A Flexible Regression Model for Count Data

Kimberly F. Sellers
Department of Mathematics, Georgetown University

Galit Shmueli
University of Maryland - Department of Decision, Operations & Information Technologies

Decision, Operations & Information Technologies Department
Robert H. Smith School of Business
University of Maryland
College Park, MD 20742-1815

This Paper can be downloaded without charge from the Social Science Research Network Electronic Paper Collection
http://ssrn.com/abstract=1127359
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

Abstract

Poisson regression is a popular tool for modeling count data and is applied in a vast array of applications from the social to the physical sciences and beyond. Real data, however, are often over- or under-dispersed and, thus, not conducive to Poisson regression. We propose a regression model based on the Conway-Maxwell-Poisson (CMP) distribution to address this problem. The CMP regression generalizes the well-known Poisson and logistic regression models, and is suitable for fitting count data with a wide range of dispersion levels. With a GLM approach that takes advantage of exponential family properties, we discuss model estimation, inference, diagnostics, and interpretation, and present a test for determining the need for a CMP regression over a standard Poisson regression. We compare the CMP to several alternatives and illustrate its advantages and usefulness using four datasets with varying dispersion.

Keywords: Conway-Maxwell Poisson distribution, dispersion, generalized linear models (GLM), generalized Poisson
1 Introduction

Regression models are the most popular tool for modeling the relationship between a response variable and a set of predictors. In many applications, the response variable of interest is a count, i.e. takes on non-negative integer values. For count data, the most widely used regression model is Poisson regression while, for binary data, the logistic (or probit) regression is most applied. Poisson regression is limiting in its variance assumption, namely that \( \text{Var}(Y) = E(Y) \). Even with the best of intent, however, count data often demonstrate over- or under-dispersion compared to the Poisson model.

One way to model over-dispersed count data is to use mixture models. An example is the gamma-Poisson mixture, where Poisson variables have means \( \mu_i \) that follow a gamma distribution. This yields a negative binomial marginal distribution of the form

\[
P(Y_i = y_i \mid \mu_i, r) = \left( \frac{r}{r + \mu_i} \right)^r \frac{\Gamma(r + y_i)}{\Gamma(y_i + 1) \Gamma(r)} \left( \frac{\mu_i}{r + \mu_i} \right)^{y_i}, \quad y_i = 0, 1, 2, \ldots
\]

The negative binomial likelihood can be expressed in the form of a generalized linear model for constant \( r \), and a log-link function (log \( \mu_i = \beta' X_i \)) is typically used. Although negative binomial regression is available in many statistical software packages, it is limited to modeling only over-dispersed data. In addition to its inability to fit under-dispersed data, McCullagh and Nelder (1997) note that this procedure is “an unpopular option with a problematic canonical link”.

An alternative model which can capture both over- and under-dispersion is the restricted generalized Poisson regression (RGPR) model by Famoye (1993). The model is given by

\[
P(Y_i = y_i \mid \mu_i, \alpha) = \left( \frac{\mu_i}{1 + \alpha\mu_i} \right)^{y_i} \frac{(1 + \alpha y_i)^{\mu_i - 1}}{y_i!} \exp \left( -\frac{-\mu_i(1 + \alpha y_i)}{1 + \alpha\mu_i} \right), \quad y_i = 0, 1, 2, \ldots
\]

where log \( \mu_i = \beta' X_i \). It is called a “restricted” model, because the dispersion parameter \( \alpha \) is restricted to \( 1 + \alpha\mu_i > 0 \) and \( 1 + \alpha y_i > 0 \) (Cui et al., 2006). When \( \alpha = 0 \), the model reduces to the Poisson case; \( \alpha > 0 \) indicates over-dispersion; and \( -2/\mu_i < \alpha < 0 \) indicates
under-dispersion. While this model allows for under- or over-dispersion in the data (albeit a limited degree of under-dispersion), it belongs to an exponential family only for a constant dispersion parameter, $\alpha$. Thus, a more general model with observation-specific dispersion ($\alpha_i$) will no longer belong to the exponential family. In short, for count data that are not binary nor follow a Poisson distribution, readily available, computationally efficient, flexible regression models are scarce. The need for such a model exists in many fields where count models are routinely fit to an array of datasets of varying dispersion.

In this paper, we propose using a more general count distribution that captures a wide range of dispersion and includes, as special cases, the popular Poisson and binary cases. A two-parameter generalized form of the Poisson distribution, called the Conway-Maxwell-Poisson (CMP) distribution (Shmueli et al., 2005), is sufficiently flexible to describe a wide range of count data distributions. It includes as special cases the Poisson, Bernoulli, and geometric distributions, as well as distributions with dispersion levels between these three well-known cases (governed by the dispersion parameter). The CMP distribution belongs to the exponential family and therefore possesses advantages in terms of estimation, conjugate priors, etc. These advantages have proven useful in several applications, such as using the CMP sufficient statistics for purposes of data disclosure (Kadane et al., 2006), in marketing applications (Boatwright et al., 2003; Borle et al., 2005), and online auctions (Borle et al., 2006). We describe the CMP distribution in Section 2.

In Section 3, we use the CMP distribution to formulate a regression model. We discuss model estimation, inference, interpretation, and diagnostics; obtaining fitted values; and testing for dispersion. A Bayesian regression formulation using CMP has been used in a marketing applications by Borle et al. (2005, 2006, 2007); Boatwright et al. (2003); Kalyanam et al. (2007). In each of these studies $\log(\lambda)$ was modeled as a linear function of predictors, and MCMC was used for estimation. Each of the datasets included a few thousand observations. For each model estimation time was between 2-24 hours. Lord et al. (2008), motivated by traffic modeling, used a slightly different Bayesian formulation.
with \( \log(\lambda^{1/\nu}) \) as the link function. They too use MCMC with long estimation time. prior
information and yields a good fit, the formulation used does not take full advantage of the
exponential family features of the CMP distribution and, in particular, requires computa-
tionally expensive MCMC for estimation. We, instead, approach the CMP distribution
from a GLM perspective, carefully choosing a link function (namely \( \log \lambda \)) that is advan-
tageous in terms of estimation, inference, and diagnostics. Our formulation also creates
a generalization of the ordinary Poisson regression as well as logistic regression, thereby
including and bridging two very popular and well-understood models.

To show the practical usefulness of the CMP regression, we compare its performance to
a few alternative regression models: Poisson, negative binomial, linear, logistic, and RGPR.
Section 4 considers three datasets of different size and with different levels of dispersion.
Using these data, we illustrate the advantages of the CMP model in terms of model fit,
inference, and wide applicability. In Section 5 we consider the Lord et al. (2008) motor
vehicle accidents example. We compare the five models along with our CMP formulation
to the Bayesian formulation. Section 6 concludes with discussion and future directions.

2 The Conway-Maxwell-Poisson (CMP) Distribution

The CMP probability distribution function (Shmueli et al., 2005) takes the form

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

for a random variable \( Y_i \), where \( Z(\lambda_i, \nu) = \sum_{s=0}^{\infty} \frac{\lambda_i^s}{(s!)^\nu} \) and \( \nu \geq 0 \). In this setting, \( \lambda_i \) denotes
the expected value under a Poisson distribution associated with observation \( i \), while \( \nu \) is
the dispersion parameter. The ratio between the probabilities of two consecutive values is
then \( \frac{P(Y_i = y_i - 1)}{P(Y_i = y_i)} = \frac{y_i^{\nu}}{\lambda_i} \). The CMP distribution generalizes the Poisson distribution in that
the ratio is not necessarily linear in \( y_i \), thereby leading to longer or shorter tails for the dis-
tribution. The CMP distribution includes three well-known distributions as special cases:
Poisson \((\nu = 1)\), geometric \((\nu = 0, \lambda_i < 1)\), and Bernoulli \((\nu \to \infty \text{ with probability } \frac{\lambda_i}{1+\lambda_i})\).

