Paper 202-2017

Fitting a Flexible Model for Longitudinal Count Data Using the NLMIXED Procedure

Darcy Steeg Morris, U.S. Census Bureau[1]; Kimberly F. Sellers, Georgetown University and U.S. Census Bureau[1]; Austin Menger, Georgetown University

ABSTRACT

Longitudinal count data arise when a subject's outcomes are measured repeatedly over time. Repeated measures count data have an inherent within-subject correlation that is commonly modeled with random effects in the standard Poisson regression. A Poisson regression model with random effects is easily fit in SAS® using existing options in the NLMIXED procedure. This model allows for over-dispersion via the nature of the repeated measures; however, departures from equi-dispersion can also exist due to the underlying count process mechanism. We present an extension of the cross-sectional Conway-Maxwell-Poisson (COM-Poisson) regression model established by Sellers and Shmueli (2010) – a generalized regression model for count data in light of inherent data dispersion – to incorporate random effects for analysis of longitudinal count data. We detail how to fit the COM-Poisson longitudinal model via a user-defined log-likelihood function in PROC NLMIXED. We demonstrate the model flexibility of the COM-Poisson longitudinal model via a real data example.

INTRODUCTION

Count data are prevalent in a variety of fields of study including health (e.g. number of hospitalizations), risk analysis (e.g. number of traffic accidents, number of insurance claims), and marketing (e.g. number of purchases of a product). The Poisson model is commonly used to analyze integer-valued count data, $Y = 0,1,2,\dots$; however, it makes the strong assumption of equi-dispersion such that the variance, $Var(Y)$, is equal to the mean, $E(Y)$. In many real data applications, this equi-dispersion assumption is violated as the data may exhibit either over-dispersion, $Var(Y) > E(Y)$, or under-dispersion, $Var(Y) < E(Y)$.

Positive correlation between responses is one cause of over-dispersion in count data (Hilbe 2008). Longitudinal count data – count data that arise from repeated measurements taken on a subject – have an inherent correlation within subject, leading to the violation of the Poisson equi-dispersion assumption. This subject-level correlation can be accounted for by extending the Poisson model to include a random effect. Specifically, the random intercept Poisson model loosens the equi-dispersion assumption to capture additional variability induced by the correlation of measurements within a subject. The model assumptions of the random intercept Poisson model are:

$$y_{it}|u_i \sim Poi(\lambda_{it}^*)$$

$$\log(\lambda_{it}^*) = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_p x_{itp} + u_i$$

$$u_i \sim N(0, \sigma^2)$$

where y_{it} is the count outcome for subject i in time period t for $i = 1, \dots, n$ and $t = 1, \dots, T_i$; $x_{it1}, \dots, x_{itp}$ are the p covariates for subject i in time period t; and u_i is the random intercept (Winkelmann 2008, Cameron and Trivedi 2013). This model is easily fit via the NLMIXED procedure using existing options. Specifically, (1) the `poisson` distribution is called in the MODEL statement, and (2) the random intercept u is defined in the RANDOM statement along with the cluster variable. The following code fits a random intercept Poisson model with one covariate:

[1] This paper is intended to inform interested parties of ongoing 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.

```
PROC NLMIXED data=mydata;
  parms logsig=0 beta0=1 beta1=1;
  lambda = exp(beta0 + beta1*x1 + u);
  model y ~ poisson(lambda);
  random u ~ normal(0,exp(2*logsig)) subject=ID;
RUN;
```

In this code, the PARMS statement starts the procedure at arbitrarily chosen initial parameter values, the variable ID uniquely identifies a subject, and the specification of the variance of the random effect ensures that the positive variance constraint is satisfied ($\sigma^2 > 0$). PROC NLMIXED implements maximum likelihood estimation of nonlinear mixed models with default options of adaptive Gaussian quadrature for numerical integration and a dual quasi-Newton algorithm for optimization. The procedure assumes that the specified random effect follows a normal distribution. For details about PROC NLMIXED, see the chapter titled “The NLMIXED Procedure” in the SAS/STAT User’s Guide (2015). The MCMC procedure and the GLIMMIX procedure can be used to implement alternative techniques - Bayesian and pseudo-likelihood, respectively – for fitting a random intercept Poisson model.

