Kraus, Florian; Steiner, Viktor

Working Paper

Modelling heaping effects in unemployment duration models - with an application to retrospective event data in the German socio-economic panel

ZEW Discussion Papers, No. 95-09

Provided in cooperation with:
Zentrum für Europäische Wirtschaftsforschung (ZEW)

ZEW
Zentrum für Europäische Wirtschaftsforschung GmbH
Centre for European Economic Research

Discussion Paper

Discussion Paper No. 95-09

Modelling Heaping Effects in Unemployment Duration Models – With an Application to Retrospective Event Data in the German Socio–Economic Panel

Florian Kraus
Viktor Steiner

ZEW
Zentrum für Europäische Wirtschaftsforschung GmbH
Labour Economics, Human Resources and Social Policy Series

Discussion Paper No. 95-09

Modelling Heaping Effects in Unemployment Duration Models – With an Application to Retrospective Event Data in the German Socio–Economic Panel

Florian Kraus
Viktor Steiner

Modelling Heaping Effects in Unemployment Duration Models –With an Application to Retrospective Event Data in the German Socio–Economic Panel

by

Florian Kraus and Viktor Steiner*)

Zentrum für Europäische Wirtschaftsforschung (ZEW)

January 1996

Abstract

Unemployment duration data derived from retrospective surveys often show an abnormal concentration of responses at certain durations. This common kind of measurement error is known as "heaping" in the statistical literature. Although heaping effects may lead to severe biases in estimated coefficients of duration models, in applied work researchers have either neglected them altogether or tried to account for them in an ad hoc way. This is also the case for recent microeconometric research based on unemployment duration data derived from the retrospective calendar information in the German Socio-Economic Panel, where a very high proportion of all unemployment spells beginning in January or end in December of each year. We show how this kind of heaping can be modelled within a maximum likelihood framework using external validation information and demonstrate for this particular data set how parameter estimates in discrete-time proportional hazard models of unemployment duration are affected by alternative specifications of the heaping mechanism. Our main result is that parameter estimates are generally rather insensitive to whether or not heaping is explicitly taken into account and to different assumptions about the heaping mechanism, but may be substantially affected by ad hoc procedures to control for heaping which tend to pick up selectivity effects and censoring.

JEL: C41, C51, J64

Acknowledgement

*) Financial support from the German Science Foundation (DFG) is gratefully acknowledged. We received helpful comments on previous versions of this paper from participants of the Applied Econometrics seminar at Mannheim University and the 1995 meeting of the Econometrics Section of the German Statistical Association held in Dresden.

1 Introduction

In labour force surveys individuals are usually asked to state the duration of their current or previous unemployment spell on a retrospective basis. It is a well–known phenomenon (see, e.g., Bowers and Horvath, 1984; Porterba and Summers, 1986; Torelli and Trivellato, 1987, 1993a, b) that, due to memory effects, unemployment durations derived from these answers are contaminated with measurement errors. Empirically, these errors show up in the abnormal concentration of responses at certain durations, which is termed the "heaping effect" in the literature. Heaping is a special case of data coarsening (see Heitjan and Rubin, 1991, for a general theory of coarse data, and Holt, McDonald and Skinner, 1991, for some applications in the context of event history studies). It is mainly due to "rounding off" at particular values of the variable of interest and is arguably the most prevalent source of measurement error in retrospective unemployment duration data. Since the average duration of the unemployment flow is quite short in most countries, these measurement errors may be relatively large and thus lead to severely biased parameter estimates in unemployment duration models.

Somewhat more formally, heaping may be defined as follows: Let the true duration, T, a non–negative random variable with density $f(t,\theta)$, where θ is a vector of unknown parameters to be estimated, be measured with error. Further assume that, with probability $G(t,\alpha)$, the true duration t will be reported as $t_h=t+\delta(t)y(t)$, with $y(t)=1$ being the realization of a Bernoulli random variable, $Y(t)$; $\delta(t)$ is a measurement error, and α is a vector of unknown parameters. $G(t,\alpha) = P(Y(t)=1)$ is called the heaping function and gives the probability that an observed duration is a heaping point instead of the true duration. The measurement errors are not purely random, but are concentrated at certain observed durations which comprise the set of heaped values or heaping points, $H = \{t_h\}$. An important issue in this kind of model is the choice of the heaping points, t_h, and the heaping pattern, $\delta(t)$, which maps the true durations into the set of heaped values. The heaping pattern defines the set of possibly heaped durations and on which of the heaping points they are heaped. It must be known a priori or be derived from external validation data. Given these definitions, the probability of observing a duration t_b is $f(t,\theta)\cdot G(t,\alpha)$, integrated over the set of values possibly heaped on t_b, if t_b is a heaping point, and $f(t_b,\theta)\cdot(1–G(t,\alpha))$ otherwise. Thus, the heaping mechanism is ignorable when drawing likelihood inferences on the vector θ if, and only if, heaping occurs purely at random and the parameter vectors θ and α are distinct. Obviously, these conditions may often be violated in practice.

Although ignoring heaping effects may lead to severely biased parameter estimates of micro–econometric models of unemployment duration data, they have so far received little attention in applied research. To the best of our

knowledge, Torelli and Trivellato (1993a) were the first authors who tried to combine the basic heaping model described above with various specifications of a continuous–time duration model of unemployment. On the basis of a simulation study they show that the effects of heaping on estimated parameters of the duration model depend on the incidence and pattern of heaping as well as the specification of the distribution of "true" durations, and that ignoring the heaping effect altogether or taking it into account in an ad-hoc way may severely affect parameter estimates. However, in contrast to these theoretical results, they also show on the basis of an empirical application using retrospective data on unemployment durations from the Italian labour force survey with quarterly rotating design that, empirically, differences in parameter estimates may not be much affected by completely ignoring heaping effects in their model. They do find, however, that ad hoc procedures like accounting for heaping by simply including dummy variables for the corresponding months in a duration model results in somewhat different parameter estimates. In a follow–up study, Torelli and Trivellato (1993b) speculate that one reason for the relatively small heaping effects observed in their empirical model may be due either to the specific observation scheme in the Italian labour force survey or to the assumed simple form of heaping. They suggest to consider different observation schemes and more general heaping models to establish if their insensitivity result holds more generally.

For Germany, most recent microeconometric research on duration of unemployment has used the monthly calendar information on an individual's labour force state in the previous year contained in the German Socio–Economic Panel (GSOEP). In principle, unemployment duration data derived from the GSOEP should be much more reliable than that obtained from labour force surveys because (i) the calendar in each wave of the panel only refers to the previous year, and (ii) the design of the calendar requires the respondent to explicitly code his or her labour force status in each month of the previous calendar year. These features of the GSOEP calendar information should as far as possible reduce potential errors of misunderstanding and memory effects on the respondents' side. However, a very high proportion of all unemployment spells calculated from these calendar data apparently beginning in January or end in December of each year. As the comparison with aggregate unemployment flow data published by the Federal Labour Office shows, this strong concentration of flows in January and December cannot be explained by cyclical factors alone.

Although there seems to be a strong presumption of heaping effects in the calendar data of the GSOEP, there has hardly been any investigation about the potential effects on parameter estimates in microeconometric models of the duration of unemployment based on this widely used data set. Hujer and Schneider (1989), Hujer, Löwenbein and Schneider (1990), and Hunt (1995) at

least acknowledge these effects and try to account for the disproportionate number of spells ending in December of each year by including a dummy variable for this month in their sets of regressors. In all three applications the coefficient on the December dummy shows a very strong positive effect on the hazard rate from unemployment, which is interpreted as evidence for heaping effects by these authors. This procedure can be criticized as inadequate in the light of the results in Torelli and Trivellato (1993a). In this paper, we extend their analysis and show how heaping effects in the calendar data of the GSOEP can be modelled within a maximum likelihood framework, and how specification issues may affect parameter estimates in standard microeconometric models of unemployment duration. In particular, we demonstrate how estimated coefficients of standard explanatory variables in the duration model and the estimated baseline hazard function are affected by alternative specifications of the heaping pattern. We also show how external validation data can be used for estimating the heaping mechanism and how its specification affects estimation results for our empirical application.

The next section describes in some detail important features of the calendar data of the GSOEP and their relationship to the data published by the Federal Labour Office, which provides the required a priori knowledge to identify the heaping mechanism. In section 3, the statistical model is described and the relevant equations for estimation are derived, while issues in estimation are discussed in section 4. The results of our empirical study are presented and discussed in section 5, and section 6 concludes.

2 Heaping Effects in the Monthly Calendar Data of the German Socio–Economic Panel

Unemployment duration data analysed in this paper are derived from the calendar information contained in the Socio–Economic Panel for West Germany (GSOEP) which is a representative sample of the resident population aged 16 years or older. The GSOEP for West Germany has been running on a yearly basis since 1984 when about 12,000 persons in some 6,000 households were surveyed (the structure of the GSOEP is described by, e.g., Burkhauser, 1991, and Wagner, Schupp and Rendtel, 1991). At the date of interview in each wave, comprehensive information on individual and household characteristics, income variables and labour force participation is obtained. In addition, detailed information on an individual's labour force status in each month of the previous calendar year is coded in the so–called calendar. The duration of individual unemployment spells can be derived from this calendar information by merging subsequent waves of the GSOEP as follows.

