The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$115	1-year subscription	$145
2-year subscription	$210	2-year subscription	$270
3-year subscription	$285	3-year subscription	$375
1-year student subscription	$ 85	1-year student subscription	$115
1-year institutional subscription	$345	1-year institutional subscription	$375
2-year institutional subscription	$625	2-year institutional subscription	$685
3-year institutional subscription	$875	3-year institutional subscription	$965
Electronic only		**Electronic only**	
1-year subscription	$ 85	1-year subscription	$ 85
2-year subscription	$155	2-year subscription	$155
3-year subscription	$215	3-year subscription	$215
1-year student subscription	$ 55	1-year student subscription	$ 55

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.


The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, STATA, Stata Press, Mata, MATA, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2015)
15, Number 2, pp. 457–479

Modeling heaped count data

Tammy H. Cummings
Institute for Families in Society
University of South Carolina
Columbia, SC
harris68@mailbox.sc.edu

James W. Hardin
Institute for Families in Society
Department of Epidemiology and Biostatistics
University of South Carolina
Columbia, SC
jhardin@sc.edu

Alexander C. McLain
Department of Epidemiology and Biostatistics
University of South Carolina
Columbia, SC
mclaina@mailbox.sc.edu

James R. Hussey
Department of Epidemiology and Biostatistics
University of South Carolina
Columbia, SC
jhussey@mailbox.sc.edu

Kevin J. Bennett
Department of Family and Preventive Medicine
University of South Carolina
Columbia, SC
kevin.bennett@uscmed.sc.edu

Gina M. Wingood
Department of Behavioral Sciences and Health Education
Emory University
Atlanta, GA
gwingoo@sph.emory.edu

Abstract. We present motivation and new commands for modeling heaped count data. These data may appear when subjects report counts that are rounded or favor multiples (digit preference) of a certain outcome, such as the number of cigarettes reported. The new commands for fitting count regression models (Poisson, generalized Poisson, negative binomial) are also accompanied by real-world examples comparing the heaped regression model with the usual regression model as well as the heaped zero-inflated model with the usual zero-inflated model.

Keywords: st0388, heapcr, ziheapcr, heapr, ziheapr, count data, heaping, Poisson, generalized Poisson, negative binomial, zero-inflation, interval censored, mixture, rescaled

1 Introduction

Heaped data result when subjects who recall the frequency of events demonstrate a preference for reporting from a limited set of rounded responses or preferred digits over reporting exact counts. Examples of these rounded responses and digit preferences (also referred to as data coarsening) can be characterized by reported frequencies (or counts) favoring multiples of 20 (for example, number of cigarettes smoked), reporting counts ending with 0 or 5, or a preference for reporting an even number over an odd number. This mixture of exact and coarsened values is a type of measurement error (pattern of misreporting) that induces increased variance and can lead to biased estimation and imprecision in predicted probabilities for discrete quantitative data.

Sometimes this pattern in data can be explained, but its effect on the statistical inference is harder to anticipate. Researchers will need to notice the way survey questions are worded. For instance, in response to the question "How many cigarettes per day did you smoke in the past 30 days?", there might be heaping on the 5s because of the number of days in a normal work week or heaping on the 7s because of the total number of days in a week. Another example of survey question wording is "In the past 12 months, how many days did you miss work?" In response, there might be heaping on the 5s because of the number of days in a normal work week or heaping on the 12s because of the number of months in the question itself. If asked "What year did you learn to drive?", most respondents would round to years ending in 5 or 0 because they lack recall of the exact year.

A visual representation of heaped data is illustrated in a frequency distribution (histogram or spikeplot), where the heaps are represented as periodic peaks or spikes within the overall data layout. However, the researcher would need to evaluate the survey questions and data carefully to investigate the presence of heaping.

Heaped counts are reported in cigarette cessation studies (Wang and Heitjan 2008; Klesges, Debon, and Ray 1995; Lewis-Esquerre et al. 2005). Participants in these types of smoking studies tend to round their cigarette counts to multiples of 20, 10, or 5, which reflects a preference for heaping counts into "packs" or fractions of a pack. Another study type where heaping (observer bias) can occur is health studies that collect blood pressure measurements. These reported measurements often display terminal digit preference (Nietert et al. 2006), where blood pressure readings tend to be recorded in measurements ending in 0 or 5 and even numbers are preferred over odd numbers. Other examples of heaped data include unemployment duration (Wolff and Augustin 2003), reported age (Pardeshi 2010), reported weight, frequency of sexual intercourse, number of months breastfeeding (Roberts and Brewer 2001), number of menstrual cycles before pregnancy (Ridout and Morgan 1991; McLain et al. 2014), and reported birth weight (Channon, Padmadas, and McDonald 2011).

We develop two statistical models for heaped count data using a mixture of likelihood functions for the heaped and nonheaped count data. In the first method, we assume that the reported outcome is not exactly known but is actually censored over the half-width of the heaping multiple. Simultaneously, we assume that nonheaped (not censored)

data follow the same count distribution's likelihood for exact counts. For example, count data that are heaped at multiples of 10 have a probability function as follows: $P\{Y \in (y - \lfloor 10/2 \rfloor, y + \lfloor 10/2 \rfloor)\}$, where $\lfloor x \rfloor$ is the greatest integer that is less than or equal to x. The count data that are reported at nonmultiples of 10 will be treated as exact results using $P(Y = y)$ for exact counts.

For the second proposed method, we again assume that the reported outcome is not exactly known but is a mixture of rescaled distributions. For this mixture, the reported outcome rates are equivalent, so the count distributions [Poisson, generalized Poisson (GP), negative binomial (NB)] differ only in the reciprocal of the heaping number on the specified time period. For both methods, the investigator specifies the heaping multiples in the count data via the interval-regression approach or the mixture of rescaled distributions approach.

In terms of model selection criteria (Akaike information criteria and Bayesian information criteria), the interval-censored approach and rescaled-mixture approach described herein will have approximately the same preference over the standard model as does a zero-inflated version of the standard model. Both approaches, however, will produce predicted probabilities that are closer to the data than will be produced by the standard model.

Herein, we propose two methods for modeling heaped count data using the Poisson, GP, and NB distributions along with their zero-inflated versions. In section 2, we review appropriate count-data regression models for both approaches. In section 3, syntax is presented for each new command, followed by real-world data examples in sections 4 and 5. Finally, the summary and conclusions are presented in section 6.

2 The methods

2.1 Interval-censored method

For a random variable Y_i, we have a response vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, where n is the sample size and Y_i, Y_j are independent and identically distributed for $i \neq j$. In this section, y_{Li} and y_{Ri} denote the right and left endpoints, respectively, of the interval-censored count observations.

$$\begin{aligned} y_{Li} &= \max(0, y_i - \Delta_i) \\ y_{Ri} &= y_i + \Delta_i \\ \Delta_i &= \max_{j=1,\ldots,H} \{\lfloor h_j/2 \rfloor \times I(y_i \bmod h_j = 0)\} \end{aligned}$$

where $I(\ell)$ is an indicator function equal to 0 if ℓ is false and equal to 1 if ℓ is true, $\lfloor h_j/2 \rfloor$ is the half-width of the heaping interval, $h_1 = 1$, and H is the total number of heaping intervals. If all observations are exact (no heaping), then $H = 1$ and these formulas simplify to that of Poisson, GP, and NB regression, respectively.

