Statistical Analysis of Heaped Duration Data

By K. Petoussis[1] R.D. Gill[2] C. Zeelenberg[3]

December 24, 1997

Abstract

This paper shows how heaping of duration data, e.g. caused by rounding due to memory effects, can be analyzed. If the data are heaped Cox's partial likelihood approach, which is often used in survival analysis, is no longer appropriate. We show how this problem can be overcome by considering the problem as a missing data problem. A variant of Cox's Proportional Hazard Model is constructed that takes heaping into account, and is estimated by maximum likelihood using the EM algorithm. with many nuisance parameters, simultaneously for all parameters. Ingredients of our method are application of the EM algorithm, Cox regression and nonparametric maximum likelihood calculation with 'predicted' data in each M step. An example from practice, where jackknife is used to estimate the variances, illustrates the power of the new methodology.

Keywords and phrases: *heaping, duration data, survival analysis, PHM, profile likelihood, EM.*

1 Introduction

Heaping occurs in many kinds of retrospectively obtained duration data. For example, heaping may occur in unemployment data obtained by periodic labour force surveys (LFS), see [13] for a discussion on the Italian LFS. Anthropometric data on children's age from Tanzania suffer from another kind of heaping, see [8]. Heaping is important in statistical analysis, for

[1] Postal address: Department of Psychology, Vrije Universiteit Amsterdam, de Boelelaan 1081c, 1081 HV Amsterdam, The Netherlands (e-mail: Kos_Petoussis@psy.vu.nl).

[2] Postal address: Mathematical Institute, University of Utrecht, P.O. Box 80,010, 3508 TA Utrecht, The Netherlands (e-mail: gill@math.ruu.nl).

[3] Postal address: Department of Statistical Methods, Statistics Netherlands, P.O. Box 4000, 2270 JM Voorburg, The Netherlands (e-mail: kzlg@cbs.nl).

it may affect results if data are wrongly assumed to be measured without errors.

The following example serves to illustrate how heaping may arise in practice. The relation between unemployment duration and covariates is studied in [6] using standard Cox regression on data from the Netherlands Socio-economic panel (SEP) survey. Unemployment spells are derived from this SEP and linked to the covariates of the participating respondents. However, a peculiarity appears in the frequency table of these unemployment spells. A suspicious 'peaking' appears at multiples of six months. For a typical plot, see Figure 1, taken from Gorter and Hoogteijling (1990a).

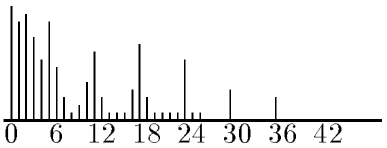

Figure 1. Typical plot of heaped data frequency table.

We suspect there has been some rounding of these data. The reported data seem 'heaped' on some months.

The SEP has been conducted from April 1984 onwards by Statistics Netherlands (CBS), resulting in a longitudinal cross-section of a representative sample of the Dutch population, among 5000 households, see [1]. The respondents were followed in time, and every six months, in April and October, they were asked, among other questions, whether they had found a job or not, and if so when. Keeping — for practical reasons — only those unemployment spells that started between April 1984 and October 1987 leads to a data set of 784 unemployment spell records. The first month is April 1984 and the last month is October 1987. Each spell is linked with a data record of 43 fields: a reported beginning date, a reported ending dates (in months), a censoring indicator (indicating whether the spell has been observed to end) and a list of covariates. In the SEP case (the problem at hand), reported durations are derived from ending and beginning month data. This suggests that heaping is here caused by rounding of beginning or ending dates of unemployment spells.

A frequency table of beginning dates indeed reveals a spiking on multiples of six months in the problem at hand; uncensored and right censored ending date frequency tables show similar features. The ending date frequency table shows a huge spike at the last month, but this is mainly due to the right censored ongoing spells. This suggests that modelling beginning and ending heaping is a more natural approach than modelling the heaping of durations directly. The idea is that *some* reported dates are true; *others* are rounded to the next or previous April or October. Such months (to which dates may be rounded) are called 'heaping months'. This concept of 'heaping months' is the basis of our heaping model. The methodology in the next section is therefore based on this approach. Note that even seemingly accurate durations, e.g. of eleven months, may occur from this rounding.

