probability mass function for a single observation i takes the form

$$P(Y = y_i \mid \lambda, \nu) = \frac{\lambda^{y_i}}{(y_i!)^{\nu} Z(\lambda, \nu)}, \quad y_i = 0, 1, 2, \ldots \tag{1}$$

for a random variable Y, where $Z(\lambda, \nu) = \sum_{s=0}^{\infty} \frac{\lambda^s}{(s!)^{\nu}}$ is a normalizing constant. In this setting, $\lambda = E(Y^{\nu})$, where $\nu \geq 0$ is the dispersion parameter such that $\nu = 1$, $\nu > 1$, and $\nu < 1$ signify equi-dispersion, under-dispersion, and over-dispersion, respectively. The moments of the COM-Poisson distribution are not of closed form, however, Shmueli et. al. (2005) note that assuming an asymptotic approximation for $Z(\lambda, \nu)$ leads to a close approximation for the mean:

$$E(Y) = \lambda \frac{\partial \log Z(\lambda, \nu)}{\partial \lambda} \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \texttt{ for } \nu \leq 1 \texttt{ or } \lambda > 10^{\nu}. \tag{2}$$

The COM-Poisson distribution includes three well-known distributions as special cases: Poisson with rate parameter λ ($\nu = 1$); geometric with success probability $1 - \lambda$ ($\nu = 0, \lambda < 1$); and Bernoulli with success probability $\frac{\lambda}{1+\lambda}$ ($\nu \to \infty$). See Shmueli et. al. (2005) and Sellers et. al. (2012) for details regarding this distribution.

Sellers and Shmueli (2010) extend the COM-Poisson distribution to the regression context allowing varying λ for each observation i. This GLM approach assumes a link function $\eta(E(Y_i)) = \log \lambda_i$ that indirectly models the relationship between the mean and the linear predictor. The relationship between λ_i and the observations is encapsulated in the row vector of covariates, $\mathbf{x}_i$, with the log-likelihood for observation i now

$$\log L_i(\beta, \nu | y_i, \mathbf{x}_i) = y_i \log \lambda_i - \nu \log y_i! - \log Z(\lambda_i, \nu), \tag{3}$$

where

$$\log \lambda_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \doteq \mathbf{x}_i \beta. \tag{4}$$

Sellers and Shmueli (2010) describe maximum likelihood estimation of this model. Guikema and Coffelt (2008) propose Bayesian estimation of a re-parameterized version of this model. Both approaches allow modeling of the dispersion parameter ν, however, we assume a constant ν in the development of the COM-Poisson mixed model for clustered data.

Poisson regression can model equi-dispersed data with one less parameter, however, the additional dispersion parameter in the COM-Poisson regression model flexibly handles equi-, over-, or under-dispersion without bias in inference. Given that most real count data exhibit some form of data dispersion, the COM-Poisson model is a viable alternative to both the negative binomial and Poisson regression models for count data.

3. COM-Poisson Mixed Model

We extend the Sellers and Shmueli (2010) COM-Poisson regression model to include a random intercept in order to model clustered data. The random intercept COM-Poisson model assumes:

$$y_{ij} | u_i \sim \text{CMP}(\lambda_{ij}^*, \nu) \tag{5}$$

$$\log(\lambda_{ij}^*) = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + u_i \doteq \mathbf{x}_{ij} \beta + u_i \tag{6}$$

$$u_i \sim N(0, \sigma^2) \tag{7}$$

where y_{ij} is the count outcome for cluster i at occurrence j, $i = 1, \ldots, N$, $j = 1, \ldots, J_i$; $\mathbf{x}_{ij}$ is a vector of explanatory variables; and u_i is the cluster-specific term for cluster i. Under these assumptions, the count outcome for cluster i at occurrence j follows a COM-Poisson distribution conditional on a cluster-specific random effect, a set of covariates, and a vector of regression parameters $(\beta_0, \ldots, \beta_p)$ that are common to all clusters. The clusters are assumed to be independent from one another and observations within a cluster are independent conditional on the random effect ($y_{ij} \perp\!\!\!\perp y_{ik} | u_i$ for $j \neq k$). We make the common assumption that the random effects are normally distributed. We obtain maximum likelihood estimates in `R` using (1) numerical integration (the `integrate` function) to obtain an approximation of the marginal loglikelihood:

$$\begin{aligned} \log L(\beta, \nu, \sigma^2 | y_{i1}, \ldots, y_{iJ_i}) &= \sum_{i=1}^{N} \sum_{j=1}^{J_i} y_{ij} \log(\lambda_{ij}) - \nu \sum_{i=1}^{N} \sum_{j=1}^{J_i} \log(y_{ij}!) - \sum_{i=1}^{N} \log(\sigma\sqrt{2\pi}) \\ &\quad + \sum_{i=1}^{N} \log \left(\int_{u_i} e^{u_i \sum_{j=1}^{J_i} y_{ij} - u_i^2/2\sigma^2} \left(\prod_{j=1}^{J_i} Z(e^{u_i} \lambda_{ij}, \nu) \right)^{-1} du_i \right), \end{aligned} \tag{8}$$

and (2) optimization (the `nlminb` function) to maximize the approximate marginal loglikelihood. Maximum likelihood estimates of the COM-Poisson mixed model can similarly be obtained in SAS® using the NLMIXED procedure (Morris et. al. 2017).

