Wolff, Augustin:

Heaping and its Consequences for Duration Analysis

Sonderforschungsbereich 386, Paper 203 (2000)

Online unter: http://epub.ub.uni-muenchen.de/

Heaping and its Consequences for Duration Analysis

Joachim Wolff[*] Thomas Augustin[†]

Abstract

This paper analyses the consequences of heaping in duration models. Heaping is a specific form of response error typical to retrospectively collected labor force status data. Respondents round-off the spell length, when duration data is collected by episode-based questionnaires. Calendar-based questionnaires instead may lead to abnormal concentrations of the start and/or end of spells at specific calendar months. The investigation concentrates on this latter type of heaping, which Kraus and Steiner [1995] identified for the unemployment spell data from the German Socio-Economic Panel (GSOEP). In the special case of an exponential model heaping with a symmetric zero-mean measurement error does not bias the parameter estimate. In the Weibull model with duration dependence, however, it is proven that even such a symmetric heaping would lead to inconsistent estimation. We discuss the bias for general heaping patterns and derive from this a proposal for bias correction. In a number of simulation studies we check the theoretical results. The Monte Carlo simulations also show that an amount of heaping, that characterizes the GSOEP-West does not lead to considerably biased parameter estimates of a Weibull model. However, it clearly leads to spurious seasonal effects. Finally, some directions of future work are indicated.

[*] Seminar für empirische Wirtschaftsforschung, Ludwig-Maximilians University of Munich, Ludwigstr. 28 RG, D 80539 Munich, Joachim.Wolff@lrz-muenchen.de

[†] Seminar für Ökonometrie und Statistik, Ludwig-Maximilians University of Munich, Akademiestr.1/I, D 80799 Munich, thomas@stat.uni-muenchen.de

1 Introduction

Retrospectively collected information of household surveys is likely to be characterized by errors of recall. In particular, when respondents are asked to provide event histories of labor force states or duration of some labor force state, substantial response errors may emerge. 'Heaping' or rounding-off is a particular form of such errors. There are two types of heaping depending on the design of the questionnaire. One may emerge from calendar-based questionnaires as for example in the German Socio-Economic Panel (GSOEP). Its calendarium requires respondents to tick (at least) one out of usually 12 possible labor force states for each month of the calendar year prior to the interview. Heaping implies that respondents round-off or use rules of thumb when reporting the calendar date of a transition from one labor force state to another. Consequently, we would find abnormal concentrations of entry and exit months. E.g., one rule of thumb could be that some respondents who become unemployed in February or March just report the start of their spell as January in the same year. Kraus and Steiner [1995] revealed this heaping pattern for unemployment duration data of the GSOEP-West.[1] Another rule of thumb emerges in duration data that stem from responses to episode-based questionnaires. When respondents recall the entire duration of unemployment, abnormal concentrations of spell lengths at multiples of six or twelve months arise. Torelli and Trivellato [1993] found that unemployment spells from Italian Labor Force Survey data are subject to this heaping pattern.

Both the study of Torelli and Trivellato as well as that of Kraus and Steiner proposed how to adjust econometric duration models, in order to achieve consistent estimates of the parameters in the presence of heaping. Kraus and Steiner who dealt with heaping in calendar-based questionnaires provided no information about the general consequences for duration analysis of the type of heaping that they identified. Torelli and Trivellato analyzed by Monte Carlo simulations possible effects of heaping on the parameter estimates of an exponential model, a Weibull and a log-logistic model. For the last two distributions the specific heaping pattern lead to parameter estimates that

[1] There is some evidence that the east-German unemployment spell data of the GSOEP are also characterized by heaping (Wolff [1998]). However, compared with the findings of Kraus and Steiner for the west-German unemployment spell data, this amount of heaping is low.

differed substantially from the parameters of the data generating process (DGP).

It is a well-known result, that a zero-mean measurement error of the dependent variable that is independent of the covariates does not lead to biased parameter estimates in linear regression models. However, response errors like heaping, even if they leave the average duration of spells unaltered, may adversely influence the estimation results of standard duration models. Applying such highly non-linear models to heaped duration data without an appropriate correction may yield inconsistent and less precise parameter estimates. The aim of this paper is to show by simulation studies and theoretical considerations whether and when specific forms of heaping have such consequences. This is particularly important to know if there is no sufficient outside information to determine the exact heaping pattern. We concentrate on heaping in calendar-based questionnaires. The simulation studies are based on an amount of heaping which was characterized as typical for the GSOEP-West.

The paper is structured as follows: Section two discusses previous work on heaping in labor force surveys. Section three shows whether heaping, even as a symmetric zero-mean measurement error of the spell length, should alter parameter estimates if the true spell lengths are drawn from the exponential or the Weibull distribution. Section four presents results from Monte Carlo simulations to highlight the consequences of heaping for parameter estimates. Section five summarizes our very preliminary results and indicates extensions of this work.

2 Previous work

Torelli and Trivellato [1993] studied recall errors in unemployment duration data. They showed that there is a strong presence of rounding-off effects in job-search duration data from a small matched sample of the Italian Labor Force Survey for Lombardy.[2] They matched the information of two consec-

[2] The Italian Labor Force Survey is a quarterly survey of rotating panel design. Each family is interviewed for two consecutive surveys and then dropped for two surveys and interviewed again for two final surveys. People who consider themselves as unemployed job-seekers are asked for how many months they have already been looking for a job.

utive surveys and combined responses to the retrospective question on the length of an unemployment spell in progress at the date of the second survey.

The authors derived their results from a sample that combined this LFS over the first and second quarter of 1986. It consists of 678 individuals aged between 14 and 29 years who are unemployed in the first survey. This sample provided evidence for abnormal concentrations of unemployment duration at certain values (heaping). The percentage distribution of the reported spell length showed spikes at multiples of six months and very strong spikes at multiples of 12 months for men and women, respectively. It is thus clear that respondents used a rule of thumb to report the length of their unemployment spells: If their duration of unemployment was close to multiples of six or 12 months, they rounded-off the duration to these values.

A heaping pattern as above may be a considerable problem for studies of unemployment duration. The authors pointed out that there are true behavioral reasons that could lead to such spikes at similar spell lengths. They may result from the time until the exhaustion of unemployment benefits or from seasonal effects. This kind of recall error implies an identification problem for studies of unemployment duration.

The authors developed a model that yields consistent estimates of a continuous-time parametric duration model in the presence of heaping. Let T be a continuous random variable of duration of unemployment, with probability density $f(t, \theta)$, where the parameter vector θ is unknown. The authors assume that completed spell durations, ${}_HT_i$, are observed for $i = 1, ..., n$ individuals. The observed duration is subject to heaping and is related to the true spell length by:

$$ {}_HT_i = T_i + K_i \cdot Y_i \tag{1}$$

where $K_i = H_{(m)} - T_i$, $H_{(m)} \in \mathcal{H}$ and such that $|H_{(m)} - T_i|$ is minimum. $\mathcal{H}$ is the set of heaped values $h_{(m)}$ $(m = 1, \ldots, M)$ which are known and arranged in increasing order. Y_i is a Bernoulli random variable, which equals one for heaped values and zero otherwise. Let further $p(t_i) = G(t_i, \gamma)$ be the probability that $Y_i = 1$, where $G(\cdot, \gamma)$ is a parametric function, known up to the finite parameter vector γ, mapping $t_i \in I\!R^+$ onto $[0, 1]$. Thus, $G(t_i, \gamma)$, the heaping function, describes the probability of heaping as a function of the true spell length. The inferential problem is to estimate θ and γ from the observed data, that are subject to heaping and given a known set of heaped values. The likelihood of a random sample is then

$$\prod_{i \in I} [f(t_i, \theta) \cdot (1 - G(t_i, \gamma))] \cdot \prod_{j \in J} \left[\int_{_lt_j}^{_ut_j} f(z, \theta) \cdot G(z, \gamma) \, dz \right] \qquad (2)$$