In Shmueli et al. (2005), the moments are given in the form

\[
E(Y_i^{r+1}) = \begin{cases} 
\lambda_i E^1(Y_i + 1) & \text{if } r = 0 \\
\lambda_i \frac{\partial}{\partial \lambda_i} E(Y_i^r) + E(Y_i)E(Y_i^r) & \text{if } r > 0
\end{cases}
\]

and the expected value is approximated by

\[
E(Y_i) = \lambda_i \frac{\partial \log Z(\lambda_i, \nu)}{\partial \log \lambda_i} \approx \lambda_i^{1/\nu} - \frac{\nu - 1}{2\nu}. \tag{2}
\]

Note that the expected value and variance can also be written in the form

\[
E(Y_i) = \frac{\partial \log Z(\lambda_i, \nu)}{\partial \log \lambda_i} \tag{3}
\]

\[
\text{Var}(Y_i) = \frac{\partial E(Y_i)}{\partial \log \lambda_i} \tag{4}
\]

We apply the results from Equations (3) and (4) to formulate the estimating equations (Appendix A) and the Fisher Information matrix (Section 3).

3 Regression Formulation

Our proposed CMP regression formulation begins as a generalization of an ordinary Poisson regression. McCullagh and Nelder (1997) view Poisson regression as a special case of loglinear models taking the form

\[
\log E(Y_i) = \log \mu_i = \eta_i = \beta'X_i = \beta_0 + \beta_1X_{i1} + \cdots + \beta_pX_{ip}, \quad i = 1, \ldots, n,
\]

where \(\text{Var}(Y_i) = \sigma^2 E(Y_i)\), and where \(\sigma^2\) denotes the dispersion parameter \([\sigma^2 > 1 \text{ (1)}\) for over- (under) dispersion]. Further, they argue that the link function is more important than the variance assumption. We will show that, while in some cases dispersion might not significantly affect mean predictions, it does affect the conditional distributions and can
affect inference. We can write a similar approximate type of relationship between the mean and variance via the CMP distribution. Using Equations (1)-(2), we write (suppressing subscript $i$)

$$\text{Var}(Y) = E(Y^2) - E^2(Y)$$

$$= \left[ \lambda \frac{\partial}{\partial \lambda} E(Y) + E^2(Y) \right] - E^2(Y) = \lambda \frac{\partial}{\partial \lambda} E(Y)$$

$$\approx \lambda \frac{\partial}{\partial \lambda} \left( \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \right) = \frac{1}{\nu} \lambda^{1/\nu} \approx \frac{1}{\nu} E(Y),$$

in accordance with McCullagh and Nelder (1997). Thus, we can see the relationship between $\nu$ (or $\frac{1}{\nu}$) and the direction of data dispersion.

In the following, we take a more direct approach to modeling the dispersion by extending the GLM formulation to the CMP case and modeling the relationship between $Y_i$ and the predictors $X_i$ via a function of $E(Y_i)$, $i = 1, \ldots, n$. Although typical link functions are direct functions of $E(Y_i)$ [e.g., $E(Y_i)$, $\log E(Y_i)$, logit($E(Y_i)$)], the most natural link function for a CMP regression is $\eta(E(Y)) = \log \lambda$. This function models the relationship between $E(Y)$ and $X$ indirectly. This choice of function is useful for two reasons. First, it coincides with the link function in two well-known cases: in Poisson regression, it reduces to $E(Y) = \lambda$; in logistic regression, where $p = \frac{\lambda}{1+X}$, it reduces to logit$(p) = \log \lambda$. The second advantage of using $\log \lambda$ as the link function is that it leads to elegant estimation, inference, and diagnostics. This result highlights the lesser role that the conditional mean plays when considering count distributions of a wide variety of dispersion levels. Unlike Poisson or linear regression, where the conditional mean is central to estimation and interpretation, in the CMP regression model, we must take into account the entire conditional distribution.

3.1 Model estimation

We write the log-likelihood for observation $i$ as

$$\log L_i(\lambda_i, \nu|y_i) = y_i \log \lambda_i - \nu \log y_i! - \log Z(\lambda_i, \nu).$$

(5)
Summing over \( n \) observations, the log-likelihood is given by

\[
\log L = \sum_{i=1}^{n} y_i \log \lambda_i - \nu \sum_{i=1}^{n} \log y_i! - \sum_{i=1}^{n} \log Z(\lambda_i, \nu).
\] (6)

Maximum likelihood coefficient estimates can be obtained by directly maximizing Equation (6) under the constraint \( \nu \geq 0 \), using a constrained nonlinear optimization tool (e.g., \texttt{nlminb} in \( R \)). One alternative is to write the log-likelihood as a function of \( \log \nu \), i.e.

\[
\log L = \sum_{i=1}^{n} y_i \log \lambda_i - e^{\log \nu} \sum_{i=1}^{n} \log y_i! - \sum_{i=1}^{n} \log Z(\lambda_i, e^{\log \nu}),
\]

and then maximize it using an ordinary nonlinear optimization tool (e.g., \texttt{nlm} in \( R \)). Unconstrained optimization is a better option because it is computationally easier and more widely available in software packages. A third option for obtaining the maximum likelihood estimates is to use the GLM framework to formulate the likelihood maximization as a weighted least squares procedure (see Appendix A) and to solve it iteratively.

The GLM formulation is also used for deriving standard errors associated with the estimated coefficients. The latter are derived using the Fisher Information matrix. For estimating \( \beta \) and \( \nu \), we have a block Information matrix of the form

\[
I = \begin{pmatrix}
I^\beta & I^{\beta, \nu} \\
I^{\beta, \nu} & I^\nu
\end{pmatrix},
\] (7)

where \( I^\beta \) pertains to the estimated variances and covariances of \( \hat{\beta} \), \( I^\nu \) contains the estimated variance for \( \hat{\nu} \), and \( I^{\beta, \nu} \) contains the componentwise estimates of the covariance between \( \hat{\beta} \) and \( \hat{\nu} \). Details regarding the information matrix components are provided in Appendix B. \( R \) code for estimating CMP regression coefficients and standard errors is available at \texttt{www9.georgetown.edu/faculty/kfs7/research}. 

7
3.2 Testing for Dispersion

How much data dispersion should exist to warrant deviation from Poisson regression? The set of hypotheses $H_0 : \nu = 1$ vs. $H_1 : \nu \neq 1$ ask whether the use of Poisson regression is reasonable versus the alternative of fitting CMP regression. Note that $H_1$ does not specify the dispersion direction (over vs. under) in the data. This can easily be assessed, however, via exploratory data analysis and the dispersion estimate, $\hat{\nu}$, from the fitted CMP regression.

We derive the test statistic, 
\[ C = -2 \log \Lambda = -2 \left[ \log L \left( \hat{\beta}^{(0)}, \hat{\nu} = 1 \right) - \log L \left( \hat{\beta}, \hat{\nu} \right) \right], \]
where $\Lambda$ is the likelihood ratio test statistic, $\hat{\beta}^{(0)}$ are the maximum likelihood estimates obtained under $H_0 : \nu = 1$ (i.e., the Poisson estimates); and $(\hat{\beta}, \hat{\nu})$ are the maximum likelihood estimates under the general state space for the CMP distribution. Under the null hypothesis, $C$ has an approximate $\chi^2$ distribution with 1 degree of freedom. For small samples, the test statistic distribution can be estimated via bootstrap.

3.3 Computing Fitted Values

Once a CMP regression model has been estimated, we can obtain fitted values ($\hat{y}_i$) in one of two ways:

1. Estimated means: We can use the approximation in Equation (2) and obtain fitted values by
\[ \hat{y}_i | x_i = \hat{\lambda}_i^{1/\hat{\nu}} - \hat{\nu} - 1 \]
where $\hat{\lambda}_i = \exp(x'_i \hat{\beta})$. Note that this approximation is accurate for $\nu \leq 1$ or $\lambda_i > 10^\nu$ (Minka et al., 2003).