Poisson model

Poisson regression analysis is the most common approach to modeling response variables comprising count data. This distribution describes the probability of the number of event occurrences, and the parameter is the expected number of occurrences that can be modeled through explanatory variables. Covariates are included in the regression model by an invertible link function describing the relationship of the linear predictor $\mathbf{x}_i\boldsymbol{\beta} = \eta_i$ to the expected value of the responses μ_i. The probability mass function of the Poisson distribution is

$$f(y_i;\mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}, y_i = 0, 1, 2, \ldots, \quad \mu_i > 0$$

The expected outcome in terms of the inverse of the log link function is given by $\mu_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$, where $\mathbf{x}_i$ is a row vector of covariates for the ith observation and $\boldsymbol{\beta}$ is a vector of regression parameters to be estimated. The Poisson model assumes equidispersion; that is, it assumes that the mean (μ_i) and variance (μ_i) of the outcomes are equal for a given set of covariates. For a random sample of observations $y_1, y_2, \ldots, y_n$, the Poisson regression log-likelihood function is given by

$$\mathcal{L} = \sum_{i=1}^{n} \{y_i \ln(\mu_i) - \mu_i - \ln\Gamma(y_i + 1)\}$$

For the interval-censored regression method, the log likelihood is given in terms of the log of the probability of being in an interval. The probability of being in the interval is calculated using the survival probabilities

$$p_{1i} = P(Y > y_{Li} - 1|Y \sim \text{Poisson}) = \Gamma_I(y_{Li}, \mu_i) = 1 - P(Y \leq y_{Li} - 1|Y \sim \text{Poisson})$$
$$p_{2i} = P(Y > y_{Ri}|Y \sim \text{Poisson}) = \Gamma_I(y_{Ri} + 1, \mu_i) = 1 - P(Y \leq y_{Ri}|Y \sim \text{Poisson})$$

where

$$\Gamma_I(y, \mu) = \frac{1}{\Gamma(y)} \int_0^y t^{\mu-1} e^{-t} dt$$

is the regularized incomplete gamma function, and p_{1i} and p_{2i} are survival probabilities. (Their difference represents the probability of the event occurring in the interval.) Hence, the interval-censored Poisson regression model has a log likelihood of

$$\begin{aligned} \mathcal{L} &= \sum_{i=1}^{n} \ln P\left\{Y \in (y_{Li}, y_{Ri}) \,|y_i \sim \text{Poisson}\right\} \\ &= \sum_{i=1}^{n} \ln(p_{1i} - p_{2i}) \end{aligned}$$

We will use the command `heapcr` with the `poisson` option to refer to our proposed interval-censored regression method for Poisson regression of heaped data.

GP model

We consider a regression model based on the GP distribution for equidispersed or for possibly overdispersed or underdispersed count data Y. This model assumes the response variable Y_i has probability mass function

$$f(y_i;\mu_i,\alpha) = \frac{\mu_i(\mu_i+\alpha y_i)^{y_i-1}e^{-\mu_i-\alpha y_i}}{y_i!}, \quad y_i = 0,1,2,\ldots$$

where α is the dispersion parameter, $\mu_i > 0$, $\max(-1,-\mu_i/4) < \alpha < 1$, and $\mu_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$. The mean and variance for the GP distribution is as follows (also see Hardin and Hilbe [2012]):

$$E(Y_i) = \frac{\mu_i}{1-\alpha} \quad \text{and} \quad \text{Var}(Y_i) = \frac{\mu_i}{(1-\alpha)^3}$$

For a random sample of observations $y_1, y_2, \ldots, y_n$, the GP log-likelihood function is

$$\mathcal{L} = \sum_{i=1}^{n} \{\ln\mu_i + (y_i-1)\ln(\mu_i+\alpha y_i) - \mu_i - \alpha y_i - \ln\Gamma(y_i+1)\}$$

Consul and Famoye (1992) and Consul (1989) illustrated that covariates can be introduced into a regression model via the relationship

$$\ln\frac{\mu_i}{1-\alpha} = \sum_{r=1}^{p} x_{ir}\beta_r$$

where x_{ir} is the ith observation of the rth covariate, p is the number of covariates in the model, and β_r is the rth regression parameter. The probability of being in the interval is calculated using the survival probabilities

$$\begin{aligned} p_{1i} &= P(Y > y_{Li}-1|Y \sim \text{Gen. Poisson}) = \Gamma_I(y_{Li}\alpha,\mu_i) \\ p_{2i} &= P(Y > y_{Ri}|Y \sim \text{Gen. Poisson}) = \Gamma_I\{(y_{Ri}\alpha)+1,\mu_i\} \end{aligned}$$

Therefore, the log-likelihood function suitable for heaped data under a GP model is

$$\mathcal{L} = \sum_{i=1}^{n} \ln(p_{1i}-p_{2i})$$

where p_{1i} and p_{2i} are survival probabilities for which the difference represents the probability of the event occurring in the interval. We will use the command `heapcr` with the `gpoisson` option to refer to our proposed method for GP regression of heaped data. Even addressing some of the overdispersion through the censored approach of the heaped data regression model, the likelihood-ratio test (LRT) of the dispersion parameter is still important, and when significant, it indicates a preference for the heaped GP model over the heaped Poisson model.

NB model

If the probability of success in each trial is given by p_i and the probability of failure is given by $(1-p_i)$, then the general probability mass function of the NB distribution is given by

$$f(y_i;\alpha,p_i)=\frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\alpha}\right)}p_i^{1/\alpha}(1-p_i)^{y_i},\quad y_i=0,1,2,\ldots$$

where α is the dispersion parameter. When $\alpha\to 0$, this reduces to the Poisson distribution. The mean and variance for the NB distribution are as follows:

$$\begin{aligned}E(Y_i) &= \frac{1-p_i}{\alpha p_i}\\ \text{Var}(Y_i) &= \frac{1-p_i}{\alpha p_i^2}=\frac{p_i(1-p_i)+(p_i-1)^2}{\alpha p_i^2}\end{aligned}$$

The NB can be parameterized using the inverse of the log-link specification $g(\mathbf{x}_i;\boldsymbol{\beta})=\exp(\mathbf{x}_i\boldsymbol{\beta})$ (Lawless 1987), where $\mathbf{x}_i$ is the $p\times 1$ vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of regression parameters. Lawless (1987) states that a Poisson model would stipulate that the distribution of $Y|\mathbf{x}$ is Poisson with a mean equal to $\mu_i(\mathbf{x}_i)=T\{g(\mathbf{x}_i;\boldsymbol{\beta})\}$. Consequently, the NB regression model is

$$f(y_i;\alpha,\mu_i)=\frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i},\quad y_i=0,1,2,\ldots$$

where α is the dispersion parameter. The common reparameterization $p_i=(1+\alpha\mu_i)^{-1}$, where p_i then depends on the covariates $\mathbf{x}_i$, results in the mean and variance of Y_i as

$$\begin{aligned}E(Y_i) &= \mu_i\\ \text{Var}(Y_i) &= \mu_i+\alpha\mu_i^2\end{aligned}$$

