Tilastokeskus
Statistikcentralen
Statistics Finland

Tutkimuksia
Forskningsrapporter
Research Reports 260

Statistical Analysis of Survey-based Event History Data

with Application to Modeling of Unemployment Duration

Marjo Pyy-Martikainen

Tilastokeskus
Statistikcentralen
Statistics Finland

Tutkimuksia
Forskningsrapporter
Research Reports 260

Statistical Analysis of Survey-based Event History Data

with Application to Modeling of Unemployment Duration

Marjo Pyy-Martikainen

Päätoimittaja – Chefredaktör – Principal editor:
Timo Alanko

Taitto – Ombrytning – Layout:
Hilkka Lehtonen

ISSN 0355–2071
= Research Reports
ISBN 978–952–244–456–1 (pdf)
ISBN 978–952–244–455–4 (print)

Edita Prima Oy

Helsinki – Helsingfors 2013

Acknowledgements

I am most indebted to my primary supervisor Leif Nordberg, Professor (emeritus) of Statistics and Econometrics, Åbo Akademi University, for his support and patience during the many years that the completion of this dissertation took. His experience in and insight into conducting scientific work has guided me along the way. I am also deeply grateful to my second supervisor Ulrich Rendtel, Professor of Statistics, Free University Berlin. I admire his enthusiasm for panel surveys. Collaboration with him was both fruitful and fun.

The reviewers of this thesis, Peter Lynn, Professor of Survey Methodology, University of Essex, and Risto Lehtonen, Professor of Statistics, University of Helsinki, are warmly thanked for their comments and constructive criticism.

I thank Statistics Finland for the possibility to construct and use longitudinal micro-merged survey-register data, for giving me leaves of absence and for publishing the dissertation. Markku Koivula kindly helped me with the construction of data. The help of Ari Toikka and Seppo Suomalainen with SAS programming was essential for the progress of my work. Mia Kokkinen, Pia Maljanen and Jari Tolonen revised the language of the thesis. The layout of the publication was skilfully done by Hilkka Lehtonen. Hannele Sauli, Veli-Matti Törmälehto and Tuula Kuula explained me various features of ECHP and EU-SILC. To crown it all, Hannele delighted my days by lending me excellent literature from her bookshelves.

I thank my colleagues at the Statistical Methods unit for creating a working atmosphere that is both inspiring and relaxed. This gave me the energy necessary to finish the thesis alongside work.

I am grateful to my parents Kaija and Seppo Pyy for their love and unconditional support. I thank my sister Paula Pallasaho for leading by example and showing me that it is possible to combine successfully family, work and PhD studies. To my children Antti and Emilia I owe my warmest thanks. By bringing so much love and joy into my life they have helped me clarify the priorities of life.

My dear friends: how much more mirthless life would be without you! I wish we will have many hilarious moments ahead.

I thank the School of Statistical Information, Inference and Data Analysis and the Finnish Graduate School in Stochastics and Statistics for the financial support. I am also grateful to the Finnish Work Environment Fund for a grant that allowed me to complete this dissertation.

Helsinki, September 2013

Marjo Pyy-Martikainen

Abstract

Longitudinal surveys are increasingly used to collect event history data on person-specific processes such as transitions between labour market states. Survey-based event history data pose a number of challenges for statistical analysis. These challenges include survey errors due to sampling, non-response, attrition and measurement.

This study deals with non-response, attrition and measurement errors in event history data and the bias caused by them in event history analysis. The study also discusses some choices faced by a researcher using longitudinal survey data for event history analysis and demonstrates their effects. These choices include, whether a design-based or a model-based approach is taken, which subset of data to use and, if a design-based approach is taken, which weights to use.

The study takes advantage of the possibility to use combined longitudinal survey register data. The Finnish subset of European Community Household Panel (FI ECHP) survey for waves 1–5 were linked at person-level with longitudinal register data. Unemployment spells were used as study variables of interest.

Lastly, a simulation study was conducted in order to assess the statistical properties of the Inverse Probability of Censoring Weighting (IPCW) method in a survey data context.

The study shows how combined longitudinal survey register data can be used to analyse and compare the non-response and attrition processes, test the missingness mechanism type and estimate the size of bias due to non-response and attrition. In our empirical analysis, initial non-response turned out to be a more important source of bias than attrition. Reported unemployment spells were subject to seam effects, omissions, and, to a lesser extent, overreporting. The use of proxy interviews tended to cause spell omissions. An often-ignored phenomenon, classification error in reported spell outcomes, was also found in the data. Neither the Missing At Random (MAR) assumption about non-response and attrition mechanisms, nor the classical assumptions about measurement errors, turned out to be valid. Both measurement errors in spell durations and spell outcomes were found to cause bias in estimates from event history models. Low measurement accuracy affected the estimates of baseline hazard most. The design-based estimates based on data from respondents to all waves of interest and weighted by the last wave weights displayed the largest bias. Using all the available data, including the spells by attriters until the time of attrition, helped to reduce attrition bias. Lastly, the simulation study showed that the IPCW correction to design weights reduces bias due to dependent censoring in design-based Kaplan-Meier and Cox proportional hazard model estimators.

The study discusses implications of the results for survey organisations collecting event history data, researchers using surveys for event history analysis, and researchers who develop methods to correct for non-sampling biases in event history data.

Key words: longitudinal surveys, survey errors, event history analysis

Tiivistelmä

Pitkittäisillä surveytutkimuksilla kerätään yhä useammin yksilöitä koskevia tapahtumahistoriatietoja, kuten esimerkiksi tietoja siirtymistä eri työmarkkinatilojen välillä. Tällaisten tietojen tilastollisessa analyysissa tulee ottaa huomioon surveytutkimuksen virhelähteet, joita ovat muun muassa otannasta, kadosta ja attritiosta sekä mittaamisesta johtuvat virheet.

Tässä tutkimuksessa tarkastellaan kadosta, attritiosta ja mittaamisesta johtuvia virheitä surveyaineistoon perustuvissa tapahtumahistoriatiedoissa sekä niiden aiheuttamaa harhaa tapahtumahistoria-analyysissa. Tutkimus käsittelee surveyaineistoon perustuvien tapahtumahistoriatietojen tilastollisessa analyysissa vastaantulevia valintoja ja niiden vaikutusta tuloksiin. Tällaisia valintoja ovat: valitaanko asetelma- vai malliperusteinen lähestymistapa; mitä osaa aineistosta hyödynnetään sekä valittaessa asetelmaperusteinen lähestymistapa, mitä painoja käytetään.

Tutkimuksessa hyödynnetään yhdistettyä pitkittäistä survey-rekisteriaineistoa. Eurooppalaisen elinolotutkimuksen (ECHP, European Community Household Panel) Suomea koskeva, tutkimuskerrat 1–5 kattava aineisto yhdistettiin henkilötasolla rekisteripaneeliaineistoon. Tutkimusmuuttujina olivat työttömyysjaksojen kestot.

Tutkimuksessa tarkastellaan lisäksi simulointimenetelmin käänteisen sensurointitodennäköisyyden painotusmenetelmän (IPCW, Inverse Probability of Censoring Weighting method) tilastollisia ominaisuuksia surveyaineistoon perustuvassa elinaika-analyysissa.

Tutkimuksessa näytettiin, kuinka yhdistettyä pitkittäistä survey-rekisteriaineistoa voidaan hyödyntää kadon ja attrition analysoinnissa, puuttuneisuuden mekanismin testaamisessa sekä kadosta ja attritiosta johtuvan harhan estimoinnissa. Ensimmäisen tutkimuskerran kato osoittautui tutkimuksen empiirisissä analyyseissa attritiota merkittävämmäksi harhan lähteeksi. Surveyvastauksiin perustuvissa työttömyysjaksoissa esiintyi jaksojen alkujen ja loppujen kasautumista viiteajankohtien ääripäihin, jaksojen raportoimatta jättämistä ja jossain määrin myös yliraportointia. Raportoimatta jättämisen todennäköisyys oli yhteydessä sijaisvastaajan käyttöön. Työttömyysjaksojen päättymissyissä esiintyi luokitteluvirheitä. Empiiristen analyysien perusteella klassiset oletukset mittausvirheistä tai oletukset puuttuneisuuden satunnaisuudesta (MAR, Missing At Random) eivät pitäneet paikkaansa. Sekä työttömyysjaksojen kestoon että päättymissyihin liittyvät mittausvirheet aiheuttivat harhaa tapahtumahistoria-analyysin tuloksiin. Työttömyysjaksojen alhainen mittaustarkkuus aiheutti eniten harhaa perushasardifunktion estimointiin. Empiiristen analyysien perusteella harhaisimpia olivat kaikkiin tutkimusaaltoihin vastanneiden henkilöiden osa-aineistoon perustuvat, viimeisen tutkimusaallon painoilla painotetut asetelmaperusteiset estimaatit. Aineiston laajentaminen kattamaan kaikki vähintään ensimmäiseen tutkimusaaltoon vastanneet henkilöt pienensi harhaa. Simulointitutkimuksen tulosten perusteella asetelmapainojen IPCW-korjaus pienentää asetelmaperusteisten Kap-

lan-Meier- ja Coxin verrannollisten hasardien mallin kovariaattivaikutusten estimaattorien informatiivisesta sensuroinnista aiheutuvaa harhaa.

Tutkimuksen tulosten merkitystä arvioidaan tapahtumahistoriatietoa surveytutkimuksilla keräävien organisaatioiden ja tietoja käyttävien tutkijoiden sekä menetelmäkehittäjien näkökulmasta.

Avainsanat: Pitkittäiset surveytutkimukset, surveytutkimuksen virhelähteet, tapahtumahistoria-analyysi

Sammanfattning

Longitudinella surveyundersökningar används mer och mer för att samla in händelsehistorik för individer, såsom t.ex. uppgifter om förändringar i arbetsmarknadsstatus. Vid statistisk analys av sådana uppgifter bör man beakta felkällorna i surveyundersökningar, dvs. fel som beror på användningen av urval, förekomsten av bortfall och attrition samt dålig mätprecision.

I denna undersökning granskas förekomsten av fel i händelsehistorikdata som beror på bortfall, attrition och mätfel samt den bias som orsakas av dessa i analysen av datat. I undersökningen analyseras vilka val man ställs inför vid statistisk analys av survey-baserat händelsehistorikdata och vilken inverkan olika lösningar har på resultaten. Bland de frågor som forskaren ställs inför kan nämnas huruvida man ska välja ett designbaserat eller ett modellbaserat betraktelsesätt, vilken del av materialet som skall användas samt, vid val av ett designbaserat betraktelsesätt, vilka vikter som skall användas.

I undersökningen utnyttjas ett kombinerat longitudinellt survey-registermaterial. Det finländska materialet i den europeiska undersökningen om levnadsförhållanden (ECHP, European Community Household Panel) i omgångarna 1–5 kombinerades på individnivå med longitudinellt registermaterial. Undersökningsvariabler var längden på arbetslöshetsperioderna.

I undersökningen granskas även effekterna av att utnyttja vikter som bygger på censureringssannolikheterna i olika faser av datainsamlingsperioden (IPCW, Inverse Probability of Censoring Weighting).