I is the set of nonheaped observations, where $_Ht_i = t_i$, since $_Ht_i \neq h(m)$. J instead is the set of heaped observations taking on values that are element of $\mathcal{H}$. $_lt_j$ and $_ut_j$ are the lower and upper limits of the interval of the jth heaped spell length. With $G(t_i, \gamma)$ being constant over the intervals $[_lt_j, _ut_j]$, the likelihood becomes

$$\prod_{i \in I} [f(t_i, \theta) \cdot (1 - G(t_i, \gamma))] \cdot \prod_{j \in J} [F(_ut_j, \theta) - F(_lt_j, \theta)] \cdot G(_Ht_j, \gamma) \qquad (3)$$

$F(\cdot, \theta)$ is the distribution function of T. By factorising the latter likelihood, one component involving only the parameter vector of the duration model and one involving the parameter vector of the heaping function emerge. So, the parameters could be consistently estimated disregarding the heaping process.

In order to reveal the effects of heaping, the authors generated duration data that stem from the exponential, the Weibull and the log-logistic distribution. They did not introduce covariates. The resulting spell lengths were altered by a heaping pattern that is close to the one that they revealed from the Italian Labor Force Survey. The exponential distribution function was chosen as the heaping function. With this data, they estimated the parameters using the likelihood functions (2), (3) and the likelihood of the standard duration models. Their Monte Carlo simulations revealed the following: Ignoring heaping leaves parameter estimates of the exponential model still close to their true values. However, this does not apply to the Weibull and even less to the log-logistic duration models. In these cases parameter estimates were considerably apart from their true values, where the bias appears to be positively related to the amount of heaping. Next they concluded that a crude way of handling heaping according to the likelihood function (3), did not improve parameter estimates considerably. In the case of a large sample size ($n = 500$), it was even dominated by duration models that ignore heaping. Applying the likelihood function (2) to the data instead yielded parameter estimates close to their true values.

Kraus and Steiner [1995] analyzed unemployment spells drawn from the GSOEP-West with respect to heaping. The retrospective labor force status questionnaire of the GSOEP relies on a calendar-based design. The respondent has to code his/her labor force status separately for each month of the previous calendar year. Heaping may hence occur as abnormal concentrations for entry and exit at certain calendar months; i.e., people date their entry into or exit out of unemployment too frequently and incorrectly at certain calendar months and not frequently enough at the months that are close to them. The study of Kraus and Steiner identified such abnormal concentrations. They analyzed an inflow sample to registered unemployment of the GSOEP-West from January 1983 to December 1991. People who worked in the construction sector prior to unemployment were discarded from this sample of uncensored and right-censored unemployment spells. These people have been excluded since their exit behaviour from unemployment follows a strong seasonal pattern which may interfere with heaping effects. The definition of registered unemployment in the GSOEP and the register data of the German Federal Labor Office is the same. Thus, for west-Germany, the authors compared the aggregate monthly inflow rate into and outflow rate from registered unemployment calculated for both data sources[3] over the observation period. The striking findings were two abnormal concentrations:

- The gross inflow rates into unemployment in January as estimated by the GSOEP-West are roughly twice as high as their population values over most of the observation period. In contrast, the inflow rates in February and March are often considerably smaller than the corresponding population values.
- Compared with their population values, the outflow rates of the GSOEP-West in December are usually four times higher, while those in the neighboring months October and November tend to be somewhat lower.

These two heaping patterns imply that spell lengths are reported as too long for spells that start in the first quarter of the year and/or terminate during its last quarter. With standard econometric duration models negative duration dependence may be overstated and biased coefficients for seasonal effects may be the outcome.

[3] The authors accounted for sample attrition of the GSOEP-West relative to the register data by appropriate weighting factors calculated on a yearly bases.

Kraus and Steiner did not carry out simulations to reveal the consequences of the identified heaping pattern for standard duration models. Nor did they deal with this topic from a theoretical perspective. They adjusted the model of Torelli and Trivellato, in order to incorporate the specific heaping pattern, that they identified for the unemployment spells of the GSOEP-West into a discrete-time proportional hazards framework.

This model was then applied to the unemployment spells of the GSOEP-West, in order to estimate simultaneously the parameters of the duration model and those of a parametrically specified heaping function. However, they ran into numerical problems, due to a small number of observations for some groups of spell length. To resolve this problem, they estimated the heaping probabilities using outside information[4]. Kraus and Steiner proceeded then by comparing the estimates of several duration models that take heaping into account or completely ignore it. All models included a standard set of covariates[5] as well as a baseline hazard.

Let us summarize their results. The authors found hardly any difference between the estimated parameters of a proportional hazard rate model with a flexible baseline hazard with and without their correction for heaping. Next, regardless of whether heaping is incorporated into the likelihood, the coefficients of the hazard model with a parametric baseline hazard[6] are by and large the same as for those of the hazard model with a flexible baseline hazard. Naturally, the baseline hazard was smoother. Yet, the estimates of a number of covariates of a proportional hazard rate model that accounts for heaping by including dummies for January and December yielded quite different coefficients of some covariates and of the baseline hazard. However, this may reveal a simple omitted variable bias of the models that excluded these dummies and may not be due to heaping.

[4] The authors again used information about the monthly population inflow and outflow as published by the German Federal Labor Office.

[5] These included age, foreigner, disability, marital status, education, household income and the regional unemployment rate.

[6] They chose a logit transformation of time, namely $\exp(t + t^2 + 1/t) \cdot (1 + \exp(t + t^2 + 1/t))^{-1}$, as their baseline hazard function.

3 A simplified theoretical look at the Weibull model

3.1 Preliminaries

In this section we take a simplified look at non-corrected maximum likelihood estimation under heaping in the Weibull model. A variable T_i is Weibull distributed with parameters λ and α if its density has the form

$$f_{T_i}(t_i) = \alpha \cdot \lambda \cdot (\lambda \cdot t_i)^{\alpha - 1} \cdot \exp\left(-\left(\lambda \cdot t_i\right)^{\alpha}\right) .$$

The hazard rate

$$\mathrm{r}_{T_i}(t_i|\lambda, \alpha) = \alpha \cdot \lambda^{\alpha} \cdot t_i^{\alpha-1}, \qquad \alpha > 0\,, \tag{4}$$

depends on time by a power of t_i, its monotonicity remains unchanged over time. The direction of time dependency is governed by the *duration dependence parameter* α, providing easy ways to test the hypothesis of increasing or decreasing risk. $\alpha < 1$ leads to monotonely decreasing hazard, while $\alpha > 1$ corresponds to monotonely increasing hazard containing the Rayleigh distribution with linear hazard ($\alpha = 2$). The special case of constant hazard ($\alpha = 1$) is the exponential model.

If one introduces covariates x_i one usually parameterizes

$$\lambda = \exp(-x_i'\beta)\,. \tag{5}$$

Then the Weibull model is a special accelerated failure time model (Kalbfleisch & Prentice [1980, p.34]), i.e. a model of the form

$$\ln T_i = x_i'\beta + \sigma \cdot \epsilon_i$$

with ϵ_i independently and identically distributed. Here, to obtain the Weibull model, ϵ_i is taken as extreme-value distributed and σ is set equal to α^{-1}.

