Flexible Modeling of Hurdle Conway-Maxwell-Poisson Distributions with Application to Mining Injuries

Emiliano A. Valdez, PhD, FSA
University of Connecticut

Joint work with S. Yin, D. Dey, G. Gan and X. Li

27th International Congress on Insurance: Mathematics and Economics
Chicago, IL United States
8-11 July 2024

Count Data Regression Models

- Count data regression models are used to analyze data that represent count events.
- Essential in situations where the response variable, N, is a non-negative integer, such as the number of accidents, number of doctor visits, or the number of claims filed by a policyholder in a given period.
- Commonly used count data regression models include:
  - Poisson regression: based on Poisson distribution, suitable for data with equidispersion
  - Negative binomial regression: based on NB distribution, used when data exhibit over-dispersion (variance $>>$ mean)
  - Zero-inflated models: models used to address the excess zeros in the data
- Boucher, Denuit, Guillen (2008), Katrien, Valdez (2012), Cameron and Trivedi (2013), Frees, Derrig, Meyers (2014)

The issue of over-dispersion and under-dispersion

- Two frequent characteristics that exist in count data:
  - over-dispersion: the presence of greater variability of the target data than is expected to be
  - zero inflation: the presence of excess of zeros
- The issue of under-dispersion is less frequently addressed in the literature:
  - Less practical
  - Model complexity
  - Less severe implications for decision making
  - Sometimes there is focus on simplicity
- While under-dispersion is less frequently encountered than over-dispersion, there are situations when it may be crucial to identify and address it when it does occur:

Conway-Maxwell-Poisson (CMP) Distribution

The count r.v. N follows a CMP distribution if its pmf has the form

$$P_N(n \mid \lambda, \nu) = \frac{1}{Z(\lambda,\nu)} \frac{\lambda^n}{(n!)^\nu}, \text{ for } n = 0, 1, \ldots,$$

where

$$Z(\lambda,\nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}, \quad \text{and} \quad \nu \geq 0.$$

- No explicit expression for normalizing constant $Z(\lambda, \nu)$; has to be numerically evaluated.
- We write $N \sim \text{CMP}(\lambda, \nu)$.
- When $\nu = 1$, we have the ordinary (standard) Poisson distribution.
- Conway and Maxwell (1962)

Dispersion Parameter

- Parameter ν governs the level of dispersion in the CMP distribution.
- Recall that for Poisson, $\frac{P_N(n-1|\lambda)}{P_N(n|\lambda)} = \frac{n}{\lambda}$. For CMP distribution, we have

$$\frac{P_N(n-1 \mid \lambda, \nu)}{P_N(n \mid \lambda, \nu)} = \frac{n^\nu}{\lambda}.$$

- We see that when $\nu = 1$, the CMP distribution becomes the ordinary Poisson, which describes no dispersion.
- When $\nu < 1$, the rate of decay decreases less than Poisson and has a longer tail; this is the case of over-dispersion.
- When $\nu > 1$, the rate of decay increases more in a nonlinear function, thus shortening the tail of the distribution; this is the case of under-dispersion.

UCONN

Hurdle Count Regression Models

- Data set is described as $D = (\boldsymbol{y}, \boldsymbol{x}, m)$ where $\boldsymbol{y}$ is a vector of responses, $\boldsymbol{x}$ is a matrix of predictor variables, and m is the number of observations.
- Define binary $\boldsymbol{y}^+ = I(\boldsymbol{y} > 0)$ and denote positive subset or zero-truncated count as $\boldsymbol{y}_{zt}$.
- $\boldsymbol{x}_{zt}$ is the subset of the predictor space $\boldsymbol{x}$ corresponding to zero-truncated $\boldsymbol{y}_{zt}$.
- Hurdle count regression model for $\boldsymbol{y}$ is a two-part model:

$$P(y = n) = \begin{cases} 1 - p, & \text{if } n = 0 \\ pP_{zt}(n \mid \gamma, \boldsymbol{x}_{zt}), & \text{if } n = 1, 2, \ldots \end{cases}$$

  where γ is a vector of coefficient parameters.
- The pmf of the modified positive count r.v. $\boldsymbol{y}_{zt}$ is $P_{zt}(n \mid \gamma, \boldsymbol{x}_{zt}) = \frac{P_{zt}(n|\gamma, \boldsymbol{x}_{zt})}{1 - P_{zt}(0|\gamma, \boldsymbol{x}_{zt})}$
- Easy to deduce that $p = P(y > 0) = P(y^+ = 1)$, with complement $1 - p = P(y = 0) = P(y^+ = 0)$.
- p also depends on set of predictors $\boldsymbol{x}$ with coefficients β.