2. Estimated medians: When the mean approximation is inadequate (or in general), we can obtain percentiles of the fitted distribution by using the inverse-CDF for $\hat{y}_i | x_i$ and $\hat{\nu}$. In particular, we use the estimated median to obtain fitted values.
3.4 Model Inference

Due to the GLM formulation, the statistical significance of individual predictors can be obtained by using the asymptotic standard normal distribution of $\frac{\hat{\beta}_j}{\hat{\sigma}_j}$. In the case of small samples, however, where the asymptotic normality might not hold (as in other count data regression models), bootstrap can be used to estimate the distributions of the coefficients of interest. With small samples, CMP model estimation is very fast, thereby being practically useful for bootstrap.

3.5 Coefficient Interpretation

When comparing coefficients from a CMP regression model to those from an ordinary Poisson model, we must first scale the CMP coefficients to that of the Poisson coefficients. A simple approximate transformation is to divide the coefficients by $\hat{\nu}$.

There are two main approaches for interpreting coefficients in regression models (Long, 1997). One examines changes in the conditional mean for a unit increase in a single predictor, e.g. $E(Y|X_j = x_j, X_{i \neq j} = x)$ and $E(Y|X_j = x_j + 1, X_{i \neq j} = x)$. In additive models, such as a linear regression, the difference between the two conditional means (or the derivative of $E(Y|X)$ with respect to $X_j$) is used for interpretation (“a unit increase in $X_j$ is associated with a $\beta_j$ increase in $E(Y)$”); in multiplicative models, such as the Poisson or logistic regressions, the ratio of the two conditional means is used for interpretation (“a unit increase in $X_j$ is associated with a factor of $e^{\beta_j}$ increase in $E(Y)$ or the odds”). The second approach, which is used for coefficient interpretation in other types of non-linear regression models (e.g., probit regression), is to directly examine the relationship between fitted values and changes in a predictor. This can be done via graphical plots for less than two predictors while, for more than two predictors, there are various solutions such as fitted value consideration at selected values of the predictors.

In the CMP regression case, we cannot use the first approach that compares conditional means, because the relationship between the conditional mean and the predictors is neither
additive nor multiplicative (except for the special cases of Poisson and logistic regressions).

To illustrate, consider the case with a single predictor (without loss of generality). Using
the approximation in Equation (2), the ratio of the conditional means is given by

$$\frac{E(Y|X_j = x_j + 1, X_{i\neq j} = x)}{E(Y|X_j = x_j, X_{i\neq j} = x)} = \frac{e^{(x'\beta + \beta_j)/\nu - \nu^{-1}/2\nu}}{e^{(x'\beta)/\nu - \nu^{-1}/2\nu}},$$

which leads to a complicated non-linear relationship between a unit increase in $X$ and the
effect on $E(Y|X)$. We can therefore use only the second approach for interpreting the effect
of individual predictors on the conditional mean (or median).

Note that we can also examine the effect of a unit increase of a predictor on the function
$\gamma_y = \frac{P(Y=y)}{P(Y=y-1)}$, which determines the tail of the distribution. The relationship between $\gamma_y$
and the predictors is multiplicative, namely

$$\frac{\gamma_y|X_j=x_j+1, X_{i\neq j}=x}{\gamma_y|X_j=x_j, X_{i\neq j}=x} = e^{\beta_j}. \quad (8)$$

The interpretation is therefore “a unit increase in $X_j$ is associated with a factor of $e^{\beta_j}$
increase in $\gamma_y$”. Note that, for the Poisson case, Equation (8) reduces to $\lambda_{x_j+1} = e^{\beta_j}\lambda_{x_j}$,
which is equivalent to the multiplicative effect of the predictor on the conditional mean.

In the logistic regression case, $\gamma_y$ is in fact equal to the odds \( \gamma_y = \frac{P(Y=1)}{P(Y=0)} \), thereby
providing the usual type of interpretation. Further, for the geometric regression case, $\gamma_y$
is equal to the probability of failure, thereby measuring the marginal impact of a predictor
on the probability of failure. Note that, although the “linear probability model” (where
the outcome variable is a probability) is inadequate for binary outcomes (see, e.g., Long,
1997), it is appropriate in the case of heavily over-dispersed data.

### 3.6 Model Diagnostics

Due to the GLM formulation, and in particular the IWLS framing (see Appendix A), stan-
dard GLM diagnostics can be used for residual analysis of a fitted CMP regression model.
We use the matrices $W$ and $X$ as defined in Appendix A for computing leverage, and the popular Pearson and Deviance residuals. Leverage can be computed from the hat matrix $H$, defined here as $H = W^{1/2}X(XX'X)^{-1}X'W^{1/2}$. An observation with an unusually high value of $h_i$ is suspect of having influence (although $H$, like other non-linear models, depends on the estimated parameters). Meanwhile, using ordinary GLM formulations, we can write the Pearson residual for observation $i$ (Davison and Tsai, 1992) as $r_{P,i} = \frac{Y_i - \hat{\mu}_i}{\sqrt{w_i(1-h_i)}}$, where $\hat{\mu}_i = E(Y_i)$, and the standardized deviance residual for observation $i$ can be written as $r_{D,i} = \text{sgn}(Y_i - \hat{\mu}_i) \frac{d_i}{\sqrt{1-h_i}}$, where $d_i = -2[\log L(\hat{\mu}_i; y_i; \hat{\nu}) - \log L(y_i; y_i; \hat{\nu})]$. These two types of residuals can be computed directly or approximated using the mean approximation in Equation (2). In particular, for deviance residuals, the approximation leads to

$$d_i = 2 \left( y_i \hat{\nu} \log \left( \frac{y_i + \frac{\hat{\nu} - 1}{2\hat{\nu}}}{\hat{\mu}_i + \frac{\hat{\nu} - 1}{2\hat{\nu}}} \right) + \log \left( \frac{Z \left( \left( \hat{\mu}_i + \frac{\hat{\nu} - 1}{2\hat{\nu}} \right)^{\hat{\nu}}, \hat{\nu} \right)}{Z \left( \left( y_i + \frac{\hat{\nu} - 1}{2\hat{\nu}} \right)^{\hat{\nu}}, \hat{\nu} \right)} \right) \right). \quad (9)$$

Equation (9)’s existence is constrained in that $Y > k$ for $\hat{\nu} < \frac{1}{2k+1}; k \in \mathbb{N}^+$. We can, however, modify Equation (9) in order to obtain valid results for $d_i$. For example, when $\nu < 1$ and $Y = 0$, we set $Z \left( \left( y_i + \frac{\hat{\nu} - 1}{2\hat{\nu}} \right)^{\hat{\nu}}, \hat{\nu} \right) = 1$. Another option is to use the exact deviance equations supplied above, though this is computationally more expensive. Finally, while the approximation is accurate for $\lambda > 10^\nu$ or $\nu < 1$, we have found that deviance residuals computed using Equation (9) are quite accurate even outside that range (e.g., for under-dispersed data with low counts).

A QQ plot of the deviance residuals as well as a log $\hat{\lambda}$ versus deviance residuals plot can assess model adequacy and detect outliers. Although normal probability plots are common, deviance residuals for non-linear models can be far from normally distributed (Ben and Yohai, 2004). One alternative is to ignore the fit to normality on the normal probability plot, and use it just to detect outliers. Another option is to use bootstrap to estimate the distribution of deviance residuals, and then to create a QQ plot of the deviance residuals against their estimated distribution.
4 Examples

In the following, we fit regression models to two authentic small datasets: one characterized by over-dispersion, and the other by under-dispersion. We also fit the models to a third dataset with binary outcomes (i.e. extreme under-dispersion). We fit five types of regression models which would be popular choices in cases of count data: Poisson, negative binomial (NB), linear with log\(Y\), restricted generalized Poisson (RGPR), and CMP. For the binary dataset we also fit a logistic regression. The goal of this section is to compare the CMP to the other models in terms of fit, inference, and flexibility. The small sample size and dimension of the first two datasets is useful for directly observing the effect of dispersion. In particular, we show the effect of dispersion (over or under) on the conditional distribution of fit. We evaluate goodness-of-fit by examining the fitted values and comparing values of MSE and \(\text{AIC}_C\) (the Akaike Information Criterion\(^1\) corrected for small sample size) across models.