Therefore, we have $Y\sim\text{NB}(\mu,\alpha)$. When $\alpha\to 0$, this reduces to the Poisson model. For a random sample of observations $y_1,y_2,\ldots,y_n$, the NB log-likelihood function is

$$\begin{aligned}\mathcal{L} = \sum_{i=1}^{n}\Bigg[&\ln\{\Gamma(y_i+1/\alpha)\}-\ln\{\Gamma(y_i+1)\}-\ln\{\Gamma(1/\alpha)\}\\ &+(1/\alpha)\ln\left(\frac{1}{1+\alpha\mu_i}\right)+(y_i)\ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)\Bigg]\end{aligned}$$

To obtain the log-likelihood function (interval-censored regression) for heaped NB data, we define interval regression using the survival probabilities

$$\begin{aligned}p_{1i} &= P(Y\geq y_{Li}|Y\sim\text{Neg. Binomial})=B_I\{y_{Li},\alpha,1/(1+\alpha\mu_i)\}\\ p_{2i} &= P(Y\geq y_{Ri}+1|Y\sim\text{Neg. Binomial})=B_I\{y_{Ri}+1,\alpha,1/(1+\alpha\mu_i)\}\end{aligned}$$

where $B_I(\cdot)$ is the three-parameter incomplete beta function

$$B_I(m;\psi,\phi) = \int_0^m t^{\psi-1}(1-t)^{\phi-1}dt$$

The resulting log likelihood can be succinctly written as

$$\mathcal{L} = \sum_{i=1}^{n} \ln(p_{1i} - p_{2i})$$

We will use the command `heapcr` with the `nbreg` option to refer to our proposed method for NB regression of heaped data.

Zero-inflated models

In applications with an excess of 0s in count response data, Poisson (and other) distribution models may not be appropriate to use. Ridout, Demétrio, and Hinde (1998) summarized some literature and cited examples from agriculture, econometrics, manufacturing, patent applications, road safety, species abundance, medical consultations, use of recreational facilities, and even sexual behavior. Hardin and Hilbe (2012) describe the two origins of zero outcomes:

1. individuals who do not enter into the counting process, and
2. individuals who enter into the counting process and have a zero outcome.

Therefore, the model must be separated into different parts, one consisting of a zero count $y_i = 0$ and the other consisting of a nonzero count $y_i > 0$.

$$P(Y_i = y_i) = \begin{cases} w_i + (1-w_i)f(y_i) & y_i = 0 \\ (1-w_i)f(y_i) & y_i = 1, 2, \ldots \end{cases} \tag{1}$$

where w_i is the probability of 0s (binary distribution), $0 \leq w_i < 1$, and $f(y_i)$ is the discrete probability function. For our interval-censored approach, the zero-inflated heaped count-data log likelihoods for Poisson, GP, and NB distributions can be shown as

$$\mathcal{L} = \sum_{i \in Z} \ln\{w_i + (1-w_i)f(0)\} + \sum_{i \notin Z} \ln\{(1-w_i)(p_{1i} - p_{2i})\}$$

where Z is the set of 0 outcomes, and p_{1i} and p_{2i} are from our interval-regression equations above. Zero-inflation models for heaped data are comprised under the `ziheapcr` command using options `poisson`, `gpoisson`, and `nbreg` for zero-inflated Poisson, zero-inflated GP, and zero-inflated NB distributions, respectively.

2.2 Mixture of rescaled distributions method

For the proposed mixture of rescaled distributions, we consider two behaviors of subjects. Behavior 1 consists of those subjects who report an exact count of the requested

frequency. Behavior 2 consists of those subjects who remember the requested frequency over $1/k$th of a specified period of time and then report k times that amount. Under behavior 1, we consider covariates and parameters $\boldsymbol{\beta}$ are associated with the mean $\mu_i^{[1]}$ under the log-link function $\ln(\mu_i^{[1]}) = \mathbf{X}_i\boldsymbol{\beta}$. In the same way, for behavior 2, we use the same covariates and parameters for the associated mean, $\mu_i^{[2]}$, as $1/k$ times $\mu_i^{[1]}$ (mean of behavior 1) under the log-link function $\ln(k\mu_i^{[2]}) = \mathbf{X}_i\boldsymbol{\beta}$, which is $\ln(\mu_i^{[2]}) = \mathbf{X}_i\boldsymbol{\beta} - \ln(k)$. Here, notice the only difference in the means for the two behaviors is the offset term $\ln(k)$. We exponentiate the coefficients to show rates that will remain constant over all time periods because of the reparameterization of the covariates and parameters.

We will denote a binary model B to represent a subject choosing behavior 2 as a function of covariates $\mathbf{J}$ and coefficients $\boldsymbol{\phi}$. This model fits the likelihood of a subject choosing behavior 2. This likelihood multiplies the likelihood of the unscaled outcome (the reported outcome divided by the heaping number), where the reciprocal of the heaping number serves as the exposure $\{\ln(k)\}$, which is included as part of the linear predictor. Therefore, the mixture probability model for a reported outcome is given by

$$\begin{aligned} P(Y_i = y_i) &= P_B(b_i = 2|\mathbf{J}_i, \boldsymbol{\phi}) P_{[2]}\left[Y_i = \frac{y_i}{k} | \mu_i = \exp\{\mathbf{X}_i\boldsymbol{\beta} - \ln(k)\}\right] I_{(y_i \bmod k = 0)} \\ &\quad + P_B(b_i = 1|\mathbf{J}_i, \boldsymbol{\phi}) P_{[1]}\{Y_i = y_i | \mu_i = \exp(\mathbf{X}_i\boldsymbol{\beta})\} \end{aligned} \tag{2}$$

where

$$\begin{aligned} P_B(b_i = 2|\mathbf{J}_i, \boldsymbol{\phi}) &= \exp(\mathbf{J}_i\boldsymbol{\phi})/\{1 + \exp(\mathbf{J}_i\boldsymbol{\phi})\} \quad \text{and} \\ P_B(b_i = 1|\mathbf{J}_i, \boldsymbol{\phi}) &= 1 - P_B(b_i = 2|\mathbf{J}_i, \boldsymbol{\phi}) = 1/\{1 + \exp(\mathbf{J}_i\boldsymbol{\phi})\} \end{aligned}$$