Link functions for the binary component

- Binary component will be described as binary regression model based on its latent variable interpretation.
- y_i^+, for observation i, is related to an unobserved z_i, called latent variable, as $y_i^+ = I(z_i > 0)$, directly linked to predictor as a linear model with an error component as $z_i = \boldsymbol{x}_i\beta + u_i$, where the error component $u_i \sim F$, its distribution function.
- Given $\boldsymbol{x}_i'$, it follows that

$$\begin{aligned} \mathrm{E}(y) = p = \mathrm{Prob}(y_i^+ = 1) &= \mathrm{P}(z_i > 0) \\ &= \mathrm{P}(u_i > -\boldsymbol{x}_i'\beta) = 1 - F(-\boldsymbol{x}_i'\beta). \end{aligned}$$

- When F is the d.f. of a symmetric r.v. u_i with mean 0, we have $p(\boldsymbol{x}_i'\beta) = F(\boldsymbol{x}_i'\beta)$.
- In this case, F^{-1} determines the link function in the GLM framework.
- In this paper, we consider the commonly used logit link regression model.

Zero-truncated CMP for the positive count

- For the positive count/zero-truncated component of the Hurdle model, we can easily show that the zero-truncated CMP distribution has the form:

$$P_{zt}(y_{zt} \mid \lambda, \nu) = \frac{\frac{1}{Z(\lambda,\nu)}}{1 - \frac{1}{Z(\lambda,\nu)}} \frac{\lambda^{y_{zt}}}{(y_{zt}!)^{\nu}} = \frac{1}{Z(\lambda,\nu) - 1} \frac{\lambda^{y_{zt}}}{(y_{zt}!)^{\nu}}, \quad \text{for } y_{zt} = 1, 2, \ldots$$

- We have the zero-truncated Poisson distribution with

$$P_{zt}(y_{zt} \mid \lambda, \nu) = [1/(e^{-\lambda} - 1)]\lambda^{y_{zt}}/y_{zt}!.$$

- To incorporate predictors $\boldsymbol{x}_{zt}$, we use the log link functions $\log(\lambda_i) = \boldsymbol{x}'_{zt,i}\gamma$.

Model Estimation

- Recall that Hurdle model can be decomposed into independent zero and positive count data components.
- Given our data set $D = (\boldsymbol{y}, \boldsymbol{x}, m)$, we can run these two models in parallel because likelihood functions are independent:

$$
\begin{aligned}
L(\beta, \gamma \mid \boldsymbol{y}) &= \prod_i^m [p_i \times P_N^*(y_{zt,i} \mid \gamma, \boldsymbol{x}_i)]^{y_i^+} \times (1 - p_i)^{1-y_i^+} \\
&= \prod_i^m p_i^{y_i^+} (1 - p_i)^{1-y_i^+} \times [P_N^*(y_{zt,i} \mid \gamma, \boldsymbol{x}_i)]^{y_i^+} \\
&= \prod_i^m (1 - F(-\boldsymbol{x}_i'\beta))^{y_i^+} F(-\boldsymbol{x}_i'\beta)^{1-y_i^+} \times [P_N^*(y_{zt,i} \mid \gamma, \boldsymbol{x}_i)]^{y_i^+} \\
&= L(\beta | \boldsymbol{y}^+) \times L(\gamma | \boldsymbol{y}_{zt})
\end{aligned}
$$

UCONN

Posterior Distribution

- The parameters in all the underlying models are fully estimated using Bayesian with multivariate normal priors.
- We combine the prior distribution and the likelihood to obtain the posterior distribution:

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma} | \boldsymbol{y}) \propto L(\boldsymbol{\beta} | \boldsymbol{y}^{+}) \times L(\boldsymbol{\gamma} | \boldsymbol{y}_{zt}) \times \pi(\boldsymbol{\beta}, \boldsymbol{\gamma})$$

- To estimate the posterior, we used MCMC based on the Metropolis-Hastings algorithm.
- For the CMP distribution, we used the Exchange algorithm to handle the intractable likelihood due to the normalizing constant.