With the exception of Section 4.3 and Appendix B we only consider the special case of a homogeneous sample without individual covariates, i.e $x_i \equiv 1$, $i = 1, \ldots, n$. Then the vector β reduces to a scalar β_0. For our study we nevertheless use the reparameterized form (5) instead of λ itself, because it is the usual way of looking at the Weibull model in economics, it allows to use standard software for Weibull regression and it also should provide a basis for a generalization of our results to covariates.

Estimation of the unknown parameters α and β_0 is typically done relying on the maximum likelihood principle. The corresponding score equations are

$$\begin{aligned} \sum_{i=1}^{n} (1 - T_i^{\alpha} \cdot \exp(-\alpha \cdot \beta_0)) &= 0 \qquad (6) \\ \sum_{i=1}^{n} (1 + \alpha \cdot (\ln(T_i) - \beta_0) \cdot (1 - T_i^{\alpha} \cdot \exp(-\alpha \cdot \beta_0))) &= 0\,. \qquad (7) \end{aligned}$$

In general (6) and (7) cannot be solved analytically and some numerical procedure is needed. However, if one treats α as fixed, the first line yields an explicit solution. One obtains

$$\hat{\beta}_0 = \frac{1}{\alpha} \cdot \ln\left(\frac{\sum_{i=1}^{n} T_i^{\alpha}}{n}\right) \qquad (8)$$

as the maximum likelihood estimate of β_0 for known α. Since the regularity conditions for applying the usual maximum likelihood asymptotics are satisfied here, $\hat{\beta}_0$ is consistent and asymptotically normal. Evidently, such statements tacitly assume that the realizations of the T_i's can be precisely observed. If only inexact measurements, like heaped data, are available, additional considerations are needed.

3.2 Heaping and the heaped maximum likelihood estimates

To formally introduce a heaping mechanism typical for calendar-based questionnaires we assume that every spell may be heaped with a certain probability which is assumed to be independent of the covariates and the spell-length itself. Denote for some sufficiently large q by $\nu^{(l)}$ the probability that a spell is prolonged by l units, $l = 1, \ldots, q$, and by $\delta^{(l)}$ the probability that the spell is shortened by l units. Further assume $\xi := \sum_{l=1}^{q}(\nu^{(l)} + \delta^{(l)}) < 1$. Therefore, instead of the 'true' duration times $T_1, \ldots, T_i, \ldots, T_n$, one observes the *heaped duration times* $T_1^*, \ldots, T_i^*, \ldots, T_n^*$ with

$$T_i^* = \begin{cases} T_i + q & & \nu^{(q)} \\ \vdots & & \vdots \\ T_i + 1 & & \nu^{(1)} \\ T_i & \text{with probability} & 1 - \xi \\ T_i - 1 & & \delta^{(1)} \\ \vdots & & \vdots \\ T_i - q & & \delta^{(q)} . \end{cases} \tag{9}$$

For some applications it makes sense to allow the heaping probabilities to depend on the month of entry. As shown in Appendix A3, assuming independence of the duration and the entry month, the behaviour of the estimates studied below does only depend on the marginal distribution of the heaping probabilities. Therefore, after calculating the marginal heaping probabilities, we can proceed without loss of generality with the model described in (9).

By plugging in the heaped duration times $T_1^*, \ldots, T_i^*, \ldots, T_n^*$ in (8) we have the *heaped (or naive) maximum likelihood estimate*[7]

$$\hat{\beta}_0^* = \frac{1}{\alpha} \cdot \ln \left(\frac{\sum_{i=1}^n (T_i^*)^\alpha}{n} \right) . \tag{10}$$

Following Kraus and Steiner [1995], for the GSOEP one has good reasons to assume $\delta^{(l)} = 0$, $l = 1, \ldots, q$. However, in general, $\delta^{(l)}$ may be non-zero for some l. Then by the heaping mechanism considered some of the data T_i^* formally may become negative. Depending on the concrete design of the questionnaire this may often be unrealistic, one simply then would not have recorded these spells. In this case the heaped (or naive) maximum likelihood estimate would have the form

$$\hat{\beta}_0^{**} = \frac{1}{\alpha} \cdot \ln \left(\frac{\sum_{i=1}^n (T_i^{**})^\alpha}{K} \right) \tag{11}$$

with

$$T_i^{**} = \max\left(0, T_i^*\right)$$

[7] We tacitly assume that $\hat{\beta}_0^*$ is well defined. This is always the case for $\alpha = 2, 4, 6, \ldots$ or if $\delta^{(l)}$ equals zero for every $l = 1, \ldots, q$.

and K as the number of spells with $T_i^{**} > 0$.

In general $\hat{\beta}_0^*$ and $\hat{\beta}_0^{**}$ are not the maximum likelihood estimates of β_0 with respect to the distribution of the observable variables T_i^* and T_i^{**}. So consistency can no longer be taken for granted, the estimates may be biased. To check this a first exploration of the behaviour of $\hat{\beta}_0^*$ and of $\hat{\beta}_0^{**}$ under the sample size growing to infinity will be performed.

3.3 Bias analysis

To obtain a first impression of the asymptotic properties of $\hat{\beta}_0^*$ and of $\hat{\beta}_0^{**}$ we restrict our attention to two important special cases, namely the exponential distribution ($\alpha = 1$) and the Rayleigh distribution ($\alpha = 2$).[8] The asymptotic bias can be given in a closed form:

Proposition 3.1 *Consider a heaping mechanism as described in (9) and assume* $\delta^{(1)} \ldots, \delta^{(q)}$ *and* $\nu^{(1)}, \ldots, \nu^{(q)}$ *to be such that the heaped maximum likelihood estimate* $\hat{\beta}_0^*$ *in (10) is well-defined*[9]. *Then the following holds for the probability limit* $\operatorname*{plim}_{n\to\infty} \hat{\beta}_0^*$*:*

a) (Exponential case). If $\alpha = 1$ *and*[10] $\frac{\sum_{l=1}^{q}\left(\nu^{(l)}-\delta^{(l)}\right)\cdot l}{\exp(\beta_0)} > -1$ *then*

$$\operatorname*{plim}_{n\to\infty} \hat{\beta}_0^* - \beta_0 = \ln\left(1 + \frac{\sum_{l=1}^{q}\left(\nu^{(l)} - \delta^{(l)}\right) \cdot l}{\exp(\beta_0)}\right). \tag{12}$$

[8] By using the General Binomial theorem, the considerations given below can be easily transferred to arbitrary even values of $\alpha \in I\!N$ and can be, assuming well-definiteness, extended to odd values of α. For other values of α the procedure used leads to some trouble, if $\delta^{(l)} > 0$ for some l.

[9] This is the case for any realistic constellation.

[10] By their dependence on the unknown β_0 these additional conditions may be sometimes tricky. But note that it is always satisfied in the case of symmetric heaping as well as for the positively biased one-sided heaping pattern observed in the GSOEP.

b) (Rayleigh distribution) If $\alpha = 2$ and

$$\frac{\sum_{l=1}^{q} \left(\nu^{(l)} + \delta^{(l)}\right) \cdot l^2}{\exp(2\beta_0)} + \frac{\sqrt{\pi} \cdot \sum_{l=1}^{q} \left(\nu^{(l)} - \delta^{(l)}\right) \cdot l}{\exp(\beta_0)} > -1$$

then

$$\operatorname*{plim}_{n\to\infty} \hat{\beta}_0^* - \beta_0 = \frac{1}{2} \cdot \ln \left(1 + \frac{\sum_{l=1}^{q} \left(\nu^{(l)} + \delta^{(l)}\right) \cdot l^2}{\exp(2\beta_0)} + \frac{\sqrt{\pi} \cdot \sum_{l=1}^{q} \left(\nu^{(l)} - \delta^{(l)}\right) \cdot l}{\exp(\beta_0)} \right) . \tag{13}$$

◇

Proof: See Appendix A1.