Next, we consider s heaping numbers $k_2, \ldots, k_s, k_{s+1}$, where we assume each response is the result of one of $s+1$ behaviors, which match the s heaping numbers and the one behavior of reporting on the specified time period. This multinomial model S is then used to estimate whether a subject chooses $s+1, s, \ldots, 2$ as a function of covariates $\mathbf{J}$ and coefficients $\boldsymbol{\phi}$ using behavior 1 as the reference. Thus, the probability model for a particular outcome is

$$
\begin{aligned}
P(Y_i = y_i|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\phi}) &= P_S(b_i = s+1|\mathbf{J}_i, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_{s+1}) \\
&\quad \times P_{[s+1]}\left[Y_i = \frac{y_i}{k_{s+1}}|\mu_i = \exp\{\mathbf{X}_i\boldsymbol{\beta} - \ln(k_{s+1})\}\right] I_{(y_i \bmod k_{s+1}=0)} \\
&\quad + P_S(b_i = s|\mathbf{J}_i, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_{s+1}) \\
&\quad \times P_{[s]}\left[Y_i = \frac{y_i}{k_s}|\mu_i = \exp\{\mathbf{X}_i\boldsymbol{\beta} - \ln(k_s)\}\right] I_{(y_i \bmod k_s=0)} \\
&\quad + \cdots + P_S(b_i = 2|\mathbf{J}_i, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_{s+1}) \\
&\quad \times P_{[2]}\left[Y_i = \frac{y_i}{k_2}|\mu_i = \exp\{\mathbf{X}_i\boldsymbol{\beta} - \ln(k_2)\}\right] I_{(y_i \bmod k_2=0)} \\
&\quad + P_S(b_i = 1|\mathbf{J}_i, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_{s+1}) P_{[1]}\left\{Y_i = y_i|\mu_i = \exp(\mathbf{X}_i\boldsymbol{\beta})\right\} \qquad (3)
\end{aligned}
$$

where

$$
\begin{aligned}
P_S(b_i = k|\mathbf{J}_i, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_{s+1}) &= \exp(\mathbf{J}_i\boldsymbol{\phi}_k)/\{1 + \exp(\mathbf{J}_i\boldsymbol{\phi}_2) + \cdots \\
&\qquad + \exp(\mathbf{J}_i\boldsymbol{\phi}_{s+1})\}, \quad k = 2, \ldots, s+1 \\
P_S(b_i = 1|\mathbf{J}_i, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_{s+1}) &= 1/\{1 + \exp(\mathbf{J}_i\boldsymbol{\phi}_2) + \cdots + \exp(\mathbf{J}_i\boldsymbol{\phi}_{s+1})\}
\end{aligned}
$$

The log likelihood for a count-data model, in a general form, is the sum of the logs of the probabilities of observed outcomes given by

$$
\mathcal{L} = \sum_{i=1}^{n} \ln\left\{P(Y_i = y_i|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\phi})\right\}
$$

where $P(Y_i = y_i|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\phi})$ is from (3). When only one heaping multiple k exists and the counting process is given by the Poisson distribution, the log likelihood is given by

$$
\begin{aligned}
\mathcal{L} = \sum_{i=1}^{n} \ln\Bigg(&\left\{\frac{\exp(\mathbf{j}_i\boldsymbol{\phi})}{1 + \exp(\mathbf{j}_i\boldsymbol{\phi})}\right\} \exp\Big[- \exp\{\mathbf{x}_i\boldsymbol{\beta} - \ln(k)\} + \frac{y_i}{k}\{\mathbf{x}_i\boldsymbol{\beta} - \ln(k)\} \\
&- \ln\Gamma\left(\frac{y_i}{k} + 1\right)\Big] I_{(y_i \bmod k=0)} + \left\{\frac{1}{1 + \exp(\mathbf{j}_i\boldsymbol{\phi})}\right\} \exp\{-\exp(\mathbf{x}_i\boldsymbol{\beta}) \\
&+ y_i(\mathbf{x}_i\boldsymbol{\beta}) - \ln\Gamma(y_i + 1)\}\Bigg)
\end{aligned}
$$

Similarly, where the counting process is given by the GP distribution, the log likelihood is given by

$$\mathcal{L} = \sum_{i=1}^{n} \ln\left(\left\{ \frac{\exp(\mathbf{j}_i\boldsymbol{\phi})}{1+\exp(\mathbf{j}_i\boldsymbol{\phi})} \right\} \exp\Big[-\exp\{\mathbf{x}_i\boldsymbol{\beta} - \ln(k)\} + \frac{y_i}{k}\{\mathbf{x}_i\boldsymbol{\beta} - \ln(k)\} \right.$$
$$- \ln\Gamma\left(\frac{y_i}{k}+1\right)\Big] I_{(y_i \bmod k=0)} + \left\{ \frac{1}{1+\exp(\mathbf{j}_i\boldsymbol{\phi})} \right\} \exp[-(1-\alpha)\mathbf{x}_i\boldsymbol{\beta} + \alpha y_i$$
$$\left. + (y_i-1)\ln\{(1-\alpha)\mathbf{x}_i\boldsymbol{\beta} + \alpha y_i\} + \ln(\mathbf{x}_i\boldsymbol{\beta}) + \ln(1-\alpha) - \ln\Gamma(y_i+1)] \right)$$

And lastly, for an NB distribution counting process, the log likelihood is given by

$$\mathcal{L} = \sum_{i=1}^{n} \ln\left(\left\{ \frac{\exp(\mathbf{j}_i\boldsymbol{\phi})}{1+\exp(\mathbf{j}_i\boldsymbol{\phi})} \right\} \exp\Big[-\exp\{\mathbf{x}_i\boldsymbol{\beta} - \ln(k)\} + \frac{y_i}{k}\{\mathbf{x}_i\boldsymbol{\beta} - \ln(k)\} \right.$$
$$- \ln\Gamma\left(\frac{y_i}{k}+1\right)\Big] I_{(y_i \bmod k=0)} + \left\{ \frac{1}{1+\exp(\mathbf{j}_i\boldsymbol{\phi})} \right\} \exp\left\{ \ln\Gamma\left(\frac{1}{\alpha}+y_i\right) - \ln\Gamma\left(\frac{1}{\alpha}\right) \right.$$
$$\left. \left. + \frac{1}{\alpha}\ln\left(\frac{1}{1+\alpha\mu_i}\right) + y_i\ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \ln\Gamma(y_i+1) \right\} \right)$$

Zero-inflated models

As in (1), we use the same general definition of zero-inflation. However, for our mixture of rescaled distributions approach, the zero-inflated heaped count-data log likelihoods for Poisson, GP, and NB distributions can be shown as

$$\mathcal{L} = \sum_{i\in Z} \ln\{w_i + (1-w_i)f(0)\}$$
$$+ \sum_{i\notin Z} \ln\left((1-w_i)[P_B(b_i=1|\mathbf{J}_i,\boldsymbol{\phi})P_{[1]}\{Y_i=y_i|\mu_i=\exp(\mathbf{X}_i\boldsymbol{\beta})\}]\right)$$

where Z, again, is the set of 0 outcomes and $P_B(b_i = 1|\mathbf{J}_i,\boldsymbol{\phi})P_{[1]}\{Y_i = y_i|\mu_i = \exp(\mathbf{X}_i\boldsymbol{\beta})\}$ is from (2) above. Zero-inflation models for heaped data are comprised under the `ziheapr` command using options `poisson`, `gpoisson`, and `nbreg` for zero-inflated Poisson, zero-inflated GP, and zero-inflated NB distributions, respectively.

