A Regression Model for Count Data with Observation-Level Dispersion

Kimberly F. Sellers[1] and Galit Shmueli[2]

[1] 306 St. Mary's Hall; Department of Mathematics; Georgetown University; Washington, DC 20057; kfs7@georgetown.edu

[2] Dept of Decision, Operations & Information Technologies; 4361 Van Munching Hall; Smith School of Business; University of Maryland; College Park, MD 20742; gshmueli@rhsmith.umd.edu

Abstract: While Poisson regression is a popular tool for modeling count data, it is limited by its associated model assumptions. One assumption is that the response variable follows a Poisson distribution. However, over- or under-dispersion are common in practice and are not accommodated by Poisson regression. In addition, the dispersion is assumed fixed across observations, whereas in practice dispersion may vary across groups or according to some other factor. Recently, Sellers and Shmueli (2008) introduced the Conway-Maxwell-Poisson (CMP) regression, based on the CMP distribution. CMP regression generalizes both Poisson and logistic regression models and allows for over- or under-dispersed count data. The model structure introduced, however, assumes a fixed dispersion level across all observations. In this paper, we extend the CMP regression model to account for observation-level dispersion. We discuss model estimation, inference, diagnostics, and interpretation, and present a variable selection technique. We then compare our model to several alternatives and illustrate its advantages and usefulness using datasets with varying types and levels of dispersion.

Keywords: Conway-Maxwell Poisson distribution; generalized linear models (GLM); generalized Poisson; observation-level (varying) dispersion.

1 Introduction

Poisson regression models are most widely used to model relationships in count data; however, the model assumption [$\text{Var}(Y_i) = \text{E}(Y_i)$] is limiting. More generally, data exhibit over- or under-dispersion. Several papers offer ways to circumvent this problem, most of which focus on addressing the matter of overdispersion (McCullagh and Nelder, 1997; Famoye, 1993). Recently, Sellers and Shmueli (2008) introduced the CMP regression model, based on the Conway-Maxwell-Poisson (CMP) distribution, which allows handling over- and under-dispersed data. CMP regression also generalizes Poisson regression and logistic regression. Although it offers flexibility in terms of dispersion, it assumes that the associated level of dispersion is constant across all observations. We term this model "constant-dispersion

CMP regression". In this paper, we propose an extension of the CMP regression model which allows for observation-level dispersion. We start by describing the CMP distribution in Section 1.1, and then the constant-dispersion regression model by Sellers and Shmueli (2008) in Section 1.2. Approaching the CMP distribution from a GLM perspective, we use $\log \lambda$ as the link function and show its benefits with regard to estimation and inference. Section 2 introduces "observation-level-dispersion CMP regression", generalizing the ideas expressed in the previous section to consider a variable dispersion parameter and thus modeling the dispersion as a function of the explanatory variables. In Section 3 we describe a hypothesis testing procedure to determine the appropriateness of using a CMP regression model that allows for constant- or observation-level dispersion, and consider a variable selection construct that accounts for all subsets of main effects associated with the data relationship and the associated dispersion. Section 4 illustrates the application of the observation-level dispersion CMP regression. We compare the results of the constant- versus observation-level dispersion CMP regression in terms of fit, inference, and interpretation, and discuss the variable selection results as they relate to the examples provided.

1.1 The CMP Distribution

The CMP probability distribution function takes the form

$$P(Y_i = y_i) = \frac{\lambda^{y_i}}{(y_i!)^{\nu} Z(\lambda_i, \nu)}, \quad y_i = 0, 1, 2, \ldots, \quad i = 1, \ldots, n, \qquad (1)$$

for a random variable Y_i, where $Z(\lambda_i, \nu) = \sum_{s=0}^{\infty} \frac{\lambda_i^s}{(s!)^{\nu}}$. In this setting, $\lambda_i = E(Y_i^{\nu})$, while ν is the dispersion parameter. The CMP distribution includes three well-known distributions as special cases: Poisson ($\nu = 1$), geometric ($\nu = 0, \lambda_i < 1$), and Bernoulli $\left(\nu \to \infty \text{ with probability } \frac{\lambda_i}{1+\lambda_i}\right)$. See Shmueli et al. (2005) for details regarding this distribution.