I undersökningen visas hur ett kombinerat longitudinellt survey-registermaterial kan utnyttjas vid analys av bortfall och attrition, testning av olika antaganden om typen av bortfall samt estimering av bias på grund av bortfall och attrition. Empirisk analys visade att bortfallet vid första undersökningsomgången var en mer betydande källa till bias än attritionen. Respondenternas svar angående början och slutet på arbetslöshetsperioder tenderade att i viss mån koncentera sig till början och slutet av referensperioderna. Vissa perioder rapporterades inte alls, men å andra sidan noterades även överrapportering i någon mån. Sannolikheten för att en period blev orapporterad var större när man intervjuade en annan person istället för intervjupersonen. Klassificeringsfel förekom ifråga om orsakerna till avslutade arbetslöshetsperioder. Varken de klassiska antagandena om mätningsfelens egenskaper eller bortfallets slumpmässighet (MAR, Missing At Random) visade sig vara valida. Mätningsfelen i såväl längden av arbetslöshetsperioderna som orsakerna till att de tog slut gav upphov till bias i analysresultaten. Den låga precisionen i mätningen arbetslöshetsperiodernas längd orsakade särskilt mycket bias vid estimeringen av baslinjehasarden. Designbaserade estimat baserade på det delmaterial som omfattade endast de personer som besvarat alla undersökningsomgångar och viktade med den sista undersökningsomgångens vikter uppvisade mest bias. En utvidgning av materialet till att omfatta alla personer som besvarat minst den första undersökningsomgången minskade biasen. På basis av resultaten från simuleringsundersökningen minskar en IPCW-korrigering av designvikterna den bias som orsakas av informativ censurering vid

design-baserad estimering av parametrarna i Kaplan-Meiers modell och i Cox proportionella hasardmodell.

I undersökningen redogörs även för hur organisationer som samlar in survey-baserat händelsehistorikdata, forskare som utnyttjar uppgifterna och metodutvecklare kan dra nytta av resultaten i denna studie.

Nyckelord: longitudinella surveyundersökningar, felkällorna i surveyundersökningar, analys av händelseförlopp

Contents

List of original publications

Articles are included in the thesis with permissions from the publishers.

1 Introduction

Longitudinal surveys are increasingly used to collect event history data on person-specific processes such as transitions between labour market states. Event history data collected by longitudinal surveys pose a number of challenges for statistical analysis. These challenges include, survey errors due to sampling, non-response, attrition and measurement.

Survey errors are problematic because they diminish the accuracy of estimates. The concept of total survey error [5] provides a theoretical framework for survey errors. For empirical analysis of errors in survey data, combined survey register data are considered a valuable tool [6, 7]. However, combining survey data with register data may be time-consuming and costly. Also, legal and ethical problems may be involved. Therefore, only few studies use this method to assess errors in event history data.

This study takes advantage of the possibility to use combined longitudinal survey register data. Finnish subset of European Community Household Panel (FI ECHP) data for waves 1–5 were linked at person-level with longitudinal register data. Unemployment spells were used as study variables of interest.

The study deals with non-response and measurement errors and the bias caused by them. The study shows how longitudinal combined survey register data can be used to conduct an analysis of non-response and attrition in longitudinal survey data. The study makes a contribution to the pool of evidence on the existence, determinants and effects of non-response, attrition and measurement errors in event history data based on longitudinal surveys. The study also assesses statistical properties of the Inverse Probability of Censoring Weighting (IPCW) method in design-based survival analysis in the presence of dependent censoring.

A researcher using longitudinal survey data for event history analysis has to make several choices that affect the results of the analysis. These choices include the following: whether a design-based or a model-based approach is taken, which subset of data to use and, if a design-based approach is taken, which weights to use. These choices are discussed in [8, 9, 10]. However, the effect of these choices in event history analysis have not been assessed yet with combined survey register data. This study makes a contribution by providing empirical evidence on the effect of these choices.

2 Background

2.1 Examples of longitudinal surveys

The first longitudinal social surveys were launched during the late 1960's and early 1980's in the UK, USA and Germany. Canada launched a number of panel surveys in the early 1990's. The first EU level household panel was launched in 1994. During the 2000's, new panel surveys have been launched in Australia (2001), New Zealand (2002) and South Africa (2008) [11].

The Panel Study of Income Dynamics (PSID, www.psidonline.isr.umich.edu) run by the University of Michigan is the pioneer of household panel surveys. Launched in 1968 and still running, it is the longest running household panel survey in the world. Its original focus was on income and poverty dynamics but its study topics have been extended to cover areas such as labour force and residential dynamics.

Survey of Income and Program Participation (SIPP, www.census.gov/sipp) run by the US Census Bureau since 1984 is another long-running panel survey in the USA. It has a rotating design with panels ranging from 2.5 to 4 years. It was mainly designed to measure the effectiveness and future costs of government transfer programs such as the food stamps program.

German Socio-Economic Panel (SOEP, www.diw.de/en/soep) launched in 1984 and run by the German Institute for Economic research is the European pioneer of household panel surveys. A special feature of SOEP is that it follows all persons ever interviewed, regardless of their relationship to the original sample persons [12].

The British Household Panel Survey (BHPS, www.iser.essex.ac.uk/bhps) was run by the Institute for Social and Economic Research during 1991–2008, with 18 yearly data collection waves. The BHPS sample was incorporated in 2010 in the second round of a new household panel survey, *Understanding Society* (http://www.understandingsociety.org.uk).

In the early 1990's, Statistics Canada launched several longitudinal surveys, including the *Survey of Labour and Income Dynamics* (SLID, see www.statcan.gc.ca/imdb-bmdi/3889-eng.htm). SLID is a rotating panel survey with new six-year panels beginning every three years [8]. One of the main aims of SLID is to support analyses of income mobility and labour market dynamics.

An ambitious multicountry panel survey, *the European Community Household Panel* (ECHP, epp.eurostat.ec.europa.eu/portal/page/portal/microdata/echp), was launched in 1994. The key features of ECHP are its comparability across countries, achieved by input-harmonisation, as well as the wide range of topics covered. The ECHP was carried out by national data collection units, mostly national statistical institutes, with the Statistical Office of the European Communities (Eurostat) providing centralised support and coordination [13]. Finland started compiling ECHP data in 1996, a year after becoming a member of the EU. The ECHP was designed for the analysis of individual change over time and in this respect, it can be claimed to be the first real Finnish longitudinal social

survey. The panel was run until 2001, resulting in 6 annual waves. The Finnish ECHP survey is described in [14].

The European Union Statistics on Income and Living Conditions (EU-SILC, epp.eurostat.ec.europa.eu/portal/page/portal/microdata/eu_ silc) was launched in 2003. Like ECHP, it is a multicountry household panel survey designed and coordinated by Eurostat. The design is however output harmonised, giving national data collection units more freedom with respect to implementing the survey. Most EU-SILC countries implement a rotating panel design with a new four-year panel beginning each year, reflecting the fact that cross-sectional estimates are considered as of primary importance. EU-SILC has been compiled in Finland since 2003. Finland has recently adopted the recommended 4-year rotating panel design.

The Millennium Cohort Study (MCS, www.cls.ioe.ac.uk) run by the Centre for Longitudinal Studies is the most recent of UK's four ongoing national longitudinal birth cohort studies. The study has been tracking the Millennium children born in UK through their early childhood years and plans to follow them into adulthood.

2.2 Collection of event history data by longitudinal surveys

Many longitudinal surveys collect event history data related to person-specific processes such as fertility, income and labour market dynamics. Event history data consists of information about durations of spells in a state of interest (such as poverty, unemployment, having no children), the outcome of the spell (transition to non-poverty, to employment or out of labour force, birth of first child), as well as a set of covariates explaining the durations and outcomes. Event history data can be collected retrospectively by using either a multi-state or an event occurrence framework [15]. In the multi-state framework the time period of interest is split into shorter time intervals and for each interval, the state occupied by the person is determined. The event occurrence framework asks for dates of specific events such as transitions between the states of interest.

PSID uses an event occurrence framework to collect information on residence and labour force status histories. The timing of transitions between different states are recorded at the accuracy of one third of a month. These data are converted into month level information in the public release dataset. [16]. SLID uses the event occurrence framework for information on job and jobless spells during the year preceding the interview. SIPP collects information about spells on food stamps program and spells without health insurance by using a multistate framework where the 4-month reference period is split into time intervals of one month. EU-SILC uses a multi-state framework very similar to that used in ECHP to collect month-level labour market state information for the year preceding the interview.

2.3 Errors in longitudinal surveys

A major objective in the design of any survey is to maximise the accuracy of key estimates, given cost and time constraints [17]. The concept of total survey error [5] provides a theoretical framework for assessing accuracy of survey estimates. The following discussion on total survey error bases on [17]. Total survey error refers to the accumulation of all errors that may arise in the design, collection, processing and analysis of data. Survey errors are problematic because they diminish the accuracy of estimates. The objective of maximising accuracy is equivalent to minimising total survey error.

Total survey errors can be decomposed into sampling errors and non-sampling errors. Sampling errors arise because a survey measures only a subset of the population of interest. Even if the total population was measured, the estimates would contain errors due to survey non-response and deficiencies in the specification of survey questions, frame, measurement or data processing. These errors are called non-sampling errors. Non-sampling errors can be viewed as mistakes or unintentional errors that can be made at any stage of the survey process whereas sampling errors are intentional in the sense that their magnitude can be controlled [18]. Each of the error sources may contribute a variable error, a systematic error, or both. Variable errors are reflected in the variance and systematic errors in the bias of an estimate.

Total survey error is usually measured in terms of mean squared error (MSE). Each estimate has a corresponding MSE reflecting the effects of all error sources [18]. The mean squared error of an estimate $\hat{\beta}$ is defined as the expected squared difference between the estimate and the value of the target parameter β, the expectation being taken over all possible realisations of the survey process:

$$\mathrm{MSE}(\hat{\beta}) = \mathrm{E}(\hat{\beta} - \beta)^2.$$

Mean squared error can be decomposed into squared bias and variance:

$$\mathrm{MSE}(\hat{\beta}) = \left[\mathrm{E}(\hat{\beta}) - \beta\right]^2 + \mathrm{E}\left[(\hat{\beta} - \mathrm{E}(\hat{\beta}))^2\right] = \left[Bias(\hat{\beta})\right]^2 + Var(\hat{\beta}).$$

Both the bias and the variance components can be further decomposed according to the error source. Biemer and Lyberg [18] use the following decomposition reflecting the most important sources of bias and variance:

$$\mathrm{MSE}(\hat{\beta}) = \left[Bias_{spec}(\hat{\beta}) + Bias_{nr}(\hat{\beta}) + Bias_{fr}(\hat{\beta}) + Bias_{meas}(\hat{\beta}) + Bias_{dp}(\hat{\beta})\right]^2 + Var_{samp}(\hat{\beta}) + Var_{meas}(\hat{\beta}) + Var_{dp}(\hat{\beta}),$$

the subscripts *spec, nr, fr, meas, dp* and *samp* referring to errors due to specification, non-response, frame, measurement, data processing and sampling.