3 Syntax

The accompanying software includes the command files as well as supporting files for prediction and help. In the following syntax diagrams, unspecified options include the usual collection of maximization and display options available to all estimation commands. In addition, all zero-inflated commands include the `ilink(`*linkname*`)` option to specify the link function for the inflation model.

The syntax for specifying an interval-censored regression model for heaped count data is given by

```
heapcr depvar [indepvars] [if] [in] [weight] [, noconstant
    exposure(varname_e) offset(varname_o) constraints(constraints)
    collinear uniform gplrtest heap(numlist) width(numlist)
    {poisson|gpoisson|nbreg} vce(vcetype) level(#) irr nocnsreport
    display_options maximize_options coeflegend]
```

with options `poisson`, `gpoisson`, and `nbreg` for each of the supported discrete distributions, respectively.

The syntax for heaped zero-inflated count data (using the interval-censored regression method) is given by

```
ziheapcr depvar [indepvars] [if] [in] [weight],
    inflate(varlist[, offset(varname)]|_cons) [ilink(linkname)
    exposure(varname_e) offset(varname_o) constraints(constraints)
    collinear uniform gplrtest vuong heap(numlist) width(numlist)
    {poisson|gpoisson|nbreg} vce(vcetype) level(#) nolrtest irr
    nocnsreport display_options maximize_options coeflegend]
```

with options `poisson`, `gpoisson`, and `nbreg` for each of the supported discrete distributions, respectively. Specific half-widths may be specified via the `width()` option, or default values are equal to the heaping factors. For example, for a model that specifies `heap(6,11)`, the widths are also (6,11) and the half-widths are (3,5).

The syntax for specifying a mixture of rescaled distributions regression model for heaped count data is given by

```
heapr depvar [indepvars] [if] [in] [weight] [, noconstant
    exposure(varname_e) offset(varname_o) constraints(constraints)
    collinear gplrtest hvars(string) heap(numlist) {poisson|gpoisson|nbreg}
    vce(vcetype) level(#) nolrtest irr vuong nocnsreport display_options
    maximize_options coeflegend]
```

with options `poisson`, `gpoisson`, and `nbreg` for each of the supported discrete distributions, respectively.

The syntax for heaped zero-inflated count data (using the mixture of rescaled distributions regression method) is given by

```
ziheapr depvar [indepvars] [if] [in] [weight],
    inflate(varlist[, offset(varname)] | _cons) [ilink(linkname)
    exposure(varname_e) offset(varname_o) constraints(constraints)
    collinear gplrtest vuong heap(numlist) hvars(varlist)
    {poisson|gpoisson|nbreg} vce(vcetype) level(#) nolrtest irr
    nocnsreport display_options maximize_options coeflegend]
```

with options `poisson`, `gpoisson`, and `nbreg` for each of the supported discrete distributions, respectively.

A Vuong test (see Vuong [1989]) evaluates whether the regression model with zero-inflation or the regression model without zero-inflation is closer to the true model. A random variable ω is defined as the vector $\log L_{\mathrm{Z}} - \log L_{\mathrm{S}}$, where L_{Z} is the likelihood of the zero-inflated model evaluated at its maximum likelihood estimate and L_{S} is the likelihood of the standard (nonzero-inflated) model evaluated at its maximum likelihood estimate. The vector of differences over the N observations is then used to define the statistic

$$V = \frac{\sqrt{N}\overline{\omega}}{\sqrt{\sum_i(\omega_i - \overline{\omega})^2/(N-1)}}$$

which, asymptotically, has a standard normal distribution. A significant positive statistic indicates preference for the zero-inflated model, and a significant negative statistic indicates preference for the model without zero-inflation. Nonsignificant Vuong statistics indicate no preference for either model. A Vuong test evaluates and tests the hypothesis that nested models and nonnested models are the same distance from the true model. Therefore, we can apply this test to our proposed heaped models (nonnested) versus nonheaped models (nested). Results of this test are included in a footnote to the estimation of the model when the user includes the `vuong` option in any of the zero-inflated commands.

Replacing an exactly measured outcome with an interval-censored outcome increases the probability (and log likelihood) by definition. To enable comparisons of heaped and nonheaped models using, for example, the Akaike information criterion (see Desmarais and Harden [2013]), the probability of the interval-censored outcome is scaled so that the total contribution of the interval is that of the weighted average over the individual outcomes in the interval. By default, triangular weights are applied, but the user can request uniform weights by using the `uniform` option.

4 NHANES example

Using the National Health and Examination Survey (NHANES) 2009–2010 data,[1] we model for 1,504 participants the average number of cigarettes smoked per day during the past 30 days (`smd650`) as a function of the covariate's age (`ridageyr`), gender (`gendernew`), and race (`racenew`). We recoded the original `ridreth1` variable, now called `racenew`, that includes non-Hispanic white versus others (Mexican American, other Hispanic, non-Hispanic black, other race/multiracial). We also recoded the original `riagendr` variable, now called `gendernew`. Selected characteristics of the given variables above from the dataset are given in table 1.

Table 1. Selected characteristics from the NHANES example ($n = 1504$)

Characteristic	Frequency
Cigarettes smoked/day in the past 30 days, mean (standard deviation [SD])	11.55 (9.98)
Age, mean (SD)	40.73 (16.64)
Gender, No. (%)	
Females	669 (44.48)
Males	835 (55.52)
Race, No. (%)	
Non-Hispanic white	749 (49.80)
Other races	755 (50.20)
Cigarettes smoked/day in the past 30 days, mean (SD)	
Females	11.17 (9.13)
Males	11.85 (10.61)
Non-Hispanic white	14.81 (10.62)
Other races	8.31 (8.11)

To visually investigate where heaping may exist in the average number of cigarettes smoked per day during the past 30 days, we graph the reported data in a spikeplot in figure 1.

1. The participants in this study provided informed consent for the collection of data, and the data are freely available in de-identified format at http://www.cdc.gov/nchs/nhanes/nhanes2009-2010/nhanes09_10.htm (accessed in March 2013).

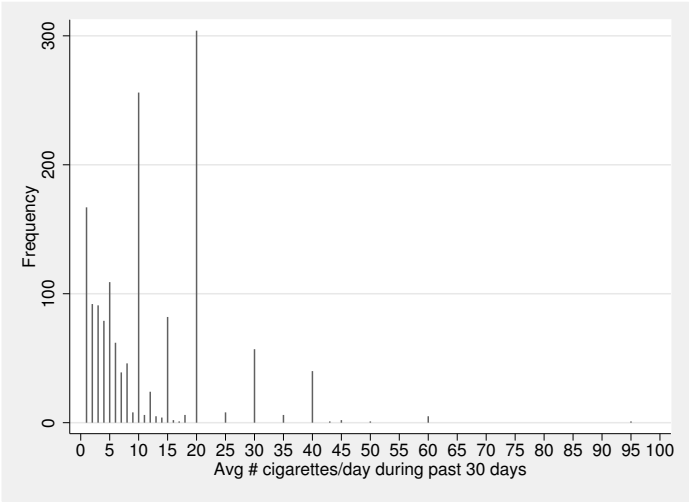

Figure 1. Average number of cigarettes smoked per day during the past 30 days

We see that heaping is present at multiples of 5 (that is, 5, 10, 15, etc.). Therefore, we specify heaping at multiples of 5 at which the outcomes are treated as being interval-censored with an interval half-width of $\lfloor 5/2 \rfloor$. We also notice that there are no 0s in our outcome variable, so the zero-inflated versions of our new commands will not be illustrated for these data.

