Predicting Censored Count Data with COM-Poisson Regression

Kimberly F. Sellers

&

Galit Shmueli

http://eprints.exchange.isb.edu/268/

Working Paper

Indian School of Business

2010

Predicting Censored Count Data with COM-Poisson Regression

Abstract

Censored count data are encountered in many applications, often due to a data collection mechanism that introduces censoring. A common example is questionnaires with question answers of the type 0,1,2,3+. We consider the problem of predicting a censored output variable Y, given a set of complete predictors X. The common solution would be to use adaptations for Poisson or negative binomial regression models that account for the censoring. We study two alternatives that allow for both over- and under-dispersion: Conway-Maxwell-Poisson (COM-Poisson) regression, and generalized Poisson regression models, each with adaptations for censoring. We compare the predictive power of these models by applying them to a German panel dataset on fertility, where we introduce censoring of different levels into the outcome variable. We explore two additional variants: (1) using the mean versus the median of the predictive count distribution, and (2) ensembles of COM-Poisson models based on the parametric and non-parametric bootstrap.

Keywords: over-dispersion, under-dispersion, predictive distribution, mean versus median predictions, ensembles

1. Introduction

Count data with censoring are encountered in various applications. One common context is in data from questionnaires, where the question of interest asks about counts of some event, but allows only a set of answers such as $0, 1, 2, 3+$ (see e.g., Terza, 1985). An example is the question in the 2000 U.S. Census individual census report: "How many people, including yourself, usually rode to work in the car, truck, or van last week?" with possible answers $1, 2, 3, 4, 5, 6, 7+$. Census block-level data based on this question could be used, for instance, by a car rental company or car sharing service (such as Zipcar) for placing the right size of vehicles in different locations. Given the census data, the company would predict the distribution of needed passengers-per-car for a certain location.

Censored data frequently occur in epidemiological studies where right censoring occurs due to

durations that extend beyond the study period. For example, Embury, Elias, Heller, Hood, Greenberg, & Schrier (1977) reported the remission length (in weeks) for acute myelogenous leukemia patients. Another context is data disclosure of sensitive information, where censoring might be used to protect individual confidentiality associated with large count information while still providing data access (see e.g., Jenkins, Burkhauser, Feng, & Larrimore, 2009). Predicting the outcome in such applications is obviously useful for many purposes.

Right-censored count data would also arise in applications where there is some capacity limitation. For example, the number of patient appointments at a clinic, the number of customers being served at a restaurant during a certain time period, or the number of cars parked at a parking lot. Although the number of arrivals in each of these cases can exceed the capacity, measurement devices might only be able to measure occupancy up to the capacity. Yet, such data could be useful for determining capacities of new operations (e.g., staffing in a new clinic, the number of tables and staffing of a new restaurant, or the expansion of a parking lot).

A variety of regression models are available for modeling censored numerical data, ranging from parametric models (known as accelerated failure time models) to the popular semi-parametric Cox model. For count data, however, the choice of regression models that can account for censoring is limited. Censored Poisson regression is the most widely used among such models (see e.g., Brannas, 1992; Terza, 1985), while for over-dispersed data, an adaptation of the negative binomial regression is used (see e.g., Chapter 9 in Hilbe (2007), or Caudill & Mixon (1995)). Famoye & Wang (2004) proposed a censored generalized Poisson (GP) model, which accommodates both over- and under-dispersion (see also the application by Mahmoud & Alderiny (2010)).

While the literature on censored count data has focused on inference, to the best of our knowledge, there has not been much in the way of prediction. Because inference and prediction are different purposes and prescribe different modeling steps (Shmueli, 2010; Shmueli & Koppius, 2010), results from studies focused on inference do not shed light on predictive power; for example, a model might be inferior in terms of parameter bias, but superior in terms of predictive accuracy. This paper is therefore focused on investigating regression models for count data with censoring in terms of **predictive behavior**.

The paper is organized as follows: In Section 2, we describe three regression models for count data with censoring. Two are existing models, namely, Poisson and GP, and one is a new adaptation

of the flexible COM-Poisson regression model by Sellers & Shmueli (2010). We describe model estimation for these models and then discuss approaches for generating point predictions for count data, and for creating ensembles using parametric and non-parametric bootstrap. In Section 3, we apply the different models and compare their predictive performance by analyzing a dataset on fertility, where we artificially add different levels of censoring to the outcome variable. We conclude with a discussion in Section 4.

2. Regression models for count data with censoring

In this section, we present three regression models for count data with censoring. In all cases, we use the following notation: $\boldsymbol{Y}$ is the count output vector, $\boldsymbol{X}$ is a matrix of predictors with a first column of 1's (the design matrix), $\boldsymbol{\beta}$ is the parameter vector, and δ_i is a censoring indicator, denoting whether observation i is censored ($\delta_i = 1$) or not ($\delta_i = 0$).

2.1. Poisson regression with right-censoring

The most common regression model for censored count data is the Poisson model. Its log likelihood can be written as

$$\begin{aligned} \log L &= \sum_{i=1}^{n}(1-\delta_i)\log P(Y_i = y_i) + \delta_i \log P(Y_i \geq y_i) \qquad (1) \\ &= \sum_{i=1}^{n}(1-\delta_i)\left[y_i \log \lambda_i - \log y_i! - \lambda_i\right] + \delta_i \log P(Y_i \geq y_i). \end{aligned}$$

Parameter estimation is obtained by numerically maximizing the likelihood function. This is implemented in various statistical software packages (e.g., the function `vglm` with family `cenpoisson`, contained in the VGAM package in *R*).