The bias grows flatter by the logarithmic function. Note further that the bias is inversely proportional to β_0, i.e. the longer the average spell-length, the smaller, ceteris paribus, is the bias. This seems quite plausible, as the error becomes smaller relative to the average spell length. The next subsection will show that (12) and (13) also provide a proposal for *bias correction* and consistent estimation of β_0.

Before discussing this, we briefly want to look at the estimate $\hat{\beta}_0^{**}$ as defined in (11) and the behaviour of both estimates in some special cases.

Since, by construction, $T_i^{**} \geq T_i^*$, and therefore

$$\hat{\beta}_0^{**} \geq \hat{\beta}_0^* ,$$

Proposition 3.1 provides immediately a lower bound for the bias of the estimate $\hat{\beta}_0^{**}$.

Corollary 3.2 *In the situation of Proposition 3.1 formulae (12) and (13) remain valid, if one replaces $\hat{\beta}_0^*$ by $\hat{\beta}_0^{**}$ and equalities by the relations 'greater or equal'.* ◇

Returning to $\hat{\beta}_0^*$, two extreme cases may be of special interest. The first one is the constellation where the heaping is *one-sided* in the sense that there is no heaping downwards but only a heaping upwards (or vice versa). This is the type of heaping Kraus and Steiner [1995] found for the GSOEP. Then we have $\beta_0^* \equiv \beta_0^{**}$ and $\delta^{(l)} \equiv 0$ for all l in the formulae above.

The second one is the *symmetric situation*, where, for every l, the proportion of the spells prolonged by l and the proportion of the spells shortened by l is the same. Note that, if the hazard rate is not constant, this 'averaging out' may nevertheless result in a bias, which, however, typically will not be very strong.

Corollary 3.3 *If in the situation of Proposition 3.1 the heaping is* symmetric*, i.e.* $\nu^{(l)} = \delta^{(l)}$*, for all* $l = 1, \dots, q$*, then* $\hat{\beta}_0^*$ *is consistent in the case of the exponential distribution (*$\alpha = 1$*), but inconsistent in the case of a Rayleigh distribution (*$\alpha = 2$*).* ◇

In the situation of Corollary 3.3 one regularly has $\hat{\beta}_0^{**} > \hat{\beta}_0^*$. Therefore, even in the exponential case a small bias can be expected using the estimate β_0^{**} based on putting negative values of T_i^* to zero.

The Torelli and Trivellato [1993] case is actually nearly symmetric. So Corollary 3.3 confirms for our heaping pattern Torelli's and Trivellato's observations discussed in Section two: In an exponential model the bias from ignoring symmetric heaping is negligible, while some care has to be taken in the Weibull model with duration dependence.

3.4 Bias Correction

(12) and (13) can be explicitly solved for β_0. If one knows the heaping probabilities $\delta^{(l)}$ and $\nu^{(l)}$, $l = 1, \dots, q$, for instance by external data, then these quantities may be used to obtain an improved estimate which has smaller bias than $\hat{\beta}_0^{**}$. Moreover, if the realizations of T_i^* are available, and therefore $\hat{\beta}_0^*$ can be calculated, even a consistent estimate for β_0 can be deduced. Therefore, in particular, in the heaping constellation noticed for the SOEP consistent estimation proves possible.

Proposition 3.4 *Consider the situation of Proposition 3.1.*

1. *If* $\alpha = 1$ *and* $\frac{\sum_{l=1}^{q}\left(\nu^{(l)}-\delta^{(l)}\right)\cdot l}{\exp(\beta_0)} > -1$ *then*

$$\widehat{{}^{(1)}\beta_0} := \ln\left(\exp\left(\hat{\beta}_0^*\right) - \sum_{l=1}^{q}\left(\nu^{(l)}-\delta^{(l)}\right)\cdot l\right)$$

estimates β_0 *consistently.*

2. *If* $\alpha = 2$ *and*

$$\sqrt{\pi}\cdot\sum_{l=1}^{q}(\nu^{(l)}-\delta^{(l)})\cdot l + \exp(\beta_0) \geq 0$$

then

$$\begin{aligned}\widehat{{}^{(2)}\beta_0} \quad := \quad & \ln\left(\frac{1}{2}\left[\sum_{l=1}^{q}\sqrt{\pi}\left(\delta^{(l)}-\nu^{(l)}\right)\cdot l + \right.\right. \\ & \left.\left. + \sqrt{\pi\cdot\left(\sum_{l=1}^{q}\left(\nu^{(l)}-\delta^{(l)}\right)\cdot l\right)^2 - 4\cdot\left(\sum_{l=1}^{q}\left(\nu^{(l)}+\delta^{(l)}\right)\cdot l^2 - \exp\left(2\hat{\beta}_0^*\right)\right)}\;\right]\right)\end{aligned}$$

estimates β_0 *consistently.* ◇

Proof: See Appendix A2.

Evidently, the results gained by these considerations are of course quite preliminary. For instance, sharpening the bound for the bias of $\hat{\beta}_0^{**}$ given in Corollary 3 would be desirable. Moreover, all the results are of asymptotic nature, naturally providing no concrete statement on finite sample bias. Additionally, one should always keep in mind that the duration dependence parameter α was assumed to be known. To get an impression whether the picture painted here changes under finite sample size or by a strong interrelation between the estimates of α and β_0 will be the task of the first part of the following simulation study.

4 How does heaping influence parameter estimates of standard duration models? Results from simulation studies

In the previous section we showed that heaping, even if the measurement error is of the zero-mean symmetric type, could asymptotically lead to biased estimates of the parameters of the Weibull model. In this section we present results from simulations in order to study the finite sample behaviour and the case of unknown α. In particular we want to answer three questions: First of all we explore whether a type of heaping typical for the GSOEP leads to considerably biased parameter estimates of duration models. We examine a situation with an amount of heaping similar to the findings of Kraus and Steiner that the January inflow of the GSOEP-West unemployment spell sample is about twice as high as one would expect from population values. Second, we attempt to show whether the parameter estimates are less precise due to heaping. Next, we investigate whether heaping introduces spurious seasonal effects or spurious duration dependence effects. In all simulations we use the estimating equations (6) and (7) derived from the ideal likelihood for parameter estimation, once with the true data, once with the heaped data. Each simulation runs the estimation procedure 200 times. We assume that spell lengths are measured in continuous time (the unit is months) and that there is no censoring.

4.1 Symmetric heaping

Let us start with heaping of the symmetric type. In the first set of simulation studies we consider different Weibull distributions as DGPs of the spell data. We assign each spell a specific calendar month at which it starts, such that there is an equal probability that a spell starts in any month of the year. Table 4.1 presents results from the 200 estimations applied to the heaped data. The heaping is such that there is 40 percent chance that the spell starts December and February are changed to January, so that the corresponding spell length is altered by one month. Next, there is a 30 percent chance that the spell starts November and March will become

January[11], so that duration is altered by two months. Thus, the percentage of heaped spells is of a reasonable size, considering that Kraus and Steiner found a January inflow in the GSOEP-West that is about twice as high as its population value. On average some 11% of the spells are affected by heaping. We ran simulations for one DGP assuming a constant (β_0) of one, and another assuming a constant of two. Next, we distinguish between the results for a sample size of 250 observations and one of 500 observations. Also the duration dependence parameter (α) is varied.