Estimation of MSE is a complex and costly process. Therefore, usually only a few of the most important components are estimated [18]. This study deals with non-response and measurements errors and the bias caused by them.

2.3.1 Non-response errors

A non-response error is caused by unsuccessful attempts to obtain the desired information from eligible units. The failure to obtain any information at all from an eligible unit results in unit non-response whereas item non-response refers to a situation where a responding unit fails to answer some questions. Our focus is on unit non-response. Hereafter, unit non-response is called simply non-response.

In longitudinal surveys, non-response may occur in three different patterns. *Total non-respondents* provide data for none of the survey waves. *Attrition non-respondents* drop permanently out of the survey at some wave after the first, while *temporary non-respondents* return to the survey after missing one or more waves [19].

Non-response errors in event history data are manifested in three different ways: due to total non-response, attrition and temporary non-response, spells may *not* be *observed* at all. Attrition and temporary non-response may cause *right-censoring* of spells. In this case the follow-up ends before the end of the spell, leaving the ending date of the spell and its outcome unknown. Temporary non-response may also cause *left-truncation* of spells. A spell is left-truncated if it has begun before the start of the follow-up period. Longitudinal surveys usually follow individuals over a fixed follow-up time with pre-specified start and end dates. Right-censoring and left-truncation may also occur because of the fixed follow-up time, a reason not related to non-response.

Figure 1 demonstrates the different non-response errors in event history data created by the different non-response patterns. The follow-up time is the time period [0,3]. Interviews are conducted at time points t=1, t=2 and t=3. At the time t interview, information about spells and covariates are collected for the time period $(t-1,t]$. Person a responds in all three interviews. He has three spells, the first being left-truncated by the start of the follow-up period and the third being right-censored by the end of follow-up period. His second spell is completely observed. The spells by person b are not observed due to total non-response. Person c attrits at wave 2 and therefore, only his first spell, the spell ongoing during (0,1] is observed. The spell is right-censored due to attrition at time t=1. His second spell is

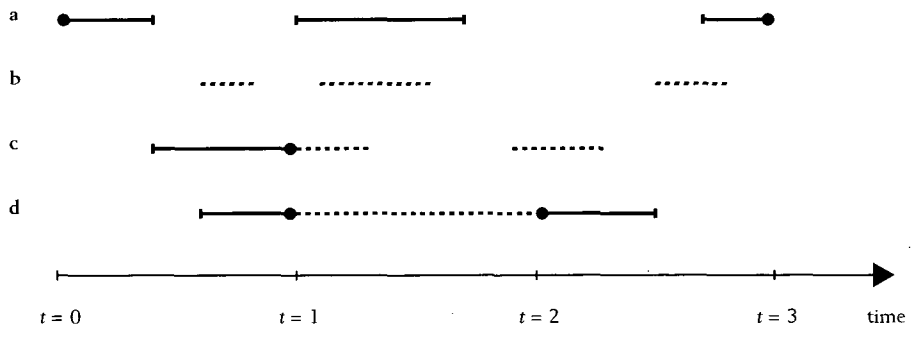


Figure 1 Non-response errors in event history data.
Solid line: part of spell observed in survey. Dashed line: Part of spell not observed by non-response. Observed starting and ending dates of spells are marked by ticks. Left-truncation and right-censoring are marked by circles.

not observed due to attrition. Person **d** misses the wave 2 interview and is, therefore, a temporary non-respondent. The spell by person **d** is observed as two spells, the first spell being right-censored and the second spell being left-truncated.

It is usually assumed that data are Missing At Random (MAR) [20]. The MAR assumption in the context of event history data is discussed in [1]. In this context, the MAR assumption means that non-response is independent of current and future events, given past events and covariates. The assumption of an independent right-censoring mechanism discussed in [3] is equivalent to the MAR assumption. If the MAR assumption does not hold, i.e. if data are Missing Not At Random (MNAR), one has to model the missing data mechanism in order to get unbiased estimates.

An example of a MNAR mechanism is a situation where persons with long unemployment spells drop out from the survey more frequently than otherwise similar persons with shorter spells. In this case, falsely assuming a MAR missingness mechanism leads to biased estimates of the distribution of unemployment duration. If, in addition, the covariate effects differ among persons with long and short spells, there will be bias in the estimated covariate effects, too. The validity of the MAR assumption is usually impossible to test because the values of study variables are unobserved from the time of non-response. [1].

Either weighting or imputation may be used to correct for non-response. Both of these methods rely on the MAR assumption. As it is very difficult to impute all items in a missing wave without distorting associations between survey variables, weighting is usually the preferred method to correct for unit non-response [19]. Survey data sets are usually equipped with weights aiming to correct for non-response. These weights are to be used with all variables included in the survey data set. It is not always clear, however, how to use weights in event history analysis, see [9, 4].

In longitudinal surveys collecting event history data, specific weights may need to be developed to account for dependent censoring that violates the independency asssumption, see e.g. [21]. Dependent censoring means that the probability of censoring is related to the length of the spells of interest. Dependent censoring may cause a bias in estimates from event history analysis. Robins [22] proposed an Inverse Probability of Censoring Weighting (IPCW) method to adjust for bias in survival analysis due to dependent censoring. Lawless [23] discussed the use of the IPCW method in a complex survey data context.

Sample selection models aim to correct for non-response that is MNAR. These models are mainly used in the analysis phase and not in the production phase of survey data. The studies by van den Berg, Lindeboom and Ridder [24] and van den Berg and Lindeboom [25] are early examples of sample selection modeling of labour market transition data.

2.3.2 *Measurement errors*

A measurement error is the discrepancy between the observed value of a variable provided by the survey respondent and its underlying true value. Measurement errors in event history data are manifested as a failure to report a spell (omission), reporting a spell that did not occur (overreporting) and misreporting the duration of a spell (misdating) [26, 27]. In event history data, misdating is typically mani-

fested as the heaping of spell starts and ends at the seam between two reference periods, a phenomenon called the seam effect. Even though spell outcomes may also be misreported, this topic has received little attention in the literature. [2].

Because of measurement errors, the true spell durations T^* are not observed in the survey. The reported durations T can be thought of as consisting of the true duration and a measurement error ϵ: $T = T^* + \epsilon$.[1] Referring to Figure 1, Figure 2 demonstrates measurement errors in reported spell durations. Thick lines show the true durations T^* and thin lines the measurement error ϵ in the spell duration. At the first interview at time t=1, person **a** reported a spell that did not occur (overreporting). At the second interview at time t=2, he misdated the start of the spell. Person **b** is a total nonrespondent and, therefore, reports no spells. Person **c** correctly reports his first spell. The first spell is right-censored and the second spell not reported due to attrition at wave 2. Person **d** omitted his first spell. The true duration T^* and measurement error ϵ cancel each other out so that T=0.

According to the classical assumptions [28, 7], measurement errors ϵ have zero mean and are independent of each other, true durations T^* and any covariates explaining T^*. Under a linear regression model, classical measurement errors in the dependent variable do not cause bias in the estimates of regression coefficients [7]. If the model specified is nonlinear or measurement errors are not classical, bias may result. The validity of the classical assumptions is usually impossible to test because the true durations T^* are not observed.

Skinner and Humphreys [29] and Augustin [30] proposed methods to correct for measurement errors in spells. A common feature of the methods proposed is that they rely on rather restrictive assumptions: that spells are generated from certain parametric duration models, there is no censoring and measurement errors satisfy the classical assumptions.

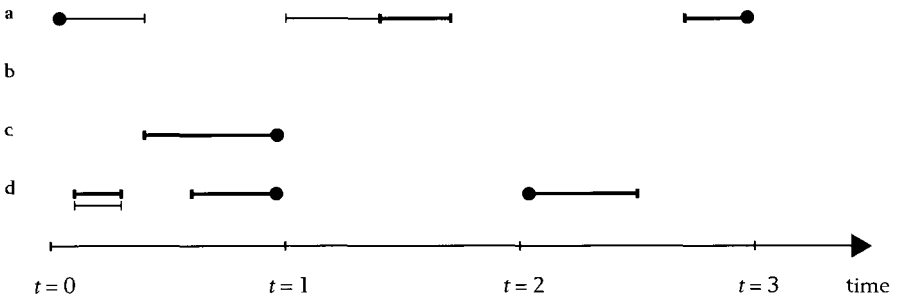


Figure 2: Measurement errors in event history data.
Thick line: true duration. Thin line: measurement error.

2.3.3 *Estimating non-response and measurement error biases*

In practice, only one realisation of the survey process is observed. Therefore, $Bias(\hat{\beta})$ is unknown. Bias may, however, be estimated using the deviation of the value of $\hat{\beta}$ obtained from the survey and the true value β. The true value is un-

1 Subscripts indicating individual and spell ignored.

known but sometimes additional, gold standard data are collected, which for evaluation purposes are considered to be the truth [18].

A reinterview study revisits respondents from the original survey sample and asks some of the questions that were asked in the original survey. Reinterview questions are designed to reference the same time period as in the original interview. The goal is to obtain higly accurate responses that can be used to estimate the true value of the parameter. Then, bias due to measurement error can be estimated as a difference between estimates from the original survey and the reinterview survey. This approach is, however, not without problems. The longer the recall period, the more erroneous the responses tend to be [7]. Also, it is likely that not all respondents from the original survey respond to the reinterview survey. The estimate from the reinterview survey may thus be plagued by non-response bias. As a consequence, reinterview data may be as erroneous as the data that is being evaluated [17].

An external validation study compares survey estimates with external estimates that are considered to be more accurate. External estimates may be obtained from administrative records or from a survey that is considered to be a gold standard for the estimate being evaluated [18]. External validation studies may suffer from differences in target populations or definitions of variables of interest in the two data sources. Moreover, external validation studies do not allow the decomposition of bias due to different sources.

Record check studies link administrative register data to survey data at individual-level. Record check studies may be classified into *prospective record check studies*, *reverse record check studies* and *complete record check studies*. Prospective record check studies link administrative records to survey respondents in order to confirm the reported behaviors. Reverse record check studies sample units from administrative records with desired characteristics and then attemp to interview them. Prospective record check studies may be used for measuring overreporting of events while reverse record check studies may be used for measuring underreporting of events. Complete record check studies with validation data for all sampled persons allow both the estimation of overreporting and underreporting. Moreover, they allow the estimation of bias due to non-response. As in external validation studies, the comparisons may be hampered by differences in the definitions of variables from the two data sources.

Even though no gold standard data are error-free, they can be very useful if the errors are small relative to errors in data being evaluated. As Biemer and Lyberg [18] point out, gold standard data provides a silver rather than a gold standard.

2.4 Analysing event history data based on a complex longitudinal survey

Surveys often use complex sampling designs involving stratification, clustering and unequal selection probabilities of units. Longitudinal surveys have an additional stage of clustering arising from the repeated observations by the same sample units. Non-response, attrition and measurement errors bring additional

challenges to the analysis of longitudinal survey data. How should these complexities be taken into account in event history analysis?