4. Analysis of Simulated Data

We conduct an analysis of simulated count regression data to demonstrate the flexibility of the random intercept COM-Poisson regression model for clustered count data. Three random intercept models (Poisson, negative binomial, COM-Poisson) are fit to clustered count datasets simulated from: Poisson, Bernoulli, geometric, under-dispersed COM-Poisson ($\nu = 5$) and over-dispersed COM-Poisson ($\nu = .75$) distributions.[3] These five simulated datasets capture various scenarios where, in addition to the clustering, the data are under-, over-, or equi-dispersed. The clustering is induced by a simulated cluster-specific term u ($u_i \sim N(0, .5)$) shared by each observation in the cluster. Each of the five simulated datasets consist of the following for a given observation: a continuous explanatory variable x ($x_i \sim N(0, .1)$), a response variable y, and the cluster identification number. Each of the 100 clusters ($N = 100$) is observed five times ($J_i = 5 \; \forall \; i$) for a total of 500 observations. The Bernoulli simulation necessarily has a binary response values (0 or 1), whereas the other four simulations all have count response values. We consider the regression model, $\log(\lambda_{ij}) = \beta_0 + \beta_1 x_i$, for all of the proposed models under consideration. For the special case of a Bernoulli/logistic model, we define $\lambda_{ij} = \frac{p_i}{1-p_i}$.

The three models – Poisson, negative binomial, and COM-Poisson – capture variations in conditional distributional assumptions of the data. Table 1 provides the dispersion parameter estimates, estimates of the random effect variance and AIC for each of the models fit to each of the simulated datasets. We find that the COM-Poisson model has the best model fit for the COM-Poisson simulated data and comparable model fit for the COM-Poisson special cases. For the Poisson and Bernoulli simulated data, we find that the COM-Poisson model accurately recognizes the special cases ($\hat{\nu} = 1.12 \approx 1$ and $\hat{\nu} = 30.0$ is large for Poisson and Bernoulli, respectively) and has AIC values roughly equivalent to the special case models (1753.8 vs. 1754.1 and 661.4 vs. 661.6). Note that the negative binomial model also recognizes the Poisson special case ($\hat{k} = 0$). For the simulated geometric data, the negative binomial model outperforms the COM-Poisson model (AIC values of 1912.2 vs. 1920.5 for the negative binomial and COM-Poisson models, respectively). Even though the COM-Poisson model accurately recognizes the geometric special case ($\hat{\nu} = 0.00$), the geometric and COM-Poisson models are only a special case of the negative binomial model, allowing it to capture variation in the geometric simulated data in a way that the COM-Poisson model cannot. Note that the negative binomial model also recognizes the geometric special case ($\hat{k} = 1.10 \approx 1$). The analysis of the special case simulated datasets illustrate the flexibility of the COM-Poisson to capture equi-dispersion and

[3]The simulated under- and over-dispersed COM-Poisson counts are generated using the `makeCMPdata` function in the `COMPoissonReg` package (Sellers and Lotze, 2015).

Table 1: Dispersion and variance estimates for random intercept Poisson, negative binomial (NB) and COM-Poisson (CMP) models for various simulated data sets. Akaike's Information Criterion (AIC) reported for model comparisons – bold values indicate the model with the lowest AIC.

Simulated		Model		
Dataset	Estimate	Poisson	NB	COM-Poisson
Poisson	Dispersion		$\hat{k} = 0.00$	$\hat{\nu} = 1.12$
	Variance	$\hat{\sigma}^2 = 0.511$	$\hat{\sigma}^2 = 0.511$	$\hat{\sigma}^2 = 0.619$
	AIC	**1753.8**	1755.8	1754.1
Bernoulli*	Dispersion		$\hat{k} = 0.00$	$\hat{\nu} = 30.0$
	Variance	$\hat{\sigma}^2 = 0.000$	$\hat{\sigma}^2 = 0.000$	$\hat{\sigma}^2 = 0.811$
	AIC	910.0	912.0	661.6
Geometric	Dispersion		$\hat{k} = 1.10$	$\hat{\nu} = 0.00$
	Variance	$\hat{\sigma}^2 = 0.692$	$\hat{\sigma}^2 = 0.456$	$\hat{\sigma}^2 = 0.042$
	AIC	2287.0	**1912.2**	1920.5
CMP	Dispersion		$\hat{k} = 0.00$	$\hat{\nu} = 5.32$
(under)	Variance	$\hat{\sigma}^2 = 0.000$	$\hat{\sigma}^2 = 0.000$	$\hat{\sigma}^2 = 0.810$
	AIC	968.6	970.6	**807.6**
CMP	Dispersion		$\hat{k} = 0.10$	$\hat{\nu} = 0.65$
(over)	Variance	$\hat{\sigma}^2 = 0.717$	$\hat{\sigma}^2 = 0.696$	$\hat{\sigma}^2 = 0.373$
	AIC	2122.6	2105.5	**2102.6**