For our empirical analysis, we have aggregated an individual's labour force status in a particular month into one of three exclusive states: employed, unemployed,

and out of the labour force. Individuals are coded as employed if they are in part–time or full–time employment, temporary employment or on vocational training schemes. By the definition used in the GSOEP, only individuals registered at the labour office are counted as unemployed; this is also the criterion used in the register–based official statistics, a fact which will gain importance later on. All others, i.e. those doing housework, attending school or higher education, doing their military service or those on (early) retirement, were aggregated into the category "out of the labour force". From this information, the months of the beginning and the end of a completed spell of unemployment and, hence, its completed duration can be derived on a monthly basis, whereas for a right–censored spell the interrupted spell duration and the censoring status are known. We selected all unemployment spells which began in January 1983 or later and are not left–censored, but which may be right–censored at the time they are observed for the last time. Unless people drop prematurely out of the GSOEP, this time is December 1991 in our sample.[1] Left–censored spells were not only excluded because they pose a severe problem in any duration model, but also for the additional reason that they would have artificially inflated the number of spells starting in January due to new entries into the panel.[2] We also excluded persons who were previously employed in the construction sector, because this sector exhibits a very pronounced seasonal pattern which would potentially interfere with heaping effects. Furthermore, we do not consider the estimated duration model adequate for purely seasonal unemployment spells.

For the following reasons, the aggregate monthly unemployment flows derived from the calendar data in the GSOEP and the register data of the Federal Labour Office (FLO) should be comparable. First, the definition of unemployment is the same in both data sources, namely beeing registered as unemployed at the labour office. Second, although foreigners are over–represented in the GSOEP, this does not affect the *relative* monthly flows into or out of unemployment of German nationals and foreigners as graphical checks have shown. Furthermore, adjusting the data by the appropriate weighting factors to account for disproportional sampling of foreigners in the GSOEP (on this see, e.g., Wagner, Schupp and Rendtel, 1991) resulted in only minor differences in the monthly flow statistics. Third, the disproportionate reduction of unemployment spells in the

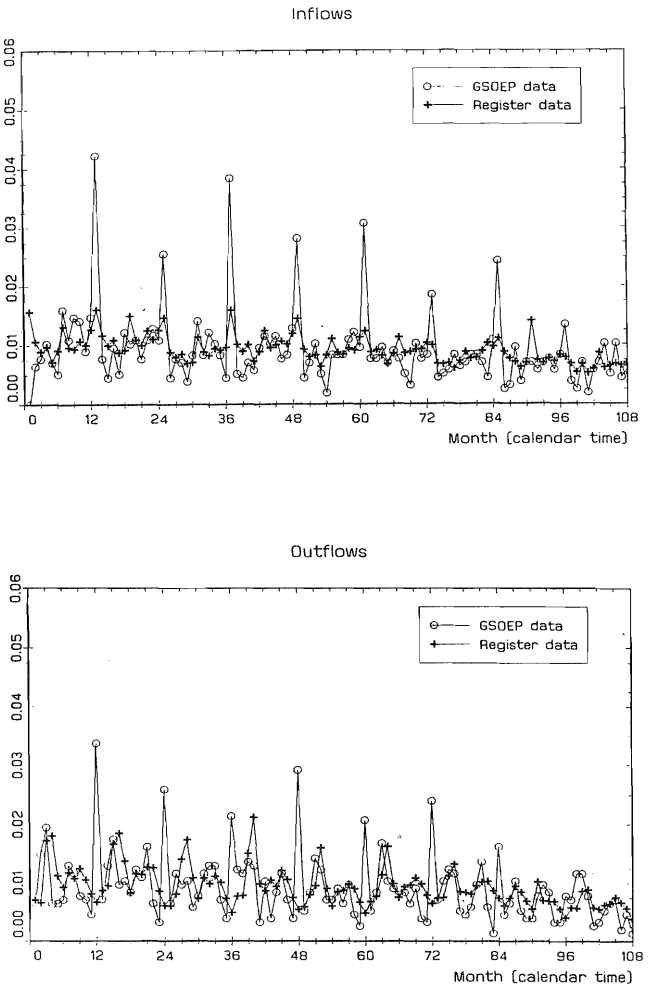
---

[1] As is usually the case with panel data, a considerable number of individuals has been lost over time due to sample attrition; about six thousand individuals took part in the first nine waves on which our study is based. On the other hand, almost 3,000 new individuals have been added to the population in this period, most of them youth already included in the panel as children in the sampled households before they had passed the age limit of 16 years.

[2] For those familiar with the peculiarities of the GSOEP it may be interesting to know that we have used the BIOSCOPE records which contain, on a yearly basis, retrospective information on an individual's labour force status since the age of fifteen to distinguish spells with an observed beginning in January from left–censored spells.

GSOEP relative to the register data due to sample attrition was taken into account by adjusting the former data using appropriate weighting factors calculated on a yearly basis.

Figure 1. Relative frequencies of inflows and outflows in the calendar date of the GSOEP and the register–based data of the FLO



Note: Inflows and outflows refer to monthly data from January 1983 to December 1991.

Source: Federal Labour Office, Amtliche Nachrichten (Official Monthly Bulletin), various issues; Socio–Economic Panel for West Germany, waves 1 – 9; own calculations.

Plots of relative frequencies of monthly unemployment inflows and outflows in the observation period derived from the calendar data in the GSOEP and the register data of the FLO show the following noticeable differences (see Figure 1):

(i) In the calendar data, there are very pronounced spikes in the relative inflows in January of each year. Although the relative inflow in January is also relatively high in the registered–based data, the difference to other months is much smaller than for the calendar data. Hence, the spikes in January are not simply due to seasonal factors, which could be an explanation for the relatively small January effect observed in the register–based data.[3]

(ii) In the neighbouring months of January in each year the relative inflows of the calendar data tend to be much smaller than those of the register–based data, which suggests a compensating effect for the concentration of inflows in January.

(iii) The relative outflows in December are much higher in the calendar data than in the register–based data for most years. There seem to be compensating effects at the neighbouring months. Differences between the two data sources with respect to outflows seem less pronounced than for inflows.

Assuming that the official register–based data represent the real pattern of the monthly unemployment flows, these results suggest that there are systematic measurement errors in the calendar–based duration data, which have to be built into the statistical model when analyzing unemployment durations derived from these data.

3 The Statistical Model

Given the pattern in the unemployment inflows and outflows derived from the calendar data as described in the previous section, it seems obvious that both the beginning and the end of an unemployment spell are affected by heaping. Hence, a model which directly builds on spell duration without distinguishing between its beginning and end seems not directly applicable for our data pattern. We therefore extend the basic model as described by Torelli and Trivellato (1993a) by taking into account this feature of our data. The following notations and formulas refer to the discrete–time duration model because it seems more appropriate for the discrete nature of the monthly calendar data used in this study and also simplifies the notation somewhat. The corresponding continuous–time duration model is a straightforward extension of the discrete case, on which we will briefly comment below.

---

[3] Note that to some extent seasonal effects have already been purged by leaving out the construction sector.

The random variables in our model are:

- T: true length of an unemployment spell with probability measure $f(t, \beta)$, where β is a vector of unknown parameters;
- B: true beginning of an unemployment spell;
- E: true end of an unemployment spell;
- T_b: observed duration of an unemployment spell;
- B_b: observed beginning of an unemployment spell, and
- E_b: observed end of an unemployment spell.

In addition to these variables, we define two Bernoulli random variables, $Y_B(b)$ and $Y_E(e)$, as follows: If $Y_B(b)$ is equal to one, the true beginning of a spell is measured with a heaping effect, else it is assumed to be measured correctly; an analogous definition holds for $Y_E(e)$.

The heaping models provide the link between the measured and true variables. The following derivation refers to a relatively simple heaping pattern as appropriate for our unemployment duration data. For every true beginning, b, measured with error, there is one heaping point b_h, with $\delta_B(b) = b_h - b$ defining the distance between b and the heaping point. Likewise, to every end, e, measured with error, belongs one heaping point e_h, with $\delta(e) = e_h - e$. Given the discussion in the previous section, the set of heaped values, $H_B = \{b_h\}$ and $H_E = \{e_h\}$, in our duration data is given by: H_B = January and H_E = December. Furthermore, it is assumed that any b in {January to March} in year x is heaped to January of the same year, and that any e in {October, November, December} in year x is heaped to December of the same year (the rationale for the choice of this heaping pattern will be explained in section 4.2). The distances to the heaping points are thus uniquely defined for each of these months. Given these definitions and rules, the heaping functions $G_B(b)$ and $G_E(e)$ are defined by

$$
\begin{aligned}
G_B(b) &= P(Y_B(B)=1|B=b) = P(\delta_B(B)\cdot Y_B(B)=\delta_B(b)|B=b) \\
G_E(e) &= P(Y_E(E)=1|E=e) = P(\delta_E(E)\cdot Y_E(E)=\delta_E(e)|E=e).
\end{aligned} \tag{1}
$$

The equations for the measurement model therefore are

$$
\begin{aligned}
B_b &= B + \delta_B(B)\cdot Y_B(B) \\
E_b &= E + \delta_E(E)\cdot Y_E(E) \\
T_b &= T + \delta_E(E)\cdot Y_E(E) - \delta_B(B)\cdot Y_B(B).
\end{aligned} \tag{2}
$$

The corresponding equations for the observations are

$$b_b = b + d_b(b)$$

(2') $$e_b = e + d_b(e)$$

$$t_b = t + d_b(e) - d_b(b),$$

where $d_b(b)$ and $d_b(e)$ are the observed heaping effects with $d_b(b) = \delta_B(b)$, $d_b(e) = \delta_E(e)$ if the observation is heaped, and $d_b(b) = 0$, $d_b(e) = 0$ otherwise.