1.2 CMP Model estimation with constant dispersion

Sellers and Shmueli (2008) took a GLM approach and used the link function $\eta(E(Y)) = \log \lambda$ that indirectly models the relationship between $E(\mathbf{Y})$ and $\mathbf{X}\beta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, that allows for estimating β and constant ν via the associated normal equations. Using the Poisson estimates, $\beta^{(0)}$ and $\nu^{(0)} = 1$ (or $\gamma^{(0)} = \ln \nu^{(0)} = 0$) as the starting values, these equations can be solved iteratively via an appropriate iterative reweighted least squares procedure to determine the maximum likelihood estimates for β and ν (or β and γ, respectively). The associated standard errors of the estimated coefficients are derived using the Fisher Information matrix; see Sellers

and Shmueli (2008) for details. *R* code for estimating the CMP regression coefficients and standard errors under the constant dispersion assumption is available at `www9.georgetown.edu/faculty/kfs7/research`.

2 CMP Model estimation with observation-level dispersion

We now allow for the dispersion parameter, ν_i, to vary with observation i, and consider a relationship between ν_i and the observations encapsulated in the $(p+1)$-dimensional row vector, $\mathbf{X}_i$. Accordingly, we write the log-likelihood for observation i as

$$\log L_i(\lambda_i, \nu_i | y_i) = y_i \log \lambda_i - \nu_i \log y_i! - \log Z(\lambda_i, \nu_i), \tag{2}$$

where

$$\log \lambda_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \doteq \mathbf{X}_i \beta, \text{ and} \tag{3}$$

$$\log \nu_i = \gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_p x_{ip} \doteq \mathbf{X}_i \gamma. \tag{4}$$

Since the CMP distribution belongs to the exponential family, we can determine appropriate normal equations for β and γ. Using the Poisson estimates, $\beta^{(0)}$ and $\gamma^{(0)} = 0$, as starting values, coefficient estimation can again be achieved via an appropriate iterative reweighted least squares procedure, or by using existing nonlinear optimization tools (e.g., `nlm` in *R*) to directly maximize the likelihood function. The associated standard errors of the estimated coefficients are derived in an analogous manner to that described in Sellers and Shmueli (2008).

3 Testing for Variable Dispersion, and Performing Variable Selection

Sellers and Shmueli (2008) established a hypothesis testing procedure to determine the need for using a CMP regression model over a simple Poisson regression model. In other words, they test whether $\nu = 1$ or not. We now ask the follow-up question: is the dispersion level fixed across observations, or is it dependent on one or more of the p covariates? More formally, we consider the set of hypotheses:

$$\begin{aligned} H_0 &: \gamma_i = 0 \text{ for } i = 1, \ldots, p \quad \text{vs.} \\ H_1 &: \gamma_i \neq 0 \text{ for at least one } i \in \{1, \ldots, p\}. \end{aligned} \tag{5}$$

The likelihood ratio statistic Λ and the derived test statistic C are given by

$$\Lambda = \frac{L\left(\hat{\beta}_{(0)}, \hat{\gamma}_{0(0)}\right)}{L\left(\hat{\beta}, \hat{\gamma}\right)} \tag{6}$$

$$C = -2\log\Lambda \quad = \quad -2\left[\log L\left(\hat{\beta}_{(0)}, \hat{\gamma}_{0(0)}\right) - \log L\left(\hat{\beta}, \hat{\gamma}\right)\right], \qquad (7)$$

where $\hat{\gamma}_{i(0)} = 0$ for $i = 1, \ldots, p$. $\hat{\beta}_{(0)}, \hat{\gamma}_{(0)}$ where $\nu_{(0)} = \exp(\mathbf{X}\gamma_{(\mathbf{0})})$ are the maximum likelihood estimates obtained under H_0, i.e. they are the CMP estimates under the constant dispersion model; and $(\hat{\beta}, \hat{\gamma})$ are the maximum likelihood estimates under the variable dispersion model, obtained by Equations (3) and (4). Under the null hypothesis, C has an approximate χ^2 distribution with p degree of freedom. Therefore, we reject H_0 in favor of H_1 when $C > \chi_\alpha(p)$.
We have also created a variable selection procedure that considers all possible subsets and the associated Akaike Information Criterion corrected for small sample sizes, when necessary (AICc). Thus, we can use the AIC or AICc to determine the "best" subset for predicting the response from a set of predictors, whether using the constant or observation-level dispersion model framework.

4 Examples

We compare the constant and observation-level CMP regression models to datasets characterized by under- and over-dispersion. These datasets were analyzed in Sellers and Shmueli (2008), comparing the constant-dispersion CMP regression results to those from other potential regression models for the data, including Poisson, negative binomial (NB), linear with $\log(Y)$, restricted generalized Poisson (Famoye, 1993). In addition, we illustrate the variable selection procedure by applying it to the example datasets. We show how the procedure can be used to find the optimal subset of predictors for predictive purposes, as well as shedding light on the effect of different predictors on the dispersion.

