The Stata Journal (*yyyy*) *vv*, Number *ii*

Heap: A command for estimating discrete outcome variable models in the presence of heaping at known points

Wiji Arulampalam
University of Warwick

Valentina Corradi
University of Surrey

Daniel Gutknecht
University of Mannheim

Zizhong Yan
Jinan University

Abstract. Self-reported survey data are often plagued by the presence of heaping. Accounting for this measurement error is crucial for the identification and consistent estimation of the underlying model (parameters) from such data. This paper introduces two Stata commands. The first command, `heapmph`, estimates the parameters of a discrete-time mixed proportional hazard model with gamma unobserved heterogeneity, allowing for fixed and random right censoring, and different sized heap points. The second command, `heapop`, extends the framework to ordered probability models, subject to heaping. Suitable specification tests are also provided.

Keywords: st0001, heap, heapmph, heapop, Discrete Time Duration Model, Heaping, Measurement Error, Ordered Probability Model

Acknowledgements: We are grateful to the British Academy (grant number: SG160731 - Estimation and inference with heaped data - a novel approach), for funding this project. We also thank David M. Drukker for helpful discussions.

1 Introduction

A problem frequently encountered in survey data is the abnormal concentration of reported observations at certain values of the outcome variable. Examples include reported dates of death in neo-natal mortality data (Arulampalam, Corradi and Gutknecht, 2017) (ACG from now on), age of starting and quitting cigarette smoking (Forster and Jones 2001), or self-reported consumption expenditure data (Pudney 2008). One of the main reasons for such concentration, often referred to as heap points, is rounding. Correctly identifying and accounting for the rounding behavior is crucial for consistent estimation and valid inference on the parameters of the underlying model of interest. The paper ACG discusses identification and estimation of popular duration and ordered choice models in the presence of heaping using maximum likelihood procedures.

In this paper, we introduce the Stata command, `'heapmph'` to estimate the underlying parameters in the case of a discrete-time mixed proportional hazard (Cox 1972) duration model as proposed in ACG. More specifically, this command estimates a semi-parametric baseline hazard function, in the presence of heaping of observations at certain durations, and gamma distributed unobserved heterogeneity (frailty). In the accompa-

nying code `'heapop',` we extend the framework to a more general ordered probability model, allowing for the presence of heaping points.[1]

As shown in ACG, when some of the parameters lie on the boundary of the parameter space, the limiting distribution of the estimator is no longer a normal distribution, and more complicated subsampling procedures are required for inference. Hence, we also provide two specification tests. The first one tests for the absence of heaping effects in the model. The second specification test examines whether all heaping parameters lie inside the parameter space, which in turn will allow for inference based on asymptotic normality. We use the M out of N bootstrap method to calculate the standard errors. These tests provide a set of tools that enable applied researchers to verify the validity of different model specifications.

2 Mixed proportional hazard model specification

We start with the Mixed Proportional Hazard (MPH) model for the unobserved true durations in continuous time, and parameterize this, for individual i as

$$\lambda_i(\tau^*|z_i, u_i) = \lambda_0(\tau^*)\exp(z_i'\beta + u_i), \tag{1}$$

where $\lambda_0(\tau^*)$ is the baseline hazard at time τ^*, u_i is the individual unobserved heterogeneity (frailty), and z_i a set of time invariant covariates. In most empirical studies, time is observed on a discrete scale. We therefore, assume that a continuous duration $\tau_i^* \in [\tau, \tau+1)$ is recorded as τ, where τ denotes a discrete time period, so that the sample of (discrete) durations is given by τ_i for $i = 1, \ldots, N$. The discrete time hazard for our model can then be written as:

$$\begin{aligned} h_i(\tau|z_i, u_i) =& \Pr\left[\tau_i^* < \tau+1 | \tau_i^* \geqslant \tau, z_i, u_i\right] \\ =& 1 - \exp\left(-\int_\tau^{\tau+1} \lambda_i(s|z_i, u_i)ds\right) \\ =& 1 - \exp\left(-\exp\left(z_i'\beta + \gamma(\tau)\right) + u_i\right), \end{aligned} \tag{2}$$

where $\gamma(\tau) = \ln\int_\tau^{\tau+1}\lambda_0(s)ds$. Due to misreporting, the researcher however, does not observe τ_i directly, but t_i, a potentially mismeasured version of it.

More specifically, the form of misreporting we address is referred to as "heaping" in the literature, and describes the phenomenon of observing an over- and under-reporting of failures at certain time periods. We briefly list informally the set of assumptions for the derivation of the estimator and its properties here, and refer the readers to ACG for further details on the assumptions and identification results.[2]

1. The more general ordered probit model can be used not only in the case of duration data, but for other applications that involves discrete data such as the number of packets of cigarettes smoked per day.
2. Note, ACG discusses a more general set up which can accommodate more complex heaping mechanisms.

Assumptions:

A1 Excessive concentrations of reported failures occur at time periods that are multiples of a non-negative integer. This implies the equal distance between the heap points. In most of the empirical applications where we see heaping due to rounding, we often see the distance between heaping points to be the same. This is the scenario `heapmph` uses. However, it can be easily amended to allow for non-equally spaced heaping points. In our illustration of neonatal deaths, the reported values are heaped at points that are multiples of 5. A stylized version of this case is illustrated in Figure 1.

A2 There is no heaping at time zero. This is not an unrealistic assumption, since one would expect survey respondents to know whether the discretized duration was a zero or not.

A3 In order to identify the baseline hazard from possibly misreported observations, we need to impose a structure on the heaping process. In our neonatal mortality illustration, the heaps are found to be at $5, 10, 15$, etc days. We, therefore, assume that one period to the right and to the left of each heap point, are associated with that heap, in the illustration provided here. We denote the maximum number of time periods that a duration can be rounded to as $\overline{r}$, and in this example $\overline{r} = 1$. For example, we assume that the duration points 4, 9 and 14, will be rounded up to 5, 10 and 15 respectively.

A4 All heaping is to observed duration points only. In our example, this implies that the heaping is to the points $5, 10$, and 15 only since the data are censored at 18 days. Maximum number of heaps is assumed to be $\overline{j}$, and in our example here $\overline{j}$ =3.

A5 The censoring is exogenous, and the censored observations are correctly reported.

A6 Whenever the true duration falls onto one of the heaping points, it will be correctly reported. However, whenever the duration falls onto the non-heaping points, it is assumed to be either correctly reported or rounded (up or down) to the nearest heaping point.

A7 Let p_1, p_2,etc. denote the probabilities that, a true duration which is away by one, two, etc units from the nearest heaping point, will be rounded up to this heaping point. Similarly let q_1, q_2,etc. denote the probabilities of rounding down by one, two, etc. units to the nearest heaping point. In our illustration, a reported duration of say 10 days, includes true durations of 11 (9), which have been rounded down (up) to 10 days (see again Figure 1). In our example, p_1 is the probability that a true duration of 9 will be rounded up to 10 days. Similarly q_1 is the probability that a true duration of 11 will be rounded down to 10 days.

A8 There exists a segment in the baseline hazard that is constant from time period $\overline{k}$, and includes a known true value (i.e. there is no mis-reporting at this value). In our example, we assume $\overline{k} = 12$.

Heuristically, the assumption that the hazard is constant over a set of time periods, which includes (at least) a known true value, enables us to uniquely identify the γ parameter associated with this correctly reported time period as well as the parameters of the heaping process, i.e. the p's and the q's, in this region, from the observed data. Subsequently, we can use these identified probability parameters to pin down the rest of the baseline hazard parameters. See Figure 1.

3 Maximum likelihood estimation

Before writing down our likelihood function, we first define some notation.

Let $\underline{\theta} = \{\beta', \gamma'\}'$ with $\gamma = \{\gamma(0), \gamma(1), \ldots, \gamma(\overline{\tau} - 1)\}'$, and $\overline{\tau}$ be some finite, positive integer, and $(\overline{\tau}-1)$ represent the maximum observed time period. Define the probability of survival at least until time period $\tau < \overline{\tau}$ in the absence of misreporting as:

$$\begin{aligned} S_i\left(\tau|z_i, u_i, \underline{\theta}\right) =& \Pr\left(\tau_i \geq \tau|z_i, u_i, \underline{\theta}\right) \\ =& \prod_{s=0}^{\tau-1} \exp\left(-\exp\left(z_i'\beta + \gamma(s) + u_i\right)\right) \\ =& \prod_{s=0}^{\tau-1} \exp\left(-v_i \exp\left(z_i'\beta + \gamma(s)\right)\right), \end{aligned}$$

where $v_i \equiv \exp(u_i)$, and u_i is the unobserved heterogeneity.