Note that, since $E_b = T_b + B_b - 1$, it is equivalent to use the observed values of the beginning and end of a heaped spell instead of its duration in drawing inferences on parameters in the estimated duration model.

Now we make the following assumptions, where P stands for probability:

A1) The duration and the beginning are independent, i.e. $P(T=t \wedge B=b) = P(T=t) \cdot P(B=b)$.

A2) $Y_E(E)$ and B, given E, are independent, i.e. the heaping process at the end of the spell is not influenced by its beginning. Hence,

$$P(Y_E(E) = y_E(E) \wedge B=b \mid E=e) = P(Y_E(E) = y_E(E) \mid E=e) \cdot P(B=b \mid E=e)$$

A3) $Y_B(B)$ and E, given B, are independent:

$$P(Y_B(B) = y_B(B) \wedge E=e \mid B=b) = P(Y_B(B) = y_B(B) \mid B=b) \cdot P(E=e \mid B=b)$$

A4) $Y_B(B)$ and $Y_E(E)$, given B and E, are independent:

$$P(Y_B(B) = y_B(B) \wedge Y_E(E) = y_E(E) \mid B=b \wedge E=e)$$

$$= P(Y_B(B) = y_B(B) \mid B=b \wedge E=e) \cdot P(Y_E(E) = y_E(E) \mid B=b \wedge E=e)$$

A5) The mechanism of right censoring is of the type "independent censoring" in the sense of Kalbfleisch and Prentice (1980:120).

A6) The observations of different spells are independent of each other.

Note that all assumptions and formulas are to be understood as conditional on the covariables in the duration model. Assumption A1) implies that the spell duration distribution is independent of entrance time, i.e. all spells in our data set are assumed to have been generated from the same duration model. Since this assumption may be violated even after having excluded unemployment spells from the seasonally sensitive construction sector, we try to statistically control for seasonal factors by conditioning the duration model on the regional unemployment rate. Assumption A4) may be violated if a person reporting, say, the beginning of a spell with error is susceptible of also wrongly reporting its end, e.g. due to weak memory. Since information to estimate the joint probability of

$Y_B(B)$ and $Y_E(E)$ is not available, we only can estimate the marginal distributions and, hence, have to assume independence. Assumption A5) is usually made even in simple duration models without measurement errors for the sake of tractability. In any case, the detailed structure of the censoring mechanism is unknown and there seems to be no plausible model for it. Assumption 6) is also standard in microeconometric models.

Given the assumption of noninformative censoring, the likelihood function for the observed durations can, with loss in efficiency, be based on (see, e.g. Kalbfleisch and Prentice, 1980):

(3) $\quad \text{Lik}(\theta) \propto \prod f(t_{b,i},\theta) \cdot \prod (1 - F(t_{b,j},\theta)),$

where the first product term refers to completed and the second to censored observations, $F(\cdot)$ being the distribution function.

In our discrete–time model the likelihood contributions are $P(B_{b,i}=b_{b,i} \wedge T_{b,i}=t_{b,i})$ for a completed observation and, since censoring means $T_{b,i} \geq t_{b,i}$, $P(B_{b,i}=b_{b,i} \wedge T_{b,i} \geq t_{b,i})$ for the censored case. Hence,

(3') $\quad \text{Lik} \propto \prod P(B_{b,i}=b_{b,i} \wedge T_{b,i}=t_{b,i}) \cdot \prod P(B_{b,j}=b_{b,j} \wedge T_{b,j} \geq t_{b,j})$

Although assumptions A2) through A4) above give conditional independence, it is clear that heaping at the beginning or the end of a spell depends on its true beginning and end, respectively. This dependence is described by the heaping functions built into the likelihood function which requires the derivation of the connection between the probability function of the observable and the true random variables in our model.

To start with a completed spell, one can show (see the appendix) that

(4) $\quad P(B_b = b_b \wedge E_b = e_b) =$

$$= \sum_b \sum_e P(d_B(B) \cdot Y_B(B) = b_b - b \mid B = b) \cdot P(d_E(E) \cdot Y_E(E) = e_b - e \mid E = e)$$
$$\cdot P(T = e - b + 1) \cdot P(B = b)$$

where the index for individual i has been dropped for expositional convenience.

For a right-censored spell the derivation of the likelihood contribution is somewhat more complicated (see the appendix). Let $u = \max\{\delta_e(n) \mid n \in \text{IN} \text{ and } n + \delta_e(n) = e_b\}$ and $o = \min\{\delta_e(n) \mid n \in \text{IN} \text{ and } n + \delta_e(n) = e_b\}$; then, $[e_b - u, e_b - o]$ is the set of values which may become heaped on e_b, and

$$(5) \qquad P(B_b = b_b \wedge E_b \geq e_b)$$

$$= \sum_b P(d_B(B) \cdot Y_B(B) = b_b - b \mid B = b) \cdot P(B = b)$$

$$\cdot \left(\sum_{e=e_b-u}^{e_b-o} \left(P(d_E(E) \cdot Y_E(E) \geq e_b - e \mid E = e) \cdot P(T = e - b + 1) \right) + P(T > e_b - b - o + 1) \right)$$

The components of the right-hand side of these equations consist of probabilities of the true variables or the measurement errors, where $P(\delta_B(B) \cdot Y_B(B) = b_b - b | B=b)$ is equal to the heaping function for the beginning, if $b_b - b = \delta_B(b)$ and one minus the heaping function if $b_b - b = 0$; an analogous interpretation holds for the term $P(\delta_E(E) \cdot Y_E(E) = e_b - e | E=e)$.

For the continuous–time duration model, analogous equations can be derived under the additonal assumption that the duration of unemployment and the densities of all other involved random variables are exponentially distributed within any particular month in calendar time. Compared to the equations for the discrete–time duration model given above, only the density and the distribution function of the duration T have to be changed accordingly and, for the right–censored case, integration has to be performed over the relevant set of e.

4 Estimation

Given the probability measure of the true spell duration, P(T=e–b–1), P(B=b), $G_B(b)$ and $G_E(e)$ as well as the heaping pattern are specified, maximum likelihood estimation on the basis of the equations derived above is, in principle, straightforward. Before we discuss the derivation of P(B=b), $G_B(b)$ and $G_E(e)$ we briefly present the model for the true spell duration of unemployment used in this study.

4.1 Specification of the Duration Model

Microeconometric modelling of unemployment durations focuses on the hazard rate, i.e. the conditional probability of exiting the unemployment state in a particular period of time given the individual has been unemployed until this period (for a survey of the literature see, e.g., Kiefer, 1988, Lancaster, 1990). We restrict the empirical analysis to males for whom only transitions into employment need to be modelled explicitly, because transitions into non–participation for males are of little quantitative importance in our data set. As usual, we treat transitions into other states as right–censored at the date of transition. In our application we use a discrete–time version of the proportional hazard model which is the most popular class of duration models. For this class

of models the hazard rate is defined by the product of the so–called baseline hazard, $\lambda_0(t)$, with a covariable–dependent term, $\exp(\Sigma x_k \cdot \beta_k)$, that is

(6) $h(t) = \lambda_0(t)\exp(\Sigma x_k \cdot \beta_k)$.

The probability of an unemployment spell censored at time t for this model is given by

(7) $S(t) = \Pi(1-\lambda_0(\tau)\cdot\exp(\Sigma x_k \cdot \beta_k))$,

with the product built over $\tau<t$. S(t) is the probability of still being unemployed after *t* periods and is therefore called the "survivor function". The probability of an unemployment spell with completed duration *t* in terms of the hazard rate is given by

(8) $P(T = t) = h(t)S(t) = \lambda_0(t)\cdot\exp((\Sigma x_k \cdot \beta_k)\cdot\Pi(1-\lambda_0(\tau)\cdot\exp(\Sigma x_k \cdot \beta_k))$.

For completed and right–censored unemployment spells measured without error the contribution to the likelihood function is simply given by P(T=t) and S(t), respectively.

For estimation purposes we have to specify the baseline hazard and the vector of covariables. Since the specification of $\lambda_0(t)$ in proportional hazard rate models is often assumed to affect parameter estimates considerably, we model it semi–parametrically here. In particular, we estimate one parameter for each month *t*, restricted by the equations: $\lambda_0(14) = \lambda_0(15)$, $\lambda_0(16) = \lambda_0(17) = \lambda_0(18)$, $\lambda_0(19) = \lambda_0(v)$ for $19 < v \leq 24$ and $\lambda_0(25) = \lambda_0(v)$ for $v>25$. The restrictions have been chosen in such a way that the number of completed durations in each month does not become too small for estimation purposes. Although unavoidable given the available data, these restrictions are admittedly somewhat arbitrary. To test the sensitivity of estimation results with respect to the specification of the baseline hazard rate we also use an alternative parametric specification which allows for a relatively flexible form of duration dependence.