Table 4.1 presents the average parameter estimates of the constant and the duration dependence parameter that result from estimation with the heaped data. The corresponding average standard errors and those that are achieved by ML estimation with the original data (prior to heaping) are also displayed.

Panel a) shows the simulation results without any duration dependence, i.e. for exponentially distributed spell lengths. Where β_0 equals one we find that both the constant and the duration dependence parameters estimated with the heaped data are slightly higher than their true values. For both sample sizes, according to the average standard errors the coefficients are not significantly different from their true values. Next, regard the mean estimated standard errors of the constant and of the natural logarithm of the duration dependence parameter that result from the heaped data. They are hardly different from those of the original data. Now turn to the lower part of panel a) which shows simulation results for a constant of value two, so for a higher average spell length. The mean coefficients in the presence of heaping are very close to the true parameter values.

In Panel b) the underlying DGP is a Weibull distribution with positive duration dependence (α=1.4). The mean parameter estimates in this table hardly differ from the parameters of the DGP. Neither are the simulation results for the standard errors in the case of the heaped and the original data any different. We carried out the same analysis for a Weibull DGP where α is set to 0.6, i.e. negative duration dependence. The results are displayed in panel c) of Table 4.1. Here, we clearly see that the mean estimate of the constant for the heaped data exceeds somewhat its true value. Also the average duration

[11] This type of heaping leads to some non-positive spell lengths. E.g., a spell that lasts for one month of which the start is heaped from December to January would never have been observed. So we discarded such spells in the heaped sample. This means, we analyze the behaviour of $\hat{\beta}_0^{**}$ from (11) and not that of $\hat{\beta}_0^{*}$ from (10).

Table 4.1: Simulation results for Weibull distributed spell lengths, 40 percent of the spell starts heaped from December and February to January, 30 percent of the spells starts heaped from November and March to January

a) $\alpha = 1$ **(no duration dependence i.e. exponentially distributed spell lengths)**

Number of obs.:	250			500			
	heaped data		original data	heaped data		original data	
	Estimated mean			Estimated mean			
	Coeff.	SE	SE	Coeff.	SE	SE	
true β_0=1							
β_0	1.033	0.066	0.067	1.04	0.046	0.047	
α	1.02	-	-	1.027		-	-
ln(α)	0.019	0.05	0.049	0.026	0.035	0.035	
true β_0=2							
β_0	2.008	0.067	0.066	2.017	0.047	0.047	
α	1.019	-	-	1.017	-	-	
ln(α)	0.017	0.05	0.049	0.016	0.035	0.035	

b) $\alpha = 1.4$ **(positive duration dependence)**

Number of obs.:	250			500		
	Estimated mean			Estimated mean		
	heaped data		original data	heaped data		original data
	Coeff.	SE	SE	Coeff.	SE	SE
true β_0=1						
β_0	1.021	0.048	0.048	1.025	0.034	0.034
α	1.407	-	-	1.403	-	-
ln(α)	0.34	0.05	0.049	0.338	0.035	0.035
true β_0=2						
β_0	2.005	0.048	0.048	2.005	0.033	0.033
α	1.402	-	-	1.414	-	-
ln(α)	0.337	0.05	0.049	0.346	0.035	0.035

c) α**=0.6 (negative duration dependence)**

Number of obs.:	250					500
	heaped data		original data	heaped data		original data
	Estimated mean			Estimated mean		
	Coeff.	SE	SE	Coeff.	SE	SE
true β_0=1						
β_0	1.078	0.108	0.11	1.075	0.077	0.079
α	0.625	-	-	0.618	-	-
ln(α)	-0.471	0.05	0.049	-0.482	0.035	0.035
true β_0=2						
β_0	2.064	0.109	0.11	2.055	0.077	0.078
α	0.62	-	-	0.621	-	-
ln(α)	-0.479	0.05	0.049	-0.478	0.035	0.035

Number of simulations = 200

dependence coefficient is higher than the true α. None of these differences is significant according to the average standard errors. Next, the reason for the bias may be less the measurement error due to symmetric heaping than the fact that some spells of an original length of less than two months are completely discarded due to the heaping. The proportion of such spells increases with shorter average spell length and is relatively high for the samples that underlie panel c). Due to discarding of a relatively large number of short spells the constant is likely to become somewhat higher than its true value.

Taken together, the conclusion so far is that symmetric heaping, provided a plausible share of the spells are heaped, would have no considerable effect on the parameter estimates of the Weibull model. We carried out another set of simulations for spell lengths that follow the log-logistic distribution. They lead to no different conclusion[12]. In Appendix B, we also show the results of a simulation of the effects of heaping, when covariates determine the hazards. The underlying random sample was generated such that it reflects characteristics of a real world sample of unemployment spells. These simulation results point to no considerable effect of symmetric heaping on the parameter estimates.

4.2 Heaping that leads to prolonged spell lengths only

The following simulations consider a heaping pattern that is not symmetric and so closer to what Kraus and Steiner identified for the unemployment duration data of the GSOEP-West. Suppose we have spell lengths that are exponentially distributed, so there is no duration dependence. Some respondents who start their spells between February and April would heap their spell start to January, while their reported spell end remains correct. This only leads to some prolonged spell lengths, so that the following may be expected: If one attempts to estimate the parameters of a Weibull distribution with the heaped data, the estimated constant should be upward biased. Next, presumably some spurious positive duration dependence ($\alpha > 1$) may emerge. The following tables show simulation results for such a case. This time we leave the sample size at 500 and vary the constant and the amount of heaping.

[12] These results are available on request.

Panel a) of the Table 4.2 raises on average the length of 40 % of spells that start in January by one month, while 30 % of spells that start in February are increased by two months. Thus somewhat less than 6 % of the original spells are increased in length. Again we carried out 200 simulations, estimating the parameters of a Weibull distribution. In the case of a constant of one both the average estimated constant and duration dependence parameter exceed their true values, though never to a considerable extent. This applies even more to the simulation results where the true constant is set to two. The biases have the expected sign, but are rather small and not significant for the chosen sample size (n=500).

In Panel b) of Table 4.2 we raised the measurement error substantially. Now 60 percent of the spells that start in February, 45 percent of those that start in March, and 30 percent of those that start in April were heaped (on average) to the starting month January. Thus somewhat more than 11 % of the spells had their spell lengths changed. The result is that both the average estimated constant and duration dependence parameter as they result from the heaped data increased in size as compared to Table 4.2 a). Where the true constant is equal to one, the average estimated constant is 1.086. According to the estimated mean standard error it is significantly different from its true value. There is no significant positive duration dependence and the bias of the constant becomes again much smaller for a DGP where it equals two. Again the conclusion is that the Weibull model is rather robust to heaping and the more so the higher the average true spell length. The latter fact is again in line with our theoretical results. The mean standard errors of the estimated parameters remain largely unaltered from the heaping.