The probability for an exit event in $\tau_i < \overline{\tau}$ is:

$$\begin{aligned} f_i\left(\tau|z_i, u_i, \underline{\theta}\right) =& \Pr\left(\tau_i = \tau|z_i, u_i, \underline{\theta}\right) \\ =& S_i\left(\tau|z_i, u_i, \underline{\theta}\right) - S_i\left(\tau+1|z_i, u_i, \underline{\theta}\right) \\ =& \prod_{s=0}^{\tau-1} \exp\left(-v_i \exp\left(z_i'\beta + \gamma(s)\right)\right) \\ & - \prod_{s=0}^{\tau} \exp\left(-v_i \exp\left(z_i'\beta + \gamma(s)\right)\right). \end{aligned} \tag{3}$$

$f_i\left(\tau|z_i, u_i, \underline{\theta}\right)$ in the above equation denotes the probability of a duration equal to τ when there is no misreporting. However, because of the rounding, heaped values are over-reported while non-heaped values are under-reported, and this needs to be taken into account when constructing the likelihood function (see next section).

Henceforth, let:

$$\phi_i\left(t|z_i, v_i, \underline{\theta}\right) = \Pr\left(t_i = t|z_i, v_i, \underline{\theta}\right)$$

with t_i denoting the discrete *reported* duration.

The likelihood contributions depend on the following four cases.

(I) For correctly reported durations, $\phi_i\left(t|z_i, v_i, \underline{\theta}\right) = f_i\left(t|z_i, v_i, \underline{\theta}\right)$. This will include the duration point discussed in Assumption A8 earlier. Depending on the application, there might be other points too.

(II) For reported durations that are $l = 1, 2$, etc points *below* the nearest heaping point, $\phi_i\left(t|z_i, v_i, \underline{\theta}\right) = (1 - p_l) f_i\left(t|z_i, v_i, \underline{\theta}\right)$, since p_l refer to the probabilities of rounding up.

(III) Similar to (II), for reported durations that are $l = 1, 2$, etc points *above* the nearest heaping point, $\phi_i\left(t|z_i, v_i, \underline{\theta}\right) = (1 - q_l) f_i\left(t|z_i, v_i, \underline{\theta}\right)$, since q_l refer to the probabilities of rounding down.

(IV) Finally for reported durations on the heaping points:

$$\phi_i\left(t|z_i, v_i, \underline{\theta}\right) = \sum_l p_l f_i\left(t - l|z_i, v_i, \underline{\theta}\right) + \sum_l q_l f_i\left(t + l|z_i, v_i, \underline{\theta}\right) + f_i\left(t|z_i, v_i, \theta\right).$$

In summary, there are four different probabilities of exit events depending on the nature of the true duration.

We next write down the corresponding unconditional probabilities under a set of assumptions on the unobserved heterogeneity v_i. We make the standard assumptions:

(i) v_i is identically and independently distributed over i and is also independent of z_i;

(ii) The density of v is gamma with unit mean and variance $\left(\frac{1}{\sigma}\right)$;[3]

The unconditional probabilities under the above assumptions, in case (I) above are given by:

$$\begin{aligned}
\int \phi_i\left(t|z_i, v, \underline{\theta}\right) g(v;\sigma) dv &= \int \Pr\left(\tau_i = t|z_i, v, \underline{\theta}\right) g(v;\sigma) dv \\
&= \int S_i\left(t|z_i, v, \underline{\theta}\right) g(v;\sigma) dv - \int S_i\left(t+1|z_i, v, \underline{\theta}\right) g(v;\sigma) dv \\
&= \left(1 + \sigma\left(\sum_{s=0}^{t-1} \exp\left(z_i'\beta + \gamma(s)\right)\right)\right)^{-\sigma^{-1}} \\
&\quad - \left(1 + \sigma\left(\sum_{s=0}^{t} \exp\left(z_i'\beta + \gamma(s)\right)\right)\right)^{-\sigma^{-1}}
\end{aligned}$$

where the last equality uses the fact that there is a closed form expression under the gamma density assumption for v (e.g., see Meyer (1990, p. 770)). Moreover, since the

3. The assumption of gamma density for ν gives us a closed form expression for the unconditional probabilities. While the gamma density choice might appear overly restrictive at first sight, we note that this can often be rationalized theoretically (Abbring and Van Den Berg 2007). Findings by Han and Hausman (1990) as well as Meyer (1990), suggest that estimation results for discrete-time proportional hazard models where the baseline is left unspecified (as in our model), display little sensitivity to alternative distributional assumptions on v_i.

integral is a linear operator, the probabilities for the cases (II) to (IV) can be derived accordingly.

Our next goal is to obtain consistent estimators for $\theta = \{\underline{\theta}', \sigma, p_1, \ldots, p_{\overline{r}}, q_1, \ldots, q_{\overline{r}}\}'$ from the, possibly misreported durations. Before setting up the likelihood function drawing from the previous derivations for truthfully and misreported durations, we introduce censoring into our setup.

Let δ_i be an indicator equal to one if the observation is uncensored and zero otherwise. It is assumed that durations are censored at a fixed time $\overline{\tau}$, which exceeds the points that are rounded and is not one of the heaping points. Assuming that censoring is independent of the heaping process and the durations (type I censoring; Cox and Oakes 1984), we have the following unconditional likelihood contributions.[4]

The likelihood function for the observed sample is:

$$L_N(\theta) = \prod_{i=1}^{N} \int \left\{ \phi_i(t|z_i, v)^{\delta_i} S_i(t|z_i, v)^{(1-\delta_i)} \right\} g(v; \sigma) \mathrm{d}v$$

and so

$$l_N(\theta) = \ln L_N(\theta) = \sum_{i=1}^{N} \ln \int \left\{ \phi_i(t|z_i, v)^{\delta_i} S_i(t|z_i, v)^{(1-\delta_i)} \right\} g(v; \sigma) \mathrm{d}v.$$

Given the definition of $\phi_i(t|z_i, v)$ and cases (I) through (IV), it is clear that the (log) likelihood function down-weights the contribution of heaped durations, and over-weights the contribution of non heaped durations.

Under the assumptions provided in ACG, it can be shown that the limiting distribution of the estimator depends on whether some heaping probability parameters lie on the boundary of the parameter space or not. That is, if one or more of the "true" probability parameters are equal to zero. It is also the case that the limiting distribution is no longer normal as the information matrix is not blocked diagonal in general, but takes a different form. We use the M out of N bootstrap method to derive the asymptotic standard errors. Details are provided in ACG.

4 Testing for heaping

As pointed out in the previous section, if some of the heaping probability parameters lie on the boundary of the parameter space, the asymptotic distribution of the estimator is no longer normal. In addition, inference becomes more complicated, since subsampling methods are used to derive the asymptotic standard errors. In the following, we discuss two specification tests. First, a test to detect whether heaping matters in a statistical

4. For ease of exposition, we have assumed a constant censoring point. However, the program allows the censoring points to vary over i.

sense ($\mathbf{H}^{\pi_1}$). If heaping matters, a second test to discriminate between the general case that allows for probability parameters on the boundary, and the special case without parameters on the boundary ($\mathbf{H}^{\pi_2}$). That is, while the first test helps to determine whether the specified heaping model is indeed preferred over a standard duration model that does not account for heaping, the second test allows one to decide whether inference, in fact, ought to be based on subsampling methods.

Thus, collecting all heaping parameters in the vector π with $\pi = \{p_1, \ldots, p_{\overline{r}}, q_1, \ldots, q_{\overline{r}}\}'$ and $\theta = \{\underline{\theta}', \sigma, \pi'\}'$, the first test examines the existence of heaping effects through:

$\mathbf{H}^{\pi_1}$:

$$H_0^{\pi_1} : p_1 = ... = p_{\overline{r}} = q_1 = ... = q_{\overline{r}} = 0$$

vs

$$H_A^{\pi_1} : p_l > 0 \text{ and/or } q_l > 0$$

for some $l = 1, ..., \overline{r}$. The above hypothesis $H_0^{\pi_1}$ can be tested through a standard likelihood ratio test of the form (ACG):

$$LR_N = -2\left(l_N\left(\widetilde{\theta}_N\right) - l_N\left(\widehat{\theta}_N\right)\right).$$

The second specification test examines whether all heaping parameters lie inside the parameter space, which in turn allows inference based on asymptotic normality. That is, the null hypothesis of the test is that at least one rounding parameter is equal to zero versus the alternative that none is zero (and thus no boundary problem exists). Formally, let $H_{p,0}^{(j)} : p_j = 0$, $H_{p,A}^{(j)} : p_j > 0$, and let $H_{q,0}^{(j)}, H_{q,A}^{(j)}$ be defined analogously. Our objective is to test the following hypothesis:

$\mathbf{H}^{\pi_2}$:

$$H_0^{\pi_2} = \left(\cup_{j=1}^{\overline{r}} H_{p,0}^{(j)}\right) \cup \left(\cup_{j=1}^{\overline{r}} H_{q,0}^{(j)}\right)$$

vs

$$H_A^{\pi_2} = \left(\cap_{j=1}^{\overline{r}} H_{p,A}^{(j)}\right) \cap \left(\cap_{j=1}^{\overline{r}} H_{q,A}^{(j)}\right),$$

so that under $H_A^{\pi_2}$ all p's and q's are strictly positive. To discriminate between $H_0^{\pi_2}$ and $H_A^{\pi_2}$, we apply the Intersection-Union principle (IUP), see e.g. chapter 5 in Silvapulle and Sen (2005). According to the IUP, we only reject $H_0^{\pi_2}$ at level α if all single null hypotheses $H_{p,0}^{(j)}$ and $H_{q,0}^{(j)}$ are rejected at level α.