Note that, except for the Poisson and logistic regressions, all other models considered have an extra dispersion parameter that is assumed fixed across observations, but unknown. Each of the models is estimated by maximum likelihood. Poisson, NB, linear and logistic regression are estimated using ordinary GLM functions in \(R\). CMP is estimated using non-linear optimization in \(R\), and standard errors are estimated as described in Section 3.1. RGPR is estimated using constrained non-linear optimization in \(R\) and standard errors are estimated as described in Famoye (1993).

4.1 Regression with Under-Dispersed Data: Airfreight Breakage

We first consider the airfreight breakage example from Kutner et al. (2003, p. 35, Exercise 1.21), where data are given on 10 air shipments, each carrying 1000 ampules on the flight.

\(^1\)All models aside from Poisson have a penalty term in the \(\text{AIC}_C\) that takes into account the extra dispersion parameter. In the linear regression model with log-transformed response (log\(Y\)), we scaled the \(\text{AIC}_C\) to reflect the non-transformed response \(Y\) by obtaining the transformed likelihood \(L(y) = \prod_{i=1}^{n} dnorm(\log(Y_i), \frac{1}{\sigma^2})\), where \(dnorm(·)\) is the normal density function calculated using the maximum likelihood estimates of \(\mu\) and \(\sigma\) obtained from the log-transformed model.
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 (\( Y_i \)). The data are provided in Appendix C.

|                | CMP Median Fit | Poisson Fit | Linear Reg Fit |
|----------------|----------------|-------------|----------------|
| AIC \( C \)    | 47.29          | 52.11       | 49.37          |
| MSE            | 1.900          | 2.210       | 2.363          |

Figure 1: Fitted mean curves (solid lines), 5th and 95th percentile curves (broken lines) for Poisson and CMP regression models for the airfreight breakage data (dots).

We first estimated the CMP regression coefficients and tested for dispersion. The estimated dispersion parameter is \( \hat{\nu} = 5.78 \), indicating under-dispersion. However, to test for dispersion we use bootstrap rather than the dispersion test, due to the small sample size. The 90% bootstrap confidence interval for \( \nu \) is (4.414, 20.643) indicating dispersion that requires a CMP regression instead of ordinary Poisson regression. We proceed by attempting to fit the five regression models. The estimated coefficients and standard errors for three of these models (linear, Poisson, and CMP) are given in Table 2. The remaining two
Table 2: Estimated coefficients and standard errors (in parentheses) for Airfreight example, for five types of regression models

| Model             | \(\hat{\beta}_0 (\hat{\sigma}_{\hat{\beta}_0})\) | \(\hat{\beta}_1 (\hat{\sigma}_{\hat{\beta}_1})\) |
|-------------------|-----------------------------------------------|-----------------------------------------------|
| Linear (log \(Y\)) \(\hat{\sigma} = 0.141\) | 2.3273 (0.0631) | 0.2800 (0.0446) |
| Poisson           | 2.3529 (0.1317) | 0.2638 (0.0792) |
| CMP \((\hat{\nu} = 5.7818, \hat{\sigma}_\nu = 2.597)\) | 13.8247 (6.2369) | 1.4838 (0.6888) |
| Neg-Bin (no convergence) | NA             | NA             |
| RGPR (no convergence) | NA             | NA             |

models could not be estimated: the negative binomial does not converge due to its inability to capture under-dispersion (in \(R\), it gives estimates identical to the Poisson estimates, an artificial condition imposed by the software in cases of under-dispersion). Similarly, RGPR did not converge and therefore no estimated model is produced. This highlights the limited ability of RGPR to fit under-dispersed data. In general for under-dispersed data, the RGPR probability function “gets truncated and does not necessarily sum to one” (Famoye et al., 2004). This example appears to fall exactly under this limitation.

Fitted values from the three models are provided in Appendix C where, for the CMP, we use the estimated conditional median for fitted values because the approximation (2) is likely to be inaccurate (here, \(\nu > 1\) and \(\lambda />10^\nu\)). We see that the models are similar in terms of the fitted values that they generate (see also Figure 1). In terms of MSE and AIC\(_C\), the CMP shows best fit, although the differences between models for these values are not large (see Table 1). The similarity of the three regression models is also in terms of the coefficient magnitudes (after dividing the CMP coefficients by \(\hat{\nu}\)). The models differ, however, in two important ways. First, although the fitted values are similar, the conditional distribution differs markedly across the models, as can be seen by comparing the 5th and 95th percentile curves in Figure 1. Secondly, the models initially appear to differ in terms of inference. Comparing the linear, Poisson, and CMP estimated models, we find that the ratio \(\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}\) is 6.28, 3.33, and 2.15 respectively. Due to the small sample size, however, the normal approximation might not be adequate. We therefore examined the distributions of \(\hat{\beta}_0\) and \(\hat{\beta}_1\) for each of the three models, based on 1,000 bootstrapped
samples. Figure 2 displays normal plots for the estimated coefficients. As expected, except for the linear regression estimates, the distributions for the other two models are skewed.

To evaluate statistical significance of the predictor (number of transfers), we examine the percent of the distribution of $\hat{\beta}_1$ to the left of the value $\beta_1 = 0$. In all three models, this percent is zero, indicating high statistical significance. Note that run time for estimating the CMP model for each sample was less than 5 seconds, thereby deeming bootstrap a useful approach in small sample inference.

In terms of model interpretation, the linear regression indicates that a unit increase in the number of transfers is associated with an increase of 28% in the average number of broken ampules. The Poisson indicates that such an increase is associated with a factor increase of 1.3 in the average number of broken ampules. Looking at Figure 1, however, shows that interpretations in term of the average number of broken ampules is insufficient. In particular, the number of transfers seems to affect the entire distribution of the number of broken ampules, as indicated by the fitted CMP model. Indeed, the CMP curves in Figure 1 can be used directly for interpreting the relationship between number of transfers
Finally, we examine leverage and scaled deviance residuals from each of the three models. Figure 3 displays scatterplots of the deviance residuals vs. the single predictor (which is equivalent to plotting vs. $\log \hat{\lambda}$ for the Poisson and CMP models or $\hat{y}$ for the linear regression), and probability plots. Overall, there is no noticeable pattern in any of the scatterplots. All three models indicate observation #5 (with $X = 3$) as suspect of being influential, and observation #7 as an outlier (having a large negative deviance residual), particularly for the CMP model.

### 4.2 Regression with extreme under-dispersion: Book Purchases

We now consider the case where the outcome variable is binary, and where typically a logistic regression would have been considered. We use a dataset from Lattin et al. (2003) that describes the results of a direct-marketing campaign by a book club, for a certain art book. The dataset contains the results for 1,000 customers. The outcome is whether the
customer purchased the art book or not. The two predictor variables are the number of months since the customer’s last purchase (\(Months\)), and the number of art books that the customer has purchased in the past (\(ArtBooks\)). We use this dataset to show the flexibility of the CMP regression over the alternatives discussed above. In particular, we show that the CMP regression produces estimates and predictions that are identical to those from a logistic regression, and that RGPR and NB fail to converge altogether.