2.2. Generalized Poisson (GP) with censoring

Famoye & Wang (2004) introduced the censored two-parameter GP regression model, with log-likelihood given by

$$\begin{aligned} \log L &= \sum_{i=1}^{n}(1-\delta_i)\log P(Y_i = y_i \mid \mu_i, \alpha) + \delta_i \log P(Y_i \geq y_i \mid \mu_i, \alpha) \qquad (2) \\ &= \sum_{i=1}^{n}(1-\delta_i)\left[y_i(\log \mu_i - \log(1+\alpha\mu_i)) + (y_i - 1)\log(1+\alpha y_i) - \log y_i! - \frac{\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}\right] \\ &\quad +\delta_i \log P(Y_i \geq y_i \mid \mu_i, \alpha), \end{aligned}$$

where $\mu_i = E(Y_i)$ and α denotes the dispersion parameter. Parameter estimation is obtained by maximizing the log likelihood, under the constraint $\alpha > -2/\max(\mu_i)$.

2.3. COM-Poisson regression with censoring

We introduce the censored COM-Poisson regression, which is an adaption of the COM-Poisson regression described in Sellers & Shmueli (2010). This section briefly outlines the COM-Poisson distribution for count data, the COM-Poisson regression model, and the new adaptation for right-censored data.

2.3.1. The COM-Poisson Distribution

The COM-Poisson distribution is a two-parameter generalization of the Poisson distribution, which also includes the geometric and Bernoulli distributions as special cases (Shmueli, Minka, Kadane, Borle, & Boatwright, 2005). The probability distribution function of Y is given by

$$P(Y = y) = \frac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)} \quad y = 0, 1, 2, \ldots, \tag{3}$$

where $\lambda > 0$, $\nu \geq 0$, and $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}$ is a normalizing constant. The COM-Poisson distribution generalizes the Poisson distribution ($\nu = 1$), the geometric distribution ($\nu = 0$ and $\lambda < 1$), and the Bernoulli distribution $\left(\nu \to \infty \text{ with probability } \frac{\lambda}{\lambda+1}\right)$.

The mean of the COM-Poisson distribution does not have a simple closed form. It can be written as

$$E(Y) = \lambda \frac{\partial \log Z(\lambda, \nu)}{\partial \lambda} \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu}, \tag{4}$$

where the approximation holds for $\lambda \gg 10^\nu$ or $\nu < 1$; see Minka, Shmueli, Kadane, Borle, & Boatwright (2003) for further details. Closed-form formulations that relate the mean to the parameters λ, ν are

$$E(Y) = \frac{\partial \log Z(\lambda, \nu)}{\partial \log \lambda}, \tag{5}$$

$$E(Y^\nu) = \lambda, \tag{6}$$

and

$$E(Y^{r+1}) = \begin{cases} \lambda[E(Y+1)]^{1-\nu} & r = 0 \\ \lambda \frac{\partial}{\partial \lambda} E(Y^r) + E(Y)E(Y^r) & r > 0. \end{cases} \tag{7}$$

2.3.2. COM-Poisson regression model

Sellers & Shmueli (2010) introduced a COM-Poisson regression model that relates a count variable $\boldsymbol{Y}$ to a function of predictors $\boldsymbol{X}$ through the link function $\log \boldsymbol{\lambda} = \boldsymbol{X'\beta}$. The log-likelihood

function is given by

$$\log L(\boldsymbol{\lambda}, \nu|\boldsymbol{y}) = \sum_{i=1}^{n} y_i \log \lambda_i - \nu \sum_{i=1}^{n} \log y_i! - \sum_{i=1}^{n} \log Z(\lambda_i, \nu). \tag{8}$$

Maximum likelihood parameter estimates can be obtained by directly maximizing Equation (8) under the constraint $\nu \geq 0$, using a constrained nonlinear optimization tool (e.g., `nlminb` in *R*). An alternative is to write the log-likelihood as a function of $\log \nu$, and then maximize it using an ordinary nonlinear optimization tool (e.g., `nlm` in *R*). 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 and to solve it iteratively. For further details, see Sellers & Shmueli (2010).

2.3.3. COM-Poisson regression model with right-censoring

To incorporate censoring, we use the classic approach of introducing a censoring indicator variable, δ_i. The log-likelihood function is given by

$$\begin{aligned} \log L &= \sum_{i=1}^{n} (1-\delta_i) \log P(Y_i = y_i) + \delta_i \log P(Y_i \geq y_i) \\ &= \sum_{i=1}^{n} (1-\delta_i) \left[y_i \log \lambda_i - \nu \log y_i! - \log Z(\lambda_i, \nu) \right] + \delta_i \log P(Y_i \geq y_i), \end{aligned} \tag{9}$$

where $P(Y_i \geq y_i) = 1 - \sum_{s=0}^{y_i - 1} \frac{\lambda_i^s}{(s!)^\nu} Z^{-1}(\lambda_i, \nu)$.

Model estimation

As in the ordinary COM-Poisson regression model, parameter estimates can be obtained by maximizing the log likelihood in Equation (9) under the constraint $\nu \geq 0$, or by writing the log-likelihood as a function of $\log \nu$ and then maximizing it using an ordinary, unconstrained nonlinear optimization tool. A third alternative is to use the score functions (given in Appendix A) and equate them to zero.