We now introduce a rule to discriminate between $H_0^{\pi_2}$ and $H_A^{\pi_2}$.

Rule IUP-PQ: Reject $H_0^{\pi_2}$, if $\max_{j=1,\ldots,\overline{r}} \{PV_{p,j}, PV_{q,j}\} < \alpha$ and do not reject otherwise.

Thus, if one rejects $H_0^{\pi_2}$, the inference can be based on asymptotic normality, while failure to reject $H_0^{\pi_2}$ requires the use of subsampling methods as outlined before.

5 Extension to the ordered probit model

In general, there are many observed discrete outcomes (other than durations) that can exhibit heaping.[5] Here we discuss an extension to the ordered probit model, and outline the link between ordered choice models in general, and the discrete duration model we considered in the previous section, see for example Han and Hausman (1990).

To keep notational clutter to a minimum, we do not explicitly show the conditioning set in the discussions below. Let F and S respectively be the distribution and survivor functions of our duration variable. Given the relationship between a hazard function and the survivor function, we have the following probabilities in the proportional hazard model specified earlier (equation 3) :

(i) the probability of observing an uncensored duration of $\tau \in \{1, ..., \overline{\tau} - 1\}$ is :

$$\begin{aligned}\Pr(\tau_i = \tau) &= S_i(\tau) - S_i(\tau+1) = F_i(\tau+1) - F_i(\tau) \\ &= \exp(-\int_0^{\tau} \lambda_i(\tau^*)d\tau^*) - \exp(-\int_0^{\tau+1} \lambda_i(\tau^*)d\tau^*) \\ &= \exp(-\exp(z_i'\beta + \delta_{\tau+1})) - \exp(-\exp(z_i'\beta + \delta_\tau)) \end{aligned} \tag{4}$$

where δ_τ is the log of the integrated baseline hazard and is given by:

$$\delta_\tau = \log \int_0^{\tau} \lambda_0(s)ds, \qquad \tau = 1, ..., \overline{\tau} - 1, \tag{5}$$

and λ_0 is the baseline hazard function in (see equation 1):

$$\lambda_i(\tau^*) = \lambda_0(\tau^*)\exp(z_i'\beta)$$

(ii) the probability of observing an uncensored duration of $\tau = 0$ is:

$$\Pr(\tau_i = 0) = S_i(0) - S_i(1) = 1 - S_i(1) = \exp(-\exp(z_i'\beta + \delta_1)),$$

and finally,

(iii) the probability of observing a censored observation is:

$$\Pr(\tau_i = \overline{\tau}) = S_i(\overline{\tau}) = 1 - \exp(-\exp(z_i'\beta + \delta_{\overline{\tau}}))$$

We can now easily see that the above probabilities are similar to an ordered choice model probabilities.[6] Hence, we can use the ordered choice model representation and write the above as:

$$\delta_{\tau_i} = -z_i'\beta + \varepsilon_i, \tag{6}$$

5. An important example are count data (e.g., number of doctor visits or cigarette consumption in a given period of time), where heaping is often present in the observed data.

6. The model only has $\delta_1, .., \delta_{\overline{\tau}}$. The relationship between these and the baseline hazard function

with ε_i as Type-I extreme value distributed.[7] Equation 5 in Han and Hausman (1990) becomes:

$$\Pr(\tau_i = \tau) = S_\varepsilon(\tau) - S_\varepsilon(\tau+1) = F_\varepsilon(\tau+1) - F_\varepsilon(\tau) = \int_{\delta_\tau + z_i'\beta}^{\delta_{\tau+1}+z_i'\beta} f_\varepsilon(\epsilon)d\epsilon \quad (7)$$

where, f_ε is the probability density function of ε.

It is now easy to see that, assuming ε_i to be normally distributed will give us the ordered probit choice probabilities. We now expand on this point in relation to the model estimated by our Stata command `heapop`.

Consider the following latent variable model representation of an ordered choice model[8]:

$$y_i^* = -z_i'\beta^\dagger + \varepsilon_i$$

where y_i^* represents the latent outcome, z_i stands for the vector of regressors, $\beta^\dagger$ is the vector of coefficients, and ε_i is normally distributed error term with zero mean and unit variance. Assume we have duration variables coded as $\tau_i \in \{0, ..., \overline{\tau}\}$, where $\overline{\tau}$ is the right-censored duration. The relationship between the ordered outcome and the underlying latent variable y_i^* is given by,

$$\Pr(\tau_i = \tau) = \Pr(\delta_\tau \le y_i^* < \delta_{\tau+1}) = \Pr(\delta_\tau + z_i'\beta^\dagger \le \varepsilon_i < \delta_{\tau+1} + z_i'\beta^\dagger) = \int_{\delta_\tau + z_i'\beta}^{\delta_{\tau+1}+z_i'\beta} f_\varepsilon(\epsilon)d\epsilon \quad (8)$$

For the normalization, in the above, we have assumed that $\delta_0 = -\infty$, and $\delta_{\overline{\tau}+1} = +\infty$.

Hence, the cut-off points in the ordered choice model are the same as the log of the integrated baseline hazard in the proportional hazards model (equation (7) vs equation (8)).

Note, the command can be used for any ordered discrete outcome variable that exhibits heaping. Discrete duration variable is one such example.

---

parameters we saw earlier, is given by:

$$\begin{aligned}\exp(\delta_\tau) &= \int_0^\tau \lambda_0(\tau^*)d\tau^* = \int_0^1 \lambda_0(\tau^*)d\tau^* + ... + \int_{\tau-1}^\tau \lambda_0(\tau^*)d\tau^* \\ &= \exp\big(\gamma(0)\big) + \exp\big(\gamma(1)\big) + ... + \exp\big(\gamma(\tau-1)\big)\end{aligned}$$

7. The distribution function and density of a random variable X distributed as a standard Type-I extreme value distribution (or Gumbel distribution) are: $F(x) = \exp(-\exp(-x))$ and $f(x) = \exp(-(x + \exp(-x)))$, respectively.
8. Supplementary materials provided in ACG details all the identification conditions required for the estimation of this model when heaping is present in the data.

6 Implementing the `heapmph` command

This section describes the implementations of the `heapmph` command for the mixed proportional hazard model. The usage of the `heapop` command for the ordered probit model is similar and examples are provided in Appendix 2.

6.1 Basic syntax

The basic syntax of the `heapmph` command follows the standard Stata command form:

`heapmph` *depvar varlist* [*if*] [*in*] [`, options`]

where *depvar* stands for the dependent variable, and *varlist* may contain the specified covariates. The usages of various `options` to this command are listed in its Stata help file. In this paper, we demonstrate the usages of the `heap` package with examples.

6.2 Data

We introduce our methodology and the command using a generated data set.

The illustration provided here uses a generated dataset using a random sample of 250 observations from the original sample used in ACG. From the selected sample, we retain two covariates that were found to be significant in the illustration in ACG: mother's age at the time of birth, and mother's years of schooling. The outcome variable `duration,` which is the time of death of the child measured in days if the child died within the first 17 days, is generated using these two covariates in an ordered probit framework. All observations where the child survived for longer than 18 days are treated as censored.[9]

Let,

- `age_m`: mother's age in years;
- `school_m`: mother's schooling in years.