4.1 An Under-dispersed Dataset

We consider the airfreight breakage example from Kutner et al. (2003), where data are given on 10 air shipments, each carrying 1000 ampules on the flight. For each shipment i, we have the number of times the carton was transferred from one aircraft to another (X_i) and the number of ampules found broken upon arrival (Y_i). A graphical representation of the data is provided in Figure 1.
Figure 1 illustrates the (potentially observation-level) dispersion present in the count data. Thus, we consider regression models that allow for a constant dispersion or an observation-level dispersion structure. Table 1 contains the parameter estimates determined under the respective models. The standard errors associated with the respective estimates, however, bring question to the statistically significant difference between the two models. We see that the estimates for γ_0 are somewhat similar between

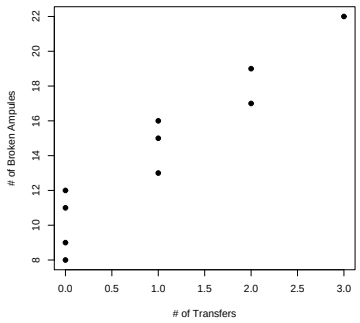

FIGURE 1. Scatter plot associated with airfreight breakage data.

TABLE 1. Estimated coefficients and standard errors (in parentheses) for the CMP regression models assuming fixed and variable dispersion for Airfreight example

Dispersion	$\hat{\beta}_0\,(\hat{\sigma}_{\hat{\beta}_0})$	$\hat{\beta}_1\,(\hat{\sigma}_{\hat{\beta}_1})$	$\hat{\gamma}_0\,(\hat{\sigma}_{\hat{\gamma}_0})$	$\hat{\gamma}_1\,(\hat{\sigma}_{\hat{\gamma}_1})$
Constant	13.8247 (6.2369)	1.4838 (0.6888)	1.7547 (0.954)	
Obs-level	15.5851 (0.7190)	4.6267 (0.3617)	1.8928 (0.3228)	0.1205 (0.1614)

the two models, and the estimate for γ_1 (in the observation-level dispersion model) has an associated standard error that allows for inclusion of 0 in the resulting confidence interval.

The dispersion test as described in Sellers and Shmueli (2008) yields a test statistic of $C = 9.10$ and associated p-value of 0.003, thus illustrating strong data (under-)dispersion and, therefore, a need to model the dataset with a more accommodating regression model, namely a CMP regression. Meanwhile, the hypothesis test for constant versus variable dispersion yields a test statistic of $C = 2.59$ and associated p-value of 0.11. Thus, one can argue that the dispersion does not statistically significantly vary by the number of aircraft transfers. We must, however, note the marginal p-value, and take the small sample size associated with this example into account. Thus, we consider variable selection under the observation-level dispersion structure to consider a broader realm of models.

Table 2 contains the variable selection results considering all relevant subsets for β and γ. Whether using the AIC or AIC$_c$ result, we see that the best predictive models explain the number of broken ampules via the number of flight transfers. Recalling the marginal statistical insignificance noted in the above hypothesis test regarding the constant versus observation-level

TABLE 2. Variable selection results for airfreight example. We consider model selection with an observation-level dispersion assumption. The AIC and AIC_C values are provided for all models under consideration.

β_0	β_1	γ_0	γ_1	AIC	AIC_C
0.000	-0.019	0.000	-3.938	179.567	181.281
2.115	0.000	-0.223	0.000	60.897	62.611
13.825	1.484	1.755	0.000	43.290	**47.290**
13.294	0.000	1.711	-0.093	44.637	48.637
15.585	4.627	1.893	0.120	**42.695**	50.695

dispersion, this dataset's small sample size further accents this result. The observation-level dispersion model is found to be optimal via AIC, while the constant dispersion model is considered the best according to the corrected AIC result (AIC_c).

Given the small sample size here, the corrected AIC is more appropriate in this setting. Focusing our attention accordingly we see that, while the constant dispersion model produces the smallest AIC_c, the observation-level dispersion models produce associated AIC_c values that are close relative to that produced by the constant dispersion model that describes the number of broken ampules in relation to the number of freight transfers. Analogously, in consideration of the AIC results, we again see these three models all producing relatively similar AIC values. Thus, while the results in Table 2 imply that the constant dispersion model is optimal, one can argue for more data to better analyze this relationship.