2.4. Generating point predictions: mean and median

In Poisson and GP regression, the mean of the response distribution ($E(\boldsymbol{Y})$) is related to the predictors via the function $E(\boldsymbol{Y}) = \exp(\boldsymbol{X}\boldsymbol{\beta})$. The common approach for generating point

predictions is therefore via $\hat{\boldsymbol{Y}} = \exp(\boldsymbol{X}\hat{\boldsymbol{\beta}})$, that is, using the mean of the (back-transformed) predictive distribution. In the COM-Poisson regression, however, the relationship between $E(\boldsymbol{Y})$ and $\log(\boldsymbol{\lambda}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is more complicated (see Section 2.3.1). One approach for obtaining the predictive mean is to use the approximation in Equation (4) by plugging in $\hat{\boldsymbol{\lambda}} = \exp(\boldsymbol{X}\hat{\boldsymbol{\beta}})$. While Minka et al. (2003) note that the mean approximation is theoretically accurate only for $\lambda \gg 10^{\nu}$, Geedipally, Guikema, Dhavala, & Lord (2008) show that the approximation is still reasonable in cases where the λ is "substantially below the lower bound suggested".

An alternative to using the predictive distribution mean as a point prediction is to use the median. The median has two advantages over the mean in count model prediction: (1) the median produces integer predictions, similar to the scale of the original Y, which in some applications might be required; and (2) the median is a better and more robust central tendency measure in skewed distributions, which are common with count models. In the case of the COM-Poisson distribution, the median has also been shown to produce better fitted values for data without censoring when the criteria for the mean approximation are not satisfied (Sellers & Shmueli, 2010).

Computing the median for the Poisson, GP, and COM-Poisson regression models is simple and fast. To compute the median for observation i, which is the inverse cumulative distribution function (CDF) at 0.5, we simply compute the individual probabilities $P(Y_i = 0), P(Y_i = 1), \ldots$ consecutively until their sum exceeds 0.5. For the COM-Poisson, we can also use the relationship $P(Y_i = y_i) = \left(\frac{\lambda_i}{y_i}\right)^{\nu} \times P(Y_i = y_i - 1)$ to compute consecutive probabilities even faster.

2.5. COM-Poisson ensembles via resampling

A popular method for increasing predictive power is via ensembles. Ensembles are achieved by combining point predictions from multiple models and/or multiple datasets. We use this motivation to generate COM-Poisson ensembles, based on resampling the training data. We can employ either parametric or non-parametric bootstrapping for generating the resamples. We use the subscript t to denote training data and h to denote holdout data.

For parametric resampling, we use the coefficients estimated from the training data ($\boldsymbol{\beta}_t$) along with the predictor values in the holdout data ($\boldsymbol{X}_h$) to simulate multiple sets (resamples) of predicted values ($\hat{\boldsymbol{Y}}_h$). For non-parametric resampling, the training data ($\{\boldsymbol{X}_t, \boldsymbol{Y}_t\}$) are resampled directly to produce multiple training sets. Then, a COM-Poisson model is fitted separately to each of the resampled training sets, and used to predict the holdout data. Note that both parametric and

non-parametric resampling create multiple predictions for each observation in the holdout set. We combine these multiple predictions into a single, improved, prediction by computing their average or their median.

Jung, Jhun, & Song (2006) note that, in the case of censored data, the two bootstrap approaches differ in terms of how the censoring operates on the resamples. In parametric bootstrap, the censoring depends on the resampled Y whereas in non-parametric bootstrap, the censoring is fixed at the same covariate pattern as in the original data. Jung et al. (2006), who used resampling to compute standard errors in the context of inference, found that the two methods produce similar results for the Poisson and negative binomial censored regression models in terms of standard errors. Nevertheless, they did not consider the effect on predictive power. In the predictive context, we note another difference between parametric and non-parametric bootstrap: in the parametric bootstrap, resampling is applied to the holdout data ($\boldsymbol{X}_h$) while, in the non-parametric bootstrap, resampling is applied to the training data ($\boldsymbol{X}_t$). In Section 3 we evaluate empirically the effect of using parametric versus non-parametric bootstrap on the predictive power of COM-Poisson censored regression models.

2.6. Evaluating Predictive Accuracy

To evaluate predictive power, if the dataset is sufficiently large, a common approach is to partition the data into a training set and a holdout set. The training set is used to estimate the model, and predictions on the holdout set are used to evaluate predictive power. In smaller datasets, cross-validation is a common alternative.

Measures of predictive power are typically based on prediction errors, which are the differences between the actual values (Y_h) and the point predictions ($\hat{Y}_h$). We use several predictive summaries, including:

1. Root-mean-squared-error (RMSE), given by $\sqrt{\frac{1}{n}\sum e_i^2} = \sqrt{\frac{1}{n}\sum (y_i - \hat{y}_i)^2}$;
2. Median-absolute-percentage-error (MdAPE), where the absolute-percentage-error (APE) for observation i is given by $\mid \frac{e_i}{y_i} \mid$ and MdAPE is the median across the APE values. We use MdAPE in place of the more common mean-absolute-percentage-error (MAPE) due to the nature of count data, which typically involve zero counts, thus leading to infinite or undefined APE values (Armstrong & Collopy, 1992).

3. Mean-absolute-scaled-error (MASE) is defined as MASE = mean($|\ q_i\ |$), where the scaled errors are $q_i = \frac{e_i}{\frac{1}{n}\sum_{j=1}^{n}|y_j - \bar{y}|}$ and the scaling in the denominator is computed from the training set (Hyndman & Koehler, 2006). Like MdAPE, MASE avoids any problems in the case of zero counts.
4. Several score metrics for count data that are based on the predictive distribution, proposed by Czado, Gneiting, & Held (2009) (averaged over the holdout data):
   (a) Logarithmic score, $\log s = -\log p_x$, where p_x is the predictive distribution at the observed count value, x
   (b) Quadratic score, $qs = -2p_x + \|p\|^2$, where $\|p\|^2 = \sum_{k=0}^{\infty} p_k^2$
   (c) Spherical score, $sphs = \frac{-p_x}{\|p\|}$
   (d) Ranked probability score, $rps = \sum_{k=0}^{\infty}\{p_k - I_{x\leq k}\}^2$
   (e) Squared error score, $ses = (x-\mu_P)^2$, where μ_P is the mean of the predictive distribution
   (f) Dawid-Sebastiani score, $dss = \left(\frac{x-\mu_P}{\sigma_P}\right)^2 + 2\log\sigma_P$, where σ_P is the standard deviation associated with the predictive distribution.