The latent dependent variable y_i^* is generated according to:

$$y_i^* = 0.1\ \texttt{age\_m}_i - 0.1\ \texttt{school\_m}_i + \varepsilon_i \quad \text{for } i = 1, ..., 250$$

where $\varepsilon_i \sim N(0,1)$. The gamma parameters are set as follows: $\exp\big(\gamma(t)\big) = 0.6$ for $t = 0, 1, ..., 11$, $\exp\big(\gamma(t)\big) = 0.6$ for $t = 12, 13, 14, 15$, $\exp\big(\gamma(16)\big) = 1.8$, and $\exp\big(\gamma(17)\big) = 3$. We keep the hazard function flat from period 12 to 15. It is readily seen that the cutoff

9. Please refer to ACG for details of the survey and the original sample used in ACG.

points $\delta_1, ..., \delta_{18}$ of the ordered probit model can be directly calculated by (see also the footnote 6):

$$\delta_t = \ln\left(\sum_{t=0}^{t-1} \exp\left(\gamma(t)\right)\right) \quad \text{for } t = 1, ..., 18$$

The discrete duration variable without heaping, for each observation $i = 1, 2, ..., 250$, for this model is then generated using the cutoff points as:

$$\texttt{duration}_{\texttt{nh},i} = t \text{ if } y_i^* \in (\delta_t, \delta_{t+1}] \quad \text{for } t = 0, ..., 18$$

where we assume $\delta_0 = -\infty$, and $\delta_{19} = \infty$ for the normalization.

Finally, we add the following heaping pattern to the dependent variable: the duration points 4, 9, and 11 are rounded up to 5, 10, and 15 with probability 0.7, respectively. Duration points 6, 10, and 16 are rounded down to 5, 10 and 15 with the same probability 0.7, respectively. Hence the heaping probability parameters are $p_1 = q_1 = 0.7$. Algebraically, the actual observed duration variable `duration` is generated by:

$$\begin{aligned}
u_i &\sim \text{Uniform}[0,1] \\
\texttt{duration}_i &= 5 \text{ if } \texttt{duration}_{\texttt{nh},i} = 4 \text{ and } u_i < 0.7 \\
\texttt{duration}_i &= 5 \text{ if } \texttt{duration}_{\texttt{nh},i} = 6 \text{ and } u_i < 0.7 \\
\texttt{duration}_i &= 10 \text{ if } \texttt{duration}_{\texttt{nh},i} = 9 \text{ and } u_i < 0.7 \\
\texttt{duration}_i &= 10 \text{ if } \texttt{duration}_{\texttt{nh},i} = 11 \text{ and } u_i < 0.7 \\
\texttt{duration}_i &= 15 \text{ if } \texttt{duration}_{\texttt{nh},i} = 14 \text{ and } u_i < 0.7 \\
\texttt{duration}_i &= 15 \text{ if } \texttt{duration}_{\texttt{nh},i} = 16 \text{ and } u_i < 0.7
\end{aligned}$$

Figure 2 plots the histograms of both observed duration variable with heaping and the true duration variable without heaping as generated from this model.

6.3 Model estimation

The analysis is restricted to modeling the hazard rate during the first 18 days after birth since the reported number of deaths is smaller after this period (see ACG). We, therefore, add `censor(18)` option to the command to fix the right-censoring period for each observation at 18. By default, the `heap` command assumes that the right-censoring period is the largest value of the dependent variable in the chosen sample. Instead of using the fixed right-censoring, it is also possible to allow for person-specific censoring points for each observation (see Section 6.5).

We next detail the values used for the four *compulsory* options to define the pattern of heaping in our example.

1. Since the heaps in the data appear to be pronounced differently at different days (cf. Table 1 of ACG), we allow for heaps at days 5, 10, and 15. In the command,

we can define the starting period (h^*) of 5 using the option `hstar(5)`. The assumption is that the heaping occurs at points that are multiples of h^*.

2. Since the heaps occur at equally spaced periods 5, 10, and 15 days, we set `jbar(3)` to indicate the multiplier of heaping periods. "$\bar{j} = 3$" indicates that there are a *maximum* of three heaping points prior to the censoring point.

3. As illustrated in our stylized example (Figure 1), the rounding probabilities are p_1, and $q1$, respectively. Hence, with the number of heaping probabilities, we have the maximum number of time periods that a duration can be rounded to is denoted as $\bar{r} = 1$. This is set by the option `rbar(1)` in the command.

4. The constant part of the baseline hazard enables us to identify the parameters of the heaping process. In this example, we set the time period after which the hazard is constant equal to 12 ($\bar{k}$), relying partially on information that the number of babies dying after 10 days after birth is relatively stable and not subject to large fluctuations. Also, we assume that the heaping is asymmetric, which suggests that constant baseline hazard parameters are at different levels for periods [12, 15] and (15,16].[10] In the command, the starting period of the flat segment can be defined by adding the option `kbar(12)`.

▷ Example

We choose `duration` as the dependent variable, and `age_m` and `school_m` as the covariates. We request Stata to implement the command using the code:

```
. heapmph duration age_m school_m, censor(18) hstar(5) jbar(3) kbar(12) rbar(1)
```

The command returns:

```
Coefficients estimation in progress (% finished approx.): 0%....1%........10%
..................20%.................30%.................40%.................50%
..............60%.................70%.................80%..................90%
.....................100%
------------------------------------------------------------------------------
 Initial temperature:       1          Final temperature:           0.000000010
 Consecutive rejections:    85         Number of function calls:    30,807
 Total final loss:          593.390    Observations:                250
------------------------------------------------------------------------------

MooN bootstrap will take approximately 24 minutes (100 replicates).
(each dot . indicates one replication)
----+--- 1 ----+--- 2 ----+--- 3 ----+--- 4 ----+--- 5
..................................................    50
..................................................   100
------------------------------------------------------------------------------
             |               Bootstrap
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
e^(gamma)    |
      gamma0 |   .0325621   .0125136     2.60   0.009     .0308278    .0342964
------------------------------------------------------------------------------
```

10. See Assumption H (iii) in ACG.

gamma1	.2090542	.07339	2.85	0.004	.1988828	.2192255
gamma2	.4862165	.1800167	2.70	0.007	.4612675	.5111656
gamma3	.5812911	.229792	2.53	0.011	.5494436	.6131387
gamma4	.5267263	.3693367	1.43	0.154	.4755389	.5779138
gamma5	.8164085	.4091477	2.00	0.046	.7597035	.8731135
gamma6	1.091646	.6770079	1.61	0.107	.997817	1.185474
gamma7	.7937079	.3725292	2.13	0.033	.742078	.8453378
gamma8	1.58863	.7767408	2.05	0.041	1.480979	1.696281
gamma9	1.680492	1.245824	1.35	0.177	1.50783	1.853154
gamma10	1.660657	1.363557	1.22	0.223	1.471678	1.849637
gamma11	1.290711	.953491	1.35	0.176	1.158564	1.422858
gamma12	2.49588	1.432036	1.74	0.081	2.297409	2.69435
gamma13	6.790965	5.201258	1.31	0.192	6.070108	7.511823
gamma14	7.92429	5.675076	1.40	0.163	7.137765	8.710816
sigma						
sigma	.7339797	.1989638	3.69	0.000	.7064047	.7615547
beta						
age_m	-.1240245	.0180542	-6.87	0.000	-.1265267	-.1215223
school_m	.1533089	.0114609	13.38	0.000	.1517205	.1548973
prob_left						
p1	.6846628	.1884005	3.63	0.000	.6585518	.7107738
prob_right						
q1	.7250183	.0980855	7.39	0.000	.7114243	.7386122

The command firstly employs a single simulated annealing algorithm (see Section 5.5.3) to solve for the point estimates. The M out of N bootstrap procedure is then conducted to yield the standard errors. The output table consists of five panels. The panel "`e^(gamma)`" reports the estimates of the baseline hazard parameters. It is worth mentioning again that we set the baseline hazard parameters γ, to be constant over periods $[12, 15]$ and over periods $(15, 17]$. Hence, the number of baseline hazard parameters we estimate is $18 - 3 - 1 = 14$. Specifically, "`gamma0`", "`gamma1`",..., "`gamma11`" in the output table correspond to the baseline hazard in period 0, 1,..., 11, respectively. "`gamma12`" corresponds to the flat baseline hazard during periods $[12, 15]$. "`gamma13`" is for period $(15, 16]$ and "`gamma14`" is for the period 17.

Panel "`sigma`" displays the estimates of σ which is the precision parameter of the gamma distributed unobserved heterogeneity variable v_i, and panel "`beta`" is for the estimates of the covariate coefficients. In panels "`prob_left`" and "`prob_right`", we report the estimated heaping probabilities p_1, and q_1.

◁

6.4 Testing for the presence of heaping effects

This command provides a subroutine to test for the presence of heaping effects ($\mathbf{H}^{\pi_1}$) via the Likelihood Ratio (LR) test described in Remark 4.2 in Section 4 of ACG, and briefly discussed in Section 4 in this paper. The null hypothesis ($\mathbf{H}_0^{\pi_1}$) states that all heaping probability parameters are zero, and the alternative ($\mathbf{H}_A^{\pi_1}$) is that at least one heaping probability parameters is greater than zero.