Model Assessment

- For model estimate and comparison, we split the data into training and testing set using 70-30 ratio.
- Measures of goodness of fit used:
  - DIC (Deviance Information Criterion): A Bayesian alternative to AIC and BIC. The model with smaller DIC generally exhibits better quality of fit to the data.
  - LPML (Log-pseudo marginal likelihood): A leave-one-out cross-validation with log likelihood as the criterion. We pick the model with largest LPML.
  - WAIC (Watanabe-Akaike information criterion): Particularly suited for Bayesian statistics. The model with better performance yields a smaller WAIC.
- All goodness of fit measures decomposes into the binary component and the zero-truncated component.

Empirical Data Investigation

- We use a dataset from the U.S. Mine Safety and Health Administration (MSHA) observed from 2013 to 2016.
- The dataset was used in the Predictive Analytics exam administrated by the Society of Actuaries in December 2018 [1].
- This dataset contains 53,746 observations described by 20 variables, including compositional variables.

[1] The dataset is available at https://www.soa.org/globalassets/assets/files/edu/2018/2018-12-exam-pa-data-file.zip.

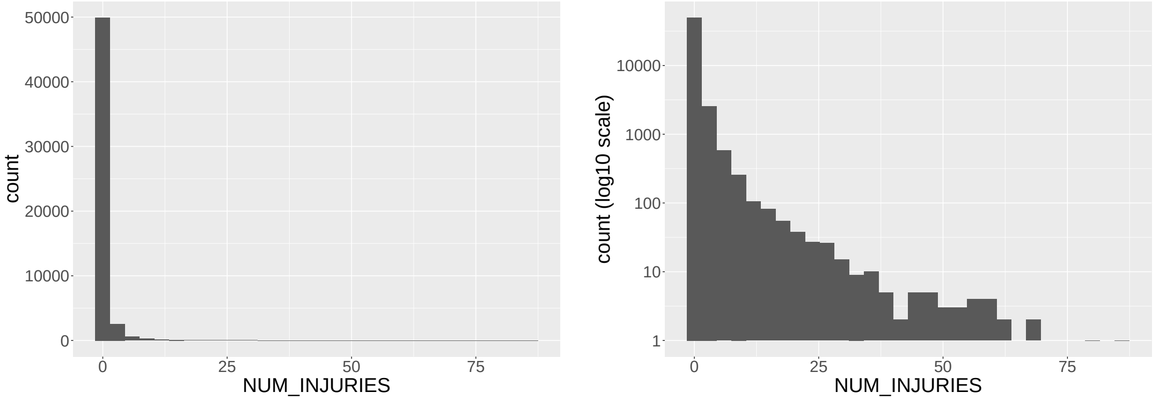
Empirical Application to Mining Injury Data

Table 1: Summary statistics of the number of injuries

Response variable	MIN	1st Q	Mean	MED	3rd Q	90th	95th	99th	MAX
Number of Injuries	0	0	0.4705	0	0	1	1	9	86
Employee Working Months	0.01	2.41	49.36	9.48	30.8	87.38	182.38	805.17	9460.21

* Employee Working Months was used as exposure or offset in the model.

Distribution of the number of mining injuries



Summary statistics of predictor variables

Table 2: Summary statistics of the predictor variables in the mining injury dataset

Categorical attribute	Description		Proportions
Mine type and status	Type of mining methods and status.	Mill&Underground&Active	7.83%
		Sand&Gravel&Active	16.76%
		Sand&Gravel&Intermittent	32.4%
		Surface&Active	27.76%
		Surface&Intermittent	15.25%

Numerical attributes	Description	Min.	1st Q	Median	Mean	3rd Q	Max.
Compositional variables:	Proportion of employee time spending in different work spaces						
Underground operations		-6.1	-4.3	-2.276	-0.428	4.253	10.355
Surface operations		-13.397	-11.6	-9.639	-7.541	-3.044	10.353
Strip mine		-8.23	0.9363	2.216	3.668	10.353	10.353
Auger mining		-5.562	-3.751	-2.094	-0.128	4.791	10.357
Culm bank operations		-4.811	-3.029	-1.332	0.619	5.541	10.361
Dredge operations		-5.9	-4.048	-1.834	-0.194	4.446	10.355
Other surface mining operations		-4.536	-2.754	-1.063	0.874	5.816	10.363
Independent shops and yards		-2.413	-0.628	1.073	3	7.94	10.439
Mills or prep plants		-7	-4.761	0.51	-0.432	3.343	10.353

* Additive logratio transformation is applied to compositional variables.