4.3 Spurious seasonal effects

The last simulation results show that even a small amount of heaping may lead to spurious seasonal effects. Let us again choose an exponential distribution as the DGP with a constant taking on the value one. The heaping pattern will be such that on average 30 percent of the spells that end in October are prolonged by two months and 40 percent of those spells that end in November will become one month longer. So, their spell end is set to December. We estimated the parameters of an exponential distribution including the parameters of covariates that should capture seasonal effects. These are

Table 4.2: Simulation results for exponentially distributed spells ($\alpha = 1$, $\beta_0 = 1$ or $\beta_0 = 2$)

a) 40 percent of the spell starts heaped from February to January, 30 percent of the spell starts heaped from March to January

	heaped data		original data	heaped data		original data
	Estimated mean			Estimated mean		
	Coeff.	SE	SE	Coeff.	SE	SE
true β_0 =1				true β_0=2		
β_0	1.033	0.046	0.047	2.015	0.046	0.047
α	1.02	-	-	1.014	-	-
$\ln(\alpha)$	0.019	0.035	0.035	0.014	0.035	0.035

b) 60 percent of the spell starts heaped from February to January, 45 percent of the spells starts heaped from March to January, 30 percent of the spells starts heaped from April to January

	heaped data		original data	heaped data		original data
	Estimated mean			Estimated mean		
	Coeff.	SE	SE	Coeff.	SE	SE
true β_0=1				true β_0=2		
β_0	1.086	0.045	0.047	2.037	0.046	0.047
α	1.036	-	-	1.036	-	-
$\ln(\alpha)$	0.035	0.035	0.035	0.035	0.035	0.035

Number of obs.: 500
Number of simulations: 200

Table 4.3: Spurious Seasonal Effects - Simulation Results for exponentially distributed spell lengths (β_0=1), 40 percent of the spell ends heaped from November to December, 30 percent of the spell ends heaped from October to December. Heaped data only

	Estimated mean	
	Coeff.	SE
jan	-0.002	0.221
feb	0.009	0.221
mar	0.002	0.221
apr	0.001	0.22
may	-0.002	0.22
jun	0.022	0.221
jul	0.002	0.221
aug	0.033	0.221
oct	0.427	0.244
nov	0.672	0.254
dec	-0.422	0.196
β_0	1.002	0.155

Number of observations: 500
Number of simulations: 200

a set of time-varying dummy variables for each month from January to August and from October to December. Thus we leave September as the base case. Table 4.3 shows our simulation results when estimating the parameters with the heaped data. They clearly suggest that there is a spurious seasonal effect. The coefficients of the October and November dummies are greater than zero. This implies that hazard rates of these months are below that of September, the base case. In contrast the December coefficient is negative and significant, so that the hazards in this month is too high.

5 Conclusions and future work

This study identified effects of heaping on the parameters estimates of some standard duration models. As far as heaping of the entry month of a size that is usual in the GSOEP-West is considered we reached the following conclusions: First, a symmetric heaping does not lead to considerably biased parameter estimates of the Weibull distribution. It hardly leads to standard errors that differ from those estimated with the original data. However, heaping patterns of the kind that Kraus and Steiner found for the unemployment duration data of the GSOEP-West are not symmetric. Their validation study found that respondents tend to place the start of their spells too often to January and not frequently enough to February and March. So, the spells were prolonged. When we consider the Weibull model, intuitively this type of heaping should lead to a higher constant and some spurious positive duration dependence. These biases should become larger the shorter the true average spell length. Our results favour all these hypotheses. Yet they also show that a great deal of heaping is necessary in order to lead to parameter estimates that are far away from those of the parameters of the Weibull DGP. If the end of spells is heaped forward to specific calendar months, we would also think that spurious seasonal effects occur. A simulation study of an exponential DGP showed, that such a heaping pattern indeed implies such an effect, even if the heaping is only of about 6% of the spells.

There are several possible extensions of this work, which we want to address theoretically as well as by simulations. First, it is plausible that heaping may depend on covariates. Suppose there are respondents who become unemployed regularly at around the same time of the year (seasonally unemployed workers). One may expect that they are more likely than others to heap the start of their spells to what they consider the month in which they usually become unemployed.

Second we want to focus on adjustment of the estimates under heaping in particular based on the use of outside information to identify the overall distribution of heaping.[13] Official statistics on unemployment are readily available in many countries. Therefore the use of outside information should generally be possible to study heaping in survey unemployment duration data.

Further theoretical analysis should also try to incorporate independent random censoring and search for direct correction of the likelihood of the observed data. Also a comparison of several correction methods is desirable.

[13] One other way to correct for heaping could be by including a set of dummy variables for the starting months as covariates. Results that are available on request suggest that this indeed improves the estimates of the constant. However it leads to a larger positive bias of the duration dependence parameter.

This should also include alternative methods to deal with the problem, for instance the treatment of heaping as a type of interval censoring.

Acknowledgement We are grateful to Hans Schneeweiß for helpful discussions and comments.

References

Kalbfleisch, J. D., and Prentice, R. L. (1980): *The Statistical Analysis of Failure Time Data.* Wiley, New York.

Kraus, F. and Steiner, V. (1995): Modelling heaping effects in unemployment duration models - with an application to retrospective event data in the German Socio-Economic Panel. *Centre for European Economic Research Discussion Paper.* **95-05**, Mannheim.

Torelli, N. and Trivellato, U. (1993): Modelling inaccuracies in job-search duration data. *Journal of Econometrics* **59**, pp. 187-211.

Wolff, J. (1998): *Essays in Unemployment Duration in two Economies in Transition: East Germany and Hungary.* Ph.D.-Thesis, European University Institute, Florence.

Appendix A1: Proof of Proposition 3.1

The proof of Proposition 3.1 is based on Slutzky's theorem, on the law of large number and on the following lemma.

Lemma A1 *Let T be Weibull distributed with parameters $\lambda = \exp(-\beta_0)$ and α. Then for every $\zeta > -\alpha$*

$$I\!\!ET_i^\zeta = \frac{\Gamma\left(\frac{\zeta+\alpha}{\alpha}\right)}{\lambda^\zeta} = \exp(\zeta \cdot \beta_0) \cdot \Gamma\left(\frac{\zeta+\alpha}{\alpha}\right) .$$

◇

To show Lemma A1 one transforms the occurring integrals by the substitution $u := (\lambda t)^\alpha$, $du = \lambda^\alpha \alpha t^{\alpha-1} dt = \frac{u}{t}\,\alpha\, dt$ and $t = u^{\frac{1}{\alpha}} \frac{1}{\lambda}$ into Gamma integrals:

$$\int\limits_0^\infty t^\zeta \alpha \lambda^\alpha t^{\alpha-1} \exp(-(\lambda t)^\alpha) dt = \int\limits_0^\infty u^{\frac{\zeta}{\alpha}} \frac{1}{\lambda^\zeta} \exp(-u) du = \frac{\Gamma(\frac{\zeta+\alpha}{\alpha})}{\lambda^\zeta}\,.$$

◇

For deriving the formulae for the bias, first introduce a random variable H_i, stochastically independent of T_i, describing the heaping such that

$$T_i^* = T_i + H_i\ .$$

According to (9) one has for $l \in \{1, \cdots, q\}$,

$$P(H_i = l) = \nu^{(l)}\,, \quad P(H_i = -l) = \delta^{(l)}\,, \quad P(H_i = 0) = \xi\,.$$

Proof of Part a)

$$\begin{aligned} I\!\!E\left(\sum_{i=1}^n \frac{T_i^*}{n}\right) &= I\!\!E(T_i^*) = I\!\!E(T_i + H_i) = I\!\!E(T_i) + I\!\!E(H_i) = \\ &= \exp(\beta_0) + \sum_{l=1}^n (\nu^{(l)} - \delta^{(l)}) \cdot l\ . \end{aligned}$$

Therefore

$$\begin{aligned} \operatorname*{plim}_{n\to\infty}(\hat\beta_0^* - \beta_0) &= \operatorname*{plim}_{n\to\infty}\left(\ln\left(\frac{\sum\limits_{l=1}^n T_i}{n}\right) - \beta_0\right) \\ &= \ln\left(\exp(\beta_0)\left(1 + \frac{\sum\limits_{l=1}^q \left(\nu^{(l)} - \delta^{(l)}\right) \cdot l}{\exp(\beta_0)}\right)\right) - \beta_0 \\ &= \ln\left(1 + \frac{\sum\limits_{l=1}^q \left(\nu^{(l)} - \delta^{(l)}\right) \cdot l}{\exp(\beta_0)}\right)\,. \end{aligned}$$