Table 3 provides the parameter estimates from the Poisson, logistic, and CMP regression models, respectively. The negative binomial and RGPR are absent from Table 3 because their estimation does not converge. The negative binomial, as in the Airfreight example, cannot capture any type of under-dispersion. The RGPR, as shown above, has limited ability to capture under-dispersion, thus it fails to capture this level of under-dispersion.

| Model               | \(\hat{\beta}_0 (\hat{\sigma}_{\hat{\beta}_0})\) | \(\hat{\beta}_{Months} (\hat{\sigma}_{\hat{\beta}_{Months}})\) | \(\hat{\beta}_{ArtBooks} (\hat{\sigma}_{\hat{\beta}_{ArtBooks}})\) |
|---------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
| Poisson             | -2.29081621 (0.18085967)                        | -0.06127833 (0.01517751)                         | 0.725029 (0.05162616)                           |
| Logistic            | -2.22563552 (0.23892413)                        | -0.07071719 (0.01922971)                         | 0.98905218 (0.13466051)                         |
| CMP (\(\nu = 30.4\), \(\hat{\sigma}_{\nu} = 10123\)) | -2.22563552 (0.23892413)                        | -0.07071719 (0.01922971)                         | 0.98905218 (0.13466051)                         |

With respect to comparing CMP with logistic regression, it is clear that the two models produce identical results in terms of coefficients and standard errors. Furthermore, comparing fitted values (or predictions), using the estimated CMP median as the fitted value (in accordance with Section 3.3) yields values that are identical to those from a logistic regression with cutoff value 0.5. To obtain fits for other cutoff values, the corresponding percentile should be used. Finally, although the Poisson model does converge, it is clearly inappropriate in terms of inference, and produces fitted values that are not binary.

4.3 Regression with Over-dispersed Data: Textile Fabric

A popular example used to illustrate data over-dispersion examines the number of yarn breaks during a textile process run, and its association with the number of completed
spindle revolutions recorded (Hinde, 1982; Famoye, 1993). To maintain consistency in the model framework to that described in Famoye (1993), we let \( Y_i \) denote the number of respective faults in the \( i \)th roll of fabric, and \( \log(X_i) \) denote the log-length of the \( i \)th fabric roll.

The estimated coefficients and standard errors for the CMP compared to the linear, Poisson, negative binomial, and RGPR models are given in Table 5. The run time was less than 5 seconds for the CMP estimation. The estimated dispersion parameter (\( \hat{\nu} = .488 \)) implies possible over-dispersion (see Table 5). Due to small sample size, we use bootstrap to test for dispersion. The 90% bootstrap confidence interval for \( \nu \) is (0.317, 0.948), indicating significant dispersion and thus supporting the use of CMP regression instead of classical Poisson regression.

We compare the CMP, Poisson and linear regression models in terms of coefficients, fitted values, and inference. Note that, in this case, \( \hat{\nu} < 1 \) and therefore we obtain CMP fitted values via the mean approximation. As in the airfreight breakage example, the estimated regression coefficients (after scaling the CMP coefficients by dividing by \( \hat{\nu} \)) are similar across the models, as are the fitted values (provided in Appendix C; the CMP and Poisson fitted values are nearly identical, with slightly different fits by the linear regression model). We see again, however, a difference in the conditional distribution of fit as shown in Figure 4. In contrast to the under-dispersion case (i.e. the airfreight breakage example) where the CMP percentile curves were tighter around the mean curve compared to the Poisson percentile curves, here the under-dispersion results in the opposite effect: the CMP percentile curves are wider than the Poisson curves around the mean curve. This illustrates the effect of the type and magnitude of dispersion on the width of band formed by the percentile curves around the mean.

In terms of inference, the ratio \( \frac{\hat{\beta} \hat{\nu}}{\hat{\beta}} \) equals 4.7 for the linear regression, 5.7 for the Poisson case, and 3.0 for the CMP. Due to the small sample size, however, inference based on a normal approximation might not be adequate for any of the five models. Figure 5 dis-
Table 4: Goodness-of-Fit statistics for fabric example, comparing the CMP with alternative regression formulations

|                | CMP fit | Poisson fit | Linear Reg fit | Neg-Bin fit | RGPR fit |
|----------------|---------|-------------|----------------|-------------|----------|
| AIC            | 184.47  | 192.23      | 180.83         | 182.24      | 181.90   |
| MSE            | 21.41   | 21.59       | 23.12          | 21.77       | 21.81    |

Figure 4: Fitted mean curves (solid lines), 5th and 95th percentile curves (broken lines) for Poisson and CMP regression models for the textile fabric data (dots). Note that the Poisson and CMP mean curves nearly overlap.

Table 5: Estimated coefficients and standard errors for each of the regression models corresponding to the textile fabric example (Hinde, 1982; Famoye, 1993).

| Model             | $\hat{\beta}_0$ ($\hat{\sigma}_{\beta_0}$) | $\hat{\beta}_1$ ($\hat{\sigma}_{\beta_1}$) | $\frac{\hat{\sigma}_{\beta_1}}{\hat{\sigma}_{\beta_1}}$ |
|-------------------|------------------------------------------|------------------------------------------|----------------|
| Linear Reg (log Y) $\sigma = 0.519$ | -3.9452 (1.2686) | 0.9427 (0.2011) | 4.69          |
| Poisson           | -4.1741 (1.1352) | 0.9971 (0.1759) | 5.67          |
| Neg-Bin ($\hat{\lambda} = 8.67, \hat{\sigma}_{\lambda} = 4.17$) | -3.7959 (1.4576) | 0.9379 (0.2279) | 4.12          |
| RGPR ($\hat{\alpha} = 0.048, \hat{\sigma}_{\alpha} = 0.021$) | -3.7365 (1.4550) | 0.9284 (0.2279) | 2.03          |
| CMP ($\hat{\nu} = 0.456, \hat{\sigma}_{\nu} = 0.135$) | -2.2735 (0.9357) | 0.5069 (0.1737) | 2.92          |
plays normal probability plots of the estimated parameters for each of the fitted regression models, based on 1,000 bootstrap samples. We can see that the distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ indeed deviate from normality for all models except linear regression (as expected), and especially for RGPR and CMP. Note that, for 132 of the 1,000, the RGPR estimate did not converge. To evaluate statistical significance of the predictor, we examine the percent of the distribution of $\hat{\beta}_1$ to the left of $\beta_1 = 0$. In all models, this percent is zero or very close to zero (for RGPR), indicating high significance.

Finally, we examine residuals and leverage measures to evaluate model fit and detect influential observations. Deviance residuals here can be computed either directly or using the mean approximation; both yield nearly identical values. Figure 6 shows QQ plots of the scaled deviance residuals from each of the five models against bootstrap samples, as well as scatterplots of the deviance residuals vs. the predictor (log($\text{RollLength}$)). For RGPR, we use unscaled deviance residuals (as $H$ is unavailable). Poisson, linear regression, and

\footnote{Thus, the normal approximation used in Famoye (1993) for RGPR is inadequate.}
negative binomial models all single out the shortest two fabric rolls (obs. #22, #24; both have the lowest number of faults) as influential observations; In contrast, these observations are not flagged by CMP and RGPR. Note that the two observations are predicted fairly accurately by all models. Examining residuals, all five models flag the fabric roll with the largest number of faults (obs. #13) and a long fabric roll with very few faults (obs. #30). All except for RGPR flag another relatively short fabric roll with an excessive number of faults (obs. #26). These three observations are easily spotted in Figure 4.

5 Modeling Motor Vehicle Crashes

The previous section has shown the flexibility of the CMP regression to capture a wide range of dispersion levels, which exceed the ability of models such as the negative binomial and RGPR. We now examine a dataset used by Lord et al. (2008) which contains crash data in 1995 at 868 signalized intersections located in Toronto, Ontario. For each intersection,
measurements included the annual number of crashes at the intersection \((Y)\) and two traffic flow variables. See Lord et al. (2008) for further details on the data.