Because a detailed economic analysis of the determinants of individual unemployment durations is not attempted here, the set of explanatory variables only includes some of the more important variables usually found in microeconometric models of unemployment durations. In addition to personal characteristics of the unemployed, such as age, nationality, health status and vocational education, we included other household income than unemployment insurance payments and the regional unemployment rate. The latter variable refers to the reported month of the beginning of the unemployment spell and is therefore particularly suspectable of being affected by reporting errors, given the large seasonal component in German unemployment dynamics. By this choice of

variables some light should be shed on the question which variables are particularly affected by heaping effects in unemployment durations, if any. Table 1 contains a description and summary statistics of the variables in the duration model.

Table 1. Description and summary statistics of variables in the duration model

Variable	Mean/Share	St. dev.
Age ≤ 25 years	0.41	–
25 < Age ≤ 30 years	0.16	–
Age ≥ 50 years	0.11	–
Foreigner	0.41	–
Severely disabled	0.08	–
Married	0.47	–
No vocational qualification	0.42	–
University degree	0.09	–
Other household income (in DM 5,000)	0.46	0.30
Regional unemployment rate/10	0.81	0.27
Beginning of spell in January	0.22	–
End of spell in December	0.29	–
Duration	8.37	11.58
Right–censored cases	0.27	
# Spells	1254	

Note: The number of spells refers to completed and right–censored spells; left–censored spells were excluded. Household income is real gross monthly income minus unemployment benefits in DM 5,000, the regional unemployment rate was divided by 10. These normalizations were used for numerical reasons.

Source: Socio–Economic Panel for West Germany, waves 1 – 9; own calculations.

4.2 Derivation of the Heaping Functions and the Entrance Probability

To take measurement errors in spell durations into account in setting up the likelihood function, $G_B(b)$ and $G_E(e)$ as well as the heaping pattern and P(B=b) must be known a priori or have to be estimated. We tried to parameterize $G_B(b)$ and $G_E(e)$ and estimate their parameters together with those of the duration models, but without much success due to numerical problems arising from the small number of observations in particular duration groups. We therefore derived estimates for the heaping probabilities and P(B=b) using external information in the following way.

External information comes from aggregate data of monthly inflows into and outflows from registered unemployment published by the Federal Labour Office (FLO), of which we assume that they are measured without error. Given the comparability of these data with the aggregate monthly flow data derived from the GSOEP, which has been established in section 2[4], we combine these two data sources to derive empirical counterparts of the functions $G_B(b)$ and $G_E(e)$ and approximate P(B=b) by the relative monthly inflows from the official register data. Since the monthly outflow data published by the FLO is not differentiated by destination[5], we have to assume that the heaping mechanism does not depend on the destination, which does not seem too restrictive an assumption.

To derive $G_B(b)$, let $A_S(b)$ denote the relative frequency of unemployment spells with observed beginning, *b*, in the GSOEP and $A_O(b)$ in the official register data, respectively. Assuming that b is not a heaping point, we have

$$
\begin{aligned}
A_S(b) &= \hat{P}(B_b = b \wedge B = b) = \hat{P}(B_b - B = 0 \wedge B = b) = \hat{P}(\delta_B(B) \cdot Y_B(B) = 0 \wedge B = b) \\
&= \hat{P}(\delta_B(B) \cdot Y_B(B) = 0 \mid B = b) \cdot \hat{P}(B = b) = (1 - \hat{G}_B(b)) \cdot \hat{P}(B = b) = (1 - \hat{G}_B(b)) A_O(b)
\end{aligned} \tag{8}
$$

The heaping function for the beginnings therefore is $\hat{G}_B = 1 - A_S(b)/A_O(b)$, its rationale being as follows. Assume that every beginning (or end) is heaped to one heaping point and the heaping points themselves are not heaped. Let b be a heaping point; then, $A_S(b)$ contains all spells with a true spell beginning in b and those with spell beginnings erroneously reported to be *b*. Therefore, $A_S(b) > A_O(b)$ and $(1 - A_S(b)/A_O(b)) < 0$. On the other hand, if *b* is not a heaping point, the value of the theoretical heaping function at that point must be positive (including the value of zero).

Using this heaping function and the observed values for $A_S(b)$ and $A_O(b)$ the heaping points for the beginnings can be inferred from the data. A similar heaping function with corresponding heaping points can also be derived for the spell endings. Then, it is a simple matter to derive the sets of heaped values under the restriction that the sum of heaped cases out of this set to the heaping point equals the surplus of cases in the GSOEP data compared to the FLO data at the heaping point itself. Since we assumed the heaping functions to have a stable pattern over time, they were calculated for each calendar month as arithmetic

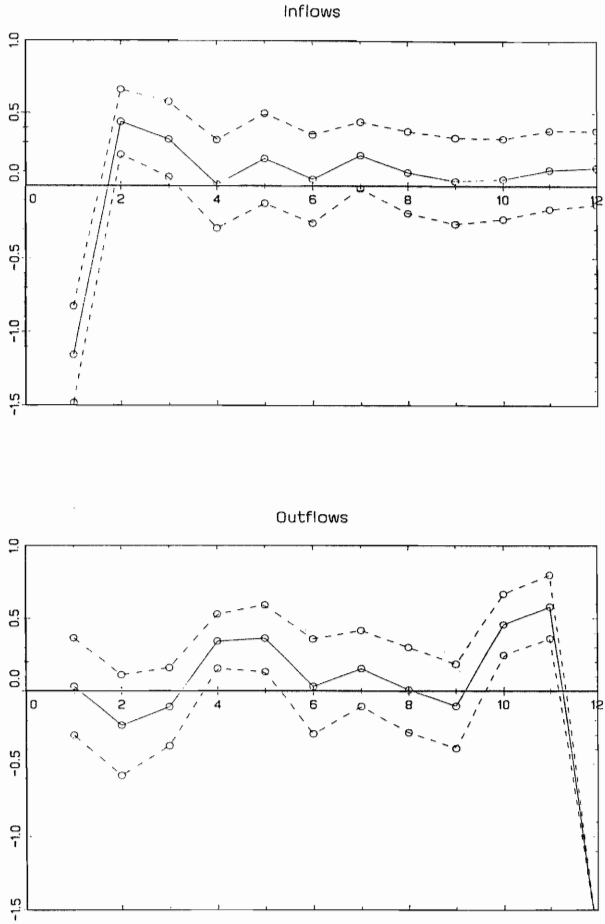
---

4 Note that, even if the levels of inflows and outflows differed between the two data sources, this is of little relevance for deriving the heaping functions if the monthly flows do not differ in relative terms, for which there is no a priori reason.

5 In fact, we had to calculate monthly unemployment outflows from the information on monthly inflows and the stock of unemployment in two consecutive months using the definitional flow–stock relation.

means averaged over all years in the observation period. The resulting heaping functions for the inflows, $G_B(b)$, and outflows, $G_E(e)$, together with their respective confidence bands are plotted in Figure 2, where we have chosen three standard errors to account for potential non–normality of the estimated heaping functions as a conservative test against the alternative hypothesis.

Figure 2. Averaged heaping functions for inflows into and outflows from unemployment

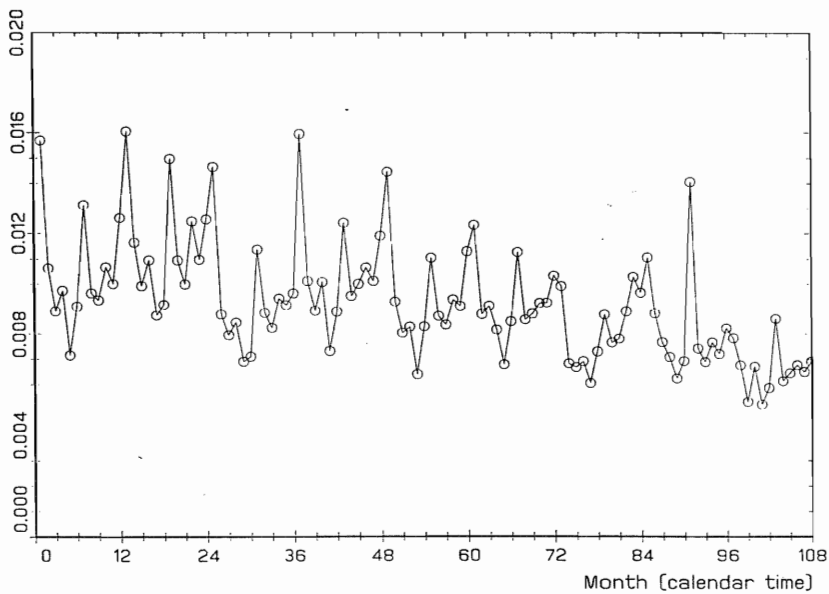


Note: The plots show the heaping functions for inflows and outflows with their respective 3σ–confidence bands (dotted lines); calculations are described in the text.