The Poisson model with random effects explicitly models the longitudinal structure by incorporating random effects to address over-dispersion induced by correlation over time. However, it is based on the assumption that the underlying count mechanism, absent the longitudinal structure, exhibits equi-dispersion. The Conway-Maxwell-Poisson (COM-Poisson) distribution is a flexible alternative count distribution that allows for under- and over-dispersion (Shmueli et. al. 2005). Cross-sectional regression formulations of the COM-Poisson distribution have been studied (Guikema and Coffelt 2008, Sellers and Shmueli 2010) and shown to be useful in many applications including in accident analysis (Lord et. al. 2010), marketing (Boatwright et. al. 2003) and auction bidding (Borle et. al. 2006). We extend the cross-sectional COM-Poisson regression model to incorporate a random intercept that allows for the modeling of additional variability due to the correlation of repeated measures as well as the over- or under-dispersion of the underlying count process. This paper describes how to fit a random intercept COM-Poisson regression model using PROC NLMIXED. The model simply requires a user-defined conditional loglikelihood – this is a straightforward programming task for any SAS user with beginner’s knowledge of PROC NLMIXED syntax and mixed models theory.

COM-POISSON DISTRIBUTION AND REGRESSION

COM-POISSON DISTRIBUTION

The COM-Poisson distribution is a two-parameter generalization of the Poisson distribution that allows for over- or under-dispersion in count data. Originally derived by Conway and Maxwell (1961) and revived by Shmueli et. al. (2005), the COM-Poisson probability mass function takes the form:

$$P(Y = y|\lambda, \nu) = \frac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)} \ \ , \ \ y = 0,1,2, \ldots$$

for a random variable Y, where $Z(\lambda, \nu) = \sum_{s=0}^{\infty} \frac{\lambda^s}{(s!)^\nu}$ is a normalizing constant. The dispersion parameter, $\nu \geq 0$, allows the COM-Poisson distribution to cover a wide range of discrete distributions as it captures all cases of count data dispersion: equi-dispersion, under-dispersion and over-dispersion for $\nu = 1$, $\nu > 1$, and $\nu < 1$, respectively. The COM-Poisson distribution reduces to commonly used discrete data distributions in three special cases governed by the assumption of the dispersion parameter. These three special cases are:

Poisson: $\nu = 1 \ \Rightarrow Z(\lambda, \nu) = e^\lambda \Rightarrow Y \sim Poi(\lambda)$,

Geometric: $\nu = 0 \ \Rightarrow Z(\lambda, \nu) = \frac{1}{1-\lambda} \Rightarrow Y \sim Geo(1 - \lambda)$ for $\lambda < 1$, and

Bernoulli: $\nu \to \infty \ \Rightarrow Z(\lambda, \nu) \to 1 + \lambda \Rightarrow Y \sim Ber\left(\frac{\lambda}{1+\lambda}\right)$.

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:

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

COM-POISSON REGRESSION

Sellers and Shmueli (2010) introduce a generalized linear models (GLM) regression formulation of the COM-Poisson distribution, taking advantage of the exponential family features to allow for elegant estimation, inference and diagnostics. Assuming a log link between the parameter λ and the linear predictor, the COM-Poisson regression model lets λ vary for each observation i:

$$y_i \sim CMP(\lambda_i, \nu)$$
$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \,.$$

This specification indirectly models the relationship between the mean and the linear predictor, capturing the special cases of logistic and Poisson regression. The maximum likelihood estimates can be obtained in SAS via the COUNTREG procedure with the DIST=`cmpoisson` and PARAMETER = `lambda` options in the MODEL statement (SAS/STAT User's Guide 2015). In this work we assume a constant dispersion parameter; however, the model, in general, and the fitting of the model in NLMIXED can additionally assume observation-level variation in ν.