Because motor vehicle crash data contain counts, Poisson and negative binomial regressions are common models in the field of transportation safety. For the Toronto dataset, which display over-dispersion, Lord et al. (2008) proposed using a Bayesian CMP regression formulation. In particular, they used non-informative priors and modeled the effect of the two traffic variables on the number of crashes via the link function \(\log \left( \lambda^{1/\nu} \right) = X\beta\). Parameter estimation was then performed via MCMC. The authors note that estimation for this dataset used 35,000 replications, requiring nearly five hours of computation. Comparing goodness-of-fit and out-of-sample prediction measures, Lord et al. (2008) showed the similarity in performance of the CMP and negative binomial regression. They then motivate the advantage of the CMP over the negative binomial regression in the ability to fit under-dispersion and low counts.

The goal of this section is two-fold: (1) to extend the model comparison in Lord et al. (2008) beyond the negative binomial model to additional models, as well as to examine a wider range of model comparison aspects, and (2) to compare the Bayesian CMP formulation to our formulation and show the advantages gained by using our formulation. Although goodness-of-fit measures might indicate similarity of the CMP performance to other models, model diagnostics provide additional information.

### 5.1 Model Estimation

Six regression models were fit to the Toronto intersection crash data. Following Lord et al. (2008), the response was the number of crashes at the intersection (except for linear regression, where the response was log-transformed), and the two covariates were the two log-transformed traffic flow variables.

Table 6 displays the six estimated models: two CMP formulations (our model and the Bayesian model of Lord et al. (2008)), and four alternative regression models (Poisson,
Table 6: Estimated models: comparing two CMP formulations (ours and Lord et al. (2008)), and four alternative models for the Toronto crash data. For ease of comparison, we report the CMP estimates and standard errors from our formulation in terms of $\hat{\nu}$ multipliers, to reflect the comparable scale with estimates from the other models. For linear regression, 0.5 was added to all values of $Y$.

| Model                | Extra parameter | $\beta_0 (\hat{\sigma}_{\beta_0})$ | $\beta_1 (\hat{\sigma}_{\beta_1})$ | $\beta_2 (\hat{\sigma}_{\beta_2})$ |
|----------------------|-----------------|-------------------------------------|-------------------------------------|-------------------------------------|
| Our formulation      | $\hat{\nu}$=0.3492 (0.0208) | -11.7027 $\nu$ (0.7501$\nu$) | 0.6559$\nu$ (0.0619$\nu$) | 0.7911$\nu$ (0.0461$\nu$) |
| Lord et al. (2008)   | $\hat{\nu}$=0.3408 (0.0208) | -11.53 (0.4159) | 0.6350 (0.0474) | 0.7950 (0.0310) |
| Linear Reg (log $Y$) | $\hat{\sigma}$=0.3491 | -9.8132 (0.5161) | 0.5966 (0.0512) | 0.6566 (0.0226) |
| Poisson              | -10.2342 (0.2838) | 0.6029 (0.0288) | 0.7038 (0.0140) |
| Neg-Bin              | $\hat{r}$=7.154 (0.625) | -10.2458 (0.4626) | 0.6207 (0.0456) | 0.6853 (0.0215) |
| RGPR                 | $\hat{\alpha}$=0.050 (0.004) | -10.2357 (0.4640) | 0.6205 (0.0451) | 0.6843 (0.0215) |

Comparing the two CMP formulations, the two are nearly identical in terms of $\hat{\nu}$ and its standard error (or the equivalent posterior credible standard error for Lord et al. (2008)) and in terms of the $\hat{\beta}$ coefficients (after scaling by a factor of $\hat{\nu}$, due the different formulation of the relationship between the covariates and the response). These similarities between the Bayesian and classic formulations indicate that the prior information does not affect the model, here most likely due to the large size of the dataset. The most dramatic difference between the two implementations is in run time: our estimation took less than three minutes, compared to five hours required by the Bayesian MCMC. This difference has significance especially since Lord et al. (2008) used noninformative priors to obtain their estimates. Thus, in the absence of strong prior information or in the presence of a large dataset, our formulation provides more efficient estimation. Even in the presence of prior information, our method is still useful for obtaining initial estimates to speed up the MCMC process.
5.2 Model Fit

Comparing goodness-of-fit measures, the two CMP formulations are practically identical in terms of $\hat{\beta}$ and thus produce nearly identical fitted values. Compared to the other regression models, the CMP model has lower MSE and AIC values, indicating better fit (see Table 7). The CMP dispersion test (with $C=518$, and associated p-value=0) indicates that the CMP model is more adequate than Poisson regression.

Table 7: Goodness of fit comparison of CMP with alternative fitted models

|            | CMP   | Poisson | Linear Reg | Neg-Bin | RGPR |
|------------|-------|---------|------------|---------|------|
| AIC        | 5073  | 5589    | 5260       | 5077    | 5092 |
| MSE        | 32.57 | 32.60   | 36.45      | 32.70   | 32.71|

We now examine model diagnostics to better understand model fit. Figure 7 displays scatterplots of the scaled deviance residuals vs. $\log \hat{\lambda}$ (bottom panels) and probability plots for the scaled deviance residuals (top panels). For RGPR, we use unscaled deviance residuals (as $H$ is unavailable).

From the residual plots and the leverage measures, we find that the linear regression marks every observation with $Y = 0$ as influential, and overall has trouble mostly with low-count $Y$ values. The negative binomial model also marks nearly half of the $Y = 0$ observations as influential, and flags mostly high-count observations. The Poisson and negative binomial models mark the observations with largest $X$ values as influential. In contrast, CMP diagnostics point out eight observations with large residuals (#15, #42, #247, #424, #494, #618, #619, #757) and three with high leverage (#133, #801, #835). Three of the large-residual intersections have a large number of crashes with relatively little traffic (small values of the covariates). The remaining large-residual intersections have a small to medium number of crashes, but less substantial traffic on one of the traffic flow covariates. All of these observations are also flagged by at least one other regression method, with observations #15 and #618 being flagged by all methods.
Figure 7: QQ plots of the scaled deviance residuals (top panels) and scatterplots of the scaled deviance residuals vs. $\log \hat{\lambda}$. Each column corresponds to a different regression model. For RGPR the deviance residuals are unscaled.

### 5.3 Inference

In terms of drawing inference about the effect of the traffic flow covariates on the number of crashes, we examine the coefficients and standard errors and assume a normal approximation. In this case, the effects are very strong across all models, resulting in p-values of zero for each of the two covariate coefficients.

For the sake of completeness, we examined the estimated coefficient distributions using 1,000 bootstrap samples, to verify the normal approximation. As can be seen in Figure 8, the normal approximation appears reasonable for all models. Note that, for 189 samples, the RGPR estimation did not converge, thus we use only the 811 remaining estimates in Figure 8.
Figure 8: Normal probability plots of $\hat{\beta}_0$ (top), $\hat{\beta}_1$ (middle), and $\hat{\beta}_2$ (bottom) for each of the five regression models (columns), based on 1,000 bootstrap samples. Note that the RGPR estimates are based on the 811 samples where estimation converged.
6 Discussion

The CMP regression model presented here provides a practical tool for modeling count data that have various levels of dispersion. It generalizes the widely-used Poisson and logistic regressions, as well as allows for other levels of dispersion. Using a GLM approach and taking advantage of the exponential family properties of the CMP distribution, we provide a straightforward, computationally efficient, and elegant framework for model estimation, dispersion testing, inference, and diagnostics. The three data examples illustrate the differences and similarities that arise in practice when using a CMP regression versus more traditional regression models such as Poisson or log-linear models. For moderate to high counts, fitted values can be similar across models but the conditional fitted distribution can differ markedly. Models also tend to diverge in terms of inference for single predictors, implying that inappropriate use of a Poisson model (instead of a CMP model) can lead to erroneous conclusions.