Different proposals have been put forward on how to cope with this phenomenon.

2 Heaping Model

In this section we introduce our heaping model in an informal way. For any duration there is a *true* beginning point, a *true* ending point, and also a *reported* beginning point, and a *reported* ending point.

A reported date coincides with a true date with some probability. However, it is also possible that the reported date is rounded forward or backward (with some restriction), i.e. a date is reported on a heaping month. The restriction we use in this paper is that any reported date may only be rounded to the nearest heaping month before the true date, or to the nearest heaping month after the true date. If such a date is rounded, then it is said to be 'heaped'.

If a true date is on a heaping month then its corresponding reported date is equal to this true date, i.e. it is not rounded.

The heaping months are fixed and known. For the problem at hand this seems reasonable, since they can be derived from the survey months: for the beginning months the set of heaping months consists of Aprils and Octobers, from October 1984 up to and including October 1986. Similarly, the set of heaping months for the uncensored ending dates consists of Marches and Septembers, from September 1985 up to and including March 1987.

3 Strategy

This section outlines the techniques needed for the construction of our heaping model. Heaping is a complex phenomenon in general. Formal definitions of heaped data are not needed here, but for a definition, see [7]. Heaping is interpreted in this paper as a kind of data coarsening, i.e. as a grouping of different kinds of data, so we consider this statistical problem as a coarse data problem.

Censoring is also a complex phenomenon in general, see for example [10]. Since the emphasis of this paper is on illustrating how to model heaping, we assume a simple censoring mechanism, see section 4.

Observe the following.

- If all variables in the model had been observed (i.e. true as well as reported durations), then (standard) maximum likelihood techniques would yield maximum likelihood estimates (MLE), along with variances, of the parameters of interest as well as of the heaping effect. However, this so-called 'full data' likelihood contains unknown, unobserved variables and therefore it cannot be calculated.

  The Expectation-Maximization (EM) algorithm provides a useful alternative for the computationally heavier method of integrating out all the unknowns. By EM the MLE is found without having ever to write down the 'actual' data full likelihood, under some regularity assumptions, see [4] for more on this topic. Therefore we will use EM to find this MLE.

- However, EM does not directly yield the variances of this MLE, because missing data increase the variances. see [12] and [11] for more on this topic.

  Jackknifing is a way to estimate the variances, see for example [14], and saves time with respect to bootstrapping, while standard large sample properties hold, t-values can be calculated etc. Therefore we will use jackknife for estimating the variances of the parameters.

- Standard survival analysis assumes no heaping and instead of the full data likelihood often the more simple Cox Partial Likelihood (PL) is used for inference on the parameter of interest. However, the problem with heaping is that it destroys the very special PL structure. The conditions that allow the use of PL techniques are not satisfied, see [3] for more on this topic.

  Now a significant contribution of Johansen [9] is that the ideas of PL are not at all necessary for estimating the parameter of interest. Johansen shows in his paper that the Cox maximum PL estimator of the regression parameters is equal to the MLE, and that the Breslow estimator of the baseline hazard is identically equal to the maximum likelihood estimator; see also subsection 4.5. Although using the full likelihood may at first glance look more complicated than using PL, in fact, we show that its implementation is not so difficult, see section 4.5. Therefore we will use profile likelihood for inference on the parameter of interest.

4 Implementation

4.1 Assumptions

The following assumptions on Cox regression are in order to compare the results of our heaping model with those of a standard Cox regression model without heaping, i.e. where the reported durations are considered true, see [6].

- All reported variables and covariates are observed correctly.
- Covariates are constant in time; for the problem at hand this implies that unemployment benefit and change of region do not affect people's attitudes and efforts.
- The true durations are independently distributed; for the problem at hand, this assumption seems reasonable. For each respondent only the first three unemployment spells during the survey period were included in the data set (most respondents lost their job at most once during the survey period).
- Cox's Proportional Hazards Model (PHM) describes the relation between true duration and covariates, see for example [10] for more on this topic.
- The baseline hazard is piecewise constant in each month; since the data we observe are all given in units of months this is not so much of a restriction on our model.