COM-POISSON LONGITUDINAL MODEL

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

$$y_{it} | u_i \sim CMP(\lambda_{it}^*, \nu)$$
$$\log(\lambda_{it}^*) = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_p x_{itp} + u_i$$
$$u_i \sim N(0, \sigma^2)$$

The generality of PROC NLMIXED for fitting nonlinear random effects models allows this model to be fit using existing options combined with a user-written model specification. The `general` loglikelihood option in PROC NLMIXED allows flexibility to fit a wide variety of random effects models. Fitting the COM-Poisson longitudinal model in PROC NLMIXED simply requires specifying the conditional loglikelihood which is:

$$\log L(\lambda_{it}^*, \nu) = y\log(\lambda_{it}^*) - \nu \log(y_{it}!) - \log(Z(\lambda_{it}^*, \nu)) \,,$$

where $Z(\lambda_{it}^*, \nu) = \sum_{s=0}^{\infty} \frac{(\lambda_{it}^*)^s}{(s!)^\nu}$ can be written as $Z(\lambda_{it}^*, \nu) = \sum_{s=0}^{\infty} \prod_{r=1}^{s} \frac{\lambda_{it}^*}{r^\nu}$. Previous work has found that truncating the infinite sum provides a good approximation (Minka et. al. 2003). The following code fits a random intercept COM-Poisson model with one covariate:

```
PROC NLMIXED data=mydata;
  parms logsig=0 beta0=1 beta1=1 nu=1;
  lambda = exp(beta0 + beta1*x1 + u);
  Z = 1;
  DO s = 1 to 100;
    Q = 1;
    DO r = 1 to s;
      Q = Q * (lambda / (r**nu));
    END;
    Z = Z + Q;
  END;
  ll = y*log(lambda) - nu*lgamma(y+1) - log(Z);
  model y ~ general(ll);
  random u ~ normal(0,exp(2*logsig)) subject=ID;
RUN;
```

The user must explicitly program the random intercept COM-Poisson conditional loglikelihood, while the rest of the PROC NLMIXED syntax is largely unchanged from the random intercept Poisson model.

APPLICATION: EPILEPSY DATA

To illustrate the performance of the random intercept COM-Poisson model for longitudinal count data, we study the epilepsy dataset originally analyzed by Thall and Vail (1990), further discussed in Diggle et. al. (1994), and generally often used as an example for longitudinal data analysis of discrete outcomes (e.g. PROC GENMOD in SAS/STAT User's Guide (2015)). This dataset concerns the number of seizures measured for 59 epileptic patients in an initial eight-week baseline period followed by four consecutive two-week treatment periods. The outcome variable of interest, y_{it}, is the number of seizures for subject i in time period t $(t = 1, 2, 3, 4, 5)$. Diggle et. al. (1994) fit a random intercept Poisson regression with the associated loglinear form:

$$\log(\lambda_{it}^{*}) = \beta_0 + \beta_1 x_{it1} + \beta_2 x_{it2} + + \beta_3 x_{it1} x_{it2} + \log(T_{it}) + u_i \ ,$$

where x_{it1} is an indicator of a period after baseline (weeks 8-16), x_{it2} indicates the receipt of an anti-epileptic drug Progabide as opposed to a placebo, the offset term T_{it} is the length of time period t in weeks, and u_i is the random intercept. Following the Diggle et. al. (1994) analysis of this data, we likewise fit this model as well as a random intercept negative binomial model (parameterized as documented for PROC NLMIXED in SAS/STAT User's Guide 2015) and a random intercept COM-Poisson (CMP) model using PROC NLMIXED as stated in the previous section. The resulting parameter estimates and associated standard errors are presented in Table 1.

