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 the individual-specific censoring, and different sized heap points. The second command, **`heapop`**, extends the framework to ordered choice outcomes, subject to heaping. Suitable specification tests are also provided.

Keywords: st0001, **`heapmph`**, **`heapop`**, Discrete time duration model, Mixed proportional hazards model, Ordered choice model, Heaping, Measurement Error.

Acknowledgements: We are grateful to the British Academy (grant number: SG160731 - Estimation and inference with heaped data - a novel approach), for funding this project. Zizhong Yan acknowledges the support from the 111 project of China (project number B18026). We also thank David M. Drukker and a referee of this journal for helpful comments and 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 et al. 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 of 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 semiparametric baseline hazard function in the presence of heaping of observations at certain durations, and gamma distributed unobserved heterogeneity (frailty). In the accompanying code `‘heapop’,` we extend the framework to an ordered choice model, allowing for the presence of heaping points.

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 so called 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.

Finally, in Appendix 1 we show how the `‘heapmph’` command can be used to test for a shift in the heaping probability and/ or baseline parameters as a consequence of a policy or regime change, while Appendix 2 outlines similar examples for the `‘heapop’` command. Finally, Appendix 3 formally links the proportional hazard model to the Extreme-Value (EV) ordered choice model (Han and Hausman 1990) outlining the implications for the interpretation of the parameters.

2 Mixed proportional hazard model with ‘heaping’

2.1 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 $i = 1, \dots, N$. 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.[1] Based on the neonatal mortality illustration from ACG, we also illustrate our command using a simulated dataset based on ACG.

Assumptions:

A1 Excessive concentrations of reported failures occur at time periods that are multiples of a positive integer. This implies 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.[2] 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. Following ACG, our illustration also assumes the .heaping to be at points that are multiples of 5.

A2 In order to identify the baseline hazard from possibly misreported observations, we need to impose a structure on the heaping process. In the illustration provided here, we assume that one period to the right and to the left of each heap point are associated with that heap. We denote the maximum number of time periods that a duration can be rounded to as $\overline{r}$, and in this example $\overline{r} = 1$. That is, we assume that the duration points 4, 9, and 14, will be rounded up, while 6, 7, and 16 will be rounded down to 5, 10 and 15, respectively.

A3 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, as we assume that the outcome variable is censored at 18 days. The maximum number of heaps is assumed to be $\overline{j}$, and in our example $\overline{j}$ =3.

A4 The censoring is exogenous, and the censored observations are correctly reported.

A5 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. Let p_1, p_2, etc. denote the corresponding rounding up probabilities when a true duration is lower by one, two, etc units from the nearest heaping point. Similarly let q_1, q_2, etc. denote the rounding down probabilities when a true duration is higher by one, two, etc. units from the nearest heaping point. In our illustration, a reported duration of say 10 days, includes true durations of 11 (9) days, which have been rounded down (up) to 10 days (see again Figure 1).

[1] Note, ACG discusses a more general setup which can accommodate more complex heaping mechanisms.

[2] It is noteworthy that the theoretical setup can in principle be straightforwardly amended to allow for non-equally spaced heaping points, see the paper ACG.

Hence, p_1 is the probability that a true duration of 9 will be rounded up to 10 days. Analogously q_1 is the probability that a true duration of 11 will be rounded down to 10 days.

A6 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 ps and the qs, in this region, from the observed data. Subsequently, we can use these identified probability parameters to pin down the rest of the baseline and other hazard parameters. See Figure 1.

2.2 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 uncensored 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 below).

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 A3 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:

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

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. More specifically, we impose the following assumptions on the properties and the distributional form of v_i, which are standard in the duration literature:

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

(ii) The density of v is the Gamma with unit mean and variance σ^2.[3]

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

[3] The assumption of Gamma distribution for v_i gives us a closed form expression for the unconditional probabilities. While the choice of the Gamma distribution might appear overly restrictive at first sight, we note that this can often be rationalized theoretically (Abbring and Van Den Berg 2007). In addition, 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, display little sensitivity to alternative distributional assumptions.

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 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 $\theta = \{\underline{\theta}', \sigma, p_1, \ldots, p_{\overline{\tau}}, q_1, \ldots, q_{\overline{\tau}}\}'$ from the possibly misreported durations. Before setting up the likelihood function, 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, we have the following unconditional likelihood contributions.[4]

The likelihood function for the observed sample is:

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

and so

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

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”

[4] For ease of exposition, we have assumed a constant censoring point (type I censoring; Cox and Oakes 1984). However, the program allows the censoring points to vary over i.

probability parameters are equal to zero. In this case, the limiting distribution is no longer normal as the information matrix is not block 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.

3 Ordered probit model with heaping: specification and estimation

In general, there are many observed discrete outcomes (other than durations) that can exhibit heaping. For instance, survey data on the number of doctor visits or on cigarette consumption in a given period of time is often subject to this phenomenon. Here we discuss the estimation of an ordered probit model allowing for heaping. In Appendix 3, we provide a discussion on the link between the discrete duration model derived from the proportional hazard specification and the Type 1 Extreme Value (EV) ordered choice model (Han and Hausman 1990).[5] To keep notational clutter to a minimum, we do not explicitly show the conditioning set in the discussions below and in Appendix 3.