* The random intercept logistic model results/estimates for the simulated Bernoulli data are: AIC = **661.4** and $\hat{\sigma}^2 = 0.704$.

extreme over-/under-dispersion, while the superior performance of the COM-Poisson model on the COM-Poisson simulated datasets shows its flexibility to also capture intermediate levels of over- or under-dispersion.

With respect to the estimates of the random intercept variance, the COM-Poisson model estimates non-zero variance in all cases, reflecting the variability due to the clustered nature of the simulated data.[4] For the cases of equi- and over-dispersion (i.e. Poisson, geometric and CMP (over)), both the Poisson and negative binomial models estimate a non-zero variance. However, because the Poisson and negative binomial models cannot handle under-dispersion, in the under-dispersed simulated data cases (Bernoulli and CMP (under)) these models estimate zero within-cluster variability. In all cases, the estimate of the random intercept variance must be cautiously compared across models – the linear predictor is directly linked to the mean in the Poisson and negative binomial models, whereas it is indirectly linked to the mean in the COM-Poisson model (as defined in equations (2) and (6)).

5. Discussion

The COM-Poisson regression model is a flexible model for count data in light of data dispersion. We extend the cross-sectional COM-Poisson model of Sellers and Shmueli (2010) to include a random intercept to address cluster-level correlation in clustered count data. The flexibility of the random intercept COM-Poisson

[4]Estimates of the regression coefficients, β_0 and β_1, are excluded for brevity as the focus of this paper is on model fit and cluster effects.

model allows deviations from equi-dispersion due to the nature of the clustering as well as the nature of the underlying count mechanism. The analysis of five simulated clustered count datasets with varying degrees of underlying dispersion illustrates the flexibility of the COM-Poisson mixed model to provide a good model fit for special cases of well-known count distributions, as well as cases with intermediate levels of over- or under-dispersion. More complex model specification, e.g. random slopes and mixed modeling of the dispersion parameter, can naturally be incorporated but are left to future research.

References

Cameron, A.C. and Trivedi, P.K. (1998). Regression Analysis of Count Data. Cambridge University Press.

Choo-Wosoba, H., Levy, S. M., and Datta, S. (2016). "Marginal regression models for clustered count data based on zero-inated Conway-Maxwell-Poisson distribution with applications." *Biometrics*, 72: 606-618.

Conway, R.W. and Maxwell, W.L. (1962). "A Queuing Model with State Dependent Service Rates." *Journal of Industrial Engineering*,12: 132-136.

Guikema, S.D. and Coffelt, J.P. (2008). "A Flexible Count Data Regression for Risk Analysis." *Risk Analysis*, 28: 213-223.

Khan, N. M. and Jowaheer, V. (2013). "Comparing Joint GQL Estimation and GMM Adaptive Estimation in COM-Poisson Longitudinal Regression Model." *Communications in Statistics - Simulation and Computation*, 42: 755-770.

Morris, D.S., Sellers, K.F., and Menger, A. (2017) "Fitting a Flexible Model for Longitudinal Count Data using the NLMIXED Procedure." In *SAS Global Forum Proceedings*, Cary, NC. SAS Institute.

Sellers, K.F., Borle, S., and Shmueli, G. (2012). "The COM-Poisson Model for Count Data: A Survey of Methods and Applications." *Applied Stochastic Models in Business and Industry*, 28: 104-116.

Sellers, K.F. and Lotze, T. (2015). "`COMPoissonReg`: Conway-Maxwell-Poisson Regression." *Version 0.3.5.* http://cran.r-project.org/web/packages/COMPoissonReg/index.html.

Sellers, K.F., & Shmueli, G. (2010). "A Flexible Regression Model for Count Data." *Annals of Applied Statistics*, 4: 943-961.

Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S., and Boatwright, P. (2005). "A Useful Distribution for Fitting Discrete Data: Revival of the Conway-Maxwell-Poisson Distribution." *Applied Statistics*, 54: 127-142.

Winkelmann, R. (2008). Econometric Analysis of Count Data. Springer.