We use the following assumptions for our heaping model.

- The sets of heaping months are known in advance and independent of the data; for the problem at hand this seems quite reasonable, considering the survey structure, since the dates of the survey are exogeneously determined.
- Heaping only depends on the true point, whether it is on the according set of heaping months or not; this seems reasonable, if precisely reported variables are considered reliable.

- In case a true point is on the corresponding set of heaping months, it depends on the structure of the survey what decision rule to use. In case a true point is not on the according set of heaping months, its heaping is described by fixed, unknown probabilities.

- Censored ending dates are not heaped; for the problem at hand this assumption seems reasonable from the survey structure.

- Censoring does not depend on the parameters of interest, so it plays no role in the full data likelihood; for the problem at hand, however, various causes of censoring are present, e.g. end of survey, panel attrition or selectivity of nonresponse, so this assumption seems quite strong.

- The distribution of beginning dates is uniform; for the problem at hand this implies that the effects of season are neglected and therefore this assumption seems quite strong.

4.2 Notations

Let V denote the set of calendar months of the total survey period. Let H_b denote (for the beginning dates) the set of heaping months. For any $s \in V$ we denote the last heaping month before s by $t^{b-}(s) \equiv \max\{h \in H_b : h < s\}$, we denote the first heaping month after s by $t^{b+}(s) \equiv \min\{h \in H_b : h > s\}$ and we denote also $t^{b=}(s) \equiv \{s\}$. Denote by H_e (for the uncensored ending dates) the set of heaping months. For any $u \in V$, $t^{e-}(u), t^{e+}(u)$ and $t^{e=}(u)$ are defined similarly.

For each of the n spells we define the following variables regarding beginning and ending of the spell: $s_i \equiv$ true beginning date, $s_{i,r} \equiv$ reported beginning date, $u_i \equiv$ true ending date (if observed), $u_{i,r} \equiv$ reported ending date, $c_i \equiv$ last observed ending date (if i is right censored), $\delta_i \equiv$ censoring indicator (with value 1 if i's ending date is observed), and $x_i \equiv (s_i, s_{i,r}, u_i, u_{i,r}, c_i, \delta_i)$, $i = 1, \ldots, n$. For any true duration τ_i we have $\tau_i = u_i - s_i + 1$ if the ending date is observed, and $\tau_i \geq c_i - s_i + 1$ if the observation is right censored. Denote the covariate vector of item i by z_i. Let ϕ denote the vector of parameters.

Beginning heaping is described by its distribution given the true beginning dates:

$$\pi_b(\cdot \mid s_i) \equiv P(S_{i,r} = \cdot \mid S_i = s_i, z_i;\ \phi),$$

where an upper case letter denotes the random variable, and a lower case letter the corresponding realization (data). Similarly, uncensored ending date heaping is described by

$$\pi_e(\cdot \mid u_i) \equiv P(U_{i,r} = \cdot \mid s_i, \tau_i = \cdot - s_i + 1, z_i, \delta_i = 1;\ \phi).$$

Denote the censoring mechanism by

$$\pi_c(\cdot) \equiv P(U_{i,r} \geq \cdot \mid z_i, \delta_i = 0).$$

Denote the distribution of right censored and uncensored true durations by

$$\pi_d^{\geq}(\cdot) \equiv P(\tau \geq \cdot \mid z_i, \delta_i = 0;\ \phi)$$

and

$$\pi_d(\cdot) \equiv P(\tau = \cdot \mid z_i, \delta_i = 1;\ \phi),$$

respectively.

Finally, denote beginning date distribution by

$$\pi_s(\cdot) \equiv P(S_i = \cdot \mid z_i).$$

4.3 Stochastic Specification

- Heaping of beginning dates is specified as follows (cf. section 2).

