IFAU - OFFICE OF LABOUR
MARKET POLICY EVALUATION

Attrition and misclassification of drop-outs in the analysis of unemployment duration

Johan Bring, Kenneth Carling

WORKING PAPER 2001:3

This report appeared in the Journal of Official Statistics (2000) Vol. 16:4, pp. 321–330. Permission to reprint it was granted by the journal 17 April 2001.

Attrition and misclassification of drop-outs in the analysis of unemployment duration

JOHAN BRING[∇] and KENNETH CARLING[*]

September, 1998

ABSTRACT

Carling et al (1996) analyze a large data set of unemployed workers in order to examine, inter alia, the effect of unemployment benefits on the escape rate to employment. In this paper we take a closer look at the 20 per cent of workers who were drop-outs and check the empirical justification for modeling attrition as independent right censoring in the analysis of unemployment duration. It may very well be that dropping out, i.e. attrition, often occurs due to employment. In the analysis, we refer to these individuals as misclassified in that they are typically treated as if their unemployment spell went beyond the time of attrition. We propose to follow up the drop-outs by a supplementary sample and apply a Multiple Imputation approach to incorporate the supplementary information. Our follow-up study revealed that 45% dropped out due to employment. The escape rate to employment was as a consequence under-estimated by 20 per cent, implying that the effect of unemployment benefits on the escape rate is likely to be much greater than reported in Carling et al (1996).

KEY WORDS: Follow-up study; Informative censoring; Multiple imputation; Register data; Survival models.

[∇] Department of Statistics, Uppsala University, and [*] IFAU, Office of Labor Market Program Evaluation, P.O. Box 513, SE-751 20 Uppsala, Sweden. This research was partly funded by The Expert Group on Labor Market Policy (EFA). The authors thank Staurt G. Baker, Per-Anders Edin, Tor Jacobson, Roderick J. A. Little, Donald B. Rubin, Adam Taube, Anders Ågren and two anonymous referees for helpful and stimulating comments. Anders Harkman supplied data.

1. INTRODUCTION

Assessment of the probability for an unemployed person to find employment after a certain length of time in unemployment, and the variation in this probability, is currently of great interest. The impact of demographic characteristics, as well as the benefit system is often in focus. Several studies over the last decade deal with this research area. For an early example, see Lancaster (1979), more recent contributions are Meyer (1990) and Narendranathan and Stewart (1993).

A fundamental problem in unemployment duration modeling, known as attrition, is that some workers in the survey drop out for unknown reasons prior to termination of the study. A common approach to circumvent this problem is to assume that the stochastic process underlying exit to employment is independent of attrition and to treat attrition as right censoring (Lagakos, 1979). We suspect that the independence assumption is false because some workers drop out at the time of employment. We refer to them as misclassified in that they are typically modeled as independently right censored while they should be modeled as employed at the time of attrition.

Van Den Berg, Lindeboom and Ridder (1994) and Carling and Jacobson (1995) consider the problem of attrition by applying mixed competing risks models to allow for dependence between unemployment duration and attrition. Hausman and Wise (1979) provide a general discussion of sample attrition bias in econometric longitudinal data analysis. Common to the three papers is that the proposed methods rely on untestable assumptions, or in the words of Fitzmaurice, Heath, and Clifford (1996, p. 249); ”*In general, informative or non-ignorable drop-out models are non-identifiable and arbitrary constraints on the drop-out model must be imposed before carrying out a statistical analysis*”. Baker, Wax, and Patterson (1993) propose instead, in a biostatistical application, to follow up the drop outs by means of a supplementary sample, and incorporate the supplementary sample in the estimation procedure.