In all cases, lower scores denote higher predictive accuracy.

3. Example: Fertility Data

We compare the predictive performance of the three models and the two ensembles by applying them to a dataset on womens' fertility, introduced by Winkelmann (1995) and used in McShane, Adrian, Bradlow, & Fader (2008). The dataset is from the 1985 German Socio-Economic Panel, and includes information on 1,243 women over age 44 who are in their first marriage. The outcome variable is the woman's number of children, while predictors include demographic information such as religion, general education, and age at marriage. Our goal is to accurately predict the number of children that a woman has, based on her demographic information. These data were found to be underdispersed.

We start by partitioning the dataset into a training set (994 records, i.e. approximately 80% of the full dataset) and holdout set (249 records, or approximately 20% of the full dataset). The training set is used for fitting all models and the holdout set is used to evaluate prediction accuracy.

3.1. Performance on Uncensored Data

We first fit the ordinary Poisson, COM-Poisson, and GP regression models to the uncensored training data, and generate predictions of the holdout data. The predictive accuracy summaries

are given in Table 1 (Columns 2-4), and the corresponding prediction error distributions are shown in Figure 1. We do not consider censored negative binomial regression here because of the underdispersion in the data, which would result in the estimated model coinciding with a censored Poisson regression. The estimated models are given in Appendix B (Table B.7). We defer the results to the Appendix due to our focus on prediction rather than inference.

We also generate predictions using COM-Poisson ensembles produced via both parametric and nonparametric bootstrapping. To summarize the ensemble predictions, we consider both the associated average and median values across the 1000 predictions (i.e., the row mean and row median). The resulting predictive accuracy summaries are given in Table 1 (last two columns) and the scores are in Table 2. The full distributions of prediction errors for each of the four models are compared in Figure 2.

The predictive accuracy summaries provide a mixed message. The summaries indicate that the three models perform almost equally well when comparing median or mean predictions. The scores are slightly in favor of the COM-Poisson model (four out of six). Meanwhile, examining the boxplots in Figure 1 indicate that there is essentially no difference between the three models when comparing their respective errors, based on the median or mean prediction[1].

We note that for all models, the median predictions produced better RMSE and MdAPE values compared to mean predictions, while the opposite is true for the MASE. The ensemble results also appear to indicate similar performance, with a slight advantage for the non-parametric ensemble over the parametric ensemble in terms of RMSE and MASE, but a disadvantage via larger prediction error variance when using the mean (Figure 2).

3.2. Censored Data

Next, we artificially censor the number of children ($\boldsymbol{Y}$) in the training data to some degree. We introduce two different levels of censoring in the following subsections.

3.2.1. Censoring to 5+

The first censoring level that we introduce is by censoring the training data such that 5 or more children are recorded as $Y = 5+$. This leads to 4.93% of the training data being censored.

[1]The holdout data contain two outlying subjects not captured via any model considered here. One subject (#159) is a Protestant, German woman, educated for 9 years with no post-secondary education (either vocational or university), living in a rural area, 58 years old, and married at age 23. She has 10 children. The other subject (#245) is a Muslim, non-German woman, educated for 8 years with no post-secondary education (either vocational or university), also living in a rural area, and is 52 years old. She married at age 17, and has 8 children.

Table 1: Comparing Predictive Accuracy with No Censoring Across Five Models. Performance across all models appears similar.

	Poisson Med/Mean	COM-Poisson Med/Mean	GP Med/Mean	Param boot COM-Poisson Ensemb Med/Mean	Non-Param boot COM-Poisson Ensemb Med/Mean
RMSE	1.260/1.235	1.261/1.236	1.260/1.234	1.274/1.235	1.260/1.231
MdAPE	0.333/0.292	0.333/0.290	0.333/0.290	0.333/0.289	0.333/0.293
MASE	0.681/0.762	0.684/0.765	0.681/0.762	0.705/0.762	0.681/0.697

Table 2: Scoring Rule Comparisons Where No Censoring Across Several Models. Bold numbers indicate the best score value across models.

	logS	QS	SphS	RPS	DSS	SES
Poisson	1.6043	-0.2479	-0.5008	**2.2260**	1.9574	1.5240
COM-Poisson	**1.5666**	**-0.2617**	**-0.5128**	2.2516	**1.3919**	1.5275
GP	1.5905	-0.2529	-0.5051	2.2406	1.7933	**1.5226**

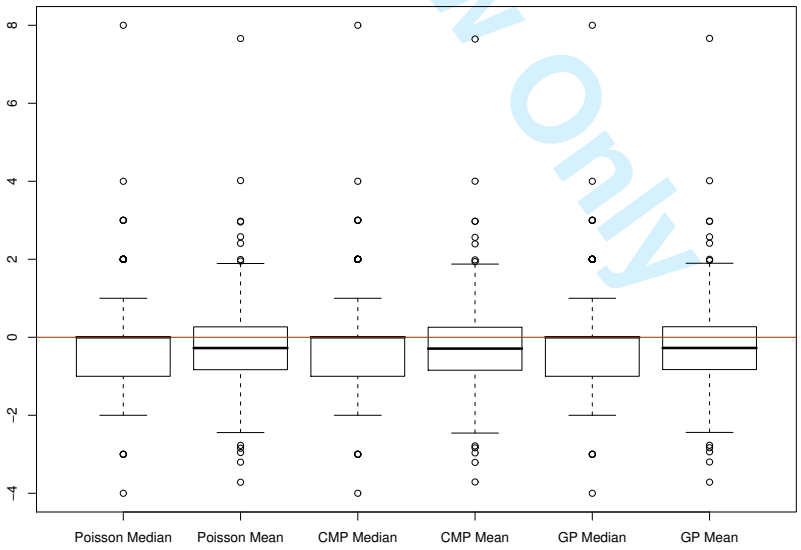

Figure 1: Predicting Uncensored Data: Side-by-side boxplots of prediction errors produced via median and mean computations, respectively, from Poisson, COM-Poisson, and GP models. Distributions are nearly identical across the three models; Median prediction distributions have lower variance.