  1. $s \in H_b$: s is correctly reported: $S_{i,r} = s$.
  2. $s \notin H_b$: s between two beginning heaping months may be rounded backward, forward or not at all: $S_{i,r} = t^{b-}$, $S_{i,r} = t^{b+}$ or $S_{i,r} = t^{b=}$:

  $$\pi_b(\cdot \mid s) = p_b^{1\{t^{b-}(s)=\cdot\}} \times q_b^{1\{t^{b+}(s)=\cdot\}} \times r_b^{1\{t^{b=}(s)=\cdot\}}, \tag{1}$$

  where p_b, q_b, r_b denote the probabilities for beginning heaping backward, forward or not at all, respectively, with the constraint $p_b + q_b + r_b = 1$, and where $1\{E\}$ is the indicator function of the event E.

- Heaping of uncensored ending dates is specified similarly, with corresponding parameters p_e, q_e, r_e, where $p_e + q_e + r_e = 1$.

- The uniform distribution of beginning dates implies $\pi_s(\cdot) = 1/n_V$, where n_V is the number of months.

- Independent censoring reduces $\pi_c(\cdot)$ to a constant in the full data likelihood.
- Cox's PHM describes the relation between true durations and covariates. Using nonparametric baseline hazard notation, $\lambda_0(\cdot) = \sum_{t \geq 0} \lambda_{0t} 1\{t = \cdot\}$, this relation can be written out explicitly as

  $$\pi_d^{\geq}(\cdot) = \exp(-\exp(\beta' z_i) \sum_{t > \cdot} \lambda_{0t});$$

  an expression for $\pi_d(\cdot)$ can be derived similarly.
- The data for the problem at hand are in discrete time, while our heaping model uses continuous time survival theory, see section 4.1. In continuous time survival theory the hazard rate follows directly from the integrated hazard, by differentiating the latter. However, for the discrete version this is not quite the same, see [10] for more on this topic. This discrepancy in the stochastic specification, see section 4.3, facilitates computation and is not expected to affect seriously the validity of our analysis.

4.4 Likelihood

In order to apply EM later on, we need to write down the full data likelihood using all the random variables in the model, whether these are observed or not. Independence of the durations allows the 'full data' likelihood to be written as a product of individual likelihoods. If we accept the previously stated assumptions, then the contribution to the likelihood of an uncensored observation is

$$P(s_i, s_{i,r}, \tau_i, u_i, u_{i,r} \mid z_i;\ \phi) = \pi_s(s_i) \times \pi_b(s_{i,r} \mid s_i) \times \pi_d(\tau_i) \times \pi_e(u_{i,r} \mid u_i)$$

and that of a right censored observation is

$$P(s_i, s_{i,r}, \tau_i, c_i, \mid z_i;\ \phi) = \pi_s(s_i) \times \pi_b(s_{i,r} \mid s_i) \times \pi_d^{\geq}(\tau_i) \times \pi_c(c_i).$$

The full data log likelihood reduces to

$$\sum_{i=1}^{n} \log \pi_b(s_{i,r} \mid s_i) + \sum_{i=1}^{n} (1 - \delta_i) \log \pi_d^{\geq}(\tau_i) +$$

$$\sum_{i=1}^{n} \delta_i \log \pi_e(u_{i,r} \mid u_i) + \sum_{i=1}^{n} \delta_i \log \pi_d(\tau_i), \qquad (2)$$

up to a constant, since π_s and π_c do not depend on ϕ.

4.5 Estimation

The principle of EM is elementary. However, application of EM on (2) is elaborate and involves a lot of careful bookkeeping but is essentially routine. Some nice features of the practical implementation are worth mentioning.