Consider the following latent variable model representation of an ordered choice model:[6]

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

where y_i^* represents the latent outcome, z_i stands for the vector of regressors, $\beta^\dagger$ is the vector of coefficients, and let the distribution function of the error term ε_i be $G(\cdot)$. Assume we have an ordered discrete outcome variable coded as $y_i \in \{0, ..., J\}$. That is, we have:

$$y_i = j \;\text{ if and only if }\; \kappa_j \leq y_i^* = z_i'\beta^\dagger + \varepsilon_i < \kappa_{j+1},$$

where $\kappa_0, ...\kappa_J$ are the threshold parameters that divide the real line into a finite number of intervals. Here, we have assumed the normalizations $\kappa_0 = -\infty$, $\kappa_{J+1} = +\infty$, and $\kappa_j < \kappa_{j+1}$. In addition, note that we require a scale normalization and so z_i may not contain a constant. For any $j \in \{0, \ldots, J\}$, the probabilities of interest are given by:

$$\begin{aligned} \Pr(y_i = j) &= \Pr(\kappa_j \leq y_i^* < \kappa_{j+1}) \\ &= \Pr(\kappa_j - z_i'\beta^\dagger \leq \varepsilon_i < \kappa_{j+1} - z_i'\beta^\dagger) \\ &= G(\kappa_{j+1} - z_i'\beta^\dagger) - G(\kappa_j - z_i'\beta^\dagger) \\ &= \Phi(\kappa_{j+1} - z_i'\beta^\dagger) - \Phi(\kappa_j - z_i'\beta^\dagger), \end{aligned} \tag{4}$$

where $\Phi(\cdot)$ denotes the distribution function of the standard normal distribution function. Of course, other distributions such as e.g. the logistic distribution giving rise to an

[5] The distribution function of the standard Type I EV distribution (or Gumbel distribution) is given by: $G(\varepsilon) = \exp(-\exp(-\varepsilon))$.

[6] Supplementary material provided in ACG sketches the key identification conditions required for the estimation of this model when heaping is present in the data. A class of ordered choice models known as generalized ordered choice models, extends the standard model in different ways to incorporate unobserved heterogeneity (Greene 2014). Our Stata command estimates the standard ordered probit model with heaping, but without unobserved heterogeneity.

ordered logit model can be chosen for $G(\cdot)$ in principle. Finally, as outlined in Appendix 3, the link between the slope coefficients of the `heapmph` and the `heapop` command is given by $\beta = -\beta^{\dagger}$.

In the presence the heaping data, the term $\Pr(y_i = j)$ depends on the four cases:

(I) For correctly reported outcomes, $\Pr(y_i = j) = \Phi(\kappa_{j+1} - z_i'\beta^{\dagger}) - \Phi(\kappa_j - z_i'\beta^{\dagger})$.

(II) For reported outcomes that are $l = 1, 2$, etc. points *below* the nearest heaping point, $\Pr(y_i = j) = (1 - p_l)\big(\Phi(\kappa_{j+1} - z_i'\beta^{\dagger}) - \Phi(\kappa_j - z_i'\beta^{\dagger})\big)$.

(III) Similar to (II), for reported outcomes that are $l = 1, 2$, etc. points *above* the nearest heaping point, $\Pr(y_i = j) = (1 - q_l)\big(\Phi(\kappa_{j+1} - z_i'\beta^{\dagger}) - \Phi(\kappa_j - z_i'\beta^{\dagger})\big)$.

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

$$\begin{aligned}
\Pr(y_i = j) \quad &= \quad \sum_l p_l \big(\Phi(\kappa_{j+1} - z_i'\beta^{\dagger}) - \Phi(\kappa_j - z_i'\beta^{\dagger})\big) \\
&\quad + \sum_l q_l \big(\Phi(\kappa_{j+1} - z_i'\beta^{\dagger}) - \Phi(\kappa_j - z_i'\beta^{\dagger})\big) \\
&\quad + \big(\Phi(\kappa_{j+1} - z_i'\beta^{\dagger}) - \Phi(\kappa_j - z_i'\beta^{\dagger})\big)
\end{aligned}$$

Note that when the outcome is duration data and for right-censored data at $y_i = \bar{\tau}$, the likelihood function can be written as:

$$L_N(\theta^{\dagger}) = \sum_{i=1}^{N} \left(\sum_{j=1}^{\bar{\tau}-1} \Pr(y_i = j) \right)^{d_{ij} \cdot \delta_i} \big(1 - \Phi(\kappa_{\bar{\tau}} - z_i'\beta^{\dagger})\big)^{(1-\delta_i)}, \tag{5}$$

where $\theta^{\dagger} = \{\beta^{\dagger'}, \kappa', p_1, \ldots, p_{\bar{\tau}}, q_1, \ldots, q_{\bar{\tau}}\}'$ and d_{ij} is an indicator equal to one when $t_i = j$ and zero otherwise.

4 Testing for 'heaping'

As pointed out in Section 2.2, 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 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 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 $\pi = \{p_1, \ldots, p_{\bar{\tau}}, q_1, \ldots, q_{\bar{\tau}}\}'$ 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 (ACG).

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). Therefore, if we reject this hypothesis, we are able to make inference based on standard normal critical values, while if we fail to reject we ought to rely on subsampling methods for inference.

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 ps and qs 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 $\max_{j=1,...,\overline{r}} \{PV_{p,j}, PV_{q,j}\} < \alpha$ and, do not reject otherwise.

Thus, as pointed out above, 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 Command Implementation

As discussed in the earlier section, if one or more of the probability parameters lie on the boundary of the parameter space, the asymptotic distribution of the estimator is no longer normal. We provide two tests that can be used to detect this. Hence, the output provides the usual asymptotic standard errors along with the standard errors calculated using the M out of N bootstrap method, where M denotes an integer strictly smaller than N (see ACG).