4.2 An Over-dispersed Dataset

Lord et al. (2008) model crash data in 1995 at 868 signalized intersections located in Toronto, Ontario using a Bayesian formulation of a CMP regression for modeling the relationship between traffic variables and motor vehicle crashes. Sellers and Shmueli (2008) compare their CMP regression results to those of Lord et al. (2008) to find that the parameter estimates are identical under the constant dispersion construct, and compare them to other potential regression models. Meanwhile, Table 3 contains the resulting parameter estimates under the constant and variable dispersion models for comparison. We see here that the corresponding estimates for β and γ do not appear to be statistically significantly different, given their associated standard errors. We will pursue this hypothesis further in the hypothesis test for the existence of statistically significant variable dispersion.

The constant dispersion test yielded a test statistic of 518.37 with an associated p-value < 0.001, thus noting the significant dispersion that exists in the data and thus the need to perform a CMP regression as opposed to a classical Poisson approach. Meanwhile, the test for variable dispersion yielded a test statistic of 1.02 with an associated p-value equaling 0.60.

TABLE 3. Estimated coefficients and standard errors (in parentheses) for the CMP regression models assuming fixed and variable dispersion

Dispersion	$\hat{\beta}_0\,(\hat{\sigma}_{\hat{\beta}_0})$	$\hat{\beta}_1\,(\hat{\sigma}_{\hat{\beta}_1})$	$\hat{\beta}_2\,(\hat{\sigma}_{\hat{\beta}_2})$
Constant	-4.0862726 (0.2619)	0.2290205 (0.0216)	0.2762342 (0.0161)
Obs-level	-3.8390 (0.3858)	0.2340 (0.0661)	0.2444 (0.0324)
Dispersion	$\hat{\gamma}_0\,(\hat{\sigma}_{\hat{\gamma}_0})$	$\hat{\gamma}_1\,(\hat{\sigma}_{\hat{\gamma}_1})$	$\hat{\gamma}_2\,(\hat{\sigma}_{\hat{\gamma}_2})$
Constant	-1.0522 (-3.8714)		
Obs-level	-0.7849 (0.2447)	0.0096 (0.0271)	-0.0385 (0.0064)

TABLE 4. Variable selection results for Toronto crash example. Note that we consider model selection with a constant dispersion assumption because the associated hypothesis test determined that a model with constant dispersion is adequate.

β_0	β_1	β_2	ν	AIC
0.055	0.000	0.000	0.049	6008.582
-1.806	0.000	0.262	0.271	5216.373
-1.527	0.166	0.000	0.094	5797.992
-4.086	0.229	0.276	0.349	**5072.950**

Given the large sample size, this demonstrates that the dispersion does not statistically significantly vary over predictor levels, and thus the assumption of a constant dispersion parameter is reasonable.
Because the constant dispersion model is sufficient, we pursue the question of variable selection under the constraint of constant dispersion. Table 4 provides the resulting parameter estimates and associated AIC for all possible subsets for consideration. As a result, the full model appears to provide the optimal choice for predicting the number of motor vehicle crashes, as demonstrated by the smallest AIC. All of the models have $\nu < 1$, which reflects the data overdispersion.

4.3 Summary

We have illustrated how the observation-level dispersion CMP model can be fitted to datasets with varying levels of dispersion. In both cases, statistical testing indicated that a constant dispersion level across all observations is better than a model that varies the dispersion level based on the covariates. We then used model selection to detect the predictor combination that yields the best predictive model. For the first example, this proved to be quite interesting because the marginal p-value associated with the hypothesis test followed with variable selection options that provided close AIC_c results where constant- or observation-level dispersion models could seem reasonable.

References

Famoye, F. (1993) Restricted generalized Poisson regression model. *Communications in Statistics - Theory and Methods*, **22**(5), 1335-1354.

Guikema, S. D. and Coffelt, J. P. (2008) A flexible count data regression model for risk analysis. *Risk Analysis*, **28**(1), 213-223.

Lord, D., Guikema, S. D., and Geedipally, S. R. (2008) Application of the Conway- Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis & Prevention*, **40**(3), 1123-1134.

Kutner, M.H., Nachtsheim, C.J., and Neter, J. (2003) *Applied Linear Regression Models*, 4th edition. McGraw-Hill.

McCullagh, P. and Nelder, J. A. (1997) *Generalized Linear Models*, 2nd edition. Chapman & Hall/CRC.

Puig, P. and Valero, J. (2006) Count Data Distributions: Some Characterizations with Applications, Journal of the American Statistical Association, **101** (473), 332-340.

Sellers, K. F., and Shmueli, G. (2008) A Flexible Regression Model for Count Data. Robert H. Smith School Research Paper No. RHS 06-061. Available at SSRN: http://ssrn.com/abstract=1127359

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.