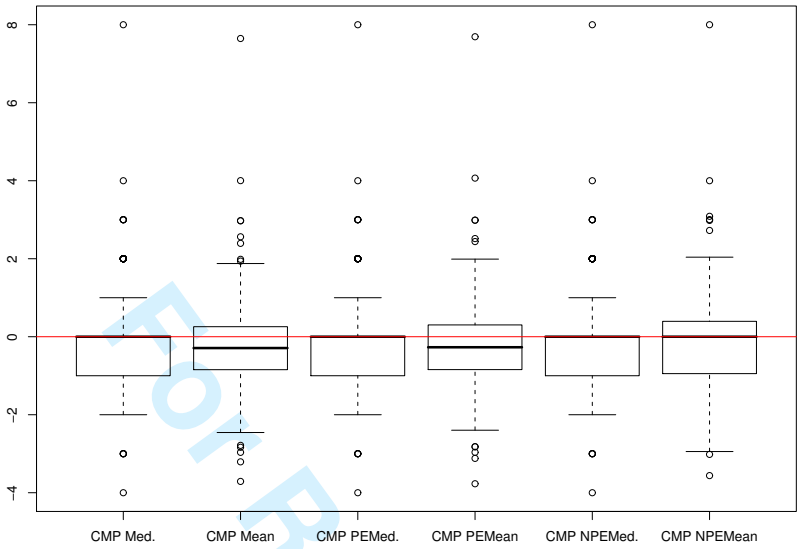

Figure 2: COM-Poisson Predictions for Uncensored Data: Side-by-side boxplots of prediction errors produced via median and mean computations, respectively, from COM-Poisson, and COM-Poisson ensembles produced via parametric and non-parametric bootstrapping. Median prediction appears nearly identical across models; Mean prediction is slightly different for nonparametric ensemble.

We then estimate the censored-Poisson, censored-COM-Poisson, and censored-GP models, and evaluate their predictive accuracy. The estimated models are given in Appendix B (Table B.7). Boxplots, summary predictive measures, and predictive scores are given in Figure 3, and Tables 3-4, respectively. The picture that emerges is very similar to the uncensored case: the prediction error distributions and predictive accuracy summaries are approximately equal across the different models; the median predictions produced better RMSE and MdAPE values compared to mean predictions, while the opposite is true for the MASE; predictive scores are slightly in favor of COM-Poisson (four out of six), and median predictions perform slightly better than mean predictions in terms of error variance.

3.2.2. Censoring to 4+

The second, heavier, censoring that we perform is obtained by censoring the training data such that 4 or more children are recorded as $Y = 4+$. This leads to 8.65% of the training data being censored. We then estimate the censored-Poisson, censored-COM-Poisson, and censored-GP models, and evaluate their predictive accuracy. The estimated models are given in Appendix

Table 3: Comparing Predictive Accuracy with 5+ Censoring Across Five Models. Performance across all models appears similar.

	Poisson Med/Mean	COM-Poisson Med/Mean	GP Med/Mean	Param boot COM-Poisson Ensemb Med/Mean	Non-Param boot COM-Poisson Ensemb Med/Mean
RMSE	1.237/1.092	1.105/1.229	1.247/1.226	1.260/1.227	1.222/1.223
MdAPE	0.333/0.276	0.333/0.283	0.333/0.277	0.333/0.285	0.333/0.317
MASE	0.674/0.751	0.684/0.759	0.681/0.754	0.701/0.756	0.664/0.683

Table 4: Scoring Rule Comparisons Where 5+ Censoring Applied Across Several Models. Bold numbers indicate the best score value across models.

	logS	QS	SphS	RPS	DSS	SES
Poisson	1.6005	-0.2487	-0.5017	**2.2166**	1.9413	**1.5024**
COM-Poisson	**1.5558**	**-0.2659**	**-0.5164**	2.2514	**1.3426**	1.5107
GP	1.5792	-0.2580	-0.5095	2.2468	1.6378	1.5029

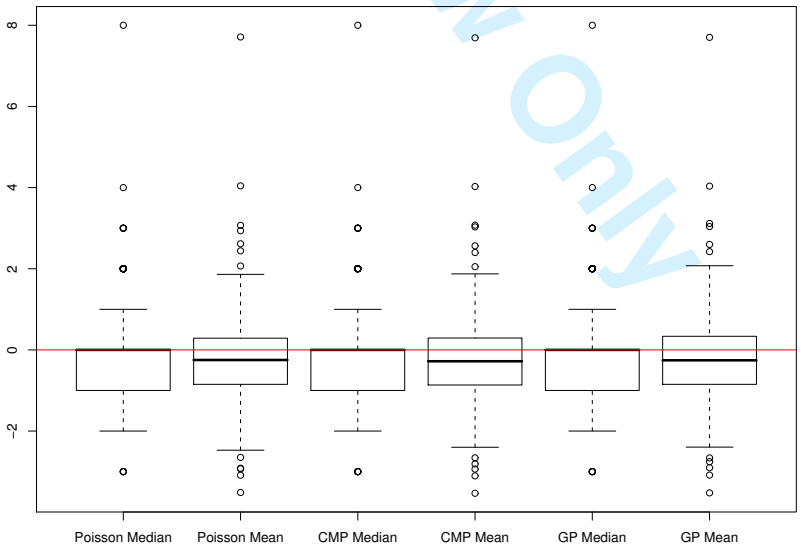

Figure 3: Predicting 5+ Censored Data: Side-by-side boxplots of prediction errors produced via median and mean computations, respectively, from Poisson, COM-Poisson, and Generalized Poisson regressions. Distributions are nearly identical across the three models; Median prediction distributions have lower variance.