5.1 Data

We illustrate the use of the `heapmph` and `heapop` commands using generated data based on 250 observations drawn randomly from the original data. More specifically, we retain two covariates of these observations that were found to be significant: 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. As shown in Appendix 3, the ordered choice model where the underlying error term in the latent variable model is Type I Extreme Value distributed, is equivalent to the Cox's proportional hazard (PH) model. We use this to generate our outcome variable. All observations where the child survived for longer than 18 days are treated as censored.[7]

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 we use two different schemes to generate ε_i for demonstrating `heapmph` and `heapop` commands, respectively. In detail:

(i) For `heapmph` command, we characterize a proportional hazard model data example by generating i.i.d. ε_i from a Type I extreme value distribution. The baseline gamma parameters are set as follows: $\exp\big(\gamma(t)\big) = 0.3$ for $t = 0, 1, 2, 3$, $\exp\big(\gamma(t)\big) = 0.6$ for $t = 4, .., 7$, $\exp\big(\gamma(t)\big) = 1.2$ for $t = 8, .., 11$, $\exp\big(\gamma(t)\big) = 2.5$ for $t = 12, ..., 15$, $\exp\big(\gamma(16)\big) = 8$, and $\exp\big(\gamma(17)\big) = 10$.

(ii) For the data example used to demonstrating `heapop`, we draw ε_i from a standard normal distribution. We set the baseline gamma parameters for `heapop` as follows: $\exp\big(\gamma(t)\big) = 0.6$ for $t = 0, 1, ..., 11$, $\exp\big(\gamma(t)\big) = 1.5$ for $t = 12, ..., 15$, $\exp\big(\gamma(16)\big) = 1.8$, and $\exp\big(\gamma(17)\big) = 3$.

Note that we keep the function flat from period 12 to 15. 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.

[7] Please refer to ACG for details of the survey and the original sample used in ACG.

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:

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

We have not included the unobserved heterogeneity in the generation of the above data. Figure 2 plots the histograms of both observed duration variable with heaping and the true duration variable without heaping as generated from the ordered probit model.

5.2 `heapmph` command

This section describes the implementation of the **heapmph** command for the mixed proportional hazard model.

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.

Model estimation

As discussed in Section 5.1, 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 the `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 5.4).

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

1. Since we have generated the data with heaps at days 5, 10, and 15, we 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. We set option `jbar(3)` (i.e., $\bar{j} = 3$) to indicate that there are a *maximum* of three heaping points prior to the censoring point (see point 1 above).
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 ($\bar{k}$). Also, we assume that the heaping is asymmetric, which suggests that constant baseline hazard parameters are at different levels for periods $\{12, 13, 14, 15\}$.[8] 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:    10         Number of function calls:   35,277
 Total final loss:          626.285    Observations:               250
------------------------------------------------------------------------------

MooN bootstrap will take approximately 24 minutes (100 replicates).
(each dot . indicates one replication)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
------------------------------------------------------------------------------------
                |                 Bootstrap              Std. Normal        Bootstrap
                |         Coef.   Std. Err.        z        P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------------
exp(gamma)      |
         gamma0 |   .3057518    .0735412       4.16     0.000        .2955595   .3159441
         gamma1 |   .1539733    .0485058       3.17     0.002        .1472508   .1606959
         gamma2 |   .2757727    .0833442       3.31     0.001        .2642218   .2873236
         gamma3 |    .244938    .0663895       3.69     0.000        .2357369   .2541391
```

[8] See Assumption H (iii) in ACG.

```
      gamma4 |  .2434391   .3819543      0.64    0.524      .1905029   .2963752
      gamma5 |  .5821127    .517579      1.12    0.261      .5103799   .6538455
      gamma6 |  .5240925   .9355118      0.56    0.575      .3944372   .6537478
      gamma7 |  .5396087   .1250595      4.31    0.000      .5222763    .556941
      gamma8 |  1.012341   .2346417      4.31    0.000      .9798218   1.044861
      gamma9 |  .6691035   1.008273      0.66    0.507       .529364    .808843
     gamma10 |  1.239431   .7855968      1.58    0.115      1.130553   1.348309
     gamma11 |  1.046086   1.855192      0.56    0.573      .7889691   1.303202
     gamma12 |   1.73946   .7507658      2.32    0.021      1.635409   1.843511
     gamma13 |  3.974719   5.473584      0.73    0.468      3.216119   4.733319
     gamma14 |  6.835998   2.162892      3.16    0.002      6.536237    7.13576
-------------+----------------------------------------------------------------
sigma        |
       sigma |  .0000838   .0041016      0.02    0.984     -.0004847   .0006522
-------------+----------------------------------------------------------------
beta         |
       age_m |  -.087153   .0098087     -8.89    0.000     -.0885124  -.0857936
    school_m |  .1161123   .0108713     10.68    0.000      .1146056    .117619
-------------+----------------------------------------------------------------
prob_left    |
          p1 |  .6780751   .3751911      1.81    0.071      .6260763    .730074
-------------+----------------------------------------------------------------
prob_right   |
          q1 |  .6848905   .9971728      0.69    0.492      .5466894   .8230916
------------------------------------------------------------------------------
```

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 "exp(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 $\{12, 13, 14, 15\}$. 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 $\{12, 13, 14, 15\}$. "gamma13" is for period 16 and "gamma14" is for the period 17.

Panel "sigma" displays the estimate of σ which is the standard deviation 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. The value of sigma coefficient can be seen to be very close to zero numerically. This does not come unexpected since the data generating process does not feature any unobserved heterogeneity.[9]

◁

Testing for the presence of heaping effects

This command provides a subroutine to test null hypothesis via the Likelihood Ratio (LR) test described in Remark 4.2 in Section 4 of ACG, and briefly discussed in Section

[9] To test this formally, note that this is a test for a parameter on the boundary which requires an adjustment of the critical value or the p-value. Alternatively, for a formally valid likelihood ratio test, see Gutierrez et al. (2001).

4 in this paper. We provide a test (`testpi1`) that can be implemented by addition of an option to the main command. `testpi1` tests the null hypothesis ($\mathbf{H}_0^{\pi_1}$) that all heaping probability parameters are zero, and the alternative ($\mathbf{H}_A^{\pi_1}$) is that at least one heaping probability parameter is greater than zero. Applying the Intersection-Union principle (IUP), we could test the null hypothesis ($\mathbf{H}_0^{\pi_2}$) that at least one heaping probability parameter is equal to zero, and the alternative ($\mathbf{H}_A^{\pi_2}$) is that none is 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 41 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
─────────────────────────────────────────────────────────────────────────
                   │
                   │          [ The Bootstrap Critical Values ]
 QLR Statistic     │               90%          95%          99%
─────────────────────────────────────────────────────────────────────────
                   │
   24.981152       │             25.4834      25.5324      26.1653
─────────────────────────────────────────────────────────────────────────
```