In a study on Swedish unemployment register data, Carling, Edin, Holmlund and Harkman (1996) examine, inter alia, the effect of unemployment benefits on the escape rate to employment by comparing workers who receive with workers who do not receive benefits. They report an effect of small magnitude and conclude, partially on basis of this finding, that the Swedish Unemployment Insurance system, including generous benefits in conjunction with a job guarantee for workers who run out of benefits, does not severely distort the job acceptance decision among the unemployed. However, the question arises to what extent the estimates are sensitive to the assumption of uninformative attrition.

Using the same data set, we examine the robustness of this assumption by calculating the escape rate to employment under a set of assumptions about misclassification rate. Furthermore, we seek additional information on the misclassification rate by tracing a random sample of 200 drop-outs drawn from the unemployment register. The selected drop-outs were asked one question; Were you employed (or becoming employed) at the time of attrition? In addition to their replies we have register background information about the drop-outs.

The supplementary sample is used to estimate a misclassification model and the model is then used for imputation in the primary sample. As the final step, an unemployment model is estimated using the original primary sample as well as the imputed version and the resulting estimates are contrasted.

2. DATA AND SENSITIVITY ANALYSIS

The data set consists of a random sample of 12098 unemployed individuals registered at the public employment agencies in Sweden (see Carling et al. 1996). Registration at the agency is compulsory for persons who receive unemployment compensation. For persons not entitled to such compensation registration is voluntary, although registration is necessary for those who want full service from the agency (including access to labor market programs). Survey evidence shows that a large majority of the unemployed does register at the agencies (Statistics Sweden, 1993).

The sample is drawn from the inflow to the unemployment register over a six month period in 1991. The observation period lasted from the time of registration until the end of the first unemployment spell or at the most until September 1993. The spells are measured in days (and later aggregated to four week periods). Background variables, as well as the information about unemployment benefits, are registered at the beginning of the spell. The unemployment rate is obtained for each of the 24 regions in Sweden by taking the average unemployment rate in the region under the period 1991-1993. The regional unemployment rate ranges from 2.1 % in major city regions (i.e. regions including Stockholm, Göteborg and Malmö) to 5.3 % in the northern most region of Sweden.

Table 1. *Sample means of background variables. The means are given for the full sample as well as for subsamples defined by the observed exit or censoring state.*

Variables	*Total sample*	*Individuals observed employed*	*Censored individuals due to attrition*	*Censored individuals due to other exits*[a]
Age 16-24	0.613	0.635	0.555	0.616
Age 25-34	0.231	0.213	0.279	0.227
Age 35-44	0.107	0.098	0.114	0.113
Age45-54	0.049	0.054	0.052	0.044
Female	0.554	0.559	0.554	0.548
Foreign citizenship	0.136	0.098	0.167	0.162
Previous work experience	0.511	0.547	0.505	0.476
Completed high school	0.625	0.689	0.534	0.599
Regional unemployment rate	3.03%	2.97%	2.96%	3.17%
Total number	5340	2237	1013	2090

[a] Other exit states are labor market programs and to leave the labor market. 859 workers entered a labor market program whereas the remaining 1231 decided to leave the labor market.

To be eligible for unemployment benefits and for labor market programs, which are targeted to receivers of unemployment compensation, it is mandatory to remain in the register. As a consequence, receivers tend to remain in the register (attrition is less than 10 %). For non-receivers, on the other hand, the incentives are weaker and the attrition rate is also high (19 %). For this reason, we will focus on attrition amongst non-receivers and have therefore reduced the sample to include only non-receivers, yielding a total number of 5340 individuals. In *Table* 1 we present characteristics for these.

By comparing the two right most columns, we note that, e.g., high school educated have a lower propensity for attrition than those without high school education. Thus, we may expect that we would over-estimate the parameter if misclassification were present.

For an unemployed there are three competing destinations; employment, labor market program, and out of labor force. Unemployment spells ending in one of the latter two destinations will be modeled as independent right censoring for the exit of interest, i.e. for employment (cf. competing risks models). This is a necessary and untestable assumption, which is always required in the empirical analysis of unemployment duration. We will briefly comment on this assumption in the discussion in the end of the paper.