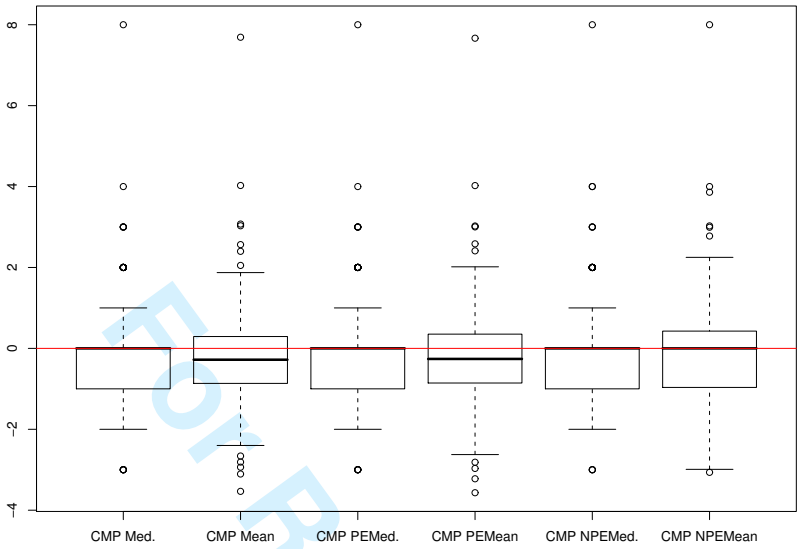

Figure 4: COM-Poisson Predictions for Censored Data at 5+: Side-by-side boxplots of prediction errors produced via median and mean computations, respectively, from COM-Poisson, and COM-Poisson ensembles produced via parametric and non-parametric bootstrapping. Median prediction appears nearly identical across models; Mean prediction is slightly different for nonparametric ensemble.

B (Table B.7). From the prediction error distribution plots (Figure 5), the predictive accuracy summaries (Table 5), and predictive scores (Table 6), it is clear that the COM-Poisson and GP models significantly outperform the Poisson model. The heavier censoring has increased the amount of underdispersion such that Poisson model produces severely biased predictions (over-predictions), as illustrated in Figure 5.

Comparing the COM-Poisson and GP models, the predictive measures and error distributions appear nearly identical, and the predictive scores are slightly in favor of the COM-Poisson model. A close look at the predictive summaries reveals that for median predictions the COM-Poisson performance is consistently equal or better than the GP model, while the opposite is true with respect to mean predictions. The slight under-performance of the COM-Poisson compared to GP when using mean predictions is likely due to the mean approximation constraints ($\lambda \gg 10^{\nu}$ or $\lambda < 1$) not being satisfied; see Equation (4). Yet, we note that the difference in performance is very small even in this case.

In terms of the ensemble performance, both ensembles appear not to add any benefit over the

Table 5: Comparing Predictive Accuracy with 4+ Censoring for Five Models. Poisson severely under-performs.

	Poisson Med/Mean	COM-Poisson Med/Mean	GP Med/Mean	Param boot COM-Poisson Ensemb Med/Mean	Non-Param boot COM-Poisson Ensemb Med/Mean
RMSE	2.720/2.885	1.043/1.223	1.248/1.213	1.244/1.219	1.237/1.233
MdAPE	1.000/1.031	0.333/0.277	0.333/0.270	0.333/0.271	0.333/0.332
MASE	1.861/1.978	0.677/0.750	0.691/0.735	0.684/0.745	0.667/0.676

Table 6: Scoring Rule Comparisons Where 4+ Censoring Applied Across Several Models. Bold numbers indicate the best score value across models.

	logS	QS	SphS	RPS	DSS	SES
Poisson	2.3026	-0.1217	-0.3431	2.8214	3.1844	7.9335
COM-Poisson	**1.5477**	**-0.2692**	**-0.5192**	2.2472	**1.3484**	1.4968
GP	1.5736	-0.2623	-0.5130	**2.2330**	1.4932	**1.4710**

ordinary COM-Poisson model in this example. The parametric ensemble appears identical to the ordinary COM-Poisson model for both mean and median predictions. The nonparametric ensemble even performs worse compared to the ordinary COM-Poisson (and parametric) model, as depicted by the wide error variance for both mean and median predictions.

4. Conclusions and Discussion

We compared the Poisson, COM-Poisson, and Generalized Poisson (GP) models for count data with right-censoring, in terms of predictive accuracy, using a data with underdispersion. We examined three levels of censoring: none, light (5+), and heavy (4+). Two types of point predictions were used: mean and median. We also examined two ensemble methods based on parametric and non-parametric bootstrapping of the COM-Poisson model. Evaluation included predictive measures, predictive scores, and prediction error distributions.

The results show that for no censoring or light censoring there is not much difference between the different models, with perhaps a small advantage for the COM-Poisson model. In general, median point predictions appear to have smaller error variance. With heavy censoring, however, the Poisson severely under-performs, producing predictions that are much too high. GP and COM-Poisson perform very similarly in such cases.