- Consider the following interpretation, under the assumption of independent durations. Denote by J the space of all admissible realizations of a duration. Denote the contribution to the likelihood of realization j by $f_j(\cdot)$. Denote the event 'observation i has realization j' by $A_j(i)$. Now
$$\sum_{j \in J} \sum_i 1\{A_j(i)\} f_j(\phi) \tag{3}$$
denotes the full data log likelihood. Taking conditional expectations on (3), given the incomplete data F and given a parameter value $\phi^{(0)}$ is equivalent to replacing the indicators in (3) by appropriate conditional probabilities. The resulting expression,
$$\sum_j \sum_i P(A_j(i) \mid F, \phi^{(0)}) f_j(\phi),$$
can be maximized over ϕ for each M step. Some realizations may be taken together, thanks to the assumptions for the problem at hand. The effects of beginning heaping, ending heaping, censored and uncensored durations are isolated, see (2). Thanks to heaping specification (1), the heaping effect splits up into six simpler sums.

- Rewriting the full data log likelihood (2) in terms of nonparametric hazards, omitting those terms that do not vary with λ_0 and maximizing with respect to λ_{0t} for given β leads to the so-called 'Breslow estimator'
$$\hat{\lambda}_{0t} \; = \; \frac{A_t}{E_t(\beta)} \tag{4}$$
as an estimator of the baseline hazard, with $A_t \equiv \sum_{i=1}^n \delta_i a_{i,t}$ and $E_t \equiv \sum_{i=1}^n b_{i,t} \exp(\beta' z_i)$ where $a_{i,t} \equiv 1\{t = \tau_i\}$ and $b_{i,t} \equiv 1\{t \leq \tau_i\}$.

  However, (4) can not be calculated, because β is unknown. As an estimator of β we use the well-known Cox PL estimator. Thanks to Johansen's result this yields the MLE.

  For the EM algorithm we apply the analogue. Omitting from (2) those terms that do not vary with λ_0 leads to (4), where $a_{i,t} \equiv P(\tau_i = t \mid$

$F_i;\ \phi)$ and $b_{i,t} \equiv P(\tau_i \geq t \mid F_i;\ \phi)$, where F_i represents the incomplete data on i.

Following the same procedure we obtain in each E step estimators of β and λ, which can be maximized in the following M step.

- In each E step, the beginning heaping parameters have to be estimated simultaneously, because of their interdependence. Ending heaping is similarly estimated, and by the censoring assumptions it is correct to ignore right censored ending dates for the estimation of the ending heaping parameters. Thanks to (1) we may use in each M step the classic trinomial MLE for estimating the proportion of the beginning date data that are heaped backward, forward and not at all.

- From (2) it is easy to see that the full data log likelihood of our heaping model has the so-called regular exponential form, see [4] for more on this topic. So we expect no problems on the choice of initial values for the parameter vector nor any problems in finding a global maximum using standard methods.

5 Results

The results of implementing our heaping model on the data can be read from table 1. The nine most important covariates from Gorter and Hoogteijling's report have been used for our heaping model. The first and third columns give the estimates of the Cox and heaping model, respectively. The variances of the standard model are obtained by standard methods and can be read from the second column. The fourth column is the result of a k-sample jackknife (where $k = 28$). The standard errors from the heaping model are better approximations than those from the standard model. We have used jackknife since using the standard errors from the last EM step is incorrect. Comparing the results of our heaping model and the model without heaping leads to the following conclusions.

- Heaping exists. All heaping parameters are significantly different from zero.

- For both models, the factors having the most important influence on unemployment duration are the same, although the coefficients slightly differ.

- Respondents tend more to round backward than forward; this holds for beginning dates as well as for ending dates.

Table 1: Comparison of estimates

SEP unemployment data

756 observations, 9 covariates

	Cox regression		Heaping model	
variable	coefficient	st.error	coefficient	st.error
age	−0.017	0.005	−0.023	0.030
northeast	−0.391	0.093	−0.329	0.146
midvoc	0.146	0.091	0.171	0.229
foreign	−0.272	0.184	−0.274	0.840
bigtown	−0.318	0.110	−0.338	0.214
before	−0.566	0.096	−0.539	0.215
sex	0.014	0.097	0.065	0.292
earner	0.024	0.113	−0.032	0.368
married	0.016	0.125	0.156	0.389
p_b			0.209	0.018
q_b			0.095	0.012
r_b			0.696	0.025
p_e			0.166	0.018
q_e			0.132	0.017
r_e			0.702	0.025