Source: Federal Labour Office, Amtliche Nachrichten (Official Monthly Bulletin), various issues; Socio–Economic Panel for West Germany, waves 1 – 9.

As Figure 2 shows, except for the heaping point in January the values of the heaping function for the inflows are only significantly different from zero for February and March. This heaping pattern also makes sense intuitively since one would expect heaped values to be distributed in the neighbourhood of the heaping points. The same also holds for the significant values of the outflow heaping function in the months October and November with respect to the heaping point in December. On the other hand, there seems to be no natural heaping point for its significant values in April and May. For estimation purposes we therefore ignore these latter values and only use the other significant point estimates of the heaping functions, that is January = 1, February = 0.4394, March = 0.3207 for the inflows, and October = 0.457, November = 0.5805, December = 1 for the outflows, while the values of the heaping functions are set to zero for all other months.

Figure 3. Monthly entrance probabilities approximated by the relative frequency of spell beginning in the register data



Note: Data are weighted by the relative yearly inflows into unemployment in the GSOEP to adjust for sample attrition. Months refer to the period from January 1983 to December 1991.

Source: Federal Labour Office.

Relative frequencies of monthly inflows into unemployment from the FLO register data as our measure for the entrance probability within the observation period are plotted in Figure 3. To adjust for sample attrition in the GSOEP, the register data have been weighted by the ratio of inflows into unemployment within a given month to all inflows within the observation period. As Figure 3 shows, the entrance probability is slightly decreasing over time and fluctuates considerably between months where the volatility seems to have become somewhat smaller at the end of the observation period.

Given the specification of the hazard function in equation (6) and estimates for $G_B(b)$, $G_E(e)$, and P(B=b), they can be plugged into the respective formulas given by equations (4) and (5) to obtain parameter estimates by maximizing the resulting likelihood function in equation (3') using standard optimization methods.

5 Results

As it is obvious from our derived likelihood function, estimated coefficients in the duration model may be affected by (i) whether the heaping effect is modelled at all and, if so, how the heaping functions are specified, and (ii) how the probability of the true spell beginning, P(B=b), is approximated. Apart from these effects, parameter estimates in the duration model are also supposed to be affected by the way the baseline hazard function is modelled. We have therefore estimated the following discrete–time models:

1A) a model with heaping with P(B=b), $G_B(b)$ and $G_E(e)$ derived as described in the previous section;

1B) a model with the heaping functions $G_B(b)$ and $G_E(e)$ as above, but the values for P(B=b) given by the rectangular distribution over all possible entry months in the observation period;

2A) a model without heaping;

2B) a model with heaping accounted for by two dummies included in the set of covariables indicating if the beginning of an unemployment spell was in January or its end was in December;

2C) a model with an arbitrarily defined heaping function;

3A) a model with heaping as in 1A), but the baseline hazard modelled as a logit transformation of a polynomial in duration, t, i.e.
$\lambda_0(t) = \exp(t + t^2 + 1/t)/(1 + \exp(t + t^2 + 1/t))$;

3B) same as 3A) but without heaping.

We have also estimated the continuous-time analogues of these models. Since estimation results hardly differ from the discrete-time versions reported below, they are not documented here (but are available on request).

Table 2. Estimation results for Models 1A and 1B

Variable	Model 1A		Model 1B	
	Coefficient	t–value	Coefficient	t–value
Baseline (month)				
1	0.1650	6.36	0.1644	6.54
2	0.1853	6.31	0.1871	6.55
3	0.1824	6.24	0.1862	6.31
4	0.1659	5.77	0.1673	5.90
5	0.1510	5.44	0.1520	5.67
6	0.1390	5.03	0.1378	5.24
7	0.1732	5.16	0.1736	5.24
8	0.1149	4.34	0.1159	4.43
9	0.1196	4.13	0.1219	4.21
10	0.0619	3.02	0.0625	3.12
11	0.1423	4.07	0.1416	4.27
12	0.1289	3.71	0.1307	3.70
13	0.1376	3.76	0.1404	3.76
14–15	0.0841	3.76	0.0845	3.79
16–18	0.0959	4.06	0.0966	4.10
19–24	0.0558	3.57	0.0560	3.63
>24	0.0488	3.88	0.0492	3.93
Age ≤ 25 years	0.4088	4.39	0.4013	4.41
25 < Age ≤ 30 years	0.1950	1.98	0.1832	1.85
Age ≥ 50 years	–1.2372	–7.75	–1.2423	–7.81
Foreigner	–0.3381	–4.86	–0.3358	–4.90
Severely disabled	–0.7488	–5.07	–0.7593	–5.21
Married	0.1103	1.35	0.1072	1.36
No vocational qualif.	–0.0970	–1.42	–0.0980	–1.40
University degree	0.4357	3.86	0.4409	3.94
Other household income	0.2245	2.06	0.2300	2.10
Regional unemployment rate	–0.4100	–3.38	–0.4157	–3.60
Log–likelihood	–8618.72		–8687.10	

Note: For Model 1A the heaping functions and the entrance probability into unemployment were derived as described in the text; for Model 1B a rectangular distribution is assumed for the entrance month of a spell.

Estimation results for models 1A) and 1B) are shown in Table 2, for models 2A) to 2C) in Table 3, and for models 3A) and 3 B) in Table 4. As mentioned above, we also tried to parameterize the heaping functions $G_B(b)$ and $G_E(e)$ and estimate their parameters jointly with the parameters of the duration model. While estimates for the latter did hardly change, due to numerical problems (the Hessian became always singular) arising from too small numbers of observations in certain duration categories, we were unable to calculate standard errors for the parameter estimates of the heaping functions; results for this model are therefore not reported here.

As to the specification of P(b), we find hardly any difference in estimated parameters both for the baseline hazard and the explanatory variables in Models 1A and 1B in Table 2. This is a surprising result given that the distribution of actual beginnings of unemployment spells shows pronounced spikes at certain months and also some variation within the observation period. Since this substantial divergence from the assumption that spell beginnings are equally distributed has so little effect on parameter estimates, we may conclude that the specification of P(b) is of secondary importance in duration models with heaping.

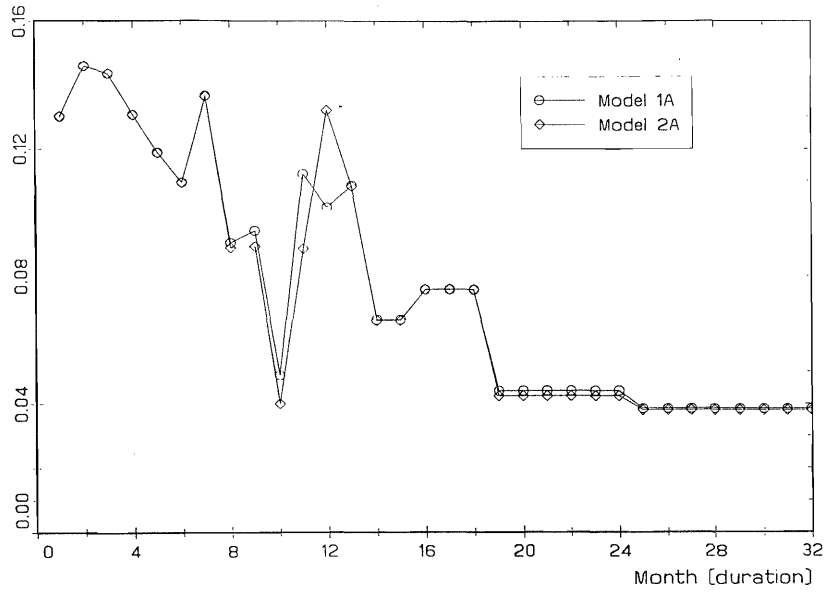
What is even more striking is the fact that there are also hardly any differences in parameter estimates when Model 1A is compared to Model 2A in Table 3, where no heaping at all is taken into account. Not only remain estimated coefficients of variables which are more or less constant within relatively short periods of time, such as personal characteristics and vocational qualification, virtually unaffected by taking into account heaping effects. Estimated coefficients of explanatory variables measured on a monthly basis, such as other household income and the regional unemployment rate, are also hardly affected by accounting for the assumed heaping mechanism in the duration data. There is also hardly any difference in the estimated coefficients of the monthly dummies for the baseline hazards in the two models, the only noticeable exception being the coefficient estimates for the 12th duration month. As shown in Figure 4, which plots the estimated hazard rates for a reference person (as defined in the note to the figure), this is in fact the only noteworthy differerence between the estimates in the two models.