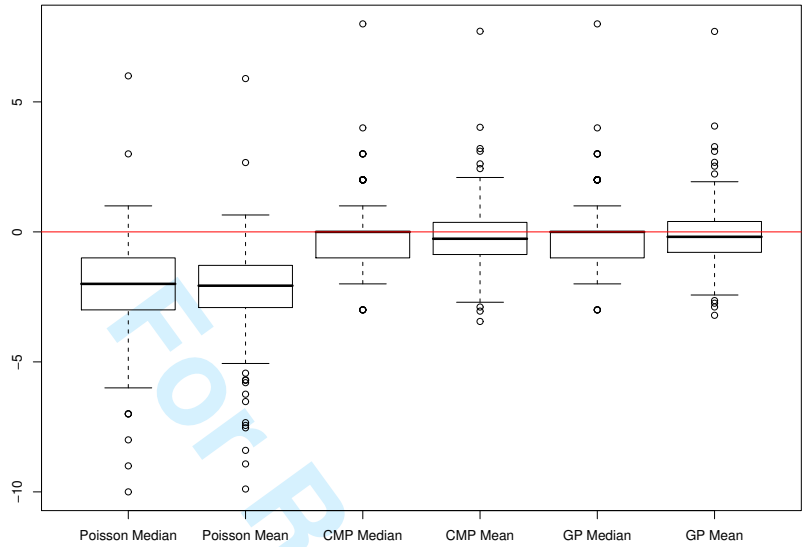

Figure 5: Predicting 4+ Censored Data: Side-by-side boxplots of prediction errors produced via median and mean computations, respectively, from Poisson, COM-Poisson, and Generalized Poisson regressions. Poisson severely under-performs.

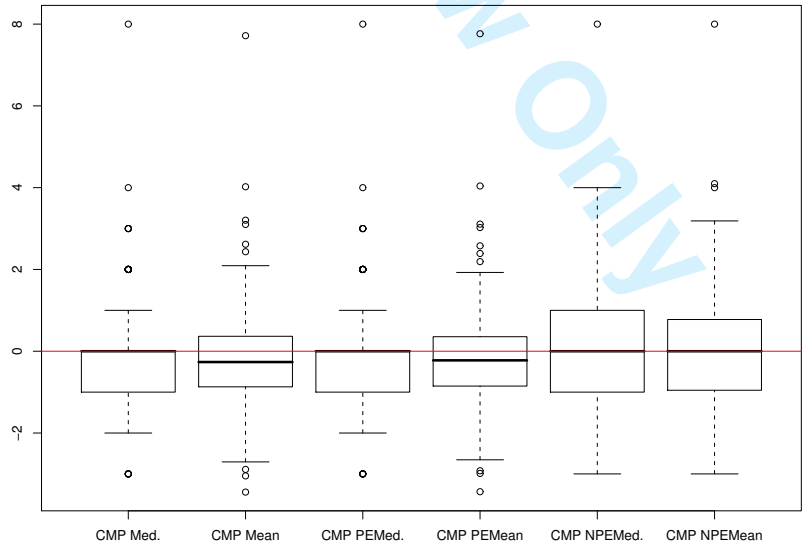

Figure 6: COM-Poisson Predictions for Censored Data at 4+: Side-by-side boxplots of prediction errors produced via median and mean computations, respectively, from COM-Poisson, and COM-Poisson ensembles produced via parametric and non-parametric bootstrapping. Neither ensemble offers any benefit over the ordinary COM-Poisson model.

Two outlying subjects significantly impacted our ability to assess the effect of censoring on prediction. For both individuals, no model was able to accurately predict the true number of children birthed by each respective mother.

We also found that the ensembles did not appear to add any value under any scenario. In the case of heavy censoring, the nonparametric ensemble even appeared to perform worse than the ordinary COM-Poisson (in terms of error variance). We initially found this result surprising, but conjecture that this is due to the added variability introduced via the non-parametric bootstrap for prediction. Taking this approach, we introduce variability associated with the maximum likelihood estimates that are determined for the bootstrapped training set; this variability is, in turn, carried forth when computing predicted values. While another option is to take ensembles of COM-Poisson and GP models, the fact that they produce very similar predictions means that there will likely not be much benefit from combining them. It remains an open question what models can be combined to produce improved predictive accuracy for censored count data.

Because this dataset exhibits underdispersion even in its raw form, negative binomial regression (whether censored or uncensored) is inappropriate. While negative binomial regression is useful when analyzing overdispersed data, it does not effectively capture the tighter variation contained here. Censoring the data, in effect, acts as a Winsorization procedure, shrinking the amount of perceived variation in the data. As a result, we see in this data example, that the obtained dispersion estimates increase as the amount of censoring increases. This phenomenon is likewise recognized in the generalized Poisson dispersion estimates for α: $\hat{\alpha}$ decreases as the censoring amount increases (i.e., the dispersion level decreases, or the data become increasingly underdispersed).

In terms of computation, given the large number of predictors in the fertility dataset, the optimization scheme used to compute the maximum likelihood estimates (whether working with training data in its raw or censored form) worked over a slowly-changing surface to locate the estimates. Accordingly, the choice of initial value for the optimization scheme, and the extension of the default number of iterations to perform in R were crucial to this operation's success. The results were obtained using `nlminb` and/or `optim` in R.

While the models used and developed in this paper were used to accommodate right-censoring, it is straightforward to derive censored Poisson, GP, and COM-Poisson models for left- or interval-censored data. Such types of censoring, however, might affect predictive power differently and are

therefore an interesting direction for further research.

References

Armstrong, J. S. & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, *8*, 69–80.

Brannas, K. (1992). Limited dependent Poisson regression. *The Statistician*, *41*, 413–423.