4.1 Poisson

By fitting a regular Poisson model to the outcome, the results are given by

```
. poisson smd650 gendernew racenew ridageyr, nolog
Poisson regression                              Number of obs     =      1,504
                                                LR chi2(3)        =    2107.84
                                                Prob > chi2       =     0.0000
Log likelihood = -7782.0546                     Pseudo R2         =     0.1193

------------------------------------------------------------------------------
      smd650 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   gendernew |  -.1100815   .0154432    -7.13   0.000    -.1403497   -.0798134
     racenew |   .6051288   .0158992    38.06   0.000     .573967    .6362906
   ridageyr  |   .0114867   .0004495    25.56   0.000     .0106057    .0123677
       _cons |    1.66475   .0246423    67.56   0.000     1.616452    1.713049
------------------------------------------------------------------------------
```

Using our interval-censored method to model the outcome with heaping at multiples of 5, with a half-width of $\lfloor 5/2 \rfloor$, the results are

```
. heapcr smd650 gendernew racenew ridageyr, heap(5) poisson nolog
Cens. heaped Poisson regression                 Number of obs   =       1504
Heaping interval(s)  = 5                        LR chi2(3)      =    2051.49
Heaping halfwidth(s) = 2                        Prob > chi2     =     0.0000
Log likelihood = -7599.224                      Pseudo R2       =     0.1189
```

smd650	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gendernew	-.1157292	.0161764	-7.15	0.000	-.1474344	-.084024
racenew	.6255269	.0166857	37.49	0.000	.5928236	.6582302
ridageyr	.0117264	.0004674	25.09	0.000	.0108104	.0126424
_cons	1.621005	.0257251	63.01	0.000	1.570585	1.671425

We see a slight difference in the `heapcr` model coefficients and an increase in the standard errors of the estimated coefficients.

Using our mixture of rescaled distributions method to model the outcome with heaping at multiples of 5, the results are

```
. heapr smd650 gendernew racenew ridageyr, heap(5) poisson vuong nolog
Heaped Poisson regression                       Number of obs   =      1,504
                                                LR chi2(3)      =     851.63
                                                Prob > chi2     =     0.0000
Log likelihood = -5321.7125                     Pseudo R2       =     0.0741
```

smd650	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
smd650						
gendernew	-.1205029	.0269503	-4.47	0.000	-.1733245	-.0676812
racenew	.6766456	.0269041	25.15	0.000	.6239144	.7293768
ridageyr	.0119632	.0007341	16.30	0.000	.0105244	.0134021
_cons	1.213202	.0389623	31.14	0.000	1.136837	1.289566
modulo_5						
_cons	.0557379	.0579067	0.96	0.336	-.0577571	.1692329

```
Vuong test of heap versus non-heap:              z =    16.53   Pr>|z|=0.0000
    Bias-corrected (AIC) vuong test:             z =    16.53   Pr>|z|=0.0000
    Bias-corrected (BIC) vuong test:             z =    16.51   Pr>|z|=0.0000
```

Again, we see a slight difference in the model's coefficients and an increase in the standard errors of the estimated coefficients. We also see a statistically significant Vuong test of the `heapr` model versus the nonheaped Poisson model.

4.2 GP

The results of fitting a regular GP model[2] to the outcomes are given by

```
. gpoisson smd650 gendernew racenew ridageyr, nolog
Generalized Poisson regression                  Number of obs   =       1504
                                                LR chi2(3)      =     289.54
Dispersion     =  .6439007                      Prob > chi2     =     0.0000
Log likelihood = -5052.9281                     Pseudo R2       =     0.0279
```

smd650	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gendernew	-.0611733	.0363914	-1.68	0.093	-.1324991	.0101526
racenew	.5461656	.0369862	14.77	0.000	.4736739	.6186573
ridageyr	.0101732	.0009927	10.25	0.000	.0082276	.0121188
_cons	1.738053	.0559818	31.05	0.000	1.628331	1.847775
/atanhdelta	.7648088	.0156008			.7342318	.7953858
delta	.6439007	.0091326			.6256475	.6614493

```
Likelihood-ratio test of delta=0:  chi2(1) = 5458.25       Prob>=chi2 = 0.0000
```

In the regular GP model, we see a statistically significant LRT of $\delta = 0$ (dispersion factor), which indicates that the GP model is more appropriate to use than the regular Poisson model. Using our interval-censored method to model the outcomes with heaping at multiples of 5, with a half-width of $\lfloor 5/2 \rfloor$, the results are

```
. heapcr smd650 gendernew racenew ridageyr, heap(5) gpoisson gplrtest nolog
Cens. heaped Gen. Poisson regression            Number of obs   =       1504
Heaping interval(s)  = 5                        LR chi2(3)      =     287.84
Heaping halfwidth(s) = 2                        Prob > chi2     =     0.0000
Log likelihood =  -5041.18                      Pseudo R2       =     0.0278
```

smd650	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gendernew	-.062372	.0367806	-1.70	0.090	-.1344607	.0097167
racenew	.55046	.0373951	14.72	0.000	.4771669	.6237532
ridageyr	.0102485	.0010034	10.21	0.000	.008282	.0122151
_cons	1.719522	.0565989	30.38	0.000	1.60859	1.830454
/atanhdelta	.7637911	.0157335			.7329541	.7946282
delta	.6433046	.0092223			.6248693	.6610229

```
Likelihood ratio test of delta=0                 x =  5116.09   Pr>x = 0.0000
```

2. The gpoisson command was created by Harris, Yang, and Hardin (2012).

We see a slight difference in the coefficients, an increase in the standard errors of the estimated coefficients, and a statistically significant LRT. Using our mixture of rescaled distributions method with heaping at multiples of 5, the results are

```
. heapr smd650 gendernew racenew ridageyr, heap(5) gpoisson gplrtest vuong nolog
Heaped Gen. Poisson regression                  Number of obs     =      1,504
                                                LR chi2(3)        =     349.00
                                                Prob > chi2       =     0.0000
Log likelihood = -4800.7669                     Pseudo R2         =     0.0351
```

smd650	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
smd650						
gendernew	-.0710181	.0382555	-1.86	0.063	-.1459974	.0039612
racenew	.6293441	.0383598	16.41	0.000	.5541602	.7045279
ridageyr	.0113049	.0010148	11.14	0.000	.0093159	.0132939
_cons	1.329487	.056593	23.49	0.000	1.218566	1.440407
modulo_5						
_cons	-.2441573	.0716583	-3.41	0.001	-.384605	-.1037096
/atanhdelta	.4486798	.0189708	23.65	0.000	.4114977	.4858619
delta	.4208132	.0156114			.3897437	.4509259

```
Likelihood ratio test of delta=0                 x =  1041.89   Pr>x = 0.0000
Vuong test of heap versus non-heap:              z =    10.16   Pr>|z|=0.0000
    Bias-corrected (AIC) vuong test:             z =    10.12   Pr>|z|=0.0000
    Bias-corrected (BIC) vuong test:             z =    10.01   Pr>|z|=0.0000
```

Again, we see a slight difference in the model's coefficients, an increase in the standard errors of the estimated coefficients, and a statistically significant LRT of $\delta = 0$ (dispersion factor). We also see a statistically significant Vuong test of the `heapr` model versus the nonheaped GP model.