The Stata output table reports the test statistic along with the corresponding bootstrapped critical values at 10%, 5% and 1% levels.[10] 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.

◁

In addition, we employ the IUP rule to test the null that at least one heaping probability parameter is equal to zero ($\mathbf{H}_0^{\pi_2}$). In detail, we sort the p-values of all heaping parameters (p_1 and q_1) displayed in the regression output. The largest p-value

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

is 0.492 in our example, so we do not reject the null at any conventional significance level, hence we have to continue to use M out of N subsampling scheme. Otherwise, if the null hypothesis was rejected, one could simply do inference based on the standard normal distribution.

5.3 `heapop` command for the ordered probit model with heaping

Basic syntax

The syntax and corresponding options of Stata command `heapop` are identical to those of the `heapmph` command (see both Section 5.2 and Appendix 2).

Model estimation

The `heapop` command estimates an ordered probit model with heaping, and can be also employed to deal with the duration outcome data. The `heapop` command requires also four compulsory options to define the pattern of heaping, i.e., `kbar()`, `jbar()`, `hstar()`, and `rbar()`, as introduced in Section 5.2 for the `heapmph` command. In the case of ordered choice or count data, the "`censor()`" option can be used to indicate the maximum number of possible choices or counts. If `censor()` is left unspecified, Stata by default uses the maximum value of the dependent variable as `censor()`.

This section attaches example usages of the `heapop` command under the same specification of the heaping pattern as used in Section 5.2.

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:    71        Number of function calls:    35,562
 Total final loss:          597.325   Observations:                250
───────────────────────────────────────────────────────────────────────────────

MooN bootstrap will take approximately 24 minutes (100 replicates).
(each dot . indicates one replication)
────┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50
..................................................   100

───────────────────────────────────────────────────────────────────────────────────
                │             Bootstrap            Std. Normal          Bootstrap
                │      Coef.   Std. Err.      z       P>|z|       [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
exp(gamma)   |
      gamma0 |  .0243883   .0105039     2.32   0.020     .0229325   .0258441
      gamma1 |  .1492473    .042799     3.49   0.000     .1433157    .155179
      gamma2 |  .3347875   .0983331     3.40   0.001     .3211593   .3484158
      gamma3 |  .3668383   .1037009     3.54   0.000     .3524661   .3812105
      gamma4 |  .3753884   .5571335     0.67   0.500     .2981737   .4526032
      gamma5 |  .6102455    .683616     0.89   0.372     .5155012   .7049899
      gamma6 |  .2802701   1.314076     0.21   0.831     .0981484   .4623918
      gamma7 |  .3632628   .1110072     3.27   0.001      .347878   .3786476
      gamma8 |  .6712402   .1979958     3.39   0.001     .6437994    .698681
      gamma9 |  .8168302   1.139751     0.72   0.474     .6588688   .9747915
     gamma10 |  .6543553   .4722391     1.39   0.166     .5889063   .7198043
     gamma11 |  .2246072   1.079989     0.21   0.835     .0749283   .3742861
     gamma12 |  .8498438   .7121108     1.19   0.233     .7511502   .9485373
     gamma13 |  .9648113   3.916519     0.25   0.805     .4220095   1.507613
     gamma14 |  1.629185   .5038433     3.23   0.001     1.559356   1.699014
-------------+----------------------------------------------------------------
beta         |
       age_m | -.0955762   .0119992    -7.97   0.000    -.0972392  -.0939132
    school_m |  .0995847    .004243    23.47   0.000     .0989966   .1001727
-------------+----------------------------------------------------------------
prob_left    |
          p1 |  .7545335   .4834233     1.56   0.119     .6875344   .8215325
-------------+----------------------------------------------------------------
prob_right   |
          q1 |  .4625635   2.109773     0.22   0.826     .1701639   .7549632
------------------------------------------------------------------------------
```

◁

Testing for the presence of heaping effects

▷ 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 42 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
-------------------------------------------------------------------------
                   |
                   |           [ The Bootstrap Critical Values ]
 QLR Statistic     |               90%          95%          99%
-------------------------------------------------------------------------
  23.808835        |             24.8310      24.8691      24.8918
-------------------------------------------------------------------------
```

◁

Also in this illustration we are, as in the previous example, not able to reject the null of heaping parameters at the boundary at any conventional significance level. Hence, we can use the standard asymptotic theory for the tests of the coefficients.

5.4 Further options

Here, we elaborate on a few additional options, which are available for both commands `heapmph` and `heapop`.

Bootstrap options

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 5.2 , 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.

When choosing the M in the M out of N bootstrap, users can set the option `moon(`*real*`)` to select the share of M observations to be randomly drawn from the sample of size N. Bickel and Sakov (2008) provide an in-depth discussion on the choice of the M parameter. The `heap` packages, by default, set `moon` at 0.8 so that in each MooN bootstrap iteration, 80% of the original sample are randomly kept.