Caudill, S. B. & Mixon, F. G. J. (1995). Modeling household fertility decisions: Estimation and testing of censored regression models for count data. *Empirical Economics*, *20(2)*, 183–96.

Czado, C., Gneiting, T., & Held, L. (2009). Predictive model assessment for count data. *Biometrics*, *65*, 1254–1261.

Embury, S. H., Elias, L., Heller, P. H., Hood, C. E., Greenberg, P. L., & Schrier, S. L. (1977). Remission maintenance therapy in acute myelogenous leukemia. *The Western Journal of Medicine*, *126(4)*, 267–272.

Famoye, F. & Wang, W. R. (2004). Censored generalized Poisson regression model. *Computational Statistics and Data Analysis*, *46*, 547–560.

Geedipally, S. R., Guikema, S. D., Dhavala, S. S., & Lord, D. (2008). Characterizing the performance of the Bayesian Conway-Maxwell-Poisson generalized linear model. Tech. rep., Texas A&M University.

Hilbe, J. M. (2007). *Negative Binomial Regression*. Cambridge University Press, 5th edn.

Hyndman, R. J. & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22(4)*, 679688.

Jenkins, S. P., Burkhauser, R. V., Feng, S., & Larrimore, J. (2009). Measuring Inequality Using Censored Data: A Multiple Imputation Approach. *SSRN eLibrary*.

Jung, B. C., Jhun, M., & Song, S. H. (2006). Testing for overdispersion in a censored Poisson regression model. *Statistics*, *40*(6), 533543.

Mahmoud, M. & Alderiny, M. (2010). On estimating parameters of censored generalized Poisson regression model. *Applied Mathematical Sciences*, *4(13)*, 623–635.

McShane, B., Adrian, M., Bradlow, E. T., & Fader, P. S. (2008). Count models based on Weibull interarrival times. *Journal of Business & Economic Statistics,*, *26(3)*, 369–378.

Minka, T. P., Shmueli, G., Kadane, J. B., Borle, S., & Boatwright, P. (2003). Computing with the COM-Poisson distribution. Tech. Rep. 776, 776, Dept. of Statistics, Carnegie Mellon University.

Sellers, K. & Shmueli, G. (2010). A flexible regression model for count data. *Annals of Applied Statistics*, *4(2)*, 943–961.

Shmueli, G. (2010). To explain or to predict? *Statistical Science, forthcoming*.

Shmueli, G. & Koppius, O. (2010). Predictive analytics in information systems research. *MIS Quarterly, forthcoming*.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., & Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *Applied Statistics*, *54*, 127–142.

Terza, J. (1985). A tobit-type estimator for the censored Poisson regression model. *Economics Letters*, *18(6)*, 361–365.

Winkelmann, R. (1995). Duration dependence and dispersion in count-data models. *Journal of Business & Economic Statistics,*, *13(4)*, 467–474.

Appendix A. Derivatives of the likelihood function

Before proceeding, we note the following equivalence: for any parameter θ,

$$\begin{aligned}\frac{\partial \log P(Y \geq y_i)}{\partial \theta} &= -\frac{1}{P(Y \geq y_i)} \sum_{s=0}^{y_i-1} \frac{\partial P(Y=s)}{\partial \theta} \\ &= -\frac{1}{P(Y \geq y_i)} \sum_{s=0}^{y_i-1} P(Y=s) \frac{\partial \log P(Y=s)}{\partial \theta}. \qquad \text{(A.1)}\end{aligned}$$

Accordingly, the first derivative of the likelihood function given in Equation (9) with respect to β_j $(j = 1, 2, \ldots k)$ can be written as

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta_j} &= \sum_{i=1}^{n} x_{ij} \left\{ (1-\delta_i)(y_i - E(Y_i)) - \frac{\delta_i}{P(Y \geq y_i)} \sum_{s=0}^{y_i-1} P(Y=s)\,(s - E(Y_i)) \right\} \\ &= \sum_{i=1}^{n} x_{ij} \left\{ (1-\delta_i)(y_i - E(Y_i)) - \frac{\delta_i}{P(Y \geq y_i)} \left(\sum_{s=0}^{y_i-1} sP(Y=s) - P(Y < y_i)E(Y_i) \right) \right\}.\end{aligned}$$

Similarly, the first derivative with respect to ν can be written as

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \nu} &= \sum_{i=1}^{n} \left\{ (1-\delta_i)(-\log y_i! + E(\log Y_i!)) - \frac{\delta_i}{P(Y \geq y_i)} \sum_{s=0}^{y_i-1} P(Y=s)\,(-\log s! + E(\log Y_i!)) \right\} \\ &= \sum_{i=1}^{n} \left\{ (1-\delta_i)(-\log y_i! + E(\log Y_i!)) + \frac{\delta_i}{P(Y \geq y_i)} \left(\sum_{s=0}^{y_i-1} \log s! P(Y=s) - P(Y < y_i)E(\log Y_i!) \right) \right\}.\end{aligned}$$

Appendix B. Estimation Results from Uncensored and Censored Regressions

Table B.7 contains the maximum likelihood estimates obtained via Poisson, COM-Poisson, and GP regression models, respectively, whether the training data were uncensored, censored to 5+, or censored to 4+ children.

For all models, we see that the resulting coefficients for the predictors (for the uncensored or censored cases) are approximately equal. Meanwhile, the COM-Poisson and GP dispersion results illustrate the impact of censoring on the associated dispersion estimate. Censoring the outcome data ($\boldsymbol{Y}$) forces the associated variation to decrease. In this case, because the raw data are already underdispersed, censoring makes the data even more underdispersed; thus, we see the corresponding COM-Poisson dispersion estimate ($\hat{\nu}$) increase, and the GP dispersion estimate ($\hat{\alpha}$) decrease in Table B.7.