Table 3. Estimation results for Models 2A – 2C

Variable	Model 1A		Model 1B	
	Coefficient	t–value	Coefficient	t–value
Baseline (month)				
1	0.1650	6.36	0.1644	6.54
2	0.1853	6.31	0.1871	6.55
3	0.1824	6.24	0.1862	6.31
4	0.1659	5.77	0.1673	5.90
5	0.1510	5.44	0.1520	5.67
6	0.1390	5.03	0.1378	5.24
7	0.1732	5.16	0.1736	5.24
8	0.1149	4.34	0.1159	4.43
9	0.1196	4.13	0.1219	4.21
10	0.0619	3.02	0.0625	3.12
11	0.1423	4.07	0.1416	4.27
12	0.1289	3.71	0.1307	3.70
13	0.1376	3.76	0.1404	3.76
14–15	0.0841	3.76	0.0845	3.79
16–18	0.0959	4.06	0.0966	4.10
19–24	0.0558	3.57	0.0560	3.63
>24	0.0488	3.88	0.0492	3.93
Age ≤ 25 years	0.4088	4.39	0.4013	4.41
25 < Age ≤ 30 years	0.1950	1.98	0.1832	1.85
Age ≥ 50 years	–1.2372	–7.75	–1.2423	–7.81
Foreigner	–0.3381	–4.86	–0.3358	–4.90
Severely disabled	–0.7488	–5.07	–0.7593	–5.21
Married	0.1103	1.35	0.1072	1.36
No vocational qualif.	–0.0970	–1.42	–0.0980	–1.40
University degree	0.4357	3.86	0.4409	3.94
Other household income	0.2245	2.06	0.2300	2.10
Regional unemployment rate	–0.4100	–3.38	–0.4157	–3.60
Log–likelihood	–8618.72		–8687.10	

Note: For Model 2C the following heaping pattern is assumed: Inflows: January=1.0, February = 0.5, March=0.25, April=0.1, November=0.15, December=0,3; Outflows: January=0.25, February=0.1, September=0.1, October=0.25, November=0.5, December=0.1; for all other months the values of the heaping functions are set to zero.

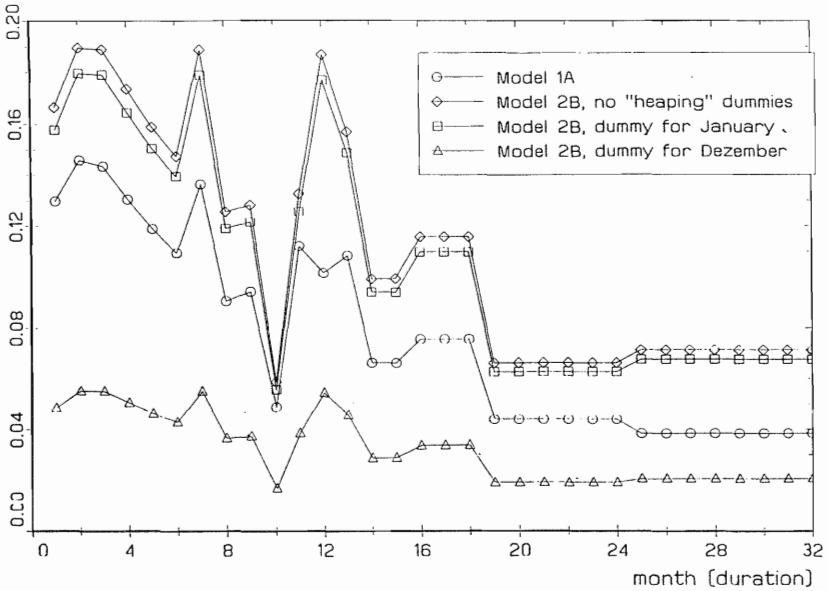
Figure 4. Estimated hazard rates for Models 1A and 2A



Note: Hazard rates are based on estimates in Tables 2 and 3 with explanatory variables evaluated at their respective base categories in case of dummy variables and at sample means for continuous variables.

In contrast, modelling heaping effects by including dummy variables for December and January has a strong impact on estimated hazard rates. While the estimated coefficient on the January dummy is insignificant, the December dummy has a very strong negative impact on the hazard. Both estimated coefficients of particularly explanatory variables and the baseline hazard change in this specification. In particular, the estimated coefficients on the dummy for nationality and the age dummies drop substantially in size or even become insignificant. The effect of the "heaping" dummies on the baseline hazard relative to Model 1A is illustrated in Figure 5. Setting both dummies to zero, implying an unemployment spell which neither beginnings in January nor ends in December, results in a hazard rate lying above the hazard derived from Model 1A throughout. While the former hazard rate is almost identical to that for a spell beginning in January but not ending in December, the hazard rate for a spell ending in that month in Model 2B lies clearly below the hazard in Model 1A throughout.

Figure 5. Estimated hazard rates for Models 1A and 2B



Note: Hazard rates are based on estimates in Tables 2 and 3, where explanatory variables are evaluated at their respective base categories in case of dummy variables and at sample means for continuous variables.

At first sight, it may seem surprising that the inclusion of dummy variables for the Janurary entrance date and especially the December exit date affects personal characteristics but has very little effect on the time–varying covariables, in particular the regional unemployment rate which refers to the beginning of an unemployment spell. This may be explained by the following observation. A comparision of spells ending in December with all other spells shows that the distribution of personal characteristics differ substantially between these two sub–populations. In particular, the shares of older workers and foreigners among spells ending in December are much higher than among spells ending in some other month. Furthermore, more than two thirds of all spells ending in December are right–censored, compared to only about ten percent of all other spells. The average duration of unemployment (including censored spells) is about 13 months in the former and about 7 months in the latter sub–population. Hence, the December dummy picks up selectivity which is not related to the heaping effect at all, as has been assumed by Hujer and Schneider (1989), Hujer, Löwenbein and Schneider (1990) and Hunt (1995) in their respective studies using the GSOEP.

To further test the sensitivity of estimation results with respect to the specification of the heaping functions, in Model 2C we have assumed particular heaping patterns for monthly inflows and outflows on priori grounds (see the note to Table 3). Note that these values differ substantially from the derived heaping pattern used in Model 1A. We would therefore expect a noticeable impact on parameter estimates if heaping were important at all. However, as a comparison of estimated coefficients for Model 2C with those for Model 1B shows, this ad hoc specification of the heaping functions affect neither the estimated coefficients of the explanatory variables nor the baseline hazard in any significant way. We may therefore conclude that for our duration model the exact specification of the heaping function is not very important for estimation purposes.

Table 4. Estimation results for Models 3A and 3B

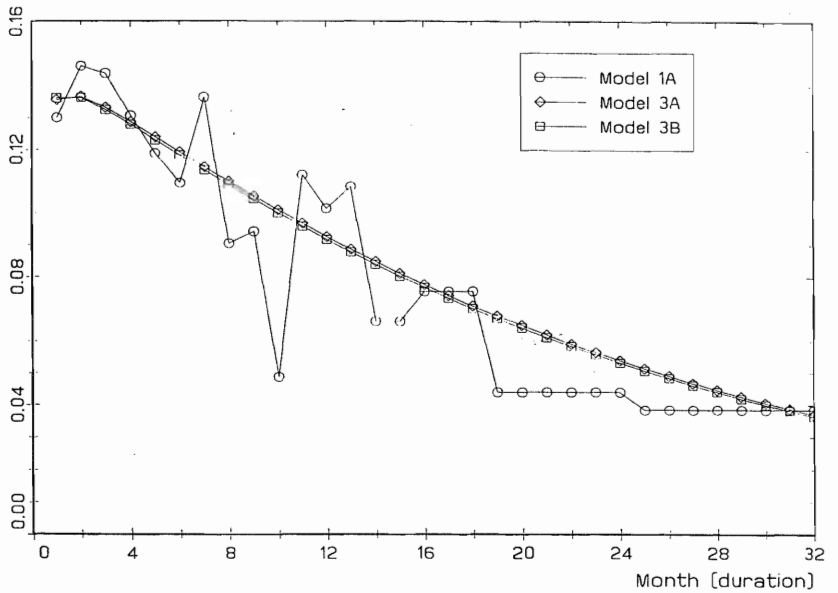
Variable	Model 3A		Model 3B	
	Coefficient	t–value	Coefficient	t–value
Duration	–1.4074	–7.13	–1.4138	–7.25
Duration squared	–0.0504	–7.36	–0.0501	–7.31
1/Duration	–0.1031	–0.70	–0.1190	–0.82
Age ≤ 25 years	0.4092	4.30	0.4092	4.33
25 < Age ≤ 30 years	0.1931	1.91	0.1935	1.90
Age ≥ 50 years	–1.2220	–7.83	–1.2226	–7.83
Foreigner	–0.3428	–4.88	–0.3413	–4.90
Severely disabled	–0.7562	–5.13	–0.7562	–5.13
Married	0.1027	1.24	0.1044	1.27
No vocational qualification	–0.0966	–1.38	–0.0958	–1.40
University degree	0.4397	3.84	0.4382	3.82
Other household income	0.2227	2.02	0.2295	2.09
Regional unemployment rate	–0.4150	–3.41	–0.4015	–3.29
Log–likelihood	–8563.93		–8622.57	

Note: The baseline hazard in Models 3A and 3B is modelled as a logit transformation of a polynomial in duration; see text. In Model 3A the same heaping mechanism as in Model 1A is assumed, in Model 3B heaping effects are not taken into account.