Optimization settings

The commands provided 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 adjusted in the `seed`(*real*) option.[11]

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.

Different censoring points for each observation

The option for variable censoring is `vcensor`(*varname*), where *varname* is a dummy variable which equals to 1 if the observation is complete and is 0 if the observation is right-censored.

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:

```
. 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.

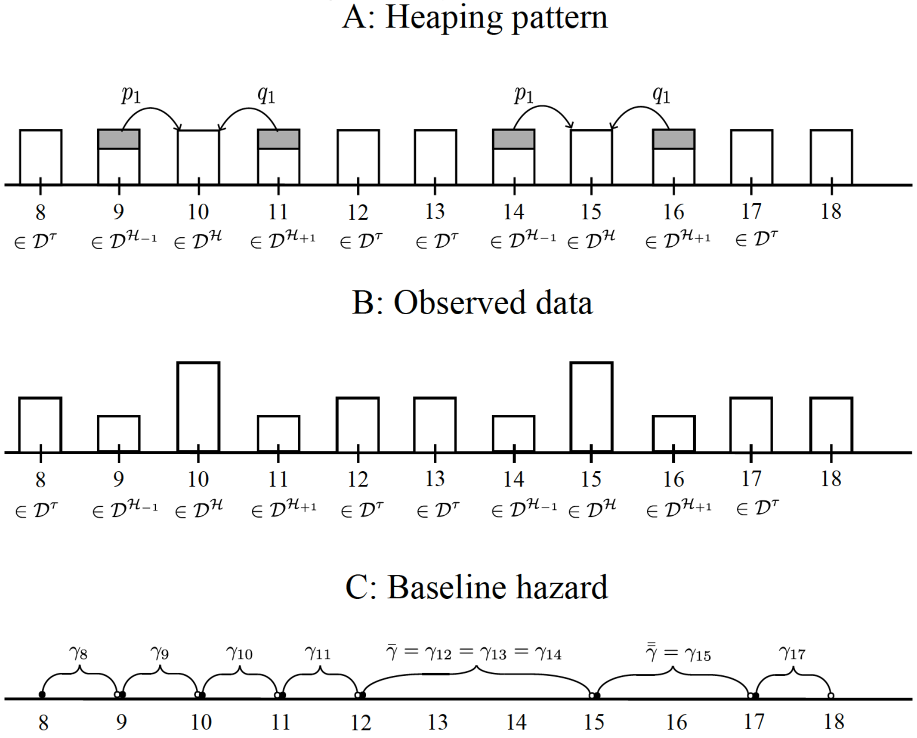
6 Conclusion

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

[11] 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 'simann' 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.

with gamma distributed unobserved heterogeneity, when the observed discrete durations exhibit abnormal concentrations at certain durations points. An accompanying code `heapop` allows for heaping in an ordered probit model. The underlying assumptions and the identification strategy used are discussed fully in ACG.

Figure 1: Stylized Example



A: Heaping pattern

p_1
q_1
p_1
q_1
8 $\in \mathcal{D}^{\tau}$
9 $\in \mathcal{D}^{\mathcal{H}_{-1}}$
10 $\in \mathcal{D}^{\mathcal{H}}$
11 $\in \mathcal{D}^{\mathcal{H}_{+1}}$
12 $\in \mathcal{D}^{\tau}$
13 $\in \mathcal{D}^{\tau}$
14 $\in \mathcal{D}^{\mathcal{H}_{-1}}$
15 $\in \mathcal{D}^{\mathcal{H}}$
16 $\in \mathcal{D}^{\mathcal{H}_{+1}}$
17 $\in \mathcal{D}^{\tau}$
18

B: Observed data

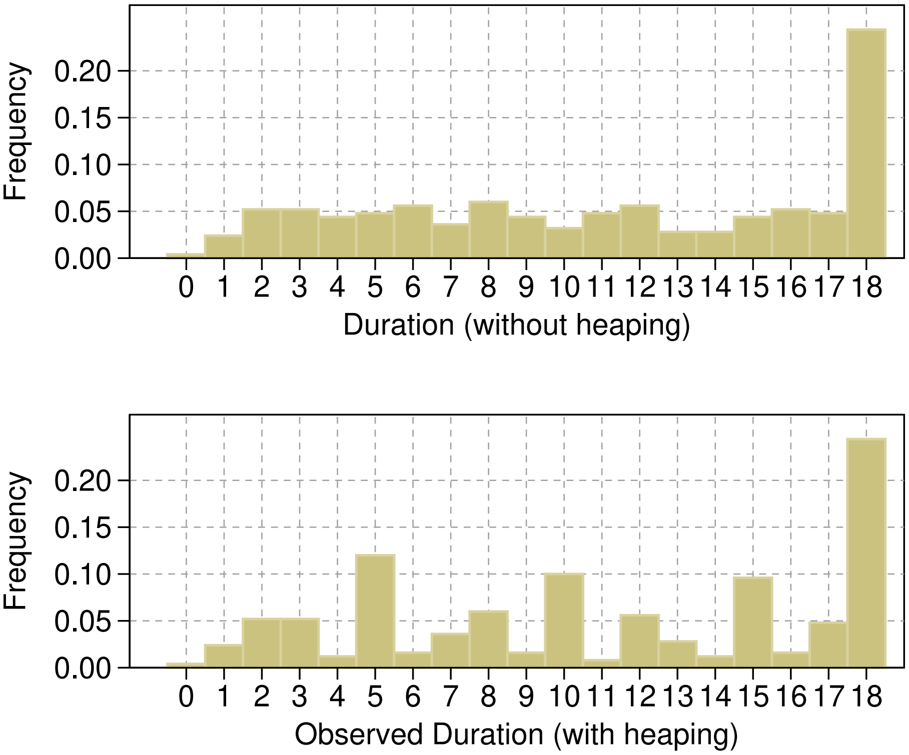
8 $\in \mathcal{D}^{\tau}$
9 $\in \mathcal{D}^{\mathcal{H}_{-1}}$
10 $\in \mathcal{D}^{\mathcal{H}}$
11 $\in \mathcal{D}^{\mathcal{H}_{+1}}$
12 $\in \mathcal{D}^{\tau}$
13 $\in \mathcal{D}^{\tau}$
14 $\in \mathcal{D}^{\mathcal{H}_{-1}}$
15 $\in \mathcal{D}^{\mathcal{H}}$
16 $\in \mathcal{D}^{\mathcal{H}_{+1}}$
17 $\in \mathcal{D}^{\tau}$
18

C: Baseline hazard

γ_8
γ_9
γ_{10}
γ_{11}
$\bar{\gamma} = \gamma_{12} = \gamma_{13} = \gamma_{14}$
$\bar{\bar{\gamma}} = \gamma_{15}$
γ_{17}
8
9
10
11
12
13
14
15
16
17
18

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 $\{12, 13, 14, 15\}$ and 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 $\{12, 13, 14, 15\}$. “gamma13” is for period 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 for demonstrating `heapmph` command (See Section 5.1)



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 of heapmph command

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)
Mother's 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 5.1 for the model that generated this data.

7 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.

Bickel, P. J., and A. Sakov. 2008. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica* 18(3): 967–985.

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.

Greene, W. 2014. Models for Ordered Choices. In *Handbook of Choice Modelling*, ed. S. Hess and A. Daly, 333–362. Edward Elgar Publishing.

Gutierrez, R. G., S. Carter, and D. M. Drukker. 2001. On boundary-value likelihood-ratio tests. *Stata Technical Bulletin* 10(60).

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), Institute for Economic and Social Research, Jinan University, Guangzhou, China. Email: helloyzz@gmail.com

Appendix to the paper

1 Estimating the policy effect

In this Appendix, we discuss an additional feature of the command `heapmph` and `heapop`, that allows the user to test for a shift in the baseline hazard in the duration model or the threshold parameters in the ordered probit model, and/or the reporting probabilities, perhaps due to a change in a binary variable.[12] 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. ACG's main focus, for example, is on whether the Janani Siraksha Yojana (JSY) program in India had any effect on neo-natal mortality, as well as on women's reporting behavior. The hypothesis being that more accurate records are available on average, compared to before, as the program encouraged women to deliver babies in health facilities.

The `treat(`*varname*`)` option of `heapmph` and `heapop` commands 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. The treatment indicator variable is the actual treatment status for the 250 children randomly chosen from the original ACG data set, and as reported in Table 1, 13.2% of the children in our sample, were born during the treatment period. Since the data set used here are the same as that discussed in Section 5.2, we would expect to not reject the null hypothesis of zero treatment effects on the gamma parameters and the misreporting probabilities.

▷ 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. Taking the example of `heapmph` command, 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:

```
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:    252        Number of function calls:    14,911
 Total final loss:          509.181    Observations:                250
-----------------------------------------------------------------------------

MooN bootstrap will take approximately 15 minutes (100 replicates).
(each dot . indicates one replication)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
-----------------+---------------------------------------------------------------
```

[12] 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.

```
                 |               Bootstrap           Std. Normal         Bootstrap
                 |      Coef.    Std. Err.      z       P>|z|       [95% Conf. Interval]
-----------------+--------------------------------------------------------------------
exp(gamma)       |
          gamma0 |   .1229606    .4263486     0.29     0.773         .0638716  .1820495
          gamma1 |   .0386019    .2927666     0.13     0.895        -.0019735  .0791773
          gamma2 |   .1268302    .3835026     0.33     0.741         .0736794  .1799809
          gamma3 |   .1383078    .3970532     0.35     0.728          .083279  .1933366
          gamma4 |   .0875479    .5449995     0.16     0.872         .0120148  .1630809
          gamma5 |   .0173866    2.212301     0.01     0.994        -.2892226  .3239959
          gamma6 |   .5470309    2.963335     0.18     0.854         .1363337  .9577282
          gamma7 |   .3435419     .733828     0.47     0.640         .2418385  .4452453
          gamma8 |   .6535276    1.153605     0.57     0.571         .4936461   .813409
          gamma9 |   .2636071    2.474182     0.11     0.915        -.0792969  .6065112
         gamma10 |   .2316901    3.614136     0.06     0.949        -.2692036  .7325839
         gamma11 |    1.36546    4.803913     0.28     0.776         .6996713  2.031248
         gamma12 |   1.093716     2.02697     0.54     0.589         .8127926   1.37464
         gamma13 |   3.628725    10.84573     0.33     0.738         2.125584  5.131866
         gamma14 |   12.31933    16.26165     0.76     0.449         10.06558  14.57308
-----------------+--------------------------------------------------------------------
exp(gamma_tr)    |
   gamma_treat0  |   1.756707    1.980838     0.89     0.375         1.482177  2.031237
   gamma_treat1  |   1.315527    2.658768     0.49     0.621         .9470402  1.684013
   gamma_treat2  |   1.342548    1.891511     0.71     0.478         1.080398  1.604698
   gamma_treat3  |   .9620915     2.02116     0.48     0.634          .681973   1.24221
   gamma_treat4  |   1.401458    3.599703     0.39     0.697         .9025645  1.900351
   gamma_treat5  |   1.691805    3.873919     0.44     0.662         1.154907  2.228703
   gamma_treat6  |   1.808029    3.684712     0.49     0.624         1.297354  2.318704
   gamma_treat7  |   .5005411    1.533699     0.33     0.744         .2879813  .7131009
   gamma_treat8  |   .6602585    1.314985     0.50     0.616         .4780109  .8425061
   gamma_treat9  |   .0465493    6.273625     0.01     0.994        -.8229308  .9160294
   gamma_tre~10  |   .5768342    4.701807     0.12     0.902        -.0748031  1.228471
   gamma_tre~11  |   .8627947    3.274777     0.26     0.792         .4089338  1.316656
   gamma_tre~12  |   1.450397    2.087993     0.69     0.487         1.161016  1.739778
   gamma_tre~13  |  -.4097875    5.926568    -0.07     0.945        -1.231168   .411593
   gamma_tre~14  |  -2.876843    3.571297    -0.81     0.421        -3.371799 -2.381886
-----------------+--------------------------------------------------------------------
sigma            |
           sigma |   .1412589     .834647     0.17     0.866         .0255827  .2569351
-----------------+--------------------------------------------------------------------
beta             |
           age_m |  -.0607541    .1016391    -0.60     0.550        -.0748406 -.0466677
        school_m |   .1220928      .24783     0.49     0.622         .0877453  .1564403
-----------------+--------------------------------------------------------------------
prob_left        |
              p1 |   .5884231    1.075745     0.55     0.584         .4393325  .7375137
-----------------+--------------------------------------------------------------------
prob_right       |
              q1 |   .8144293    2.117911     0.38     0.701         .5209018  1.107957
-----------------+--------------------------------------------------------------------
prob_left_treat  |
             p1D |  -.3078594    2.353369    -0.13     0.896        -.6340198  .0183009
-----------------+--------------------------------------------------------------------
prob_right_tr~t  |
             q1D |   .9277816    .4917421     1.89     0.059         .8596296  .9959336
--------------------------------------------------------------------------------------
```

The specifications of the heaping pattern is same as the one in Section 5. This Stata output table has the same format as the output table in Section 5.2. In particular, the panel “`exp(gamma_treat)`” in this table reports the estimated baseline parameters for the treatment group units (i.e., $\exp(\gamma^{(2)}(1))$). Panels “`prob_left_treat`” presents the estimated change of the heaping probabilities ($p_1^{(2)}$) of the treatment group. “`prob_left_right`” reports ($q_1^{(2)}$) of the treatment group.

◁

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 35 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
───────────────────┬───────────────────────────────────────────