Proof of Part b)

$$\begin{aligned}
\mathbb{E}\left(\sum_{l=1}^{n}\frac{(T_i^*)^2}{n}\right) &= \mathbb{E}((T_i^*)^2) = \mathbb{E}(T_i^2 + 2\cdot T_i\cdot H_i + H_i^2) = \\
&= \mathbb{E}(T_i^2) + 2\cdot\mathbb{E}(T_i)\cdot\mathbb{E}(H_i) + \mathbb{E}(H_i^2) = \\
&= \exp(2\beta_0)\cdot\Gamma(2) + 2\cdot\left(\exp(\beta_0)\cdot\Gamma\left(\frac{3}{2}\right)\right)\cdot\sum_{l=1}^{q}(\nu^{(l)} - \delta^{(l)})\cdot l \; + \\
&\qquad + \sum_{l=1}^{q}(\nu^{(l)} + \delta^{(l)})\cdot l^2 = \\
&= \exp(2\beta_0)\cdot\left(1 + \frac{\sum\limits_{l=1}^{q}\left(\nu^{(l)} + \delta^{(l)}\right)\cdot l^2}{\exp(2\beta_0)} + \frac{\sqrt{\pi}\cdot\sum\limits_{l=1}^{q}\left(\nu^{(l)} - \delta^{(l)}\right)\cdot l}{\exp(\beta_0)}\right).
\end{aligned}$$

Therefore

$$\lim_{n\to\infty}(\hat{\beta}_0^* - \beta_0) = \operatorname*{plim}_{n\to\infty}\left(\frac{1}{2}\cdot\ln\left(\frac{\sum\limits_{i=1}^{n}(T_i^*)^2}{n}\right) - \beta_0\right) =$$

$$= \frac{1}{2}\ln(\exp(2\beta_0)) + \frac{1}{2}\ln\left(1 + \frac{\sum\limits_{l=1}^{q}\left(\nu^{(l)} + \delta^{(l)}\right)\cdot l^2}{\exp(2\beta_0)} + \frac{\sqrt{\pi}\cdot\sum\limits_{l=1}^{q}\left(\nu^{(l)} - \delta^{(l)}\right)\cdot l}{\exp(\beta_0)}\right) - \beta_0 =$$

$$= \frac{1}{2}\ln\left(1 + \frac{\sum\limits_{l=1}^{q}\left(\nu^{(l)} + \delta^{(l)}\right)\cdot l^2}{\exp(2\beta_0)} + \frac{\sqrt{\pi}\cdot\sum\limits_{l=1}^{q}\left(\nu^{(l)} - \delta^{(l)}\right)\cdot l}{\exp(\beta_0)}\right).$$

◇

Appendix A2: Proof of Proposition 3.4

Proof of Part a)

From Proposition 3.1, Part a)

$$\begin{aligned}
\beta_0 &= \operatorname*{plim}_{n\to\infty} \hat{\beta}_0^* - \ln\left(1 + \frac{\sum\limits_{l=1}^{q}\left(\nu^{(l)} - \delta^{(l)}\right)\cdot l}{\exp(\beta_0)}\right) \\
\iff \quad \exp\left(\beta_0\right) &= \exp\left(\operatorname*{plim}_{n\to\infty} \hat{\beta}_0^*\right)\cdot\left(1 + \frac{\sum\limits_{l=1}^{q}\left(\nu^{(l)} - \delta^{(l)}\right)\cdot l}{\exp(\beta_0)}\right)^{-1} = \\
&= \frac{\exp\left(\operatorname*{plim}\limits_{n\to\infty} \hat{\beta}_0^*\right)\cdot\exp\left(\beta_0\right)}{\exp\left(\beta_0\right) + \sum\limits_{l=1}^{q}\left(\nu^{(l)} - \delta^{(l)}\right)\cdot l} \\
\iff \quad \exp\left(\beta_0\right) &= \exp\left(\operatorname*{plim}_{n\to\infty} \hat{\beta}_0^*\right) - \sum_{l=1}^{q}\left(\nu^{(l)} - \delta^{(l)}\right)\cdot l\,.
\end{aligned}$$

Therefore, by Slutzky's Theorem,

$$\begin{aligned}
\operatorname*{plim}_{n\to\infty} \widehat{{}^{(1)}\beta_0} &= \operatorname*{plim}_{n\to\infty}\left(\ln\left(\exp\left(\hat{\beta}_0^*\right) - \sum_{l=1}^{q}\left(\nu^{(l)} - \delta^{(l)}\right)\cdot l\right)\right) = \\
&= \ln\left(\exp\left(\operatorname*{plim}_{n\to\infty} \hat{\beta}_0^*\right) - \sum_{l=1}^{q}\left(\nu^{(l)} - \delta^{(l)}\right)\cdot l\right) = \\
&= \ln(\exp(\beta_0)) = \beta_0\,.
\end{aligned}$$

Proof of Part b)

From Proposition 3.1, Part b)

$$\beta_0 = \operatorname*{plim}_{n\to\infty} \hat{\beta}_0^* - \frac{1}{2} \cdot \ln \left(1 + \frac{\sum\limits_{l=1}^{q} \left(\nu^{(l)} + \delta^{(l)} \right) \cdot l^2}{\exp(2\beta_0)} + \frac{\sqrt{\pi} \cdot \sum\limits_{l=1}^{q} \left(\nu^{(l)} - \delta^{(l)} \right) \cdot l}{\exp(\beta_0)} \right)$$

$$\Longleftrightarrow \exp(2\beta_0) = \frac{\exp\left(2 \operatorname*{plim}\limits_{n\to\infty} \hat{\beta}_0^* \right)}{1 + \frac{\sum\limits_{l=1}^{q} \left(\nu^{(l)} + \delta^{(l)} \right) \cdot l^2}{\exp(2\beta_0)} + \frac{\sqrt{\pi} \cdot \sum\limits_{l=1}^{q} \left(\nu^{(l)} - \delta^{(l)} \right) \cdot l}{\exp(\beta_0)}}$$

$$\Longleftrightarrow 0 = (\exp(\beta_0))^2 + \left(\sqrt{\pi} \cdot \sum_{l=1}^{q} \left(\nu^{(l)} - \delta^{(l)} \right) \cdot l \right) \cdot \exp(\beta_0) + \\ + \left(\cdot \sum_{l=1}^{q} \left(\nu^{(l)} + \delta^{(l)} \right) \cdot l^2 - \exp\left(2 \operatorname*{plim}_{n\to\infty} \hat{\beta}_0^* \right) \right) .$$

Using (13) the assumption

$$\sqrt{\pi} \cdot \sum_{l=1}^{q} (\nu^{(l)} - \delta^{(l)}) \cdot l + \exp(\beta_0) \geq 0$$

made in Proposition 3.4 implies

$$\sum_{l=1}^{q} \left(\nu^{(l)} + \delta^{(l)} \right) \cdot l^2 < \exp\left(2 \operatorname*{plim}_{n\to\infty} \hat{\beta}_0^* \right) .$$