▷ Example

To test for the presence of heaping effects under the model specification described in the last subsection, we can simply add `testpi1` option to the command:

```
. heapmph duration age_m school_m, censor(18) hstar(5) jbar(3) kbar(12) rbar(1)
testpi1
```

and it returns:

```
-- Estimating the coefficients under the null (H0) --
Coefficients estimation in progress (% finished approx.): 0%....1%........10%
..................20%.................30%.................40%.................50%
..............60%..................70%.................80%..................90%
.....................100%
-- Estimating the coefficients under the alternative (H1) --
Coefficients estimation in progress (% finished approx.): 0%....1%........10%
..................20%.................30%.................40%.................50%
..............60%..................70%.................80%..................90%
.....................100%
```

```
MooN bootstrap will take approximately 44 minutes (100 replications).
(each dot . indicates one replication)
────┼─── 1 ────┼─── 2 ────┼─── 3 ────┼─── 4 ────┼─── 5
..................................................    50
..................................................   100

H0: all heaping probability parameters are zero
H1: at least one heaping probability parameters is greater than zero
```

QLR Statistic	[The Bootstrap Critical Values] 90%	95%	99%
28.693814	28.8097	28.8310	28.8604

The Stata output table reports the test statistic along with the corresponding bootstrapped critical values at 90%, 95% and 99% levels.[11] In this example, we fail to reject the null hypothesis at the 10% significance level, which suggests that there is no clear evidence of heaping.

◁

6.5 Further options

Different censoring points for each observation

The option for variable censoring is `vcensor`(*varname*), where *varname* contains the data for individual specific censoring points.

Let `uncensor_dummy` stand for a period-specific censoring indicator variable. `uncensor_dummy`=1 if the observation's spell is complete, and `uncensor_dummy`=0 if the spell is right-censored. For example, we randomly generate `uncensor_dummy` from a Bernoulli(0.1) distribution, and apply the `heapmph` command:

11. Note that the command stores 1^{st}, 5^{th}, 10^{th}, 90^{th}, 95^{th} and 99^{th} percentiles of the bootstrap empirical distribution function in `e()`. See the help file to this command for details.

```
. generate byte uncensor_dummy = uniform() <0.1
. heapmph duration age_m school_m,vcensor(uncensor_dummy) hstar(5) jbar(3) kbar(12)
rbar(2)
```

(output omitted)

Note that if neither `vcensor`(*varname*) nor `censor` (*integer*) is specified, the command by default will fix the right-censoring point at the maximum value of the dependent variable in the usable sample.

Bootstrap repetitions

The `rep`(*integer*) option allows users to specify the number of M out of N bootstrap replications for calculating the standard errors. The default value is set at 100. In the example shown in Section 6.3, it takes 24 minutes to run 100 bootstrap iterations in a 64 bit Stata 15 SE on a desktop computer with the Intel i7 quad-core processor with 4.0GHz.

Optimization settings

This paper implements the Simulated Annealing (SA) algorithm to maximize the likelihood function of the model. The SA method, proposed by Kirkpatrick et al. (1983), is a popular local search algorithm for stochastically approximating the global optimum of a given objective function. The review of the algorithm and its technical details can be found in Dowsland and Thompson (2012), for example. The SA algorithm is particularly useful for our model, and may be preferable to the conventional Newton algorithm, since SA is better at locating global maximum when the likelihood function is complex, as in our case.

The `heap` package integrates a self-contained Mata function for SA method in Kirkpatrick et al. (1983). In this function, we have designed 10 options for users to control settings of the SA algorithm. For instance, `sa_maxiter`(*integer*) allows the user to set the maximum number of total iterations (the default is 8000) and the `sa_stopTemp` (*real*) option allows one to set the temperature at which to stop the searching algorithm (the default is 1×10^{-8}). The full details about the settings are listed in the help file to this command. Besides, the seed state for initializing the random number generator is set to be 1000 by default, and can be tweaked in the `seed`(*real*) option.[12]

12. Another user-written Mata function is 'simann'. We have not used this since, we did not know how the function actually performed as the author did not disclose the source code of this function. Additionally, the command was not flexible enough, since some of the parameters were fixed in the 'sinmann' function. Based on the Matlab's simulated annealing function, one of the authors (Zizhong Yan) has programmed a more flexible Mata simulated annealing function for our heaping command.

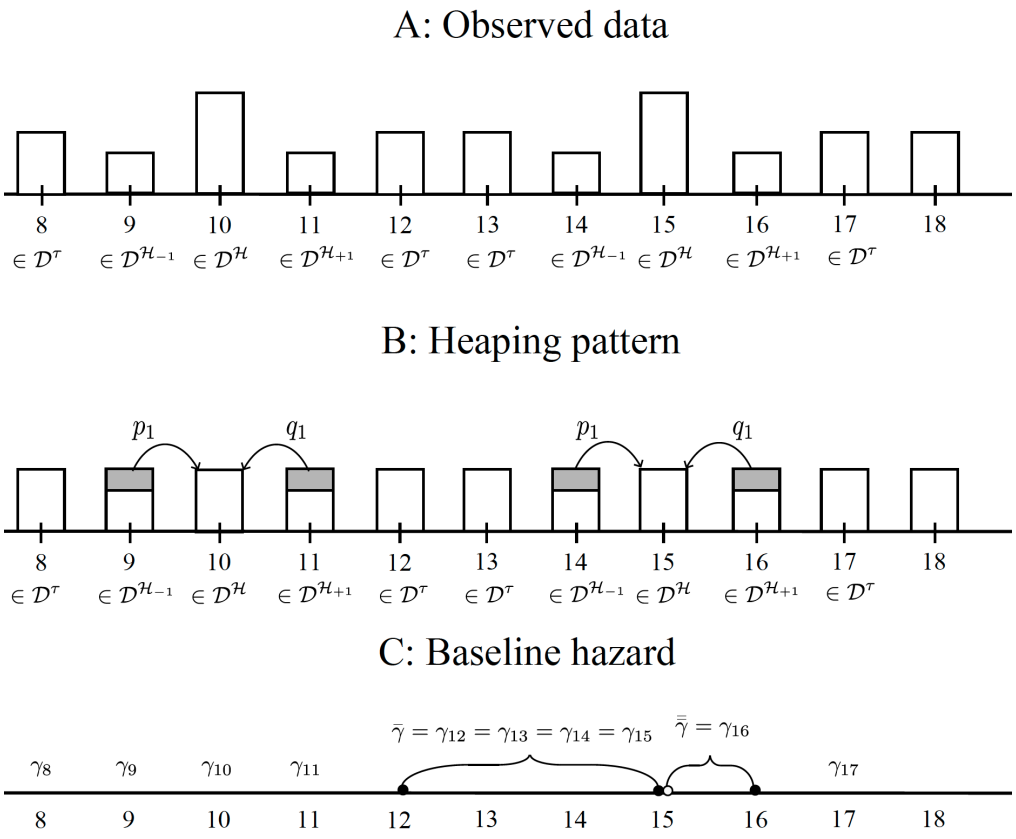
Display options

For diagnosing and monitoring purposes, we provide the following two options to display the intermediate command outputs. First, the `detail` option can be used to display a summary of heaping model specifications, and produce a table only for point estimates before conducting the bootstrap. Second, the `sa_maxiter`(*integer*) option can be set to 1 for producing the final report of the simulated annealing, and set to 2 for further displaying the temperature changes in each iteration. The default value of this option is zero which suppresses all output.

7 Conclusion

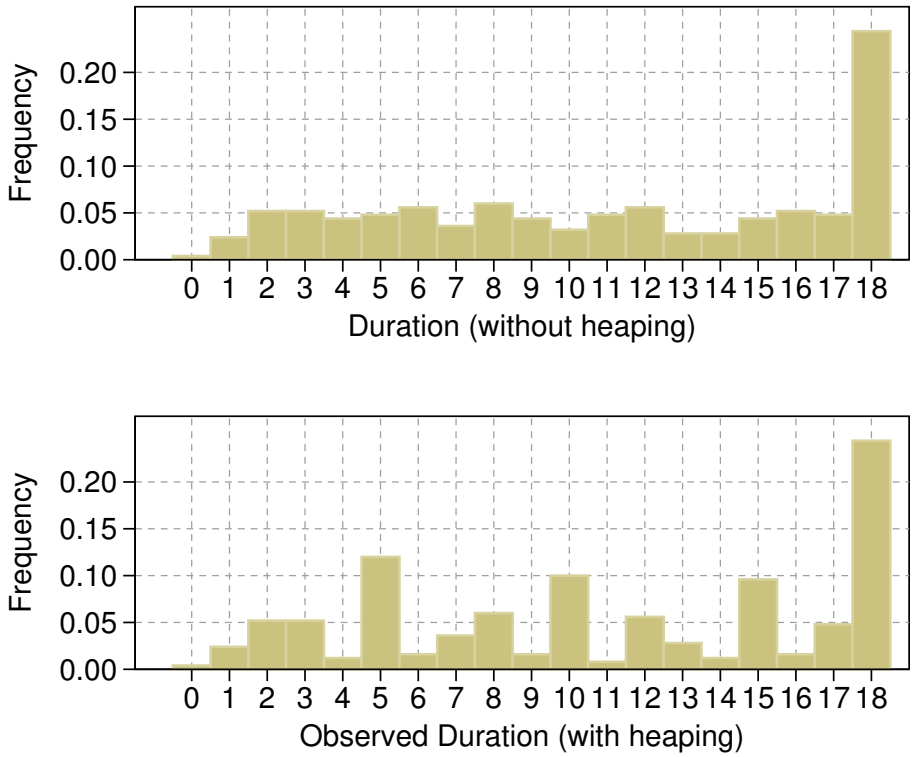
Discrete time duration models are very popular among researchers. The Stata command **heapmph** allows one to estimate a mixed proportional hazard model in discrete time with gamma distributed unobserved heterogeneity, when the observed discrete durations exhibit abnormal concentrations at certain durations points. An accompanying code **heapop** generalizes the above to an ordered probit model. The underlying assumptions and the identification strategy used are discussed fully in ACG.