Table 2. *The estimated employment rate (in per cent) at 0-12, 13-26 and 27-52 weeks for selected rates of misclassification*

	Assumed misclassification rate, in per cent			
Weeks	0	20	50	100
0-12	31.9	34.6	38.5	45.1
13-26	7.1	8.0	9.2	11.2
27-52	2.3	2.6	3.0	3.7

Let us now examine the robustness of the estimated employment rates under different assumptions of the attrition mechanism. Table 2 gives the employment rate at three selected time periods; 0-12 weeks, 13-26 weeks, and 27-52 weeks. The employment rate is calculated as the proportion that received employment during the time span under the assumption of no misclassification, 20 per cent, 50 per cent, and 100 per cent misclassification. It goes without saying that the estimates of the coefficients pertaining to the background variables would in case of misclassification also be biased.

3. THE UNEMPLOYMENT MODEL

Job search theory constitutes a framework for the empirical analysis of unemployment duration. Kiefer (1988) and Lancaster (1979) provide accounts of the connection between the theoretical model and the econometric specification in terms of hazard models. Economic theory, however, is not very informative on the precise form of the hazard function. We specify an unrestricted baseline hazard for the duration variable *T* – being the elapsed unemployment until employment - and estimate the model semi-parametrically (see Meyer 1990 and Narendranathan and Stewart 1993). The model is of the proportional hazard variety in which the hazard function of *T* for worker *i* is,

$$\lambda_i\left(t|\mathbf{x}_i\right) = \exp\left(\mathbf{x}_i\mathbf{b}'\right)\lambda_0(t)\,, \qquad (1)$$

where $\mathbf{x}_i\ (1\times k)$ the individual specific covariate vector, $\mathbf{b}\,(1\times k)$ a vector of unknown parameters and, $\lambda_0(t)$ the baseline hazard at time *t* of unknown functional form. Equation (1) is a continuous-time specification. The grouped hazard, for pre-specified units of time, is given by

$$h_i\left(t|\mathbf{x}_i\right) \equiv P\left[T_i < t+1 | t \le T_i, \mathbf{x}_i\right] = 1 - \exp\left\{-\int_t^{t+1} \exp\left(\mathbf{x}_i\mathbf{b}'\right)\lambda_0(u)du\right\}. \qquad (2)$$

The grouped hazard can be written as

$$h_i\left(t|\mathbf{x}_i\right) = 1 - \exp\left\{-\exp\left(\mathbf{x}_i\mathbf{b}'+\eta(t)\right)\right\}, \tag{3}$$

where $\eta(t) = \ln\left\{\int_t^{t+1}\lambda_0(u)du\right\}$. In the application we follow Carling et al (1996) and use four week intervals as the time unit and estimate the hazard for the first 48 weeks (the proportion of spells completed after the 48th week is only 0.5 %). Hence, we take $\eta = \left[\eta(1),...,\eta(12)\right]$. The log likelihood contribution for individual i with observed unemployment duration t_i is

$$\ln L_i(\mathbf{b},\eta) = y_i \ln\left(1-\exp\left\{-\exp\left(\mathbf{x}_i\mathbf{b}'+\eta\left(t_i\right)\right)\right\}\right) - \sum_{s=1}^{t_i-1}\exp\left(\mathbf{x}_i\mathbf{b}'+\eta(s)\right). \tag{4}$$

The log likelihood, the sum of these contributions, is maximized with respect to **b** and η to provide maximum likelihood estimates. The indicator variable y_i equals unity if the duration ended in employment and zero otherwise. Furthermore, let d_i equal unity if employment is reported and zero otherwise. It follows that $d_i = 1$ implies $y_i = 1$, i.e. a reported employment is always an actual employment at time point t_i, whereas the contrary is not necessarily true because some workers may fail to report employment. In the primary sample d_i is observed, not y_i.