	Poisson R.E.		**Negative Binomial R.E.**		**CMP R.E.**	
Variable	**Est.**	**Std. Err.**	**Est.**	**Std. Err.**	**Est.**	**Std. Err.**
Intercept	1.033*	(0.153)	1.100*	(0.176)	-0.779*	(0.178)
x_1	0.111*	(0.047)	0.016	(0.101)	-0.791*	(0.071)
x_2	-0.024	(0.211)	0.074	(0.242)	0.051	(0.106)
$x_1 * x_3$	-0.104	(0.065)	-0.312*	(0.142)	-0.177*	(0.053)
k			0.148*	(0.025)		
ν					0.421*	(0.050)
σ	0.780*	(0.075)	0.813*	(0.082)	0.379*	(0.054)
AIC	2031.4		1789.8		1754.6	

* Statistically significant estimates at the 5% significance level are marked with an asterisk.

Table 1. Epilepsy Data Results: Poisson Random Intercept, Negative Binomial Random Intercept and COM-Poisson Random Intercept Model Estimates and Standard Errors.

The random intercept COM-Poisson model has the lowest AIC, indicating the best model fit. Both the random intercept negative binomial and random intercept COM-Poisson models provide a better fit than the random intercept Poisson model providing evidence that there is additional variability beyond that captured by the subject-specific random effect. Within-subject variability is reflected through the significantly greater than zero estimates of the random intercept variance parameter, $\hat{\sigma} = 0.780$, $\hat{\sigma} = 0.813$ and $\hat{\sigma} = 0.379$, for the Poisson, negative binomial and COM-Poisson model, respectively. Furthermore, additional over-dispersion is evident through the estimates of the dispersion parameters, $\hat{k} = 0.148 > 0$ and $\hat{\nu} = 0.421 < 1$, for the negative binomial and COM-Poisson model, respectively. Both the negative binomial and the COM-Poisson models can account for dispersion beyond that induced by within-subject correlation, however the slightly better model fit for the COM-Poisson model indicates that it captures the dispersion in a way that the negative binomial model cannot.

The interaction effect $(x_1 * x_3)$ is negative in all models indicating that the Progabide drug is associated with fewer seizures in the treatment period. However, estimated parameters are not directly comparable across models. For example, the point estimates of the interaction vary as they represent different effects: on the mean for the Poisson and negative binomial models and on an indirect function of the mean for the COM-Poisson model. Similarly, the estimated random effect variance is smaller for COM-Poisson than the estimates from the Poisson and negative binomial models.

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 random effects to address subject-level correlation in longitudinal data. The flexibility of the random intercept COM-Poisson model allows modeling of additional variability in the count outcome beyond that induced by repeated measures – multiple such sources of dispersion are evident in the epilepsy data. Implementing this model in the NLMIXED procedure simply requires the user-written log-likelihood function detailed in this paper. The flexibility of PROC NLMIXED allows any SAS programmer with beginner's knowledge of the procedure to fit this general, possibly better fitting, COM-Poisson mixed model. Furthermore, the generality of PROC NLMIXED would allow additional features in a longitudinal COM-Poisson model, notably incorporating random slopes and mixed modeling of the dispersion parameter.

An alternative parameterization of the COM-Poisson regression model is available in SAS and can be extended to include random effects. In order to link on a more direct centering measure, Guikema and Coffelt (2008) re-parameterize the COM-Poisson regression model letting $\mu = \lambda^{1/\nu}$ so that:

$$y_i \sim CMP(\mu_i^{\nu}, \nu)$$

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \, .$$

In this formulation of the COM-Poisson regression model, the centering parameter μ is a decent approximation for the mean in the case of $\mu > 10$ (see Equation 1) and in all cases the integer part of μ is the mode. The Bayesian approach for estimating the COM-Poisson regression model as studied in Guikema and Coffelt (2008) can be fit with PROC MCMC. The maximum likelihood approach is implemented in SAS via PROC COUNTREG with the DIST=`cmpoisson` and PARAMETER = `mu` (the default) options in the MODEL statement (SAS/STAT User's Guide 2015).