                   │      [ The Bootstrap Critical Values ]
 QLR Statistic     │          90%          95%          99%
───────────────────┼───────────────────────────────────────────
  22.797652        │       115.5279    160.6586    276.2059
───────────────────┴───────────────────────────────────────────
```

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 26 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
─────────────────────────────────────────────────────────────────────────
                   │
                   │         [ The Bootstrap Critical Values ]
 QLR Statistic     │            90%          95%          99%
───────────────────┼─────────────────────────────────────────────────────
                   │
  101.96437        │          241.9557    272.1655    306.3246
───────────────────┴─────────────────────────────────────────────────────
```

From the output tables, we find that the null of $\mathbf{H}_0^{\gamma_2}$ cannot be rejected at a 10% significance level.

◁

2 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:     16,453
 Total final loss:          522.045    Observations:                 250
──────────────────────────────────────────────────────────────────────────────

MooN bootstrap will take approximately 30 minutes (100 replicates).
(each dot . indicates one replication)
────┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50
..................................................   100

────────────────┬───────────────────────────────────────────────────────────────────
                │                Bootstrap            Std. Normal         Bootstrap
                │      Coef.     Std. Err.      z        P>|z|       [95% Conf. Interval]
────────────────┼───────────────────────────────────────────────────────────────────
exp(gamma)      │
         gamma0 │   .0198937   .0438967     0.45     0.650         .0138099   .0259775
         gamma1 │   .1281145   .2127225     0.60     0.547         .0986327   .1575963
         gamma2 │   .2467684   .2990683     0.83     0.409         .2053197   .2882172
         gamma3 │    .170216   .3739299     0.46     0.649         .1183919     .22204
         gamma4 │   .1762838   1.031001     0.17     0.864         .0333944   .3191733
         gamma5 │    .630011   .8208959     0.77     0.443         .5162407   .7437814
         gamma6 │   .1488082   .4784919     0.31     0.756         .0824926   .2151238
         gamma7 │   .3055578   .7792804     0.39     0.695         .1975551   .4135606
         gamma8 │   .5325998    .611184     0.87     0.384          .447894   .6173055
         gamma9 │   .6474085    2.52873     0.26     0.798         .2969443   .9978726
        gamma10 │    .590414   2.103227     0.28     0.779         .2989215   .8819064
        gamma11 │   .1285976   .4670169     0.28     0.783         .0638723   .1933228
        gamma12 │   .6918678   .8269513     0.84     0.403         .5772581   .8064774
        gamma13 │   .4021108   1.662175     0.24     0.809         .1717451   .6324765
        gamma14 │   1.333671   1.699331     0.78     0.433         1.098156   1.569187
────────────────┼───────────────────────────────────────────────────────────────────
exp(gamma_tr)   │
   gamma_treat0 │   .4740139   4.257217     0.11     0.911        -.1160064   1.064034
   gamma_treat1 │  -1.400214   3.117646    -0.45     0.653        -1.832298  -.9681307
   gamma_treat2 │  -.5067254   4.012836    -0.13     0.900        -1.062876   .0494253
   gamma_treat3 │   1.711607    4.29803     0.40     0.690         1.115931   2.307284
   gamma_treat4 │   1.315424   3.435702     0.38     0.702         .8392599   1.791588
   gamma_treat5 │   .2801411   4.349913     0.06     0.949        -.3227262   .8830083
   gamma_treat6 │   1.414938    2.69613     0.52     0.600         1.041273   1.788602
   gamma_treat7 │  -2.521129   5.270579    -0.48     0.632        -3.251594  -1.790664
   gamma_treat8 │  -.3376568   2.666029    -0.13     0.899        -.7071496    .031836
   gamma_treat9 │  -.8587126   5.239897    -0.16     0.870        -1.584925  -.1324999
   gamma_tre~10 |-.0699398   4.351692    -0.02    0.987     -.6730535   .5331739
   gamma_tre~11 | .0273263   5.325095     0.01    0.996     -.7106942   .7653469
   gamma_tre~12 | .0906128   2.880496     0.03    0.975     -.3086037   .4898292
   gamma_tre~13 | .9698083    2.76759     0.35    0.726      .5862398   1.353377
   gamma_tre~14 | .1811721   1.331974     0.14    0.892       -.00343   .3657743
----------------+----------------------------------------------------------------
beta            |
          age_m |-.0826192   .0426131    -1.94    0.053     -.0885251 -.0767133
       school_m | .1064239   .0657532     1.62    0.106       .097311   .1155369
----------------+----------------------------------------------------------------
prob_left       |
             p1 | .7309284    .478333     1.53    0.127      .6646349    .797222
----------------+----------------------------------------------------------------
prob_right      |
             q1 |  .174482    1.35772     0.13    0.898     -.0136884   .3626524
----------------+----------------------------------------------------------------
prob_left_treat |
            p1D |-.1937085    2.39476    -0.08    0.936     -.5256053   .1381883
----------------+----------------------------------------------------------------
prob_right_tr~t |
            q1D |-.5026433   2.322392    -0.22    0.829     -.8245103 -.1807762
---------------------------------------------------------------------------------
```