One important insight from the CMP regression model is that, in a model that allows for different levels of dispersion, the role of the conditional mean is no longer central. Unlike linear regression or Poisson regression where the conditional mean is central to interpretation, the CMP regression uses a more general function of the response distribution. The resulting model means that, when examining goodness-of-fit or when predicting new observations, the complete conditional fitted distribution must be taken into account rather than just the conditional mean. This point was illustrated in the data examples.

The elegance of the CMP regression model lies in its ability to address applications containing a wide range of dispersion in a parsimonious way. While the negative binomial model is a popular resource for count data applications where over-dispersion exists, it cannot address problems where data are under-dispersed. The RGPR formulation offers more flexibility in its ability to handle data dispersion, yet it is limited in the level of under-dispersion that it can capture. We have shown that in such cases the CMP regression does not encounter such difficulties and produces reasonable fitted models. The CMP regression
has the flexibility even in the extreme case of a binary response where it reduces to a logistic regression in theory and in practice.

Our regression model is similar to the Bayesian formulation used by Borle et al. (2005, 2006, 2007); Boatwright et al. (2003); Kalyanam et al. (2007) and that by Lord et al. (2008) in terms of the generated estimated parameters. It differs from the Bayesian formulation, however, both conceptually (in terms of the link function of Lord et al. (2008) and the estimation method) and practically (with regard to run time). Although the Bayesian implementation allows for the incorporation of prior information in the form of prior parameter distributions (e.g., see Kadane et al. (2005)), the benefit of such information is useful only when informative priors are used and when the sample size is small. Secondly, specifying meaningful priors on the $\beta$ coefficients is not straightforward, as it requires an understanding of the function $\lambda^{1/\nu}$, which is not equal to the mean. Software implementation also differentiates these models because our formulation relies on traditional estimation methods for exponential family distributions: estimation, inference, and diagnostics can be programmed in most statistical software packages in a straightforward manner. From a computational point of view, although the $Z$ function requires approximation (because it is an infinite sum), in practice a simple truncation of the sum performs well.

A potential restricting factor in our current CMP regression formulation is that it assumes a constant dispersion level across all observations. This is similar to the classic homoscedasticity assumption in linear regression. A possible enhancement is to allow $\nu$ to be observation-dependent (and to model it as a function of covariates as well). In our CMP regression formulation such an extension still maintains the structure of an exponential family, unlike that of the generalized Poisson regression of Famoye (1993), for example.

Our choice of link function $\log \lambda$ is meaningful not only because of its usefulness for the CMP regression estimation and diagnostics, but also because it shows the relationship between the two well-known regression models for count data, namely Poisson and logistic regression, and further justifies their popular choices of link functions (log for Poisson, and
logit for logistic).

The relationship between the associated fitted mean bands and the estimated data dispersion is nicely illustrated in accordance with McCullagh and Nelder (1997). Further work is needed to investigate their impact on Type I errors associated with hypothesis testing about the slope, or slope coverage. In addition, this work introduces several questions regarding sample size, which although can be easily overcome by using bootstrap, present interesting research questions.

Acknowledgements

The authors thank Seth Guikema and Dominique Lord for supplying the Toronto crash data to allow for method comparison. The authors also thank Jay Kadane and the reviewers for their helpful and insightful comments.

References

Ben, M. G. and Yohai, V. J. (2004). Quantilequantile plot for deviance residuals in the generalized linear model. *Journal of Computational and Graphical Statistics*, 13(1):36–47.

Boatwright, P., Borle, S., and Kadane, J. B. (2003). A model of the joint distribution of purchase quantity 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(2):194–205.

Borle, S., Boatwright, P., Kadane, J. B., Nunes, J. C., and Shmueli, G. (2005). The effect of product assortment changes on customer retention. *Marketing Science*, 24(4):616–622.
Borle, S., Dholakia, U., Singh, S., and Westbrook, R. (2007). The impact of survey participation on subsequent behavior: An empirical investigation. *Marketing Science*, 26(5):711–726.

Cui, Y., Kim, D.-Y., and Zhu, J. (2006). On the generalized poisson regression mixture model for mapping quantitative trait loci with count data. *Genetics*, 174:21592172.

Davison, A. and Tsai, C.-L. (1992). Regression model diagnostics. *International Statistical Review*, 60:337–353.

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

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

Famoye, F., Wulu, J. J., and P., S. K. (2004). On the generalized poisson regression model with an application to accident data. *Journal of Data Science*, 2:287–295.

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

Hinde, J. (1982). Compound Poisson regression models. In Gilchrist, R., editor, *GLIM 82: Proc. Internat. Conf. Generalized Linear Models*, pages 109–121, Berlin. Springer.

Kadane, J. B., Krishnan, R., and Shmueli, G. (2006). A data disclosure policy for count data based on the COM-Poisson distribution. *Management Science*, 52(10):1610–1617.

Kadane, J. B., Shmueli, G., Minka, T. P., Borle, S., and Boatwright, P. (2005). Conjugate analysis of the Conway-Maxwell-Poisson distribution. *Bayesian Analysis*, 1(2):363–374.

Kalyanam, K., Borle, S., and Boatwright, P. (2007). Deconstructing each item’s category contribution. *Marketing Science*, 26(3):327–341.
Kutner, M. H., Nachtsheim, C. J., and Neter, J. (2003). *Applied Linear Regression Models, Fourth Edition*. McGraw-Hill.

Lattin, J. M., Green, P. E. J., and Carroll, D. (2003). *Analyzing Multivariate Data*. Duxbury.

Long, J. S. (1997). *Regression models for Categorical and Limited Dependent Variables*. Sage Publications.

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.

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

Minka, T. P., Shmueli, G., Kadane, J. B., Borle, S., and Boatwright, P. (2003). Computing with the com-poisson distribution. Technical Report 776, 776, Dept. of Statistics, Carnegie Mellon University.

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.

A 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 $\beta$ (Dobson, 2002, p.40):

$$\frac{\partial \log L_i}{\partial \beta_j} = \frac{\partial \log L_i}{\partial \log \lambda_i} \cdot \frac{\partial \log \lambda_i}{\partial \beta_j} = \left(y_i - \frac{\partial \log Z(\lambda_i, \nu)}{\partial \log \lambda_i}\right) x_{ij} = (y_i - E(Y_i)) x_{ij} \quad (10)$$

by Equation (3), where $j = 0, \ldots, p$. To estimate $\nu$ (which is assumed unknown and constant across observations), we use Equation (5) to consider

$$\frac{\partial \log L_i}{\partial \nu} = -\log y_i! - \frac{\partial \log Z(\lambda_i, \nu)}{\partial \nu} = -\log y_i! + E(\log Y_i!). \quad (11)$$

By summing the results in Equations (10) and (11) over all $n$ observations and equating each summation to zero, we obtain the normal equations for estimating $\beta$ and $\nu$:

$$\sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} \left\{ x_{ij} \sum_{s=0}^{\infty} s e^{x_i \beta / (s!)} / (s!)^\nu \right\}$$

$$\sum_{i=1}^{n} \log y_i! = \sum_{i=1}^{n} \left\{ \sum_{s=0}^{\infty} \log(s!) e^{x_i \beta / (s!)} / (s!)^\nu \right\}.$$

These equations are non-linear in $\beta$ and $\nu$, and therefore require an iterative solution, starting with the Poisson estimates, $\beta^{(0)}$ 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,\(^3\)

$$X'WX\theta^{(m)} = X'WT,$$  \quad (12)

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

\(^3\)Note that this is a generalization of the classic WLS formulation, as in Dobson (2002), to a two parameter case.
right column
\[
\begin{pmatrix}
-\log Y_1 + E(\log Y_1) \\
Y_1 - E(Y_1) \\
\vdots \\
-\log Y_n + E(\log Y_n) \\
Y_n - E(Y_n)
\end{pmatrix}.
\]

The weight matrix \(W\) is an \(n \times n\) diagonal matrix with elements \(W_{ii} = \text{Var}(Y_i)\); The vector \(\theta^{(m)}\) is the \(m\)th iteration of the estimated coefficient vector \([\hat{\beta}', \hat{\nu}']\); Finally, the vector \(T\) (of length \(n\)) has element \(i\) equal to
\[
t_i = \sum_{j=0}^{p} x_{ij} \beta_j^{(m-1)} + \nu^{(m-1)} \frac{\log Y_i + E(\log Y_i)}{Y_i - E(Y_i)} - \frac{Y_i - E(Y_i)}{\text{Var}(Y_i)}.
\]

This derivation is based on an initial Newton-Raphson iterative formulation of the form \(\theta^{(m)} = \theta^{(m-1)} + I^{-1} U\), where \(I\) is the Fisher Information matrix defined in (7), and \(U\) is the score vector with \(p + 2\) elements equal to the right-hand side in Equations (10)-(11), summed over \(i = 1, \ldots, n\). Left-multiplication of both sides by \(I\), and using \(I = X'\mathcal{M}X\) leads to the formulation in Equation (12).