4. THE FOLLOW-UP STUDY

The results presented in *Table 2* indicate the importance of correct classification of drop-outs. This can be done by means of external information. Therefore we conduct a follow-up study on the drop-outs from the unemployment register.

Before discussing the design and results from the supplementary study we need an operational definition of drop-out. The directives at the public employment agencies are the following; if a registered unemployed does not appear at an appointment at the agency, then he will be kept in the register for one week. If, during this week, the registered does not make contact with the agency, he will be classified as a drop-out. There is no obligation for the agency to contact the drop-out during this week, nor after. It is natural that many will drop out due to reasons such as illness, traveling, etc. and later re-enter the register. This was checked by tracking drop-outs at different time-points in the data-base. It turned out that about 15% of the drop-outs did re-enter, a surprisingly small proportion implying that a majority of drop-outs are genuinely lost for unknown reasons.

A sample of 200 drop-outs was drawn from the population of drop-outs in January and February 1994. Obviously, it would be preferable to sample from the individuals who dropped out from the 1991 sample. Unfortunately this is not possible, since the addresses of drop-outs are deleted from the register three months after attrition. This may distort the results if the behavior has changed over time. We found in the registers that the proportion of drop-outs in each month is stable over the period 1991-1994. Moreover, the distribution of the background characteristics of the drop-outs in the primary and the supplementary samples are alike.

The individuals in the study were asked the following question; *Were you employed (or becoming employed) at the time of attrition*? In addition to their replies we have, from the register, background information about the individuals.

From the register we have phone numbers to 168 individuals, and among these we obtained a response rate of 95% yielding an overall response rate of 80%. The follow-up study, using trained interviewers, was carried out under three weeks at the total cost of roughly $US 400. In *Table* 3 the result from the follow-up study is presented.

Table 3. *Responses to the question; Were you employed (or becoming employed) at the time of attrition?*

Response	*Frequency*	*Proportions of those responding*
Yes	71	44.7%
No	88	55.3%
Refused to answer	4	
No reply	5	
Phone number missing	32	
Total	200	

The results confirm that a substantial part of drop-outs becomes employed at the time of attrition. The category NO contains those out of labor force as well as those claiming to still be in the register. The proportion in the study claiming to still be in the register (16.3%) coincided well with the previous figure of 15% obtained when trying to track those returning to register.

5. MULTIPLE IMPUTATION AND THE MODEL RE-ESTIMATED

We will now use the additional information obtained from the supplementary sample to predict whether the drop-outs in the primary sample are misclassified and, thereafter, re-estimate the unemployment model. The probability of misclassification, $\pi_{\mathbf{x}} = \Pr\left[Y = 1 \middle| d = 0, \mathbf{x}\right]$, is estimated by means of a logistic regression model. The regression model yields the conditional probability $(\hat{\pi}_{\mathbf{x}i})$ of attrition due to employment for individual *i*. The estimated conditional probability is

$$\hat{\pi}_{\mathbf{x}i} \equiv \hat{P}\left[Y = 1 \middle| d_i = 0, \mathbf{x}_i\right] = \frac{1}{1 + \exp\left(-\mathbf{x}_i \hat{\mathbf{c}}'\right)}, \tag{5}$$

where, again, $\mathbf{x}_i$ is the covariate vector of individual *i* and $\hat{\mathbf{c}}$ is a vector of estimated parameters. To handle the additional uncertainty introduced by the supplementary sample we propose to use multiple imputation by applying the following algorithm (Rubin 1987, p. 170).

(a) Draw $\mathbf{c}^*$from $N\left(\hat{\mathbf{c}}, \hat{\Sigma}_{\hat{\mathbf{c}}}\right)$, and calculate $\pi^*_{\mathbf{x}i}$ based on $\mathbf{c}^*$ and equation (5).

(b) For all drop-outs, draw a random number u_i (U is uniform (0,1)), if $u_i < \pi^*$ then impute $y_i = 1$, otherwise impute $y_i = 0$.