4.3 NB

The results of fitting a regular NB model to the outcomes are given by

```
. nbreg smd650 gendernew racenew ridageyr, nolog
Negative binomial regression                    Number of obs     =      1,504
                                                LR chi2(3)        =     290.60
Dispersion     = mean                           Prob > chi2       =     0.0000
Log likelihood = -5048.0101                     Pseudo R2         =     0.0280

------------------------------------------------------------------------------
      smd650 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   gendernew |  -.0995121   .0413518    -2.41   0.016    -.1805602   -.0184641
     racenew |    .614582   .0411743    14.93   0.000     .5338819    .6952822
    ridageyr |   .0138921   .0013283    10.46   0.000     .0112887    .0164956
       _cons |   1.552952   .0658433    23.59   0.000     1.423901    1.682002
-------------+----------------------------------------------------------------
    /lnalpha |  -.6339091   .0425475                     -.7173006   -.5505176
-------------+----------------------------------------------------------------
       alpha |   .5305139    .022572                       .488068    .5766512
------------------------------------------------------------------------------
LR test of alpha=0: chibar2(01) = 5468.09                 Prob >= chibar2 = 0.000
```

In the regular NB model, we see a statistically significant LRT of $\alpha = 0$ (dispersion factor), which indicates that the NB model is more appropriate to use than the regular Poisson model. Using our interval-censored method (with the `heap()` option) to model the outcomes with heaping at multiples of 5, with a half-width of $\lfloor 5/2 \rfloor$, the results are

```
. heapcr smd650 gendernew racenew ridageyr, heap(5) nbreg nolog
Cens. heaped Neg. Binomial regression           Number of obs     =      1504
Heaping interval(s)  = 5                        LR chi2(3)        =    291.34
Heaping halfwidth(s) = 2                        Prob > chi2       =    0.0000
Log likelihood = -5036.998                      Pseudo R2         =    0.0281

------------------------------------------------------------------------------
      smd650 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   gendernew |   -.102105   .0418005    -2.44   0.015    -.1840324   -.0201776
     racenew |    .622198    .041628    14.95   0.000     .5406086    .7037873
    ridageyr |   .0140387   .0013413    10.47   0.000     .0114098    .0166676
       _cons |   1.529909   .0665855    22.98   0.000     1.399404    1.660414
-------------+----------------------------------------------------------------
    /lnalpha |  -.6275339   .0428549                     -.7115281   -.5435398
-------------+----------------------------------------------------------------
       alpha |   .5339068   .0228805                       .4908935    .5806891
------------------------------------------------------------------------------
```

A slight difference in the model's coefficients, increase in standard errors, and dispersion factor (α) is shown in the `heapcr` model. Using our mixture of rescaled distributions method with heaping at multiples of 5, the results are

```
. heapr smd650 gendernew racenew ridageyr, heap(5) nbreg vuong nolog
Heaped Neg. Binomial regression                 Number of obs     =      1,504
                                                LR chi2(3)        =     307.62
                                                Prob > chi2       =     0.0000
Log likelihood = -4577.0611                     Pseudo R2         =     0.0325
```

smd650	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
smd650						
gendernew	-.1186497	.0461925	-2.57	0.010	-.2091854	-.028114
racenew	.7160425	.046097	15.53	0.000	.6256941	.8063909
ridageyr	.0152091	.0014285	10.65	0.000	.0124092	.018009
_cons	1.338431	.0716839	18.67	0.000	1.197933	1.478929
modulo_5						
_cons	-.1512008	.0653464	-2.31	0.021	-.2792774	-.0231243
/lnalpha	-.7635081	.0516495	-14.78	0.000	-.8647392	-.662277
alpha	.4660287	.0240701			.4211614	.5156758

```
Vuong test of heap versus non-heap:              z =    16.57   Pr>|z|=0.0000
    Bias-corrected (AIC) vuong test:             z =    16.54   Pr>|z|=0.0000
    Bias-corrected (BIC) vuong test:             z =    16.44   Pr>|z|=0.0000
```

Again, we see a slight difference in the model's coefficients and an increase in the standard errors of the estimated coefficients. We also see a statistically significant Vuong test of the `heapr` model versus the nonheaped NB model.

5 Fishing example

To highlight the application of regression-modeling data that exhibit both heaping and zero-inflation, we examine a model of data on counts of fish. In these data, the drivers of each car exiting a park were questioned about the number of fish caught. It is believed that some persons did not fish and so reported 0 for a different reason than those who were simply unsuccessful. Whether a person reported 0 was modeled by whether the car had a camper and whether children were brought along. The number of fish reported was thought to be associated with the number of persons in the car and whether persons reported having purchased live bait.

```
. webuse fish, clear
. ziheapcr count persons livebait, inflate(child camper) ilink(cloglog) vuong
> gpoisson gplrtest heap(6,11) nolog
Zero-inflated heaped gen. Poisson regression      Number of obs   =        250
Heaping interval(s)  = 6 11                       LR chi2(6)      =      60.10
Heaping halfwidth(s) = 3 5                        Prob > chi2     =     0.0000
Inflation link: cloglog                           Nonzero obs     =        108
                                                  Zero obs        =        142
Log likelihood = -402.6561                        Pseudo R2       =     0.0695
```

count	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
count						
persons	.6571264	.0847749	7.75	0.000	.4909707	.823282
livebait	.8228948	.3333838	2.47	0.014	.1694746	1.476315
_cons	-.9913993	.4362422	-2.27	0.023	-1.846418	-.1363803
inflate						
child	2.45161	.6812248	3.60	0.000	1.116433	3.786786
camper	-1.858881	.7953618	-2.34	0.019	-3.417762	-.3000007
_cons	-2.524079	.7205229	-3.50	0.000	-3.936278	-1.11188
/atanhdelta	.9737121	.0791377			.818605	1.128819
delta	.750331	.0345834			.6743099	.8106147

```
Likelihood ratio test of delta=0:                 x =   886.57   Pr>x = 0.0000
Vuong test of zinbregf vs. gen neg binomial(F):   z =     6.25   Pr>|z|=0.0000
    Bias-corrected (AIC) Vuong test:              z =     5.86   Pr>|z|=0.0000
    Bias-corrected (BIC) Vuong test:              z =     5.18   Pr>|z|=0.0000
```

As can be seen in the output, the zero-inflated interval-censored regression model is preferred over the nonzero-inflated model as evidenced by the significant Vuong test. Also, even after adjusting for overdispersion due to zero-inflation and heaping, there is still evidence of overdispersion as seen by the significant likelihood ratio of the dispersion statistic.