Pfefferman and Sverchkov [31] and Pfefferman[32] discuss different approaches to the modelling of survey data. Kovačević and Roberts [10] discuss model-based and design-based approaches to the modelling of event history data. In the model-based approach, the target parameters of interest are parameters β of a superpopulation model that is assumed to have generated the variable values in the finite population. The standard model-based approach ignores the probability distribution $\boldsymbol{P(S)}$ induced by the sampling design. The only source of random variation in the superpopulation model parameter estimator $\hat{\beta}$ is due to the random component in the model. Accordingly, the model-based standard errors of parameter estimates reflect the uncertainty due to the model. Sample design variables or sample weights might be incorporated as covariates of the model in order to protect against nonignorable sample design.

The design-based approach is traditionally used for descriptive inference. However, the ideas of design-based inference can be applied to analytic inference as well. In this approach the target parameter of interest is defined as a finite population parameter $\boldsymbol{B}$ that would be obtained from the model estimation procedure if all data values in the finite population were available. In the design-based approach, the only source of random variation in the estimation procedure is the sampling distribution of the estimator $\hat{B}$. Inference about $\boldsymbol{B}$ could in principle be carried out with certainty if all elements of the population were measured [33]. In practice, there would be uncertainty in the estimates even in this case due to non-sampling errors. An analyst taking the design-based approach would conduct a weighted analysis. The design information would be used to calculate the standard errors of parameter estimates. These standard errors reflect the uncertainty due to making inferences on the basis of a sample only instead of the whole population.

The test for ignorability of sample design suggested by Pfeffermann [34] may be used to choose between the design-based and the model-based approaches for event history analysis. The test compares design-based and model-based estimates of parameters of interest and rejects the null hypothesis of ignorability of sample design if the model-based estimates are "too far" from the design-based estimates. In this case, a design-based approach for the analysis should be taken.

Longitudinal analyses often use only respondents to each wave of interest [35, 19]. Even though the available data until the time of attrition could be used, attriters are often discarded from the analysis. In an analysis using weights this can be motivated by the fact that weights are usually adjusted for non-response and attrition. However, the general purpose weights included in a survey data set may not fully correct for non-response and attrition that is selective with respect to the particular response variable of interest. The inclusion of the available data from the attriters might in this case help to reduce the bias due to attrition. It is not clear, however, which weights should be used in an analysis including attriters, see [9].

3 Aims of the study

Article [1].
To show how register data combined at person-level with survey data can be used to conduct a non-response and attrition analysis that enables to 1) study the determinants of non-response and attrition; 2) test the validity of the MAR assumption; and 3) estimate the size of bias due to non-response and attrition. To apply this analysis to unemployment spell data from FI ECHP survey in order to provide novel information on relative importance and determinants of non-response and attrition in event history data and their effects on event history analysis.

Article [2].
To conduct a complete record check validation study of retrospective reports of unemployment spells from FI ECHP survey data in order to provide novel evidence about 1) the type, magnitude and determinants of measurement errors in survey reports of event histories, 2) the validity of classical assumptions about measurement errors, 3) the size of bias due to measurement errors and low measurement accuracy in event history analysis of survey data.

Article [3].
To study statistical properties of the Inverse Probability of Censoring Weighting (IPCW) method in design-based survival analysis based on complex survey data.

Article [4].
To discuss the following choices involved in event history analysis of survey data: 1) whether to take a design-based or a model-based approach for modelling; 2) which subset of data to use; and 3) if a design-based approach is chosen, which weights to use. To demonstrate the effect of these choices by using unemployment spell data from FI ECHP survey combined at person-level with register data.

4 Data and Methods

4.1 Combined longitudinal survey register data

The Finnish subset of the European Community Household Panel (FI ECHP) survey data were combined at person-level with longitudinal register data. Empirical analyses were based on data from FI ECHP sample persons aged 16 or over at the beginning of the panel. Sample persons are defined in the ECHP as all members of the initial sample of households. The first five waves of the FI ECHP data covering the years 1996–2000 were used in the analyses. Temporary non-respondents were excluded, leaving 10,720 persons for the analysis. Unemployment spells were used as study variables of interest.

FI ECHP target population and sample design. The target population of FI ECHP consists of members of private households permanently resident in Finland. As most household panel surveys, FI ECHP aims to remain cross-sectionally representative of the household population over time. This is strived for using certain follow-up rules of the sample persons, see [14]. The FI ECHP sample is a two-phase stratified network sample. The population information system of the Population Register Centre was used as a frame. The frame population consisted of persons permanently living in Finland aged 15 and over. In the first phase, a master sample of target persons was drawn from the frame. Dwelling units were constructed by adding all the persons sharing the same domicile code as the target persons to the master sample. The master sample was merged with the most recent taxation records and their information was used to form a socio-economic group for each target person. The second phase consisted of drawing the final sample from the master sample using stratification according to socio-economic groups.

Collection of event history data on labour market states in FI ECHP. Retrospective labour market state data were collected by a multi-state framework in the form of a month-by-month main activity state calendar obtained for the year preceding the interview. The respondent was first asked whether there were changes in his/her main activity state during the preceding year. If not, the respondent was asked to choose a main activity state from a showcard. If there were changes, the respondent was asked to choose a main activity state from the showcard for each month of the year beginning from January.

Construction of combined survey-register panel data. FI ECHP survey data were merged with administrative data on unemployment spells retrieved from the Ministry of Labour's register of jobseekers. The register contains day-level information on the starts and ends of unemployment spells, as well as on spell outcomes. All register spells ongoing between 1 January 1995 and 31 December 1999 were used in the analysis. This time period corresponds with the main activity state reference periods of the first five waves of the FI ECHP. The register of jobseekers and other administrative registers such as Statistics Finland's register of completed education and degrees, the population information system of the population register centre and registers of the tax administration were also used to retrieve the background variables used in the analyses. Personal identifi-

cation numbers were used in order to merge the data files at person-level across time, across various administrative registers and across survey and register data.

4.1.1 Article [1]

Register data were used as a source of information on unemployment spells and covariates. This information is available for all sample persons irrespective of the response status. Survey data were used only to obtain, for each wave, the sample person's participation status in the FI ECHP. This way we obtained directly comparable information for respondents and non-respondents and were able to detect a pure non-response effect, free from measurement errors. The statistical analyses were conducted in a model-based framework.

Assessing determinants of non-response and attrition. Separate models were estimated for the non-response and attrition processes. The initial non-response analysis was conducted by estimating logit models for the probability of being a non-respondent at the first wave of the panel. The analysis was restricted to sample persons having at least one spell of unemployment during the follow-up period (2,956 persons). The attrition process was modeled by a discrete-time hazard model where the conditional probability of attrition at a specific year, given that the person has remained in the survey until the year in question, is explained by a set of time-varying covariates. Initial non-respondents were excluded, leaving 2,085 persons for the attrition analysis.

Testing the validity of the MAR assumption. Covariates describing number of days spent in unemployment and number of unemployment spells were used to test the MAR assumption. For the initial non-response analysis, the number of unemployment days and the number of spells were calculated both before and after the time of the interview (or time of contact, if an interview was not obtained) in wave 1. In the attrition analysis, the number of unemployment days and the number of spells were calculated for each wave before and after the last obtained interview. If the initial non-response mechanism is MAR, none of these covariates should explain probability of non-response. In the attrition model, a MAR non-response mechanism implies that covariates measured after the last obtained interview should not affect probability of non-response. The validity of the MAR assumption was tested by looking at the statistical significance of these covariates.

Estimating non-response bias. The participation behaviour in the survey is known for each sample person having one or more spells during the observation period. It was assumed that unemployment spells are observed until the time of the last interview or until the end of the observation period, whichever comes first. This creates a number of different cases:

a Spells that end before the last interview (or before 31 December 1999, whichever comes first) are *fully observed*.
b Spells ongoing at the time of the last interview, which is followed by attrition, are *right censored by attrition* at the time of the last interview.
c Spells that start after the last interview, which is followed by attrition, are *not observed by attrition*.
d Spells by persons without any interviews are *not observed by initial non-response*.

On the basis of this taxonomy, three different sets of unemployment spells were constructed:

The *full information* set of spells uses the entire register information without restrictions by initial non-response or attrition. Cases **a, b, c, d** (10,734 spells).

The *partial information* set of spells is a subset of the full information set of spells, obtained by excluding spells unobserved by initial non-response. Cases **a, b, c** (7,712 spells).

The *observed information* set of spells is a subset of the partial information set of spells, obtained by excluding spells unobserved by attrition and the remaining length of the spells censored by attrition. Cases **a, b** (6,496 spells).

The size of bias due to non-response and attrition was estimated by comparing Kaplan-Meier estimates of survival function and estimates of regression coefficients from a Cox shared frailty model based on the three sets of unemployment spells. The analyses were conducted in a cause-specific setting, the outcome of interest defined as transition from unemployment to employment. The bias due to non-response was estimated as

$$\hat{B}ias_{nr} = \hat{\beta}_{partial} - \hat{\beta}_{full},$$

the bias due to attrition was estimated as

$$\hat{B}ias_{attr} = \hat{\beta}_{obs} - \hat{\beta}_{partial},$$

and the joint effect due to non-response and attrition as

$$\hat{B}ias_{nr+attr} = \hat{\beta}_{obs} - \hat{\beta}_{full}.$$

The Hausman test [36] was used to test the statistical significance of bias.

4.1.2 Article [4]

The full information and observed information sets of spells described in the previous section were used together with a *total respondents* set of spells. The total respondents set of spells uses data from respondents who provide data on all waves of interest (4,066 spells).

Design-based estimates based on the full information data set were used as benchmark estimates $\hat{B}_{bm}$ against which estimates based on the observed information and the total respondents sets of spells were evaluated. These benchmark estimates are free from the effects of non-response and attrition. The benchmark estimates were taken to be the best available estimates of $\boldsymbol{B}$, the finite population regression parameters, which, if the model postulated is correct, in turn es-

timate the model parameters β. Model-based and design-based estimates of Cox proportional hazard models for the total respondents and observed information data sets were calculated. The outcome of interest was defined as transition from unemployment to employment.

The design-based total respondents analyses were weighted by the last wave base weights (described in [35]). The design-based observed information estimates were calculated using both first wave base weights and base weights from the starting year of the spell.

The test proposed by Pfeffermann [34] was used to test ignorability of sample design. A Mahalanobis type of distance measure was used to assess the closeness of estimated coefficients to the benchmark estimates.

4.1.3 *Article [2]*

A complete record check validation study of retrospective reports of unemployment spells from the FI ECHP survey data was conducted. The survey data consists of all unemployment spells reported by FI ECHP sample persons (2,710 spells). For each person, the validation data cover the same time span as his/her follow-up time in the survey. The validation data contains 6,050 register spells. The statistical analyses were conducted in a model-based framework.

Assessing determinants of measurement errors. To study determinants of measurement errors and test validity of the classical assumptions, measurement error variables were constructed for each person. The survey and register data can be reliably linked only at person-level and not at spell-level. Therefore, the measurement error variable was calculated as the difference between the sums of spell durations from the survey and the register. Measurement error variables were calculated separately for each person and for each panel wave in which the person was unemployed according to both survey and register. Measurement errors were modeled in two phases. In the first phase, a random effects logit model was specified for the probability of reporting no unemployment spells in a specific wave, given that at least one unemployment spell was found in the register. In the second phase, a random effects linear model was specified for the magnitude of measurement errors in the reported unemployment spells, given that at least one unemployment spell was both reported and found in the register.