### B Fisher Information Matrix Components Associated with CMP Coefficients

Due to the GLM formulation, the \(j, k\)th component in \(I^\beta\) is given by
\[
I^\beta_{j,k} = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\text{Var}(Y_i)} \left[ \frac{\partial E(Y_i)}{\partial \log \lambda_i} \right]^2 \\
= \sum_{i=1}^{n} x_{ij} x_{ik} \text{Var}(Y_i) \text{ by Equation (4)} \\
= \sum_{i=1}^{n} x_{ij} x_{ik} \left\{ \sum_{s=0}^{\infty} s^2 e^{sX_i \beta} / (s!)^\nu - \left[ \sum_{s=0}^{\infty} s e^{sX_i \beta} / (s!)^\nu \right]^2 \right\}.
\]
Similarly, the value for $I^\nu$ is given by

$$I^\nu = \sum_{i=1}^{n} \text{Var}(\log Y_i)$$

$$= \sum_{i=1}^{n} \left\{ \frac{\sum_{s=0}^{\infty} \log(s!)^2 e^{sX_i\beta}/(s!)^\nu}{\sum_{s=0}^{\infty} e^{sX_i\beta}/(s!)^\nu} - \left[ \frac{\sum_{s=0}^{\infty} \log(s!) e^{sX_i\beta}/(s!)^\nu}{\sum_{s=0}^{\infty} e^{sX_i\beta}/(s!)^\nu} \right]^2 \right\},$$

and finally,

$$I^{\beta_j,\nu} = \sum_{i=1}^{n} x_{ij} \text{Cov}(Y_i, \log Y_i)$$

$$= \sum_{i=1}^{n} x_{ij} \left\{ \frac{\sum_{s=0}^{\infty} s \log(s!) e^{sX_i\beta}/(s!)^\nu}{\sum_{s=0}^{\infty} e^{sX_i\beta}/(s!)^\nu} - \left[ \frac{\sum_{s=0}^{\infty} se^{sX_i\beta}/(s!)^\nu}{\sum_{s=0}^{\infty} e^{sX_i\beta}/(s!)^\nu} \right] \left[ \frac{\sum_{s=0}^{\infty} \log(s!) e^{sX_i\beta}/(s!)^\nu}{\sum_{s=0}^{\infty} e^{sX_i\beta}/(s!)^\nu} \right] \right\}.$$ 

### C Full Datasets and Fitted Values Under Various Regression Models

The following tables show a more complete comparison of regression models with regard to fitted values, MSE, and AIC_C for the airfreight and fabric examples, respectively.

#### Table 8: Airfreight breakage example: data and fitted values

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

AIC_C 47.29 52.11 49.37
MSE 1.900 2.210 2.363
Table 9: Fitted values for fabric example, comparing the CMP with alternative regression formulations

| Obs. | log(RollLength) (X) | Actual Faults (Y) | CMP fit | Poisson fit | Linear Reg fit | Neg-Bin fit | RGPR fit |
|------|---------------------|-------------------|---------|-------------|---------------|-------------|---------|
| 1    | 6.312               | 6                 | 8.21    | 8.33        | 7.43          | 8.36        | 8.36    |
| 2    | 6.479               | 4                 | 9.75    | 9.83        | 8.69          | 9.78        | 9.76    |
| 3    | 6.724               | 17                | 12.62   | 12.56       | 10.95         | 12.31       | 12.25   |
| 4    | 5.927               | 9                 | 5.56    | 5.67        | 5.17          | 5.83        | 5.85    |
| 5    | 6.572               | 14                | 10.76   | 10.80       | 9.49          | 10.68       | 10.65   |
| 6    | 6.766               | 8                 | 13.20   | 13.10       | 11.40         | 12.81       | 12.75   |
| 7    | 5.602               | 5                 | 4.05    | 4.10        | 3.80          | 4.30        | 4.33    |
| 8    | 6.446               | 7                 | 9.43    | 9.52        | 8.43          | 9.48        | 9.47    |
| 9    | 6.196               | 7                 | 7.29    | 7.42        | 6.66          | 7.50        | 7.51    |
| 10   | 5.919               | 7                 | 5.51    | 5.63        | 5.13          | 5.78        | 5.80    |
| 11   | 6.469               | 6                 | 9.66    | 9.74        | 8.62          | 9.69        | 9.67    |
| 12   | 6.089               | 8                 | 6.54    | 6.67        | 6.02          | 6.79        | 6.80    |
| 13   | 6.797               | 28                | 13.64   | 13.50       | 11.73         | 13.18       | 13.11   |
| 14   | 6.127               | 4                 | 6.79    | 6.92        | 6.24          | 7.03        | 7.04    |
| 15   | 6.465               | 10                | 9.61    | 9.70        | 8.58          | 9.65        | 9.63    |
| 16   | 6.198               | 4                 | 7.31    | 7.44        | 6.67          | 7.52        | 7.52    |
| 17   | 6.297               | 8                 | 8.08    | 8.21        | 7.33          | 8.25        | 8.25    |
| 18   | 6.736               | 9                 | 12.79   | 12.71       | 11.08         | 12.45       | 12.39   |
| 19   | 6.808               | 23                | 13.80   | 13.65       | 11.86         | 13.32       | 13.25   |
| 20   | 6.295               | 9                 | 8.07    | 8.19        | 7.31          | 8.23        | 8.23    |
| 21   | 6.258               | 6                 | 7.76    | 7.89        | 7.06          | 7.95        | 7.95    |
| 22   | 4.804               | 1                 | 2.02    | 1.85        | 1.79          | 2.03        | 2.06    |
| 23   | 6.488               | 9                 | 9.85    | 9.92        | 8.77          | 9.86        | 9.84    |
| 24   | 5.136               | 4                 | 2.66    | 2.58        | 2.45          | 2.78        | 2.81    |
| 25   | 6.604               | 9                 | 11.12   | 11.14       | 9.78          | 11.00       | 10.96   |
| 26   | 5.916               | 14                | 5.50    | 5.61        | 5.12          | 5.77        | 5.79    |
| 27   | 6.601               | 17                | 11.09   | 11.11       | 9.76          | 10.97       | 10.94   |
| 28   | 6.619               | 10                | 11.30   | 11.31       | 9.92          | 11.15       | 11.11   |
| 29   | 6.205               | 7                 | 7.35    | 7.48        | 6.71          | 7.56        | 7.57    |
| 30   | 6.574               | 3                 | 10.78   | 10.81       | 9.51          | 10.69       | 10.66   |
| 31   | 6.859               | 9                 | 14.57   | 14.36       | 12.44         | 13.96       | 13.89   |
| 32   | 6.033               | 2                 | 6.18    | 6.31        | 5.71          | 6.44        | 6.45    |

AIC$_C$ | 184.47   | 192.23   | **180.83** | 182.24   | 181.90   |
MSE     | **21.41** | 21.59    | 23.12    | 21.77    | 21.81    |