Figure 1: Stylized Example



Notes: (i) This stylized example allows the heaps at periods 10 and 15. (ii) $\mathcal{D}^{\mathcal{H}}$ is the set of the reported durations on the heaping points. $\mathcal{D}^{\mathcal{H}-1}$ stands for the set of the reported duration that are one period below the nearest heaping point. Similarly, $\mathcal{D}^{\mathcal{H}+1}$ stands for the set of the reported duration that are one period above the nearest heaping point. $\mathcal{D}^{\tau}$ is for the correctly reported durations. (iii) The rounding probabilities of heaping are p_1 and q_1 for $\mathcal{D}^{\mathcal{H}-1}$ and $\mathcal{D}^{\mathcal{H}+1}$, respectively. (iv) The constant part of the baseline hazard starts from period 12. By the asymmetric heaping, the constant baseline hazard parameters ($\gamma_t = \ln \int_t^{t+1} \lambda_0(s)\mathrm{d}s$) are at different levels for periods $[12, 15]$ and $(15, 16]$. (v) In Stata output table, `gamma8`", "`gamma9`",..., "`gamma11`" correspond to the baseline hazard in period 8, 9,..., 11, respectively. "`gamma12`" corresponds to the constant baseline hazard during periods $[12, 15]$. "`gamma13`" is for period $(15, 16]$ and "`gamma14`" is for the period 17. The period 18 in this example is the right-censoring date.

Figure 2: Histograms of the duration variable in the example data (see Section 6.2)



Notes: (i) The upper graph plots the unobserved true duration variable without the heaping pattern. (ii) The lower graph presents the observed duration variable. In this example, duration points 4, 9, and 11 are rounded up to 5, 10, and 15 with probability 0.7, respectively. Duration points 6, 10, and 16 are rounded down to 5, 10, and 15 with the same probability 0.7, respectively. (i.e., $p_1 = q_1 = 0.7$) (iii) The right-censoring date is the period 18 in this data.

Table 1: Summary statistics for the variables used in the illustration

Variable	**Mean (SD)**
Number of days of survival of the children excluding the censored observations	8.873 (4.828)
Proportion of censored observations at 18 days	0.244 (0.430)
Age of mother at the birth of the child, in years	24.060 (5.120)
Mothers education, in years	3.248 (4.135)
Proportion of children who were born during the treatment period	0.132 (0.339)
Total number of children	250

Notes: See Section 6.2. for the model that generated this data.

8 References

Abbring, J. H., and G. J. Van Den Berg. 2007. The unobserved heterogeneity distribution in duration analysis. *Biometrika* 94(1): 87–99.

Andrews, D. W. 2000. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* 68(2): 399–405.

Arulampalam, W., V. Corradi, and D. Gutknecht. 2017. Modeling heaped duration data: An application to neonatal mortality. *Journal of Econometrics* 200(2): 363–377.

Cox, D. R. 1972. Models and life-tables regression. *Journal of Royal Statistical Society: Series B* 34: 187–220.

Cox, D. R., and D. Oakes. 1984. Analysis of survival data. *Chapman&Hall, London* .

Dowsland, K. A., and J. M. Thompson. 2012. Simulated annealing. In *Handbook of Natural Computing*, 1623–1655. Springer.

Forster, M., and A. M. Jones. 2001. The role of tobacco taxes in starting and quitting smoking: duration analysis of British data. *Journal of the Royal Statistical Society: Series A* 164(3): 517–547.

Han, A., and J. A. Hausman. 1990. Flexible parametric estimation of duration and competing risk models. *Journal of applied Econometrics* 5(1): 1–28.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science* 220(4598): 671–680.

Meyer, B. D. 1990. Unemployment insurance and unemployment spells. *Econometrica* 58(4): 757–782.

Pudney, S. 2008. Heaping and leaping: Survey response behaviour and the dynamics of self-reported consumption expenditure. Technical report, ISER Working Paper Series.

Silvapulle, M., and P. Sen. 2005. Constrained Statistical Inference: Inequality, Order, and Shape Restrictions (Wiley Series in Probability and Statistics).

About the authors

Wiji Arulampalam, Department of Economics, University of Warwick, Coventry CV4 7AL, UK. Email: Wiji.Arulampalam@warwick.ac.uk

Valentina Corradi, Department of Economics, University of Surrey, School of Economics, Guildford GU2 7XH, UK. Email: V.Corradi@surrey.ac.uk

Daniel Gutknecht, Department of Economics, University of Mannheim, L7, 3-5, 68131 Mannheim, Germany. Email: Daniel.Gutknecht@gmx.de

Zizhong Yan (Corresponding Author), Center for Econometrics and Microdata Practice, Institute for Economic and Social Research, Jinan University, Guangzhou, China. Email: helloyzz@gmail.com

Appendix to the paper

1 Policy analysis under heaping

1.1 Estimating the policy effect

In this Appendix, we discuss an additional feature of the command `heapmph`, that allows the user to test for a shift in the baseline hazard and/or the reporting probabilities, perhaps due to a change in a binary variable.[13] For example, one might be interested in the analysis of the effects of a certain policy change on duration outcomes, and the binary indicator will then take the value of one for treated individuals. The main focus of ACG was on whether the introduction of the conditional-cash-transfer program (Janani Siraksha Yojana (JSY)), had an effect on neo-natal mortality, and also on the reporting behaviour of women. The hypothesis being that more accuarate records are available on average, compared to before, as the program encouraged women to deliver babies in health facilities.

In the `heapmph` command, the `treat`(*varname*) option allows the user to account for the effect of a policy change on duration outcomes where *varname* is the name of the binary indicator variable. We use the same data as discussed in Section 6.2. The treatment indicator variable is the actual treatment status for the 250 children randomly chosen from the original ACG dataset, and as reported in Table 1, 13.2% of the children in our sample, were born during the treatment period. The data used here is the same one as before, and thus, we would expect the null-hypothesis of zero treatment effects on the gamma parameters and the mis-reporting probabilities, to be not-rejected.

▷ Example

In the data example used in this paper, the `jsy_dummy` variable is the indicator for whether the JSY program was in place at the time of birth of the child. We code:

```
. heapmph duration age_m school_m, treat(jsy_dummy) censor(18) hstar(5) jbar(3)
kbar(12) rbar(1)
```

and the command returns:

```
------------------------------------------------------------------------------
 Initial temperature:       1          Final temperature:           0.000000010
 Consecutive rejections:    142        Number of function calls:    16,796
 Total final loss:          545.678    Observations:                250
------------------------------------------------------------------------------

MooN bootstrap will take approximately 20 minutes (100 replicates).
(each dot . indicates one replication)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
------------------------------------------------------------------------------
                 |             Bootstrap
                 |    Coef.   Std. Err.      z       P>|z|   [95% Conf. Interval]
```

13. Similar tests can also be carried out in the ordered probit model, where the treatment is allowed to shift the gamma parameters and also the mis-reporting probabilities. In order to save space, we do not report an example using the `heapop` command here.