Just as for the Sellers and Shmueli (2010) COM-Poisson regression model, the Guikema and Coffelt (2008) parameterization of the random intercept COM-Poisson model can be fit using PROC NLMIXED. For the epilepsy data, this fit yields estimates of the linear predictor roughly similar to those from the negative binomial model[2]. For example, the estimated random effect variance parameter is $\hat{\sigma} = 0.899\ (0.092)$ and the estimated interaction effect is $\widehat{\beta_3} = -0.419\ (0.118)$. Because the average estimated μ_{it}^{*} is reasonably large $\left(\overline{\hat{\mu}_{it}^{*}} = 14.06\right)$, it appears to be sensible to assume that $\mu = \lambda^{1/\nu}$ is a decent approximation of the mean. Thus, in the epilepsy data, the linear predictor is closely related to the mean and the estimated coefficients can be loosely interpreted in the usual way. This interpretation, however, depends crucially on the assumptions of a further approximation of the mean approximation detailed in Equation 1.

REFERENCES

Boatwright, P., Borle, S. and Kadane, J.B. 2003. "A Model of the Joint Distribution of Purchase Wuantity and Timing." *Journal of the American Statistical Association*, 98: 564-572.

Borle, S., Boatwright, P., and Kadane, J.B. 2006. "The Timing of Bid Placement and Extent of Multiple Bidding: An Empirical Investigation using ebay Online Auctions." *Statistical Science,* 21: 194-205.

Cameron, C. and Trivedi, P.K. 2013. *Regression Analysis of Count Data.* Econometric Society Monograph No. 53: Cambridge University Press.

[2] Note that, by definition, fitting the Guikema and Coffelt (2008) parameterization of the COM-Poisson model returns the same AIC and dispersion estimate as the Sellers and Shmueli (2010) model.

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

Diggle, P., Heagerty, P.J., Liang, K.Y., and Zeger, S.L. 1994. *Analysis of Longitudinal Data.* Oxford: Clarendon.

Guikema, S.D. and Coffelt, J.P. 2008. “A Flexible Count Data Regression Model for Risk Analysis.” *Risk Analysis*, 28(1): 213-223.

Hilbe, J.M. 2008. *Negative Binomial Regression.* Cambridge University Press.

Lord, D., Geedipally, S.R., and Guikema, S.D. 2010. “Extension of the Application of Conway-Maxwell-Poisson Models: Analyzing Traffic Crash Data Exhibiting Underdispersion.” *Risk Analysis*, 30(8): 1268-1276.

Minka, T.P., Shmueli, G., Kadane, J.B., Borle, S., and Boatwright, P. 2003. “Computing with the COM-Poisson Distribution.” *Technical Report 776, Department of Statistics, Carnegie Mellon University.*

SAS Institute Inc. 2015. *SAS/STAT® 14.1 User’s Guide*. Cary, NC: SAS Institute Inc.

Sellers, K. and Shmueli, G. 2010. “A Flexible Regression Model for Count Data.” *Annals of Applied Statistics*, 4:943-961.

Shmueli, G., Minka, T.P., Kadane, J., Borle, S. and Boatwright, P. 2005. “A Useful Distribution for Fitting Discrete Data: Revival of the Conway-Maxwell-Poisson Distribution.” *Journal of the Royal Statistical Society, Series C,* 54:127-142.

Thall, P.F. and Vail, S.C. 1990. “Some Covariance Models for Longitudinal Count Data with Overdispersion.” *Biometrics,* 46:657-671.

Winkelmann, R. 2008. *Econometric Analysis of Count Data.* Springer: Berlin.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the corresponding author at:

Darcy Steeg Morris
U.S. Census Bureau
301-763-3989
darcy.steeg.morris@census.gov