Testing the validity of the classical assumptions. Covariates related to length and number of unemployment spells and covariates of the model explaining unemployment duration were used to test the classical assumptions about measurement errors. Statistically significant effects of these covariates were taken as evidence of violation of the classical assumptions.

Estimating bias due to measurement errors and low measurement accuracy. Kaplan-Meier estimates of survival function and estimates from Cox and Weibull proportional hazards models based on survey data were compared with estimates from the validation data. The estimates based on validation data were used as benchmarks against which the bias due to measurement errors in survey-based estimates was evaluated. Both analyses ignoring spell outcome and cause-specific analyses were conducted. In the cause-specific analyses, the outcome of interest was defined as transition from unemployment to employment.

The analyses were conducted in two phases. The phase 1 analyses were concerned with measurement errors in spell durations only. Therefore, spell outcomes were ignored. The phase 1 survey data analyses were conducted using survey spells and register covariates. By using the same source of covariates as in the validation data, the differences in estimates could only be attributed to differences in spell durations. The Phase 2 analyses took measurement errors in spell outcomes into account by conducting cause-specific analyses. Phase 2 survey data analyses were conducted using survey spell durations and outcomes, and register covariates.

Differences in estimates based on survey and validation data result not only from measurement errors but also from low measurement accuracy in survey data. Survey reports on main activity state were collected at the accuracy of one month. Moreover, if a person has had various activity states during a month, employment was preferred over other states. Therefore, it is difficult to obtain information on unemployment spells shorter than one month. We aimed at separating the effects of measurement accuracy and measurement error by discretizing the register spells at the accuracy of one month and repeating the analyses with discretized data. Differences between estimates based on survey data and discretized register data (reg2) could then be taken as estimates of bias due to measurement error:

$$\hat{B}ias_{me} = \hat{\beta}_{survey} - \hat{\beta}_{reg2}.$$

Bias due to measurement accuracy could be estimated by calculating differences of estimates from original (reg) and discretised register data:

$$\hat{B}ias_{ma} = \hat{\beta}_{reg2} - \hat{\beta}_{reg}.$$

4.2 Simulation study (Article [3])

Statistical properties of the IPCW method in design-based survival analysis in the presence of dependent censoring were assessed by simulation methods. The parameters of interest were defined as the values of the finite population survival function ***S(t)*** at certain time points and the finite population regression coefficient ***B*** from a Cox proportional hazards model.

Generation of the populations. Four different populations of persons, each of size N = 10,000 and corresponding to the following scenarios were generated:

1 The variable determining the censoring mechanism is known,
2 A variable that is either a) strongly or b) weakly associated with the variable determining the censoring mechanism is observed,
3 The variable determining the censoring mechanism is unknown.

The population characteristics consist of three binary variables: social exclusion, sex and level of education and a variable describing the length of the unemployment spell. Social exclusion determines the probability of censoring but is unob-

served. Sex was used both as an auxiliary variable in the censoring model and as a stratification variable in the sampling stage. Level of education is the covariate in the survival model whose effect on the length of unemployment spells is of interest. The four populations differ by the degree of association between variables sex and social exclusion, see Table 2 in [3]. Perfect (No) association between sex and social exclusion corresponds to scenario 1 (3) above.

Unemployment spells were generated from the Weibull distribution using a value of 0.8 for the shape parameter (a decreasing hazard rate) and scale parameters depending on the level of education and social exclusion. The median duration of the unemployment spells, as well as the effect of education on the hazard of spell completion, are different among the excluded and the non-excluded. Censoring that depends on social exclusion thus biases both the estimates of survival function and the estimate of the regression coefficient.

Sampling design and estimation. From each population, 500 stratified simple random samples of size n=600 were drawn without replacement and using sex as a stratification variable. Inclusion probabilities of 0.07 for men and 0.05 for women were used. For each sample, an artificial 2-wave panel survey was conducted. It was assumed that there is no non-response at wave 1. Selective survey attrition at wave 2 was generated by stratifying the samples according to exclusion status and drawing 80% samples of respondents among the non-excluded and 20% samples of respondents among the excluded. For each sample s_j, the IPC corrected design weights (see equation 7 in [3]) were constructed using sex as an auxiliary variable in the censoring model. Estimates $\hat{B}_j$ and $\hat{S}_j(t)$ were calculated using these weights. The empirical distribution of these estimates was used as an approximation of the sampling distribution of $\hat{B}$ and $\hat{S}(t)$.

5 Results

5.1 Article [1]

Determinants of non-response and attrition. Initial non-response and attrition turned out to be different processes driven by different background variables, Tables 3 and 4 in [1]. Low level of education, high household disposable income, small family size as well as being middle-aged, living in an urban municipality or in the capital region, not being married, being unemployed or outside the labour force were associated with a high probability of initial non-response. There were far fewer strong predictors of attrition which suggests that attrition was less selective than initial non-response. Young age, low level of education, low household disposable income and living in Northern Finland were associated with high probability of attrition. The difficulties in fieldwork in 2000 due to uncertainty about the continuation of the panel, showed as a peak in the attrition hazard, Table 4 in [1].

Validity of the MAR assumption. Both the initial non-response and attrition processes were non-ignorable with respect to analysis of unemployment duration. Being in the uppermost decile with respect to the number of unemployment days after the time of the first interview, raised the odds of initial non-response by 30.5% in a model including covariates of the unemployment spell model. An increase of 100 days of unemployment after the last obtained interview increased the odds of attrition hazard by 3%.

Size of bias due to non-response and attrition. Initial non-response caused downwards bias in the estimated survival function, whereas attrition did not have a biasing effect. The Hausman tests showed that both initial non-response and attrition caused bias in the coefficient estimates of a Cox shared frailty model, Table 6, [1]. The bias due to initial non-response tended to be larger than the bias due to attrition, Table 1. The largest biases were caused to the effect of receiving earnings-related unemployment benefit.

Table 1:
Analysis of unemployment duration. Non-response and attrition bias in estimates of Cox shared frailty models.

Variable	Non-response bias %	Attrition bias %
Female	−46.9	−5.4
Age	20.9	15.8
Age squared	15.4	13.4
Upper secondary education	−56.2	71.0
Higher education	−26.1	21.3
Prop. of UE[1] time	10.5	11.7
Semi urban municipality	−6.6	−31.5
Rural municipality	−18.7	−57.5
Southern Finland	−9.1	15.7
Eastern Finland	−42.0	33.7
Central Finland	−13.3	2.2
Northern Finland	−13.0	17.5
Earnings-related UE benefit	−682.8	226.6
Year 1996	20.2	17.5
Year 1997	−657.7	67.9
Year 1998	−24.9	−26.7
Year 1999	−14.5	−3.5

1 UE Unemployment

5.2 Article [2]

Type and magnitude of measurement errors in reported unemployment spells. The retrospective reports of unemployment spells showed both omitting and overreporting of spells, omitting being much more important, Figure 1 in [2]. The starts and ends of survey spells were strongly heaped at the seams between the reference periods of consecutive panel waves, Figures 2 and 3 in [2]. Of register spells ending in subsidised work, 85% were misclassified as ending because of normal employment in the survey, Table 2 in [2].

Determinants of measurement errors in reported unemployment spells. Conducting a proxy interview instead of an interview with the person of interest increased the odds of omitting unemployment spells by 72.8%, Table 3 in [2]. During the years 1998-2000, the odds of omission were more than double compared to the year 1995. This is likely a consequence of the shifting of the fieldwork period from spring to autumn from 1998 onwards, and of the resulting prolongation of the recall period by more than six months. The fieldwork covariates did not have a clear effect on the magnitude of measurement errors.

Validity of classical assumptions. Both the probability of omission and the magnitude of measurement errors depended on variables related to unemployment spells and covariates used in the event history model, Table 3 in [2]. Moreover, both the propensity to omit reporting unemployment spells and the measurement errors were correlated across survey waves. The classical assumptions about measurement errors were thus not valid.

The size of bias due to measurement errors, effect of measurement accuracy. The survey data overestimated both the median duration of unemployment (5 months vs. 2 months) and the median time to become employed (6 months vs. 3.8 months), Figures 5 and A.6 in [2]. The effect of education and in the competing risks model also the effect of receiving earnings-related unemployment benefit were estimated with sizeable bias (biases ranging from 18 to 30 percentage points and 28 to 30 percentage points, respectively), Table 6 in [2]. The bias in the effect of education was mainly due to measurement errors. Neither dummies for the heaping months, nor a more flexible model specification, protected against bias in coefficient estimates, Table 5 in [2]. The biases in January and December dummies showed that the heaping of spell starts and ends was a measurement error and not a measurement accuracy problem, Table 6 in [2]. The lack of short spells in survey data and in discretised register data led to underestimation of the baseline hazard function from the Cox proportional hazard models for durations shorter than six months, Figure A.4 in [2]. For longer durations, the biases due to measurement accuracy and measurement error worked in opposite directions. Measurement accuracy created a small positive bias leading to overestimation of the baseline hazard. The hazard spikes were however correctly placed in time. Measurement error created a large negative bias and flattened the shape of the baseline hazard. The joint effect of measurement accuracy and measurement errors was underestimation of the baseline hazard. The low measurement accuracy and the resulting lack of short spells in survey data led to badly biased shape of the baseline hazard from the Weibull model, while measurement errors only led to slight underestimation of the level of the baseline hazard (Figure A.5 in [2]).

5.3 Article [3]

The IPC corrected design weighted estimators of $\boldsymbol{S(t)}$ and $\boldsymbol{B}$ had the smallest bias in Scenario 1, see Table 3 in [3]. Scenario 1 corresponds to a situation where the censoring mechanism is known. This is an ideal situation for the IPCW method. The bias of IPC corrected design weighted estimators growed as information on the censoring mechanism lessened but was always smaller than the bias of design weighted estimators (Scenarios 2a and 2b). When the censoring mechanism was unknown (Scenario 3), the bias of IPC corrected design weighted Kaplan-Meier estimators was equal to that of design weighted estimators. In that case, there was no gain from using IPC corrected design weights in survival curve estimation. By contrast, the IPC corrected design weighted estimators of the hazard ratio performed quite well even in this case.

5.4 Article [4]

The observed information estimates of covariate effects of Cox proportional hazard models were closer to the benchmark estimates than the total respondents estimates, Table 1 in [4]. Thus using all the available data in the analysis, including the spells by attriters until the time of attrition, helped to reduce attrition bias. Comparison of the model-based and the design-based estimates revealed that the weighting correction for attrition is not very helpful in our analysis. The weights from the last wave analysed and the weights from the starting wave of the spell produced estimates that were further from the benchmark than the corresponding unweighted estimates.

The design-based estimates with total respondents data and the last wave weights were furthest from the benchmark estimates. The design-based estimates from the observed information data and weighted by the first-wave weights were closest to the benchmark estimates. However, the tests indicated nonignorability of the sample design (Table 2 in [4]) and a model-based analysis would be valid in this case. Contrary to expectations, the inclusion of design variables moved estimates *farther* from the benchmark estimates.