e^(gamma)						
gamma0	.0119782	.099657	0.12	0.904	-.0018336	.0257899
gamma1	.3101611	1.361863	0.23	0.820	.1214165	.4989057
gamma2	.1934687	.4291931	0.45	0.652	.1339856	.2529519
gamma3	.1055331	.6029322	0.18	0.861	.021971	.1890953
gamma4	.1569196	.7616333	0.21	0.837	.0513626	.2624766
gamma5	.6022752	1.159461	0.52	0.603	.4415821	.7629683
gamma6	.2193861	.6705273	0.33	0.744	.1264558	.3123164
gamma7	.7906292	2.405641	0.33	0.742	.4572244	1.124034
gamma8	.6238689	1.22665	0.51	0.611	.4538639	.793874
gamma9	.7439532	2.45145	0.30	0.762	.4041996	1.083707
gamma10	.7713627	1.847784	0.42	0.676	.5152728	1.027453
gamma11	.2208376	2.131989	0.10	0.918	-.074641	.5163163
gamma12	1.051655	1.575731	0.67	0.505	.8332702	1.27004
gamma13	1.187741	5.920126	0.20	0.841	.3672531	2.008228
gamma14	2.575353	3.840772	0.67	0.503	2.043049	3.107656
e^(gamma_treat)						
gamma_treat0	.628858	4.295244	0.15	0.884	.0335676	1.224148
gamma_treat1	-2.195223	5.176265	-0.42	0.672	-2.912617	-1.47783
gamma_treat2	.7136727	1.167223	0.61	0.541	.5519039	.8754415
gamma_treat3	2.354873	2.405889	0.98	0.328	2.021434	2.688312
gamma_treat4	2.016853	2.59271	0.78	0.437	1.657522	2.376184
gamma_treat5	-.2498153	1.541195	-0.16	0.871	-.463414	-.0362165
gamma_treat6	1.333424	2.787956	0.48	0.632	.9470334	1.719815
gamma_treat7	-2.273401	4.296768	-0.53	0.597	-2.868903	-1.6779
gamma_treat8	.2881261	.9729039	0.30	0.767	.1532885	.4229637
gamma_treat9	.000829	5.266443	0.00	1.000	-.7290627	.7307207
gamma_tre~10	.3785186	3.832831	0.10	0.921	-.1526847	.9097219
gamma_tre~11	.5854206	6.323443	0.09	0.926	-.2909638	1.461805
gamma_tre~12	.4464092	1.193773	0.37	0.708	.2809608	.6118576
gamma_tre~13	.7901876	3.151577	0.25	0.802	.3534013	1.226974
gamma_tre~14	.3583457	.9396153	0.38	0.703	.2281217	.4885697
sigma						
sigma	.6125635	.985235	0.62	0.534	.4760169	.7491101
beta						
age_m	-.0888528	.0527379	-1.68	0.092	-.0961619	-.0815437
school_m	.2165004	.2431189	0.89	0.373	.1828059	.250195
prob_left						
p1	.6887247	.4938071	1.39	0.163	.6202865	.7571628
prob_right						
q1	.4523355	2.103406	0.22	0.830	.1608182	.7438527
prob_left_treat						
p1D	.0494005	1.420037	0.03	0.972	-.1474066	.2462076
prob_right_tr~t						
q1D	-.1049818	2.649767	-0.04	0.968	-.4722207	.2622572

The specifications of the heaping pattern is same as the one in Section 6.3. This Stata output table has the same format as the output table in Section 6.4. In particular, the panel "`e^(gamma_treat)`" in this table reports the estimated baseline parameters for the treatment group units (i.e., $\exp(\gamma^{(2)}(1))$). Panels "`prof_left_treat`"

presents the estimated change of the heaping probabilities ($p_1^{(2)}$) of the treatment group. "`prof_left_right`" reports ($q_1^{(2)}$) of the treatment group.

◁

1.2 Testing hypotheses

When estimating the policy effect, the `heapmph` command provides two options for testing hypotheses as follows.

Test for the changes in the reporting behavior after the policy introduction

As outlined in Section 5 of ACG, first we would like to rule out that changes in the reporting behavior (as a result of the policy introduction) confound any observable effect of the program. Therefore, we start by testing $\mathbf{H}^{\pi_3}$, which under the null ($(\mathbf{H}_0^{\pi_3})$) postulates that all deviations $p_1^{(2)}$ and $q_1^{(2)}$ are jointly equal to zero.

▷ **Example**

For instance, we could use the `testpi3` option:

```
. heapmph duration age_m school_m, treat(jsy_dummy) censor(18) hstar(5) jbar(3)
kbar(12) rbar(1 testpi3
```

and it returns:

```
-- Estimating the coefficients under the null (H0) --
Coefficients estimation in progress (% finished approx.): 0%....1%........10%
...................20%.................30%.................40%.................50%
..............60%.................70%.................80%.................90%
.....................100%
-- Estimating the coefficients under the alternative (H1) --
Coefficients estimation in progress (% finished approx.): 0%....1%........10%
...................20%.................30%.................40%.................50%
..............60%.................70%.................80%.................90%
.....................100%
```

```
MooN bootstrap will take approximately 38 minutes (100 replications).
(each dot . indicates one replication)
    ├── 1 ──┼── 2 ──┼── 3 ──┼── 4 ──┼── 5
..................................................    50
..................................................   100

H0: treatment has not changed the exit probability
H1: over at least one period the exit probability decreased
```

QLR Statistic	[The Bootstrap Critical Values] 90%	95%	99%
6.1086863	36.8367	43.3614	74.7000

Here, in this illustration, we cannot reject the null hypothesis that there is no change in the heaping probability parameters after the policy introduction, at the 10% level.

◁

Test for whether the treatment has changed the exit probability

The `heapmph` command provides the `testgamma2` option to test for the null hypothesis ($\mathbf{H}_0^{\gamma 2}$) that treatment has not changed the exit probability (e.g., the probability of the event happens) in any of the first ($\bar{\tau}-1$) periods against the alternative ($\mathbf{H}_A^{\gamma 2}$) that over at least one period the exit probability decreased. For the technical details of this test, see Section 5 of ACG.

▷ **Example**

In Stata, we code:

```
. heapmph duration age_m school_m, treat(jsy_dummy) censor(18) hstar(5) jbar(3)
kbar(12) rbar(1 testgamma2
```

and it returns:

```
-- Estimating the coefficients under the null (H0) --
Coefficients estimation in progress (% finished approx.): 0%....1%........10%
..................20%.................30%.................40%.................50%
..............60%.................70%................80%..................90%
.....................100%
-- Estimating the coefficients under the alternative (H1) --
Coefficients estimation in progress (% finished approx.): 0%....1%........10%
..................20%.................30%.................40%.................50%
..............60%.................70%................80%..................90%
.....................100%
```

```
MooN bootstrap will take approximately 37 minutes (100 replications).
(each dot . indicates one replication)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100

H0: no change in the heaping probability parameters after the policy (treatment) introduction
H1: a change in at least some rounding parameters
```

QLR Statistic	[The Bootstrap Critical Values] 90%	95%	99%
59.759797	95.0681	100.5209	112.4219

From the output tables, we find that the null of $\mathbf{H}_0^{\gamma 2}$ cannot be rejected at a 10% significance level.

◁

2 Examples of `heapop` command

The syntax and corresponding options of `heapop` command are identical to those of the `heapmph` command (see both Section 6 and Appendix 1). This appendix attaches

example usages of the heapop command under the same specification as used in Section 6.

2.1 Model estimation: using heapop command

We first request Stata to implement the heapop command to estimate the model:

▷ Example

```
. heapop duration age_m school_m, censor(18) hstar(5) jbar(3) kbar(12) rbar(1)
```

The command returns:

```
Coefficients estimation in progress (% finished approx.): 0%....1%........10%
..................20%.................30%.................40%.................50%
..............60%.................70%.................80%..................90%
.....................100%

 Initial temperature:       1          Final temperature:           0.000000010
 Consecutive rejections:    45         Number of function calls:    34,215
 Total final loss:          593.206    Observations:                250

MooN bootstrap will take approximately 26 minutes (100 replicates).
(each dot . indicates one replication)
────┼─── 1 ────┼─── 2 ────┼─── 3 ────┼─── 4 ────┼─── 5
..................................................    50
..................................................   100
```

	Coef.	Bootstrap Std. Err.	z	P>|z|	[95% Conf.	Interval]
e^(gamma)						
gamma0	.2721881	.075306	3.61	0.000	.2617512	.282625
gamma1	.3747939	.0963867	3.89	0.000	.3614354	.3881524
gamma2	.4633893	.1222535	3.79	0.000	.4464458	.4803327
gamma3	.4146261	.1156667	3.58	0.000	.3985955	.4306567
gamma4	.3467341	.6549929	0.53	0.597	.2559567	.4375114
gamma5	.5220416	.3889637	1.34	0.180	.4681339	.5759492
gamma6	.4701779	.9988004	0.47	0.638	.3317512	.6086046
gamma7	.3721548	.1083006	3.44	0.001	.3571451	.3871645
gamma8	.6864175	.1918713	3.58	0.000	.6598255	.7130095
gamma9	.694045	1.225619	0.57	0.571	.5241829	.8639071
gamma10	.6801856	.5940031	1.15	0.252	.597861	.7625103
gamma11	.3868165	.8571374	0.45	0.652	.2680233	.5056097
gamma12	.8342315	.5854894	1.42	0.154	.7530868	.9153762
gamma13	1.68294	2.923839	0.58	0.565	1.277717	2.088163
gamma14	1.991177	.6263514	3.18	0.001	1.90437	2.077985
beta						
age_m	-.0854586	.010742	-7.96	0.000	-.0869474	-.0839699
school_m	.0968189	.0027978	34.61	0.000	.0964311	.0972066
prob_left						
p1	.7011283	.4484472	1.56	0.118	.6389767	.7632799

```
prob_right
          q1 | .6718045   1.121392      0.60     0.549     .5163875   .8272215
```

◁

Two points worth noting here. First, the specification of the ordered probit model is in line with the framework of our proportional hazard model, hence we have $-z_i\beta$ on the right-hand side of the equation 6. When interpreting the estimated `beta` coefficients, one might need to reorient the signs. Second, the model estimated does not account for the unobserved heterogeneity, one has to change the set up of the data input into a panel form.