(c) Re-estimate the unemployment model to obtain $\hat{\mathbf{b}}$ and $\hat{V}(\hat{\mathbf{b}})$.

The algorithm is repeated M times and estimates of the parameters and their variances are obtained accordingly (Rubin and Schenker 1986),

$$\hat{b}_r = \bar{b}_r = \frac{1}{M}\sum_{j=1}^{M} \hat{b}_r(j), \qquad r = 1, 2, ..., k, \tag{6}$$

$$\hat{\sigma}^2(\hat{b}_r) = \frac{1}{M}\sum_{j=1}^{M} \hat{V}_j\left(\hat{b}_r\right) + \frac{(M+1)}{M(M-1)}\sum_{j=1}^{M}\left(\hat{b}_r(j) - \bar{b}_r\right)^2 . \tag{7}$$

We adopt the same procedure for the estimated hazard function. M is taken as 5, 10, 20, 50, 100 and 1000. The following results are based on $M = 1000$, we find however that $M = 20$ is sufficient for a practical purpose.

As explanatory variables in the logistic regression model, we use Age, Gender, Citizenship, Work experience, Education, Region in which the individual lives and three interaction terms. For the imputation results it is preferable that the model is as rich as possible. *Table* 4 presents the estimated logistic regression model.

Table 4 *Estimated logistic regression model.*

Variables	*Estimates*	*Stand. errors*	*p-value*
Constant	-1.92	0.95	0.02
Age 16-24	1.12	0.92	0.11
Age 25-34	1.52	1.03	0.07
Age 35-44	-2.16	1.23	0.04
Female	0.05	0.39	0.45
Foreign citizenship	-1.59	0.74	0.02
Work experience	1.49	0.60	0.01
Completed high school	0.65	0.56	0.12
Major city region	0.60	0.36	0.05
Age 25-34 and Work experience	-1.78	0.90	0.02
Age 16-24 and Completed high school	-1.11	0.91	0.11
Age 35-44 and Completed high school	1.67	1.54	0.14

NOTE: n = 159, The reference person is a man aged 45-54, Swedish citizen, has no previous work experience, education less than high school and lives in a non major city region

We find the misclassification rate to be low for, e.g., foreign citizens. Drop-outs from major city regions, on the other hand, have a high rate. The interpretation is twofold; it may be that individuals in major cities are less prone to report employment to agencies, but it may also be that living in a major city means facing a low unemployment rate that implies a high probability of finding a job associated with low unemployment.

In general one might expect the duration of the unemployment spell to be related to the probability of dropping out. Including the duration variable in the regression model, we found a very weak, if any, effect for it. Moreover, the inclusion of the duration variable in the imputation procedure would render the specification somewhat more complex. Given this fact and the weak empirical relevance of the variable, we settled for the model presented in *Table 4*.