6 Discussion

6.1 Discussion of methods

6.1.1 Combined longitudinal survey register data

Combining longitudinal survey data with administrative register data is time-consuming and costly. In many countries, linking of various data sources is difficult because of a lack of a variable that uniquely identifies persons. Also, as noted by Calderwood and Lessof [6], legal and ethical problems may be involved. As a consequence, there are only a few studies available on non-response bias or measurement error bias in event history analysis based on combined longitudinal survey register data. Van Den Berg, Lindeboom and Dolton [37] studied initial non-response bias in the analysis of unemployment spells. Pyy-Martikainen and Rendtel [38] tested the validity of the assumption of independent censoring in event history analysis with the same data set as in this study. Mathiowetz and Duncan [39] studied the type, magnitude and determinants of measurement errors in retrospective reports of unemployment. Jäckle [40] studied measurement error bias in analysis of benefit receipt spells. I am unaware of previous studies using combined longitudinal survey register data to assess the effects of different approaches to event history analysis.

Even though combined longitudinal survey register data are considered a valuable tool for assessing errors in survey data, there are potential problems related to the use of such data. Next I discuss the relevance of four potential problems raised by Bound, Brown and Mathiowetz [7]and Biemer and Lyberg [18] to the study:

1. *The time periods for the administrative data and the survey data may not coincide*

The time periods in the survey data and the register data used in this study have a complete overlap.

2. *The definitions of the characteristic of interest may differ in administrative data and in survey data*

In FI ECHP, a person is defined as unemployed if he/she is without a job, available for work and looking for work through the employment office or newspaper advertisements or some other way. Persons dismissed temporarily are also regarded as unemployed. In the register, an unemployed job seeker is defined as being without a job and seeking a new job. Registering at the employment office is considered as evidence of seeking a job. Persons dismissed temporarily are regarded as unemployed. The definitions of unemployment in survey and register data are thus close to each other.

3. *The micro-merging of register and survey data is often restricted to a very specific population which makes generalisation of results problematic*

Register data were merged to all FI ECHP sample persons eligible for interview. The results obtained are thus generalisable to the population aged 16 and over and residing in Finland. This is an advantage compared to the studies by Mathiowetz and Duncan [39] and by Jäckle [40], who use samples restricted to very specific populations.

4. *Administrative data can be prone to errors as well*

As unemployed persons need to register at employment office in order to utilise their services and to receive unemployment benefits, register information is likely to cover most unemployed persons. Moreover, the duration of unemployment is likely to be precisely measured as register information on unemployment is used in order to pay unemployment benefits. However, persons who get a new job do not always inform the employment office about the job. Thus, an unemployment spell in the data base may erroneously continue for some time after the true ending date of the spell.

Lastly, linking of register and survey data is virtually error-free due to personal identity codes. All Finnish citizens are registered in the Finnish Population Information System, which is a national register that contains basic information such as name, date of birth and address. As part of the registration process, citizens are issued a personal identity code that is used as a means of identifying persons. Data from the Finnish Population Information System is used throughout Finnish society's information services and management, including the production of statistics and research.

6.1.2 Simulation study

The IPC corrected design weights are time-dependent and change each time a censoring occurs in the data. To incorporate time varying weights in the analysis, the data had to be transformed into a counting process form. Each unemployment spell was split into several intervals, the splitting points being defined by the times at which censorings occurred in the sample. Time was defined as time from the beginning of the unemployment spell. The estimations were conducted by R software, which supports estimation of Kaplan-Meier survival function and Cox proportional hazard model based on counting process form data. Due to problems with computing capacity in R, the number of replicate samples had to be restricted to 500. For the same reason, the artificial data had to be generated so that all censorings occurred during first 30 days (so that there was a maximum of thirty weights per person). For applications of this method to real data, it might be useful to model the censoring process as a discrete time process where the probability of censoring changes only at the time points defined by survey interviews. The discrete-time hazard model used in [1] is one option.

6.2 Discussion of main results

The results of our study have implications for 1) survey organisations collecting event history data by longitudinal surveys; 2) researchers using longitudinal surveys for event history analysis; and 3) researchers who develop methods to correct for non-sampling errors in event history data.

6.2.1 Implications for survey organisations

Our study demonstrated a novel way to conduct a non-response analysis of longitudinal survey data. The linking of register data at person-level to survey data enables to analyse and compare the non-response and attrition processes, test the type of the missingness mechanism and estimate the size of bias due to non-response and attrition. Our study also contributed to the pool of evidence on the existence, determinants and effects of non-response, attrition and measurement errors in event history data based on longitudinal surveys. This pool may be used to provide both collectors and users of data with information on data quality, in adjusting survey estimates for non-sampling bias, and to optimise future collection of event history data by longitudinal surveys.

Our results suggest that initial non-response may be a more important source of bias than attrition in event history analysis. Other studies with different variables and different analyses have reached similar conclusions. The studies by Fitzgerald, Gottschalk and Moffitt [41] and Sisto [42] even suggested that the bias in cross-sectional estimates of income distribution and socioeconomic status caused by initial non-response may fade away over the life of the panel. These results challenge the common view of attrition being the main threat to the value of panel data [41, 43, 44], and argue in favor of conducting panel surveys in order to provide not only longitudinal but also cross-sectional data. Moreover, the existence of a fade away effect would imply that long-term panels should be preferred over short-term panels. However, a recent study with Finnish subsample of EU-SILC survey finds a clear biasing effect of panel attrition on estimates of transition probabilities between household income quintiles [45]. More research with different variables, panel surveys and countries are needed in this important issue.

According to our analysis, reported unemployment spells were subject to both omissions and, to a lesser extent, overreporting. Spell starts and ends were strongly heaped at the seams between the reference periods of consecutive panel waves. These findings are consistent with earlier studies by Mathiowetz [26], Mathiowetz and Duncan [39] and Kraus and Steiner [46]. The use of proxy interviews tended to cause spell omissions and should, therefore, be avoided in the collection of event history data.

A previously unnoticed finding was the classification error in reported spell outcomes. There was an excess of exits into employment in survey data due to the fact that exits into subsidised work were often misclassified by respondents as becoming employed. Attention needs to be paid to the definition of states in a multi-state framework in order to minimize misclassification errors.

Almost 40% of the register spells were shorter than one month. A measurement accuracy of one month used in ECHP main activity state calendar and currently in EU-SILC is clearly too coarse and leads to biased estimates. Register information about the distribution of the spells of interest should be taken into account in the questionnaire design phase in order to find an appropriate level of measurement accuracy.

6.2.2 *Implications for researchers using longitudinal surveys for event history analysis*

An unsettling result for the researchers using longitudinal surveys for the analysis of labour market transitions is that some of the key covariates such as type of unemployment benefit and level of education, had large biases due to non-response and measurement errors. Compared to the Weibull model, the more flexible Cox model did not turn out to be more robust with respect to measurement errors in estimated covariate effects. This contradicts an earlier empirical finding concerning the robustness of the Cox model with respect to initial non-response bias [37]. However, the flexibility of the Cox model was clearly advantageous in the estimation of the baseline hazard. In the light of our results, including dummies for the heaping months is not helpful in correcting measurement error bias in estimated covariate effects or distribution of spells.

As discussed in Boudreau [8], the choice of approach for analytical inference of survey data is a controversial topic. Kovačević and Roberts [10] discuss and demonstrate model-based and design-based approaches for event history analysis. The test for ignorability of sample design suggested by Pfeffermann [34] may be used to choose between these two approaches. The test compares design-based and model-based estimates of parameters of interest and rejects the null hypothesis of ignorability of sample design if the model-based estimates are "too far" from the design-based estimates. In this case, a design-based approach for the analysis should be taken. However, the use of this test may be problematic in some cases. It is not always clear in longitudinal analyses which set of weights should be used. The choice of weights may affect the result of the test. Also, our results showed that the design-based estimates may be even more biased than model-based estimates.

Longitudinal analyses often use only respondents to each wave of interest, thus discarding attriters from the analysis. In a design-based analysis using weights this can be motivated by the fact that weights are usually adjusted for non-response and attrition. However, the general purpose weights included in a survey data set may not fully correct for non-response and attrition that is selective with respect to the response variable of interest. The inclusion of the available data from the attriters might, in this case, help reduce the bias due to attrition. This is a topic shortly discussed in [9]. Our results point towards the importance of using all the available data in the analysis. The often recommended way to use survey data for longitudinal analyses; total respondents with last wave weights [19, 35] is not a modeling strategy to recommend in the light of our results.

Results from the simulation study suggest that combined design IPC weights may be useful in event history analyses based on survey data with dependent cen-

soring. These weights were effective in reducing bias due to dependent censoring even when there was little information available about the censoring mechanism. Results from recent simulation studies by Lawless and Hajducek [47, 48] are in line with our results. However, due to the very specific purpose of the design IPC weights and the fact that the weights are time-dependent, their calculation is not easily integrated in the routine production of survey data. Instead, analysts of event history data may benefit from constructing them for their own research purposes.

6.2.3 *Implications for the development of methods to correct for non-sampling errors in survey data*

The number of days unemployed after the last interview had a statistically significant effect on the probabilities of non-response and attrition in our study. Moreover, measurement errors in reported unemployment spells were shown to be correlated across survey waves, with variables related to true spells and with covariates used to explain the duration of spells. Thus, neither the MAR assumption about non-response and attrition mechanisms, nor the classical assumptions about measurement errors, were valid in our study. Our results suggest that methods that make more realistic assumptions about the mechanisms generating non-sampling errors need to be developed.

6.3 *Areas for future research*

The performance of the IPCW method has not yet been studied with real survey data. Lawless and Hajducek [47, 48] illustrated the use of the method using jobless spell durations from Statistics Canada's Survey of Labour and Income Dynamics. However, they lacked gold standard data and were thus not able to assess neither the size of bias due to censoring nor the effectiveness of the IPCW method in reducing bias. Pyy-Martikainen and Rendtel [38] showed that censoring is independent with respect to analysis of unemployment spells in FI ECHP data. There is thus no scope for the IPCW method unless dependent censoring is generated in the data. Studies with other combined longitudinal survey register data sets might shed light on the usefulness of this method for event history analysis based on survey data.

More studies with different combined longitudinal survey register data sets and different event history variables are needed to increase our understanding of the existence, determinants and effects of non-sampling errors in event history data. Also, the effects of different modelling approaches for event history analysis, the choice of the subset of data used in analysis, and the choice of weights to use in event history analysis are areas where more research is needed.

References

[1] Marjo Pyy-Martikainen and Ulrich Rendtel. Assessing the impact of initial nonresponse and attrition in the analysis of unemployment duration with panel surveys. *Advances in Statistical Analysis*, 92:297–318, 2008.

[2] Marjo Pyy-Martikainen and Ulrich Rendtel. Measurement errors in retrospective reports of event histories. A validation study with Finnish register data. *Survey Research Methods*, 3(3):139–155, 2009.

[3] Marjo Pyy-Martikainen and Leif Nordberg. Inverse probability of censoring weighting method in survival analysis based on survey data. *Statistics in Transition*, 8(3):487–501, 2007.