2.2 Testing hypotheses: using `heapop` command

▷ Example

To test for the presence of heaping effects ($\mathbf{H}^{\pi_1}$), we code:

```
. heapop duration age_m school_m, censor(18) hstar(5) jbar(3) kbar(12) rbar(1) testpi1
```

and it returns:

```
-- Estimating the coefficients under the null (H0) --
Coefficients estimation in progress (% finished approx.): 0%....1%........10%
...................20%.................30%.................40%.................50%
..............60%.................70%.................80%.................90%
.....................100%
-- Estimating the coefficients under the alternative (H1) --
Coefficients estimation in progress (% finished approx.): 0%....1%........10%
...................20%.................30%.................40%.................50%
..............60%.................70%.................80%.................90%
.....................100%
```

```
MooN bootstrap will take approximately 47 minutes (100 replications).
(each dot . indicates one replication)
----+--- 1 ----+--- 2 ----+--- 3 ----+--- 4 ----+--- 5
..................................................    50
..................................................   100

H0: all heaping probability parameters are zero
H1: at least one heaping probability parameters is greater than zero
```

QLR Statistic	[The Bootstrap Critical Values] 90%	95%	99%
25.995373	26.2054	26.2146	26.2384

◁

2.3 Policy analysis: using `heapop` command

▷ Example

One might be interested in using the ordered probit model to estimate the effects of a certain policy change on the heaping probabilities. We code:

```
. heapop duration age_m school_m, treat(jsy_dummy) censor(18) hstar(5) jbar(3)
kbar(12) rbar(2)
```

and the command returns:

```
Coefficients estimation in progress (% finished approx.): 0%....1%........10%
...................20%.................30%.................40%.................50%
..............60%.................70%................80%..................90%
.....................100%

------------------------------------------------------------------------------
 Initial temperature:       1          Final temperature:           0.000000010
 Consecutive rejections:    0          Number of function calls:    17,079
 Total final loss:          580.834    Observations:                250
------------------------------------------------------------------------------

MooN bootstrap will take approximately 20 minutes (100 replicates).
(each dot . indicates one replication)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
```

	Coef.	Bootstrap Std. Err.	z	P>|z|	[95% Conf.	Interval]
e^(gamma)						
gamma0	.2201253	.1222564	1.80	0.072	.2031814	.2370692
gamma1	.3234947	.1733278	1.87	0.062	.2994727	.3475167
gamma2	.373393	.2060417	1.81	0.070	.3448371	.4019489
gamma3	.2748416	.1529418	1.80	0.072	.2536449	.2960382
gamma4	.1865983	.5957995	0.31	0.754	.1040247	.2691719
gamma5	.4966012	.6241665	0.80	0.426	.4100962	.5831063
gamma6	.3905122	.940318	0.42	0.678	.2601907	.5208336
gamma7	.4948187	.3479875	1.42	0.155	.4465901	.5430473
gamma8	.5846685	.3741055	1.56	0.118	.5328202	.6365169
gamma9	.3797094	1.122963	0.34	0.735	.2240747	.5353441
gamma10	.6438094	.7899548	0.81	0.415	.5343273	.7532916
gamma11	.5438619	1.104397	0.49	0.622	.3908003	.6969236
gamma12	.619283	.9040273	0.69	0.493	.4939912	.7445748
gamma13	1.587706	4.084947	0.39	0.698	1.021561	2.153851
gamma14	1.736305	1.103575	1.57	0.116	1.583357	1.889252
e^(gamma_treat)						
gamma_treat0	-.0277171	3.423028	-0.01	0.994	-.5021246	.4466905
gamma_treat1	.4552177	2.410112	0.19	0.850	.1211932	.7892421
gamma_treat2	1.580163	1.246502	1.27	0.205	1.407406	1.752919
gamma_treat3	2.63764	1.641752	1.61	0.108	2.410105	2.865175
gamma_treat4	2.118725	2.607281	0.81	0.416	1.757375	2.480076
gamma_treat5	.1418619	1.802993	0.08	0.937	-.1080202	.391744
gamma_treat6	1.270889	2.182315	0.58	0.560	.9684354	1.573342
gamma_treat7	-2.791271	5.147388	-0.54	0.588	-3.504663	-2.07788
gamma_treat8	.4949781	.6722757	0.74	0.462	.4018055	.5881508
gamma_treat9	.2580867	5.021704	0.05	0.959	-.4378859	.9540594
gamma_tre~10	.8517118	3.613198	0.24	0.814	.350948	1.352476
gamma_tre~11	-.5838686	5.963755	-0.10	0.922	-1.410403	.2426656
gamma_tre~12	1.044385	.7260383	1.44	0.150	.9437617	1.145009
gamma_tre~13	.388965	3.01887	0.13	0.897	-.029429	.8073589
gamma_tre~14	.391813	.4705357	0.83	0.405	.3266001	.457026

beta						
age_m	-.075491	.0198171	-3.81	0.000	-.0782375	-.0727445
school_m	.0978651	.0102829	9.52	0.000	.09644	.0992903
prob_left						
p1	.5402499	.4846781	1.11	0.265	.4730769	.6074228
prob_right						
q1	.7557395	2.360927	0.32	0.749	.4285317	1.082947
prob_left_treat						
p1D	.1991267	1.050912	0.19	0.850	.0534778	.3447756
prob_right_tr~t						
q1D	-.5226667	2.121365	-0.25	0.805	-.816673	-.2286605

Similar to the Appendix 1, the `heapop` command provides two options for testing hypotheses for the policy analysis.

Example

First, to test for the changes in the reporting behavior after the policy introduction ($\mathbf{H}^{\pi_3}$), one can code:

```
. heapop duration age_m school_m, treat(jsy_dummy) censor(18) hstar(5) jbar(3)
kbar(12) rbar(1) testpi3
```

and it returns:

```
-- Estimating the coefficients under the null (H0) --
Coefficients estimation in progress (% finished approx.): 0%....1%.....10%
..................20%.................30%.................40%.................50%
.............60%.................70%................80%..................90%
.....................100%
-- Estimating the coefficients under the alternative (H1) --
Coefficients estimation in progress (% finished approx.): 0%....1%.....10%
..................20%.................30%.................40%.................50%
.............60%.................70%................80%..................90%
.....................100%
```

```
MooN bootstrap will take approximately 37 minutes (100 replications).
(each dot . indicates one replication)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100

H0: treatment has not changed the exit probability
H1: over at least one period the exit probability decreased
```

QLR Statistic	[The Bootstrap Critical Values] 90%	95%	99%
-2.3731951	1.9105	2.3122	4.7291

▷ Example

Second, we test for whether the treatment has changed the exit probability ($\mathbf{H}^{\gamma_2}$):

```
. heapop duration age_m school_m, treat(jsy_dummy) censor(18) hstar(5) jbar(3)
kbar(12) rbar(1) testgamma2
```

and it returns:

```
-- Estimating the coefficients under the null (H0) --
Coefficients estimation in progress (% finished approx.): 0%....1%........10%
..................20%.................30%.................40%.................50%
..............60%.................70%.................80%.................90%
.....................100%
-- Estimating the coefficients under the alternative (H1) --
Coefficients estimation in progress (% finished approx.): 0%....1%........10%
..................20%.................30%.................40%.................50%
..............60%.................70%.................80%.................90%
.....................100%
```

```
MooN bootstrap will take approximately 37 minutes (100 replications).
(each dot . indicates one replication)
----+--- 1 ----+--- 2 ----+--- 3 ----+--- 4 ----+--- 5
..................................................    50
..................................................   100
```

```
H0: no change in the heaping probability parameters after the policy (treatment) introduction
H1: a change in at least some rounding parameters
```

QLR Statistic	[The Bootstrap Critical Values] 90%	95%	99%
22.219006	25.1267	26.0856	27.8067

◁