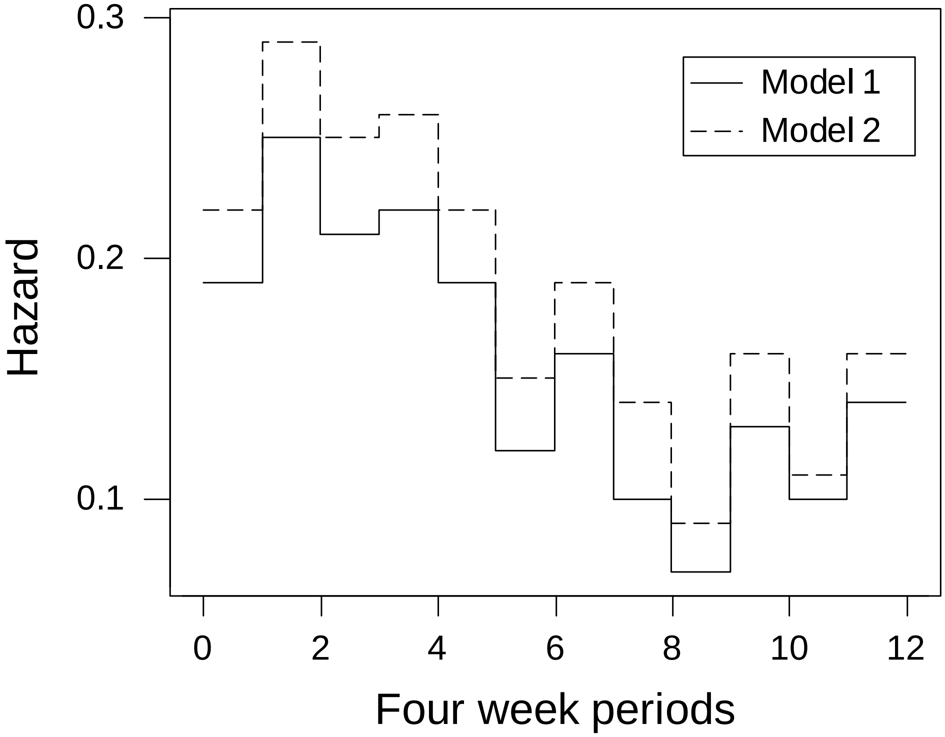
Table 5. *The unemployment model. Model 1 for the original data set and Model 2 for the imputed data set.*

	Model 1		Model 3	
Variables	*Estimates*	*Stand. errors*	*Estimates*	*Stand. errors*
Age 16-24	0.350	0.100	0.416	0.098
Age 25-34	-0.143	0.105	-0.022	0.108
Age 35-44	-0.247	0.114	-0.264	0.117
Female	0.064	0.044	0.061	0.053
Foreign citizenship	-0.328	0.073	-0.393	0.095
Work experience	0.161	0.044	0.200	0.054
Completed high school	0.241	0.048	0.152	0.063
Unemployment rate	-0.202	0.032	-0.219	0.034

NOTE: The empirical baseline hazards are shown in Figure 1.

Table 5 presents the estimates of the unemployment model. Model 1 refers to the results of the original primary sample whereas Model 2 refers to the results of the imputed data set. Comparing Model 2 with Model 1, we note that there are small changes in the estimates of the parameters associated with explanatory variables (all of them within, roughly, one standard error). In our application we have two effects that counter-act. For instance, high school education is negatively correlated with the propensity to drop out (cf. *Table* 1) but positively correlated with the likelihood of misclassification (cf. *Table* 4). Notably, though, if we compare the empirical hazard for Model 1 and Model 2 we see that the hazards differ by about 20 %, see Figure 1.

Figure 1. Empirical baseline hazard for Model 1 and Model 2.



NOTE: The empirical baseline hazards are calculated accordingly, $\hat{h}(t) = 1 - \exp\{-\exp(\hat{\eta}(t))\}$.

We find the multiple imputation approach convenient for incorporating the information of the secondary sample in the analysis. An alternative is to evaluate the likelihood function directly for the two models simultaneously. The two approaches are asymptotically equivalent, provided that the imputation approach is proper (Rubin, 1987 and Rubin and Schenker, 1991). For the complex situation of our data sets we find the imputation approach easier to implement.

6. DISCUSSION

We have discussed the problem of attrition and informative censoring in the context of unemployment duration. Any method aiming at correcting for this problem using only the available information would need to rely on untestable assumptions. We therefore propose to follow up a sub-sample of the drop outs whenever possible. This study was encouraging in that it was inexpensive, easy to conduct and the response rate was high. We believe that follow-up studies may often be conveniently used in empirical analyses of unemployment data.

Our follow up study provided some information on the misclassification rate. There are of course other potential reasons for non-independent censoring of the drop outs. It could be that the non-misclassified drop outs systematically had better or worse labor market prospectives than those remaining in the sample. If this problem is of concern, we suggest some of the models outlined in Carling and Jacobson (1995) or, if possible, a follow up of the type described in Baker et al (1993).