[4] Marjo Pyy-Martikainen. Approaches for event history analysis based on complex longitudinal survey data. *Advances in Statistical Analysis*, 97:297–315, 2013.

[5] Robert Groves. Survey Errors and Survey Costs: *An Introduction to Survey Errors. Wiley*, 1989.

[6] Lisa Calderwood and Carli Lessof. Enhancing longitudinal surveys by linking to administrative data. In Peter Lynn, editor, *Methodology of Longitudinal Surveys*, pages 55–72.Wiley, 2009.

[7] John Bound, Charles Brown, and Nancy Mathiowetz. Measurement error in survey data. In James Heckman and Edward Leamer, editors, *Handbook of econometrics*, volume 5, pages 3705–3833. Elsevier, 2001.

[8] Christian Boudreau. *Duration Data Analysis in Longitudinal Surveys*. PhD thesis, University of Waterloo, 2003.

[9] Georgia Roberts and Milorad Kovačević. New research problems in analysis of duration data arising from complexities of longitudinal sur- veys. In *Proceedings of the Survey Methods Section*, pages 111–116. SSC Annual Meeting, 2001.

[10] Milorad Kovačević and Georgia Roberts. Modelling durations of multi- ple spells from longitudinal survey data. *Survey Methodology*, 33(1):13–22, 2007.

[11] Peter Lynn. Methods for longitudinal surveys. In Peter Lynn, editor, *Methodology of Longitudinal Surveys*, pages 1–19. Wiley, 2009.

[12] Matthias Schonlau, Nicole Watson, and Martin Kroh. Household survey panels: how much do following rules affect sample size? SOEPpaper 347 on Multidisciplinary Panel Data Research, 2010.

[13] Franco Peracchi. The European Community Household Panel: A review. *Empirical Economics*, 27(1):63–90, 2002.

[14] Marjo Pyy-Martikainen, Johanna Sisto, and Marie Reijo. The ECHP study in Finland. Quality report. Living conditions 2004:1, Statistics Finland, 2004.

[15] Jerald Lawless. Event history analysis and longitudinal surveys. In Ray Chambers and Chris Skinner, editors, *Analysis of Survey Data*, pages 221–243. Wiley, 2003.

[16] Mario Callegaro. *Seam Effects Changes due to Modifications in Question Wording and Data Collection Strategies. A Comparison of Conventional Questionnaire and Event History Calendar Seam Effects in the PSID*. PhD thesis, University of Nebraska, 2007.

[17] Paul Biemer. Total survey error: design, implementation and evaluation. *Public Opinion Quarterly*, 74(5):817–848, 2010.

[18] Paul Biemer and Lars Lyberg. *Introduction to Survey Quality*. Wiley, 2003.

[19] Grahan Kalton and Michael Brick. Weighting in household panel surveys. In David Rose, editor, *Researching Social and Economic Change: the Uses of Household Panel Studies*. Routledge, 2000.

[20] Donald Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[21] John Kalbfleisch and Ross Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, 1980.

[22] James Robins. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the Biopharmaceutical Section*, pages 24–33. American Statistical Association, 1993.

[23] Jerald Lawless. Censoring and weighting in survival estimation from survey data. In *Proceedings of the Survey Methods Section*, pages 31–36. SSC Annual Meeting, 2003.

[24] Gerard van den Berg, Maarten Lindeboom, and Geert Ridder. Attrition in longitudinal data and the empirical analysis of dynamic labour market behaviour. *Journal of Applied Econometrics*, 9(4):421–435, 1994.

[25] Gerard van den Berg and Maarten Lindeboom. Attrition in panel data and the estimation of dynamic labor market models. Leiden university research memorandum, 1994.

[26] Nancy Mathiowetz. The problem of omissions and telescoping error: New evidence from a study of unemployment. In *Proceedings of the section on Survey research Methods*, pages 482–487, 1986.

[27] Daniel Holt, J McDonald, and Chris Skinner. The effect of measurement errors on event history analysis. In Paul Biemer, Rober Groves, Lars Lyberg, Nancy Mathiowetz, and Seymour Sudman, editors, *Measurement Errors in Surveys*, pages 665–685. Wiley, 1991.

[28] Anders Skrondal and Sophia Rabe-Hesketh. Generalized *Latent Variable Modeling. Chapman and Hall*, 2004.

[29] Chris Skinner and K Humphreys. Weibull regression for lifetimes mea- sured with error. *Lifetime Data Analysis*, 5:23–27, 1999.

[30] Thomas Augustin. Correcting for measurement error in parametric duration models by quasi-likelihood. Discussion paper 157, Collaborative Research Center 386, 1999.

[31] Danny Pfeffermann and Michail Sverchkov. Inference under informative sampling. In Danny Pfeffermann and C Rao, editors, *Sample Surveys: Inference and Analysis*, volume 29B, pages 455–487. Elsevier, 2009.

[32] Danny Pfeffermann. Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology*, 37(2):115–136, 2011.

[33] Ray Chambers and Chris Skinner, editors. *Analysis of Survey Data*, chapter 1. Wiley, 2003.

[34] Danny Pfeffermann. The role of sampling weights when modeling survey data. *International Statistical Review*, 61(2):317–337, 1993.

[35] Eurostat. ECHP UDB manual: *European Community Household Panel Longitudinal User's Database, waves 1 to 7, survey years 1994 to 2000*. Doc. Pan 168/2003-6, 2003.

[36] Jerry Hausman. Specification tests in econometrics. *Econometrica*, 46(6):1251–1271, 1978.

[37] Gerard van den Berg, Maarten Lindeboom, and Peter Dolton. Survey non-response and the duration of unemployment. *Journal of the Royal Statistical Society A*, 169(3):585–604, 2006.

[38] Marjo Pyy-Martikainen and Ulrich Rendtel. The effects of panel attri- tion on the analysis of unemployment spells. Chintex working paper 10, 2003.

[39] Nancy Mathiowetz and Greg Duncan. Out of work, out of mind: Re- sponse errors in retrosepctive reports of unemployment. *Journal of Business and Economic Statistics*, 6(2):221–229, 1988.

[40] Annette Jäckle. Measurement error and data collection methods: Effects on estimates from event history data. Iser working paper series 13, 2008.

[41] John Fitzgerald, Peter Gottschalk, and Robert Moffitt. An analysis of sample attrition in panel data. The Michigan panel study of income dynamics. *Journal of Human Resources*, 33(2):251–299, 1998.

[42] Johanna Sisto. Attrition effects on the design-based estimates of dispos- able household income. Chintex working paper 9, 2003.

[43] Ugo Trivellato. Issues in the design and analysis of panel studies: a cursory review. *Quality and Quantity*, 33:339–352, 1999.

[44] Greg Duncan. Using panel studies to understand household behaviour and well-being. In David Rose, editor, *Researching Social and Economic Change. The Uses of Household Panel Studies*. Routledge, 2000.

[45] Tara Junes. Initial wave nonresponse and panel attrition in the Finnish subsample of EU-SILC. Master's thesis, University of Helsinki, 2012.

[46] Florian Kraus and Viktor Steiner. Modelling heaping effects in unem- ployment duration models -with an application to retrospective event data in the German socio-economic panel. *In Jahrbücher für Nationalökonomie und Statistik*. Lucius & Lucius, 1998.

[47] Jerald Lawless and Dagmar Hajducek. Duration analysis in longitudinal studies with ascertainment of information at intermittent followup times and losses to followup. *Canadian Journal of Statistics*, 40(1):1–21, 2012.

[48] Jerald Lawless and Dagmar Hajducek. Estimation of finite population duration distributions from longitudinal survey panels with intermittent followup. *Lifetime Data Analysis*, 19:371–392, 2013.

Article I

***Pyy-Martikainen, M. & Rendtel, U.**, Assessing the Impact of initial nonresponse and attrition in the analysis of unemployment duration with panel surveys. Advances in Statistical Analysis 92, 297–318, 2008.*

Assessing the impact of initial nonresponse and attrition in the analysis of unemployment duration with panel surveys

Marjo Pyy-Martikainen · Ulrich Rendtel

Received: 2 October 2007 / Accepted: 12 March 2008 / Published online: 10 April 2008

Abstract We show how register data combined at person-level with survey data can be used to conduct a novel type of nonresponse analysis in a panel survey. The availability of register data provides a unique opportunity to directly test the type of the missingness mechanism as well as estimate the size of bias due to initial nonresponse and attrition. We are also able to study in-depth the determinants of initial nonresponse and attrition. We use the Finnish subset of the European Community Household Panel (FI ECHP) data combined with register panel data and unemployment spells as outcome variables of interest. Our results show that initial nonresponse and attrition are clearly different processes driven by different background variables. Both the initial nonresponse and attrition mechanisms are nonignorable with respect to analysis of unemployment spells. Finally, our results suggest that initial nonresponse may play a role at least as important as attrition in causing bias. This result challenges the common view of attrition being the main threat to the value of panel data.

Keywords Combined survey-register data · Nonresponse analysis · Duration of unemployment

M. Pyy-Martikainen (✉)
Department of Economics and Statistics, Åbo Akademi University and Statistics Finland, 00022 Statistics, Finland
e-mail: marjo.pyy-martikainen@stat.fi

U. Rendtel
Institute of Statistics and Econometrics, Freie Universität Berlin, Gary Str. 21, 14195 Berlin, Germany
e-mail: rendtel@wiwiss.fu-berlin.de

1 Introduction

It is usually assumed that the mechanism generating missing values in survey data is ignorable. This means that, conditional on covariates, the probability that study variables are missing is not related to their underlying values. The majority of available methods used to correct for nonresponse assume an ignorable missing data mechanism. If the assumption of ignorability is not valid, then an analysis which ignores the missing data mechanism produces biased results. However, the validity of this assumption is usually impossible to test because we do not observe the values of study variables for the nonrespondents. Comparisons of survey data estimates with external data are often conducted in order to evaluate the size of bias due to nonresponse and, in this way, to obtain indirect information on the type of the missing data mechanism. However, these comparisons may be hampered by measurement errors, differences in target populations or in the definitions of variables in two data sources.

In this article, we present a novel type of survey nonresponse analysis by making intensive use of register data. We use the first five waves of the Finnish European Community Household Panel (FI ECHP) survey data combined at person-level with longitudinal register data (for a review of the ECHP, see Peracchi (2002); the FI ECHP was documented by Pyy-Martikainen et al. (2004)). We use unemployment spells as our study variables of interest. The register data are used as a source of information on unemployment spells and covariates. This information is available both for respondents and nonrespondents. The survey data are used only to obtain the result of the interview. This way we get directly comparable information for respondents and nonrespondents, and are able to detect a pure nonresponse effect free from measurement errors (for a study on measurement errors in income data based on FI ECHP, see Hovi et al. 2000). The availability of longitudinal register data provides a unique opportunity to directly test the type of missing data mechanism—whether ignorable or nonignorable—as well as estimate the size of bias due to initial nonresponse and attrition.