The final comparison of models relates to the specification of the baseline hazard. In Table 4 estimation results for a model with the parametric baseline hazard as specified above and heaping (Model 3A) and a similar model without heaping (Model 3B) are summarized. Comparing estimated coefficients for the explanatory variables in Model 3A with those in Model 1A shows that they are hardly affected by the way the baseline hazard is specified. Furthermore, there is

also almost no difference in estimation results between Models 3A and 3B for both the coefficients of the baseline hazard and of the models' explanatory variables. Hence, estimated hazard rates in these two models are virtually indistinguishable (see Figure 6). However, although the parametric specification of the baseline hazard overall tracks negative duration dependence in the hazard rate quite well, it naturally cannot account for several pronounced spikes in the hazard. In particular, there is no way to account for the large spike at month twelve, part of which is due to heaping.

Figure 6. Hazard rates with parametric and non–parametric specification of the baseline hazard



Note: Plots are based on estimation results in Tables 2 and 4.

Having described the estimation results for the various models the question naturally springs to mind which one is preferrable on statistical grounds. Given the interpretation of the results from above, in our opinion the relevant comparison would be between Model 1A and Model 2A (in any case, there would be very little difference between the former and Model 1B on the one hand, and Model 2A and 2C on the other, whereas we consider Model 2B as clearly misspecified). Since these two models are not nested, the distribution of a standard likelihood ratio test is not known. We have therefore simulated their likelihood ratios under the null ("no heaping") and the alternative hypothesis

("heaping as in Model 1A") with respect to the data generating process. The simulations yielded the following distribution of the likelihood ratio statistics under the null and the alternative hypotheses, respectively.

Table 5. Simulation results for the distribution of the likelihood ratio statistic for Model 1A ("heaping") and Model 2A ("no heaping")

Distribution	H_0	H_1
Lowest 1 % Quantile	322.85	118.90
Lowest 5 % Quantile	340.60	127.78
Upper 95 % Quantile	427.09	184.01
Upper 99 % Quantile	448.19	213.86
Mean	384.27	156.04

Note: The test statistic is $z = 2(ln\ \text{lik}(H_0)/\text{lik}(H_1))$, the simulations are based on 100 replications. Graphical checks have shown that the distributions of the test statistics approximate the normal distribution quite closely both under H_0 and under H_1.

Since large (small) values of z indicate acceptance of H_0 (H_1) the critical region under H_0 at the 1 percent significance level lies in the intervall $]-\infty, 322.85/2]$, while under H_1 the critical region is $[213.86/2, +\infty[$. The empirical value of the test statistic obtained from a comparison of Model 1A and Model 2A is 126.8. Comparing this value with the simulated distributions and the alternative hypotheses shows that the null is rejected against the alternative hypothesis, whereas the alternative case does not lead to a rejection of H_1 at the 1 percent level. We therefore conclude that taking into account heaping as in Model 1A provides a more accurate description of the data generating process.

6 Conclusion

Since a disproportionately large share of all unemployment spells derived from the retrospective monthly calendar data of the German Socio–Economic Panel begin in January or end in December there is a strong a priori reason for the believe that microeconometric duration models estimated on these data may yield severely biased parameter estimates due to heaping effects. In applied work, researchers working with these data have either neglected potential heaping effects altogether or tried to account for them in an ad hoc way. In particular, researchers have included a dummy variable to account for what they call a "December effect" in the GSOEP. As Torrelli and Trivellato (1993a) have shown on the basis of simulation studies, this procedure is likely to severely bias estimated coefficients in continuous-time proportional hazard models. They also suggest that ignoring heaping effects altogether would be preferable to this ad hoc

treatment. In fact, their empirical results based on retrospective unemployment duration data from the Italian labour force survey seem to suggest that the effects of heaping on estimated coefficients in duration models may quantitatively be rather unimportant in practice.

This presumption is also validated by our empirical results for a discrete–time proportional hazard model explaining the duration of unemployment derived from the calendar data in the GSOEP, where we have also allowed for a somwhat more general heaping pattern than that assumed by Torrelli and Trivellato. In particular, we have shown that modelling heaping effects in a statistically consistent way by using external validation information gives a better description of the data generating process than neglecting heaping effects altogether. However, this does hardly affect estimated coefficients of explanatory variables in the duration model. Furthermore, estimation results are rather robust to different specifications of the heaping mechanism. This not only holds with respect to personal characteristics, but also for time–varying covariables and does not depend on the specification of the baseline hazard rate. We therefore conclude that in estimating proportional hazard models of unemployment durations derived from the calendar data of the GSOEP heaping effects may be ignored at relatively little cost, especially if the focus of interest is on the estimated coefficients of the explanatory variables. This strategy is somewhat less secure if spikes in the baseline hazard at particular months are of substantial interest. On the other hand, modelling heaping effects in an ad hoc way, i.e. by including dummy variables for particular months, does indeed not properly take into account heaping, but rather picks up selectivity effects with respect to the composition of outflows from unemployment and censoring.

Appendix: The derivation of the likelihood contributions

In the following, we derive the contributions to the likelihood equation referred to in the text. The numbers below the equation signs refer to the respective assumption set out in section 3 of the text.
The probability that a regular spell is observed is:

$$
\begin{aligned}
&\mathrm{P}(B_b = b_b \wedge E_b = e_b) \\
&= \sum_b \sum_e \mathrm{P}(B_b = b_b \wedge E_b = e_b \wedge B = b \wedge E = e) \\
&= \sum_b \sum_e \mathrm{P}(B_b - B = b_b - b \wedge E_b - E = e_b - e \wedge B = b \wedge E = e) \\
&= \sum_b \sum_e \mathrm{P}(B_b - B = b_b - b \wedge E_b - E = e_b - e \mid B = b \wedge E = e) \cdot \mathrm{P}(B = b \wedge E = e) \\
&= \sum_b \sum_e \mathrm{P}(\delta_B(B) \cdot Y_B(B) = b_b - b \wedge \delta_E(E) \cdot Y_E(E) = e_b - e \mid B = b \wedge E = e) \cdot \mathrm{P}(B = b \wedge E = e) \\
&\underset{4)}{=} \sum_b \sum_e \mathrm{P}(\delta_B(B) \cdot Y_B(B) = b_b - b \mid B = b \wedge E = e) \cdot \mathrm{P}(\delta_E(E) \cdot Y_E(E) = e_b - e \mid B = b \wedge E = e) \cdot \mathrm{P}(B = b \wedge E = e) \\
&= \sum_b \sum_e \mathrm{P}(\delta_B(B) \cdot Y_B(B) = b_b - b \mid B = b \wedge E = e) \cdot \mathrm{P}(\delta_E(E) \cdot Y_E(E) = e_b - e \wedge B = b \mid E = e) \cdot \mathrm{P}(E = e) \\
&\underset{2)}{=} \sum_b \sum_e \mathrm{P}(\delta_B(B) \cdot Y_B(B) = b_b - b \mid B = b \wedge E = e) \cdot \mathrm{P}(\delta_E(E) \cdot Y_E(E) = e_b - e \mid E = e) \cdot \mathrm{P}(B = b \mid E = e) \cdot \mathrm{P}(E = e) \\
&= \sum_b \sum_e \mathrm{P}(\delta_B(B) \cdot Y_B(B) = b_b - b \mid B = b \wedge E = e) \cdot \mathrm{P}(\delta_E(E) \cdot Y_E(E) = e_b - e \mid E = e) \cdot \mathrm{P}(B = b \wedge E = e) \\
&= \sum_b \sum_e \mathrm{P}(\delta_B(B) \cdot Y_B(B) = b_b - b \wedge E = e \mid B = b) \cdot \mathrm{P}(B = b) \cdot \mathrm{P}(\delta_E(E) \cdot Y_E(E) = e_b - e \mid E = e) \\
&\underset{3)}{=} \sum_b \sum_e \mathrm{P}(\delta_B(B) \cdot Y_B(B) = b_b - b \mid B = b) \cdot \mathrm{P}(\delta_E(E) \cdot Y_E(E) = e_b - e \mid E = e) \cdot \mathrm{P}(E = e \mid B = b) \cdot \mathrm{P}(B = b) \\
&= \sum_b \sum_e \mathrm{P}(\delta_B(B) \cdot Y_B(B) = b_b - b \mid B = b) \cdot \mathrm{P}(\delta_E(E) \cdot Y_E(E) = e_b - e \mid E = e) \cdot \mathrm{P}(B = b \wedge E = e) \\
&= \sum_b \sum_e \mathrm{P}(\delta_B(B) \cdot Y_B(B) = b_b - b \mid B = b) \cdot \mathrm{P}(\delta_E(E) \cdot Y_E(E) = e_b - e \mid E = e) \cdot \mathrm{P}(E - B + 1 = e - b + 1 \wedge B = b) \\
&= \sum_b \sum_e \mathrm{P}(\delta_B(B) \cdot Y_B(B) = b_b - b \mid B = b) \cdot \mathrm{P}(\delta_E(E) \cdot Y_E(E) = e_b - e \mid E = e) \cdot \mathrm{P}(T = e - b + 1 \wedge B = b) \\
&\underset{1)}{=} \sum_b \sum_e \mathrm{P}(\delta_B(B) \cdot Y_B(B) = b_b - b \mid B = b) \cdot \mathrm{P}(\delta_E(E) \cdot Y_E(E) = e_b - e \mid E = e) \cdot \mathrm{P}(T = e - b + 1) \cdot \mathrm{P}(B = b)
\end{aligned}
$$