Estimation results: Poisson, Negative Binomial, CMP

Table 3: Estimation results of model fitting to the mining injury data

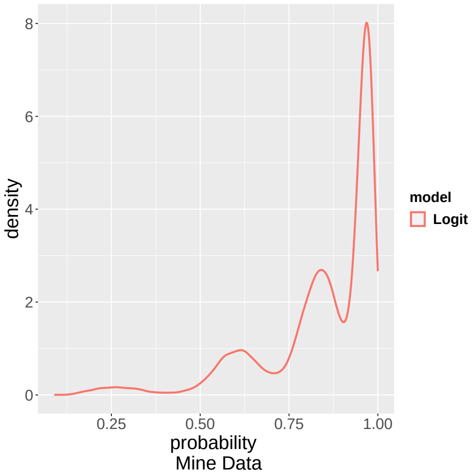
	Poisson			Negative Binomial			CMP		
Variables	estimate	95%lower	95%upper	estimate	95%lower	95%upper	estimate	95%lower	95%upper
Employee time in each type of work									
UnderGround	**0.372**	0.247	0.452	**0.313**	0.244	0.381	**0.623**	0.530	0.725
Surface	-0.132	-0.294	0.088	-0.141	-0.278	0.002	**-0.188**	-0.364	-0.058
Strip	**-0.163**	-0.286	-0.022	**-0.069**	-0.136	-0.010	**-0.191**	-0.246	-0.137
Auger	**0.455**	0.017	0.919	0.089	-0.161	0.345	**0.328**	0.102	0.555
Culm Bank	-0.466	-1.111	0.482	-0.109	-0.330	0.138	-0.028	-0.422	0.522
Mills or Prep Plants	-0.110	-0.332	0.108	-0.030	-0.112	0.062	**-0.280**	-0.430	-0.071
Dredge	0.161	-0.039	0.429	0.089	-0.151	0.332	0.126	-0.398	0.393
Other Surface	-0.012	-0.630	0.329	-0.020	-0.184	0.129	**-0.294**	-0.525	-0.113
Shops and Yards	**0.119**	0.041	0.194	**0.081**	0.037	0.131	**0.301**	0.242	0.367
Mine Status and Type									
Sand and Gravel Active	**0.522**	0.245	0.764	0.129	-0.025	0.294	**1.011**	0.776	1.267
Sand and Gravel Intermittent	**0.508**	0.249	0.758	0.111	-0.071	0.282	**1.026**	0.616	1.551
Surface Active	**0.318**	0.111	0.533	**0.129**	0.013	0.261	**0.687**	0.560	0.844
Surface Intermittent	**0.580**	0.330	0.829	**0.233**	0.066	0.409	**0.599**	0.214	0.955
1/dispersion				**1.285**	1.186	1.384			
dispersion							**0.712**	0.453	0.876
DIC	39077			35877			36140.7		
LPML	-21646.78			-17890			-18130.93		
WAIC	38213.56			35757			36078.15		

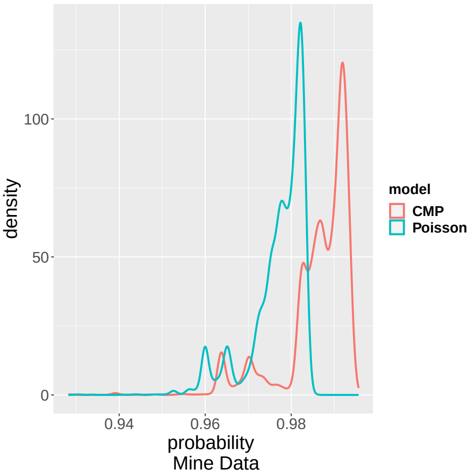
Estimation results: logit, zero-truncated Poisson, zero-truncated CMP