Meaning of variables:

age: age in years
northeast:1 if indvidual lives in the North or East of the country
midvoc: 1 if level of education is middle vocational level
foreign: 1 if individual has a foreign nationality
bigtown: 1 if individual lives in a large city (more than 100,000 inhabitants)
before: 1 if individual has been unemployed before
sex: 1 if individual is male
earner: 1 if individual is the main bread winnner of his family
married: 1 if individual is married

The nonparametric optimal baseline hazard for our heaping model is somewhat jumpy, however. The small amount of data might have caused these jumps.

Finally, a remark on computing time of the jackknife. On a 386 SX-20, computing took approximately 13 days, on a 486 DX-33, it took 4 days and on a parallel computer with 16 processors it took 13 minutes [15]. Clearly, parallel computing can be of great value for jackknife computations.

6 Remarks

We conclude with the following remarks.

- Right censored durations usually give less information than fully observed durations. If censoring is independent of heaping, then right censored durations may give more information about the ending date than fully observed durations whose ending dates are on a 'suspicious' month.

- The heaping model described in this paper is based on a separate modelling of beginning and ending date heaping. We believe it is indeed better (more realistic) to model beginning heaping and ending heaping separately, instead of modelling the heaping of durations, even when no information is available about the beginning dates and ending dates.

- We emphasize that the assumptions in this paper are only intended to keep the implementation of the model transparent; extensions of the model and relaxation of the assumptions are straightforward, but left for future research.

It is our opinion that we have established a reasonable balance between the degree of complexity of the model and the degree of realism of the assumptions, at least for the problem at hand. We hope to have illustrated a reasonable heaping model and the use of profile likelihood.

References

[1] CBS, 1991. Socio-economic panel survey, contents, concept and organization, 1991, SDU/Publishers, the Hague.

[2] Cox, D.R. and D.V. Hinkley, 1978. Problems and solutions in theoretical statistics (Chapman & Hall).

data. Chapman & Hall, London.

[3] Cox, D.R., 1975. Partial likelihood. Biometrika 62 p. 269.

[4] Dempster, A.P., N.M. Laird and D.B. Rubin, 1976. Maximum Likelihood from Incomplete Data via the EM algorithm. Journal of Royal Statistical Society B 19, pp.1-38.

[5] Gorter D. and E. Hoogteijling, 1990a. Determinanten van werkloosheidsduur, 1985-1987. Internal report (Department of Statistical Methods, Statistics Netherlands, Voorburg).

[6] Gorter D. and E. Hoogteijling, 1990b. Duur van het zoeken naar werk. Report (Department of Statistical Methods, Statistics Netherlands, Voorburg).

[7] Heijtjan, D.F. and D.B. Rubin, 1991. Ignorability and coarse data. Annals of Statistics 19, pp.2244-53.

[8] Heijtjan, D.F. and D.B. Rubin, 1990. Inference from coarse data via multiple imputation with application to age heaping. Journal of the American Statistics Association 85, pp.304-14.

[9] Johansen, S., 1983. An extension of Cox's regression model. International Statistical Review, 51, pp. 165-174.

[10] Kalbfleisch, J.G., and R.L. Prentice, 1980. The statistical analysis of failure time data, John Wiley, New York.

[11] Louis, T.A., 1982. Finding the observed information matrix when using the EM algorithm. J. R. Statist. Soc. B 44, No. 2, pp.226-233.

[12] Meilijson, I., 1989. A fast improvement to the EM algorithm on its own terms. Journal of the Royal Statistical Society B 51, pp.127-38.

[13] Torelli, N. and U. Trivellato, 1993. Modelling inaccuracies in job-search duration data. Journal of Econometrics 59, pp.185-211.

[14] Wolter, K.M., 1985. Introduction to variance estimation. Springer-Verlag, New York Inc.

[15] Zwemmer, W., 1995. Parallel computing in statistics. Report (Department of Statistical Methods, Statistics Netherlands, Voorburg).