The probability of a right-censored spell is:

$$
\begin{aligned}
&\mathrm{P}(B_b = b_b \wedge E_b \geq e_b) \\
&= \sum_b \sum_e \mathrm{P}(B_b = b_b \wedge E_b \geq e_b \wedge B = b \wedge E = e) \\
&= \sum_b \sum_e \mathrm{P}(B_b - B = b_b - b \wedge E_b - E \geq e_b - e \wedge B = b \wedge E = e) \\
&= \sum_b \sum_e \mathrm{P}(\delta_B(B) \cdot Y_B(B) = b_b - b \wedge \delta_E(E) \cdot Y_E(E) \geq e_b - e \mid B = b \wedge E = e) \cdot \mathrm{P}(B = b \wedge E = e) \\
&\underset{4)}{=} \sum_b \sum_e \mathrm{P}(\delta_B(B) \cdot Y_B(B) = b_b - b \mid B = b \wedge E = e) \cdot \mathrm{P}(\delta_E(E) \cdot Y_E(E) \geq e_b - e \mid B = b \wedge E = e) \cdot \mathrm{P}(B = b \wedge E = e) \\
&= \sum_b \sum_e \mathrm{P}(\delta_B(B) \cdot Y_B(B) = b_b - b \mid B = b \wedge E = e) \cdot \mathrm{P}(\delta_E(E) \cdot Y_E(E) \geq e_b - e \wedge B = b \mid E = e) \cdot \mathrm{P}(E = e) \\
&\underset{2)}{=} \sum_b \sum_e \mathrm{P}(\delta_B(B) \cdot Y_B(B) = b_b - b \mid B = b \wedge E = e) \cdot \mathrm{P}(\delta_E(E) \cdot Y_E(E) \geq e_b - e \mid E = e) \cdot \mathrm{P}(B = b \mid E = e) \cdot \mathrm{P}(E = e)
\end{aligned}
$$

$$= \sum_b \sum_e P(\delta_B(B)\cdot Y_B(B) = b_b - b \mid B = b \wedge E = e)\cdot P(\delta_E(E)\cdot Y_E(E) \geq e_b - e \mid E = e)\cdot P(B = b \wedge E = e)$$

$$= \sum_b \sum_e P(\delta_B(B)\cdot Y_B(B) = b_b - b \wedge E = e \mid B = b)\cdot P(B = b)\cdot P(\delta_E(E)\cdot Y_E(E) \geq e_b - e \mid E = e)$$

$$\underset{3)}{=} \sum_b \sum_e P(\delta_B(B)\cdot Y_B(B) = b_b - b \mid B = b)\cdot P(\delta_E(E)\cdot Y_E(E) \geq e_b - e \mid E = e)\cdot P(E = e \mid B = b)\cdot P(B = b)$$

$$= \sum_b \sum_e P(\delta_B(B)\cdot Y_B(B) = b_b - b \mid B = b)\cdot P(\delta_E(E)\cdot Y_E(E) \geq e_b - e \mid E = e)\cdot P(B = b \wedge E = e)$$

Now, let $u = \max\{\delta_e(n) \mid n \in IN \text{ and } n+\delta_e(n) = e_b\}$ and $o = \min\{\delta_e(n) \mid n \in IN \text{ and } n+\delta_e(n)=e_b\}$; then $[e_b - u, e_b - o]$ is the set of values which may become heaped on e_b. Hence, $P(\delta_E(E)Y_E(E) \geq e_b\text{-}e \mid E=e) = 0$ for $e_b\text{-}e > u$, because from $e < e_b\text{-}u$ the end e will never be heaped to e_b, and $P(\delta_E(E)Y_E(E) \geq e_b\text{-}e \mid E=e) = 1$ for $e_b\text{-}e < o$, because by the definition of u and o, $o \leq 0$ always holds. The above equation can therefore be written as

$$= \sum_b \left(\sum_{e=e_b-u}^{e_b-o} P(d_B(B)\cdot Y_B(B) = b_b - b \mid B = b)\cdot P(d_E(E)\cdot Y_E(E) \geq e_b - e \mid E = e)\cdot P(T = e-b+1 \wedge B = b) \right.$$

$$\left. + \sum_{e=e_b-o+1}^{\infty} P(d_B(B)\cdot Y_B(B) = b_b - b \mid B = b)\cdot P(d_E(E)\cdot Y_E(E) \geq e_b - e \mid E = e)\cdot P(T = e-b+1 \wedge B = b) \right)$$

$$= \sum_b \left(\sum_{e=e_b-u}^{e_b-o} P(d_B(B)\cdot Y_B(B) = b_b - b \mid B = b)\cdot P(d_E(E)\cdot Y_E(E) \geq e_b - e \mid E = e)\cdot P(T = e-b+1 \wedge B = b) \right.$$

$$\left. + \sum_{e=e_b-o+1}^{\infty} P(d_B(B)\cdot Y_B(B) = b_b - b \mid B = b)\cdot P(T = e-b+1 \wedge B = b) \right)$$

$$\underset{1)}{=} \sum_b P(\delta_B(B)\cdot Y_B(B) = b_b - b \mid B = b)$$

$$\cdot \left(\sum_{e=e_b-u}^{e_b-o} P(d_E(E)\cdot Y_E(E) \geq e_b - e \mid E = e)\cdot P(T = e-b+1)\cdot P(B = b) + \sum_{e=e_b-o+1}^{\infty} P(T = e-b+1)\cdot P(B = b) \right)$$

$$\underset{1)}{=} \sum_b P(\delta_B(B)\cdot Y_B(B) = b_b - b \mid B = b)\cdot P(B = b)$$

$$\cdot \left(\sum_{e=e_b-u}^{e_b-o} P(\delta_E(E)\cdot Y_E(E) \geq e_b - e \mid E = e)\cdot P(T = e-b+1) + \sum_{e=e_b-o+1}^{\infty} P(T = e-b+1) \right)$$

$$= \sum_b P(\delta_B(B)\cdot Y_B(B) = b_b - b \mid B = b)\cdot P(B = b)$$

$$\cdot \left(\sum_{e=e_b-u}^{e_b-o} \left(P(\delta_E(E)\cdot Y_E(E) \geq e_b - e \mid E = e)\cdot P(T = e-b+1) \right) + P(T \geq e_b - o + 1 - b + 1) \right)$$

$$= \sum_b P(\delta_B(B)\cdot Y_B(B) = b_b - b \mid B = b)\cdot P(B = b)$$

$$\cdot \left(\sum_{e=e_b-u}^{e_b-o} \left(P(\delta_E(E)\cdot Y_E(E) \geq e_b - e \mid E = e)\cdot P(T = e-b+1) \right) + P(T > e_b - b - o + 1) \right)$$

References

Bowers, N. and F.W. Horvath (1984), 'Keeping Time: An Analysis of Errors in the Measurement of Unemployment Durations', *Journal of Business and Economics Statistics*, **2**, 140–149.

Burkhauser, R. (1991), 'An Introduction to the German Socio–Economic Panel for English Speaking Researchers', *Cross National Studies in Ageing Program Project Paper*, **1**, Syracuse University.

Heitjan, D.F. and D.B. Rubin (1991), 'Ignorability and coarse data', *Annals of Statistics,* **19**, 2244–2253.

Holt, D., J.W. McDonald and C.J. Skinner (1991), 'The Effect of Measurements Error on Event History Analysis', in P.P. Biemer et al., Measurement Errors in Surveys, John Wiley, New York.

Hujer, R. and H. Schneider (1989), 'The Analysis of Labour Market Mobility Using Panel Data', *European Economic Review,* **33**, 530–536.

Hujer, R., O. Löwenbein and H. Schneider (1990), 'Wages and Unemployment. A Microeconometric Analysis for the FRG', in H. König, 'Economics of Wage Determination - Studies in Contemporary Economics', Springer Verlag, Berlin Heidelberg.

Hunt, J. (1995), 'The Effects of Unemployment Compensation on Unemploymet Duration in Germany', *Journal of Labor Economics*, **13**, 88-120.

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

Kiefer, N.M. (1988), 'Economic duration data and hazard functions', *Journal of Economic Literature*, **26**, 646–679.

Lancaster, T. (1992), 'The Econometric Analysis of Transition Data', Cambridge University Press, Cambridge.

Porterba, J.M. and L.H. Summers (1986), 'Reporting Errors and Labour Market Dynamics', *Econometrica*, **54**, 1319–1338.

Torelli, N. and U. Trivellato (1989), 'Youth Unemployment Duration from the Italien Labour Force Survey', *European Economic Review*, **33**, 407–15.

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

Torelli, N. and U. Trivellato (1993); 'Data Inaccuracies and Sampling Plan in a Model of Unemployment Duration', in H. Bunzel, P. Jensen and N. Westergård–Nielsen, 'Panel Data and Labour Market Dynamics', North–Holland, Amsterdam.

Wagner, G., J. Schupp and U. Rendtel (1991), 'The Socio–Economic Panel (SOEP) for Germany – Methods of Production and Management of Longitudinal Data', *DIW Discussion Paper*, **31a**, Berlin.