Table 4: Estimation results of model fitting to the mining injury data

	logit			zero-truncated Poisson			zero-truncated CMP		
Variables	estimate	95%lower	95%upper	estimate	95%lower	95%upper	estimate	95%lower	95%upper
Employee time in each type of work									
UnderGround	**1.266**	1.146	1.375	**0.222**	0.173	0.262	**0.163**	0.102	0.214
Surface	**-0.492**	-0.670	-0.292	**-0.196**	-0.280	-0.120	-0.083	-0.185	0.006
Strip	**0.164**	0.066	0.250	**-0.385**	-0.435	-0.333	**-0.404**	-0.488	-0.349
Auger	**-0.477**	-0.804	-0.101	**0.260**	0.141	0.417	**0.401**	0.034	0.776
Culm Bank	**-1.520**	-2.113	-0.939	**0.445**	0.307	0.600	**0.432**	0.266	0.673
Mills or Prep Plants	**0.181**	0.075	0.291	**-0.293**	-0.383	-0.207	**-0.408**	-0.539	-0.297
Dredge	0.152	-0.279	0.583	-0.006	-0.131	0.116	-0.058	-0.226	0.136
Other Surface	0.147	-0.065	0.407	**0.219**	0.097	0.355	**0.303**	0.094	0.471
Mine Status and Type									
Shops and Yards	**0.627**	0.566	0.699	-0.012	-0.047	0.024	-0.013	-0.083	0.019
Sand and Gravel Active	**-0.351**	-0.545	-0.155	**1.171**	1.000	1.347	**1.290**	1.119	1.447
Sand and Gravel Intermittent	**-1.960**	-2.158	-1.735	**1.900**	1.681	2.160	**2.084**	1.757	2.338
Surface Active	0.061	-0.144	0.220	**0.518**	0.397	0.629	**0.508**	0.396	0.664
Surface Intermittent	**-1.702**	-1.908	-1.483	**1.715**	1.532	1.926	**1.832**	1.592	2.079
dispersion							**0.978**	0.960	0.992
DIC				35356.3			34707.51		
LPML				-17389			-16553		
WAIC				35227.76			33578		

Density plot of $P(y_i = 0)$ for Hurdle models





Model performance: expected cost of injuries

Table 5: Model performance comparison based on predicted samples

Model		True Counts			Expected Cost
		Nonevent	1-10	11+	$ 82,291,940
Poisson	Nonevent	11932	1406	2	
	1-10	237	701	35	
	11+	1	54	73	
Negative Binomial	Nonevent	11915	1356	3	$ 76,117,733
	1-10	255	745	37	
	11+	0	60	70	
CMP	Nonevent	11821	1308	2	$ 75,958,412
	1-10	348	813	44	
	11+	1	40	64	
logit+ZTPoisson	Nonevent	8702	613	1	$ 14,725,719
	1-10	3469	1479	33	
	11+	0	69	75	
logit+ZTCMP	Nonevent	8571	594	1	$14,153,858
	1-10	3597	1513	37	
	11+	2	54	72	

* $46,400 is the average cost of nonfatal injuries. Camm, Girard-Dwyer (2005)

Concluding remarks

- Here, we introduce the flexibility of applying the Hurdle structure with Conway-Maxwell-Poisson (CMP) distribution and integrate use of binary link function as better alternative to Poisson and Negative Binomial.
- The CMP distribution introduces a parameterization than can handle a wide range of dispersion: under-dispersion, over-dispersion.
- While not presented here, we also performed simulation studies to understand the performance of the CMP models against other well-known count models such as Poisson and Negative Binomial.
- We offer methods to estimate parameters: Bayesian with MCMC.
- Count regression models will continue to be important in insurance and actuarial science.

Selected references

- Cameron, AC and Trivedi, PK (2013). *Regression Analysis of Count Data*. Cambridge University Press: Cambridge.
- Frees, EW, Derrig, RA and Meyers, G (2014). *Predictive Modeling Applications in Actuarial Science: Vol. I and II*. Cambridge University Press: Cambridge.
- Conway, RW and Maxwell, WL. (1962). A queueing model with state dependent service rate. *Journal of Industrial Engineering*. 12: 132-136.
- Camm T and Girard-Dwyer, J. (2005). Economic consequences of mining injuries. *Mining Engineering*. 57(9): 89-92.
- Boucher, JP et al. (2008). Models of insurance claim counts with time dependence based on generalization of Poisson and negative binomial distributions. *Variance*. 2(1): 135-162.
- Antonio, K and Valdez, EA (2012). Statistical concepts of *a priori* and *a posteriori* risk classification in insurance. *AStA Advances in Statistical Analysis*. 96: 187-224.
- Shuang, Y et al. (2024). Flexible Modeling of Hurdle Conway–Maxwell–Poisson Distributions with Application to Mining Injuries. *Journal of Statistical Theory and Practice*. 18:34.