This guarantees that the quadratic form in the variable $\exp(\beta_0)$ from above possesses a well-defined and unique solution, namely

$$\exp(\beta_0) = \frac{1}{2} \cdot \left(-\sqrt{\pi} \sum_{l=1}^{q} \left(\nu^{(l)} - \delta^{(l)} \right) \cdot l + \right. \\ \left. + \sqrt{\pi \cdot \left(\sum_{l=1}^{q} \left(\nu^{(l)} - \delta^{(l)} \right) \cdot l \right)^2 - 4 \cdot \left(\sum_{l=1}^{q} \left(\nu^{(l)} + \delta^{(l)} \right) \cdot l^2 - \exp\left(2 \operatorname*{plim}_{n\to\infty} \hat{\beta}_0^* \right) \right)} \right) .$$

Therefore

$$\begin{aligned}
\operatorname*{plim}_{n\to\infty} \widehat{{}^{(2)}\beta_0} &= \operatorname*{plim}_{n\to\infty} \ln\left(\frac{1}{2}\left[\sum_{l=1}^{q}\sqrt{\pi}\left(\delta^{(l)}-\nu^{(l)}\right)\cdot l+\right.\right. \\
&\left.\left.+\sqrt{\pi\cdot\left(\sum_{l=1}^{q}\left(\nu^{(l)}-\delta^{(l)}\right)\cdot l\right)^2-4\cdot\left(\sum_{l=1}^{q}\left(\nu^{(l)}+\delta^{(l)}\right)\cdot l^2-\exp\left(2\hat{\beta}_0^*\right)\right)}\;\right]\right) \\
&= \ln\left(\frac{1}{2}\left[\sum_{l=1}^{q}\sqrt{\pi}\left(\delta^{(l)}-\nu^{(l)}\right)\cdot l+\right.\right. \\
&\left.\left.+\sqrt{\pi\cdot\left(\sum_{l=1}^{q}\left(\nu^{(l)}-\delta^{(l)}\right)\cdot l\right)^2-4\cdot\left(\sum_{l=1}^{q}\left(\nu^{(l)}+\delta^{(l)}\right)\cdot l^2-\exp\left(2\operatorname*{plim}_{n\to\infty}\hat{\beta}_0^*\right)\right)}\;\right]\right)= \\
&= \ln(\exp(\beta_0)) = \beta_0\,.
\end{aligned}$$

◇

Appendix A3: Heaping Which Depends on the Entry Month

Here we consider the case where the heaping probabilities depend on the entry month. Let for every month $j \in \{1,\ldots,s_{max}\}$ the heaping probabilities $\nu^{(l,j)}$ and $\delta^{(l,j)}$ be defined analogous to Section 3.2. Further let B_i be the random variable describing the entry month of unit i.

Assume that the true duration T_i is stochastically independent of B_i and, analogous to above, that H_i and T_i are conditionally independent given B_i. Further the entry month B_i is taken to be independently and identically distributed among all units $i = 1,\ldots,n$.

We will consider explicitly only the case $\alpha = 2$; the case $\alpha = 1$ can be treated in the same way. It is shown that the expectation $I\!E((T_i^*)^2)$ depends only on the marginal probabilities

$$\begin{aligned}
\nu^{(l)} &= \sum_{j=1}^{s_{max}} \nu^{(l,j)}\cdot P(\{B_i=j\})\,, \quad l=1,\ldots,q, \\
\delta^{(l)} &= \sum_{j=1}^{s_{max}} \delta^{(l,j)}\cdot P(\{B_i=j\})\,, \quad l=1,\ldots,q,
\end{aligned}$$

of the heaping variable H_i. Then the arguments used in Appendiy A1 and Appendix A2 to discuss bias and bias correction are also valid for the case considered here.

$$\begin{aligned}
&\mathbb{E}((T_i^*)^2) = \\
= \; & \mathbb{E}\Big(\mathbb{E}((T_i^*)^2\,|B_i)\Big) = \sum_{j=1}^{s_{max}} \mathbb{E}\left((T_i^*)^2\,|B_i\right) \cdot P(\{B_i = j\}) = \\
= \; & \sum_{j=1}^{s_{max}} \left(\mathbb{E}(T_i^2\,|B_i) + \mathbb{E}(T_i\,|B_i) \cdot \mathbb{E}(H_i\,|B_i) + \mathbb{E}(H_i^2\,|B_i)\right) \cdot P(\{B_i = j\}) = \\
= \; & \mathbb{E}(T_i^2) + \mathbb{E}(T_i) \cdot \sum_{j=1}^{s_{max}} \left(\sum_{l=1}^{q} \left(\nu^{(l,j)} - \delta^{(l,j)}\right) \cdot l\right) \cdot P(\{B_i = j\}) + \\
& \qquad + \sum_{j=1}^{s_{max}} \left(\sum_{l=1}^{q} \left(\nu^{(l,j)} + \delta^{(l,j)}\right) \cdot l^2\right) \cdot P(\{B_i = j\}) = \\
= \; & \mathbb{E}(T_i^2) + \\
& + \mathbb{E}(T_i) \cdot \left(\sum_{l=1}^{q} \left(\sum_{j=1}^{s_{max}} \nu^{(l,j)} \cdot P(\{B_i = j\}) - \sum_{j=1}^{s_{max}} \delta^{(l,j)} \cdot P(\{B_i = j\})\right) \cdot l\right) + \\
& + \sum_{l=1}^{q} \left(\sum_{j=1}^{s_{max}} \nu^{(l,j)} \cdot P(\{B_i = j\}) + \sum_{j=1}^{s_{max}} \delta^{(l,j)} \cdot P(\{B_i = j\})\right) \cdot l^2 = \\
= \; & \mathbb{E}(T_i^2) + \mathbb{E}(T_i) \cdot \sum_{l=1}^{q} \left(\nu^{(l)} - \delta^{(l)}\right) \cdot l + \sum_{l=1}^{q} \left(\nu^{(l)} + \delta^{(l)}\right) \cdot l^2 .
\end{aligned}$$

Appendix B

Appendix Table B.1 shows simulation results, when a number of covariates is introduced to an exponential DGP. The heaping is again such that there is 40 percent chance for spell starts December and February to be changed to January and a 30 percent chance that the spell starts November and March will become January. The probability that a spell starts in a specific calendar month is no longer 1/12. The calendar start of the spell are drawn such that the distribution of the inflow over calendar months is by and large in line

Table B1: Simulation results for exponentially distributed spells including covariates, 40 percent of the spell starts heaped from December and February to January, 30 percent of the spells starts heaped from November and March to January

	Estimated mean		
	Coeff.	(Parameter)	SE
age	-0.026	(-0.021)	0.027
$age^2/100$	0.205	(0.2)	0.038
Education dummies			
unskilled	0.688	(0.7)	0.136
vocational training (base)			
master/craftsmen	-0.518	(-0.5)	0.166
engineer/technical university	-0.913	(-0.9)	0.184
university degree	-1.451	(-1.5)	0.158
ln(Vacancy-Unemployment Ratio)	19.661	(20)	4.473
β_0	-0.113	(-0.250)	0.482
α	1.026	(1)	-
$\ln(\alpha)$	0.025	(0)	0.035

Number of obs.: 500
Number of simulations: 200

with the distribution of the population inflow over calendar months into registered unemployment in East-Germany over the period 1991 to 1994. Values for covariates are drawn such that the sample is reflecting characteristics of a male inflow sample into registered unemployment from the GSOEP-East from 1991 to 1994 (Wolff [1998]). Table B.1 displays the mean coefficients and standard errors of 200 simulations of maximum likelihood estimation of a Weibull likelihood. The true parameter values of the covariate are displayed in brackets. There is no indication, that at the chosen sample size the simulated parameter estimates are significantly different from their true values.