A related problem is whether other competing exits can be assumed independent of the exit of interest. For instance, it is plausible that the presence of labor market programs as an alternative escape route can affected the estimated probability of entering employment, in that workers with poor chances of finding employment may choose the available alternative. This is a minor problem in this application as very few non-receivers exited to labor market programs within the first year of unemployment. However, a large fraction of the workers choose to leave the labor market altogether. We have little hope that it would be possible to address this problem since the exits are mutually exclusive, and hence, even under an ideal situation would it be difficult to distinguish the real affect from potential dependence. The problem is however inherent in survival analysis and sometimes pointed out (see Lancaster, 1990), yet little progress has been made to solve it. By following all workers who left the labor force during 1994 until March 1997, we note that only five per cent did ever return to the labor market. Hence, "out of labor force" is not compatible with employment at a later stage, but constitute an end stage. As such, it is natural to model the destination as independent of the destination of primary interest – employment – and recall that all inference about the employment rate is conditional on the omnipresence of the secondary destination.

We find that the hazard function is underestimated by 20 per cent due to misclassification. A fact that suggests that the implied affect on the employment rate of unemployment benefits has been grossly underestimated in Carling et al (1996). They found a 10 per cent difference between receivers and non-receivers of benefits in the employment rate and concluded that the affect was of marginal importance. A difference of about 30 per cent seems to be a more accurate estimate of the affect and it is doubtful that such affect would be considered marginal to its importance.

REFERENCES

Baker, S, Wax, , and Patterson,, (1993) “Regression analysis of grouped survival data: Informative censoring and double sampling”, *Biometrics*, 49, 379-389.

Carling, K., Edin, P-A, Harkman, A., and Holmlund, B. (1996), "Unemployment Duration, Unemployment Benefits, and Labor Market Programs in Sweden", *Journal of Public Economics*, 59, 313-334.

Carling, K, and Jacobson, T, (1995) “Modeling Unemployment Duration in a Dependent Competing Risks Framework: Identification and Estimation”, *Lifetime Data Analysis*, 1, 111-122.

Fitzmaurice, G. M., Heath, A. F., and Clifford, P. (1996), ”Logistic Regression Models for Binary Panel Data with Attrition”, *Journal of Royal Statistical Society*, A, 159, 249-263.

Hausman, J. A., and Wise, D. A. (1979), "Attrition bias in experimental and panel data: the Gary Income Maintenance Experiment", *Econometrica*, 46, 455-473.

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

Lagakos, S. W. (1979), ”General Right Censoring and its Impact on the Analysis of Survival Data”, *Biometrics*, 35, 139-156.

Lancaster, T. (1979), "Econometric methods for the duration of unemployment", *Econometrica*, 47, 936-956.

Little, R. J. A., and Rubin, D. B (1987), *Statistical Analysis with Missing Data*, New York: John Wiley.

Meyer, B. (1990), "Unemployment Insurance and Unemployment Spells", *Econometrica*, 58, 757-782.

Narendranathan, W., and Stewart, M. B. (1993), "Modeling the probability of leaving unemployment: Competing risks models with flexible base-line hazards", *Applied Statistics*, 42, 63-83.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

Rubin, D. B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse", *Journal of the American Statistical Association*, 81, 361-374.

Rubin, D. B., and Schenker, N. (1991), "Multiple imputation in health-care databases: An overview and some applications", *Statistics in Medicine*, 10, 585-598.

Statistics Sweden, (1993), ”A study of the unemployed according to data from the employment agencies and the labor force surveys”, SCB.

Van Den Berg, G. J., Lindeboom, M., and Ridder, G. (1994), "Attrition in longitudinal panel data, and the empirical analysis of dynamic labour market behavior", *Journal of Applied Econometrics*, 9, 421-435.