Lastly, the output for the zero-inflated mixture of the rescaled distributions model is shown below. Herein, we model the likelihood of heaping only on multiples of six as a function of having a camper.

```
. webuse fish, clear
. ziheapr count persons livebait, inflate(child camper) ilink(cloglog) vuong
> gpoisson gplrtest heap(6) hvars(camper) nolog
Obtaining LL for zero inflated heaped Poisson for LR test
Zero-inflated heaped gen. Poisson regression      Number of obs   =        250
Inflation link: cloglog                           Nonzero obs     =        108
                                                  Zero obs        =        142
Log likelihood = -403.3094                        Pseudo R2       =     0.0708
```

count	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
count						
persons	.6630002	.0844656	7.85	0.000	.4974507	.8285496
livebait	.834241	.3334796	2.50	0.012	.180633	1.487849
_cons	-1.009603	.436174	-2.31	0.021	-1.864488	-.1547178
inflate						
child	2.42767	.6613499	3.67	0.000	1.131448	3.723892
camper	-1.832131	.7738512	-2.37	0.018	-3.348852	-.3154106
_cons	-2.497034	.7021803	-3.56	0.000	-3.873282	-1.120786
modulo_6						
camper	1.25883	209.0805	0.01	0.995	-408.5314	411.0491
_cons	-13.0241	194.5259	-0.07	0.947	-394.2878	368.2396
/atanhdelta	.972532	.0788769	12.33	0.000	.8179362	1.127128
delta	.7498149	.0345305			.673945	.810034

```
Likelihood ratio test of delta=0:                 x =   860.99   Pr>x = 0.0000
Vuong test of heap versus non-heap:               z =     6.31   Pr>|z|=0.0000
    Bias-corrected (AIC) vuong test:              z =     6.04   Pr>|z|=0.0000
    Bias-corrected (BIC) vuong test:              z =     5.56   Pr>|z|=0.0000
```

These results are very similar to the interval-censored regression model, with consistent covariate coefficients and LRT. The Vuong test is statistically significant, meaning that in this case, the heaped model is preferred over the nonheaped model.

6 Discussion

In this article, we presented two new approaches for modeling heaped ("rounded") count data: with interval-censored regression and with a mixture of rescaled distributions. Both methods for heaped count data are more similar to the true model than is a regular count-data model, based on the significance of the Vuong tests for most models for the presented data. Heaped count data can lead to biased estimation and imprecision in discrete quantitative data. We also introduced supporting commands and programs that illustrate the effectiveness of our approach.

7 References

Channon, A. A. R., S. S. Padmadas, and J. W. McDonald. 2011. Measuring birth weight in developing countries: Does the method of reporting in retrospective surveys matter? *Maternal and Child Health Journal* 15: 12–18.

Consul, P. C. 1989. *Generalized Poisson Distributions: Properties and Applications*. New York: Dekker.

Consul, P. C., and F. Famoye. 1992. Generalized Poisson regression model. *Communications in Statistics—Theory and Methods* 21: 89–109.

Desmarais, B. A., and J. J. Harden. 2013. Testing for zero inflation in count models: Bias correction for the Vuong test. *Stata Journal* 13: 810–835.

Hardin, J. W., and J. M. Hilbe. 2012. *Generalized Linear Models and Extensions*. 3rd ed. College Station, TX: Stata Press.

Harris, T., Z. Yang, and J. W. Hardin. 2012. Modeling underdispersed count data with generalized Poisson regression. *Stata Journal* 12: 736–747.

Klesges, R. C., M. Debon, and J. W. Ray. 1995. Are self-reports of smoking rate biased? Evidence from the Second National Health and Nutrition Examination Survey. *Journal of Clinical Epidemiology* 48: 1225–1233.

Lawless, J. F. 1987. Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics* 15: 209–225.

Lewis-Esquerre, J. M., S. M. Colby, T. O. Tevyaw, C. A. Eaton, C. W. Kahler, and P. M. Monti. 2005. Validation of the timeline follow-back in the assessment of adolescent smoking. *Drug and Alcohol Dependence* 79: 33–43.

McLain, A. C., R. Sundaram, M. Thoma, and G. M. Buck Louis. 2014. Semiparametric modeling of grouped current duration data with preferential reporting. *Statistics in Medicine* 33: 3961–3972.

Nietert, P. J., A. M. Wessell, C. Feifer, and S. M. Ornstein. 2006. Effect of terminal digit preference on blood pressure measurement and treatment in primary care. *American Journal of Hypertension* 19: 147–152.

Pardeshi, G. S. 2010. Age heaping and accuracy of age data collected during a community survey in the Yavatmal district, Maharashtra. *Indian Journal of Community Medicine* 35: 391–395.

Ridout, M., C. G. B. Demétrio, and J. Hinde. 1998. Models for count data with many zeros. In *Proceedings of the XIXth International Biometric Conference*, 179–192. Cape Town: The International Biometric Society.

Ridout, M. S., and B. J. T. Morgan. 1991. Modelling digit preference in fecundability studies. *Biometrics* 47: 1423–1433.

Roberts, J. M., and D. D. Brewer. 2001. Measures and tests of heaping in discrete quantitative distributions. *Journal of Applied Statistics* 28: 887–896.

Vuong, Q. H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57: 307–333.

Wang, H., and D. F. Heitjan. 2008. Modeling heaping in self-reported cigarette counts. *Statistics in Medicine* 27: 3789–3804.

Wolff, J., and T. Augustin. 2003. Heaping and its consequences for duration analysis: A simulation study. *Allgemeines Statistisches Archiv* 87: 59–86.

About the authors

Tammy H. Cummings is a senior research associate in the Institute for Families in Society at the University of South Carolina in Columbia, SC.

James W. Hardin is an associate professor in the Department of Epidemiology and Biostatistics and an affiliated faculty in the Institute for Families in Society at the University of South Carolina in Columbia, SC.

Alexander C. McLain is an assistant professor in the Department of Epidemiology and Biostatistics at the University of South Carolina in Columbia, SC.

James R. Hussey is an associate professor in and Chair of the Department of Epidemiology and Biostatistics at the University of South Carolina in Columbia, SC.

Kevin J. Bennett is an associate professor in the Department of Family and Preventive Medicine and is an Affiliated Faculty in the South Carolina Rural Health Research Center at the University of South Carolina in Columbia, SC.

Gina M. Wingood is a professor in the Department of Behavioral Sciences and Health Education and the Agnes Moore Faculty in HIV/AIDS research at Emory University in Atlanta, GA.