◁

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:

```
. heapop duration age_m school_m ,treat(jsy_dummy) hstar(5) jbar(3) kbar(12) rbar(1) testpi3
-- 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 51 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
_________________________________________________________________________
                   |
                   |           [ The Bootstrap Critical Values ]
```

```
QLR Statistic  |         90%         95%         99%
---------------+-------------------------------------------
  62.247078    |       17.9649     35.6791     37.3497
---------------+-------------------------------------------
```

◁

▷ 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 50 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
---------------------------------------------------------------------------
               |        [ The Bootstrap Critical Values ]
QLR Statistic  |         90%          95%          99%
---------------+-----------------------------------------------------------
  148.89668    |      156.3855     156.8654     157.3738
---------------+-----------------------------------------------------------
```

◁

3 Proportional hazard model as an EV ordered choice model

In this Appendix we outline how the (continuous time) proportional hazard model can be represented as an Type I Extreme Value (EV) ordered choice model. This discussion draws heavily from Han and Hausman (1990).

As in equation (1), denote the conditional hazard (without unobserved heterogeneity) by:

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

and let $F(\cdot)$ and $S(\cdot)$, respectively, be the distribution and survivor functions of our duration variable τ^* specified in continuous time. Using the discretized duration variable and the relationship between a hazard function and the survivor function, the probability of observing an uncensored duration of $\tau \in \{1, ..., \overline{\tau}-1\}$ in the proportional hazard model is:

$$\begin{aligned} \Pr\left(\tau_i = \tau\right) = S_i\left(\tau\right) - S_i\left(\tau+1\right) &= F_i\left(\tau+1\right) - F_i\left(\tau\right) \\ &= \exp(-\int_0^{\tau} \lambda_i(\tau^*|z_i)d\tau^*) - \exp(-\int_0^{\tau+1} \lambda_i(\tau^*|z_i)d\tau^*) \\ &= \exp(-\exp(\delta_{\tau+1} + z_i'\beta)) - \exp(-\exp(\delta_\tau + z_i'\beta)). \end{aligned} \tag{6}$$

Here, δ_τ denotes the log of the integrated baseline hazard given by:

$$\delta_\tau = \ln \int_0^{\tau} \lambda_0(s)ds, \qquad \tau = 1, ..., \overline{\tau} - 1.^{13} \tag{7}$$

Comparing (6) with (4), we see that deriving the discrete duration model using the proportional hazard function set-up is equivalent to the ordered choice model where the corresponding latent variable structure is given by:

$$y_i^* = -z_i'\beta + \varepsilon_i, \tag{8}$$

with ε_i following a Type I EV.[14]. Using this in the above equation, we obtain:

$$\begin{aligned} \Pr(\tau_i = \tau) &= \Pr(\kappa_\tau \leq y_i^* < \kappa_{\tau+1}) \\ &= G(\kappa_{\tau+1} - z_i'\beta^\dagger) - G(\kappa_\tau - z_i'\beta^\dagger) \\ &= \exp(-\exp(\kappa_{\tau+1} - z_i'\beta^\dagger)) - \exp(-\exp(\kappa_\tau - z_i'\beta^\dagger)) \end{aligned}$$

Thus, the interpretation of the threshold parameters and also the effects of the covariates differ. In particular, a variable that has a positive effect on the duration will have a negative effect on the hazard of exiting at that point, or, in other words, $\beta = -\beta^\dagger$.

In summary, the ordered choice model derived by assuming a Type I EV distribution for the underlying latent variable equation error ε_i, is equivalent to the discrete duration model derived from a continuous time proportional hazard model.

[13] The model only contains the coefficients $\delta_1, .., \delta_{\overline{\tau}}$. The relationship between these and the baseline hazard function parameters we saw earlier, is given by:

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

[14] See footnote 5