There are few other studies of survey nonresponse that make use of register data. Sisto (2003) studied the effects of initial nonresponse and attrition on various estimates of income distribution. Her study is based on the same data set used in this paper. She found out that initial nonresponse biases most estimates more than attrition. Moreover, she found no evidence of a trend towards a growing bias during the panel. Pyy-Martikainen and Rendtel (2003) studied the effects of survey attrition in the analysis of unemployment spells using the same data set as in this paper. They tested the validity of the assumption of independent censoring of unemployment spells. According to their results, the probability of an unemployment spell being censored was not related to the (normally unobserved) remaining length of the spell. Thus, they concluded that the assumption of independent censoring was valid. They also evaluated the size of bias due to survey attrition. No attrition bias was found in the estimated survival functions. The estimates of regression coefficients of a Cox proportional hazard model were slightly biased. Van den Berg et al. (2006) evaluated the size of bias due to initial nonresponse in a model for the duration of unemployment. They used combined survey-register data that were originally gathered in order to evaluate the impact of a "Restart" policy programme for unemployed workers in

the UK. The first wave of the survey was used to obtain persons' response status, whereas data on unemployment duration and covariates were taken from the register. The data did not enable distinguishing between different destination states out of unemployment. By comparing the estimates based on a full sample with the estimates calculated only on the basis of respondents, Van den Berg et al. (2006) concluded that the baseline hazard function was underestimated and did not decrease as fast as in the population. However, the covariate effects were found to be correctly estimated on the basis of the sample of respondents.

Register data are usually considered as more precise than survey data. However, there can be errors in register data as well. Bring and Carling (2000) studied the effects of attrition from the unemployment register on the estimates of an unemployment duration model. They conducted a survey of drop-outs from the register in order to find out whether the reason for the drop-out was because the person became employed. On the basis of survey responses and register data, they estimated a logit model explaining the probability of attrition due to employment and used this model to impute outcomes of drop-outs in the register data. The unemployment duration model was then re-estimated with imputed data. They found that the baseline hazard function was underestimated under a false assumption of noninformative attrition. The effects of attrition on estimates of covariate effects were negligible.

This article extends the work by Pyy-Martikainen and Rendtel (2003) in several ways. First, we incorporate initial nonrespondents in the analysis. In FI ECHP, initial nonrespondents constitute 27% of sample persons eligible for interview (Pyy-Martikainen et al. 2004). Even though nonignorable attrition is often considered as one of the most serious disadvantages of panel data (Fitzgerald et al. 1998; Duncan 2000; Trivellato 1999), little is known about the relative importance of initial nonresponse and attrition in biasing the estimation results. It may well be that nonresponse at the start of the panel is more selective with respect to unemployment duration than nonresponse at later waves. With our combined survey-register data, we are able to show novel evidence on the relative importance of initial nonresponse and attrition in biasing estimates. We restrict our attention to transitions from unemployment to employment as this is usually the event of main interest in an econometric unemployment duration analysis. Second, we conduct an in-depth analysis of the mechanisms leading to initial nonresponse and attrition and test the assumption of an ignorable missing data mechanism by studying whether the normally unobserved values of study variables affect the probability of nonresponse. We estimate the size of bias due to initial nonresponse and attrition by comparing data sets restricted by nonresponse to a benchmark data set that uses the entire register information without any restrictions by nonresponse. Third, we extend the Cox proportional hazard model used in our earlier paper by incorporating a person-specific frailty term that allows dependency between spells by the same person. We take a model-based approach to the analysis. Consequently, no survey weights are used.

Section 2 describes the classification of missing data mechanisms introduced by Rubin (1976) in the context of longitudinal data. Section 3 describes the data and the patterns of missingness observed there. Section 4 conducts an analysis of the determinants of initial nonresponse and attrition, as well as tests for the type of the missingness mechanisms. Section 5 develops a taxonomy of unemployment spells

that classifies spells according to persons' response status. Section 6 uses this taxonomy to construct different sets of unemployment spells that are used to estimate the size of bias due to initial nonresponse and attrition. Section 7 concludes by discussing the results.

2 Classification of missing data mechanisms

Rubin's (1976) classification of missing data mechanisms helps to clarify the conditions under which missingness may be ignored. The key issue is whether the fact that variables are missing is related to the underlying values of the variables in the data set (Little and Rubin 2002).

Let $Y = (y_{ij})$ be a $(n \times K)$ data matrix of outcomes of random variables Y_{ij}, where $i = 1, \ldots, n$ indexes persons and $j = 1, \ldots, K$ indexes panel waves. The variables Y_{ij} could measure, for example, the main activity status of person i at wave j. The Y_{ij}'s may also be random vectors, as is the case when persons are interviewed every year and at year j interview, retrospective monthly information about main activity status during year $j-1$ is collected. Let $X = (x_1, \ldots, x_n)'$ be the $(n \times p)$ matrix of fixed covariates that are assumed to be fully observed. The ith row of X, $(x_{i1}, \ldots, x_{ip})$, contains the covariates of person i. Define $M = (M_{ij})$ the $(n \times K)$ matrix of missing data indicators such that $M_{ij} = 1$ if y_{ij} is missing and $M_{ij} = 0$ if y_{ij} is observed. Let $f(Y \mid X, \theta)$ be the model on the basis of which we want to make inferences of the parameters θ. Divide the data matrix Y into observed and missing parts: $Y = (Y_{\text{obs}}, Y_{\text{mis}})$. Y_{obs} corresponds to values of study variables before attrition and Y_{mis} corresponds to values at the time or after attrition. For the initial nonrespondents, all values of Y_{ij} are missing. The observed data consist of (Y_{obs}, M). The joint conditional density of the observed data, given X, can be obtained by integrating Y_{mis} out of the joint conditional density of Y and M:

$$f(Y_{\text{obs}}, M \mid X, \theta, \phi) = \int f(Y_{\text{obs}}, Y_{\text{mis}} \mid X, \theta) f(M \mid X, Y_{\text{obs}}, Y_{\text{mis}}, \phi)\, dY_{\text{mis}}.$$

The conditional distribution of M given the data (Y, X), $f(M \mid X, Y_{\text{obs}}, Y_{\text{mis}}, \phi)$, describes the missing data mechanism governed by unknown parameters ϕ. Data are said to be *missing completely at random* (MCAR) if

$$f(M \mid X, Y_{\text{obs}}, Y_{\text{mis}}, \phi) = f(M \mid \phi),$$

i.e. if the missing data mechanism does not depend on covariates or any values of the study variables, missing or observed. A more realistic assumption is that data are *missing at random* (MAR). MAR allows the missingness mechanism to be related to covariates and observed outcomes of the study variables:

$$f(M \mid X, Y_{\text{obs}}, Y_{\text{mis}}, \phi) = f(M \mid X, Y_{\text{obs}}, \phi).$$

In the context of event history data, MAR means that nonresponse is independent of current and future events, given past events and covariates. When MAR holds, the

distribution of observed data can be written

$$f(Y_{\text{obs}}, M \mid X, \theta, \phi) = f(M \mid X, Y_{\text{obs}}, \phi) \int f(Y_{\text{obs}}, Y_{\text{mis}} \mid X, \theta)\, dY_{\text{mis}}$$
$$= f(M \mid X, Y_{\text{obs}}, \phi) f(Y_{\text{obs}} \mid X, \theta).$$

If, in addition, the parameters θ and ϕ are distinct, the missing data mechanism is called *ignorable*. This means that one need not model the missing data mechanism when making inferences of θ. If MAR does not hold, i.e. if data are *missing not at random* (MNAR), one has to model the missing data mechanism in order to get valid estimates of θ. An example of a MNAR mechanism is a situation where persons with long unemployment spells drop out from the survey more frequently than otherwise similar persons with shorter spells. In this case, the estimated distribution of unemployment duration will be biased toward short spells. If, in addition, the covariate effects differ among persons with long and short spells, there will be bias in the estimated covariate effects, too.

3 Data

In the ECHP, questions related to individual labour market histories are asked in the form of a month-by-month main activity status calendar obtained retrospectively for the preceding year. By combining calendars from consecutive waves it is possible to get information on individual labour market histories for several years. However, the measurement of unemployment spells by questionnaires is plagued by several problems. Respondents tend to forget about short episodes of unemployment. The tendency gets stronger the longer the event lies in the past (Brown et al. 1990). Also, in the retrospective annual reporting scheme ranging from January to December, respondents tend to heap up end points in December and starting points in January (Steiner and Kraus 1995). Therefore, in order to detect a pure nonresponse effect free from measurement error, information about unemployment spells was taken from administrative registers. Another advantage of using only register-based unemployment spells is that this way, unemployment is defined in an identical way for both respondents and nonrespondents. Data on unemployment spells were taken from the register of job seekers compiled by the Ministry of Labour. Information about unemployment spells consists of the starting date of the spell, the ending date of the spell and the cause of ending of the spell. The background variables used in the analysis were also taken from various administrative registers. The survey data were used only to obtain information of the occurrence and timing of nonresponse.[1]

The first five waves of the FI ECHP survey data covering the years 1996–2000 were used in the analysis. The survey and register data were linked at person-level by

[1] This approach relies on the assumption that the definitions of unemployment in survey and register are close enough to each other. Indeed, this seems to be the case. In FI ECHP, an unemployed person is defined as being without a job, available for work and looking for work through the employment office or newspaper advertisements or some other way. In the register, an unemployed job seeker is without a job and seeking a new job. Registering with the employment office is considered as evidence of seeking a job.

Table 1 Distribution of missingness patterns of the 11,641 sample persons

Pattern	w1	w2	w3	w4	w5	Frequency	Percent
Total respondents	0	0	0	0	0	4,364	37.5
Attrition at wave 5	0	0	0	0	1	1,486	12.8
Attrition at wave 4	0	0	0	1	1	469	4.0
Attrition at wave 3	0	0	1	1	1	680	5.8
Attrition at wave 2	0	1	1	1	1	575	4.9
Initial non-respondents	1	1	1	1	1	3,146	27.0
Temporary drop-outs						921	7.9
All						11,641	100.0

Table 2 Distribution of number of spells among the 10,720 sample persons having a regular response pattern

# spells	Frequency	Percent
0	7,762	72.4
1	891	8.3
2	585	5.5
3	438	4.1
4	300	2.8
5	204	1.9
6–10	394	3.7
11 or more	146	1.4
All	10,720	100.0

personal identification numbers. Our analysis was based on the unemployment spells from the 11,641 sample persons aged 16 or over at the beginning of 1996. Sample persons are defined in the ECHP as all members of the initial sample of households.

Table 1 shows the distribution of missingness patterns of the 11,641 sample persons. Value 0 refers to observed data and value 1 to missing data. Of all sample persons, 37.5% responded in each of the five interviews. This group of total respondents also includes the small group of persons who exited the survey population during waves two to five. Exits from the survey population occurred because of death, moving abroad or into an institution. Attriters constituted 27.6% of sample persons. Slightly fewer, 27.0%, of sample persons did not respond in any of the survey waves. Most of these initial nonrespondents were wave one nonrespondents that were not forwarded to wave two (for the follow-up rules implemented in the FI ECHP, see Pyy-Martikainen et al. (2004)). Temporary drop-outs are persons who do not participate in one wave but re-enter the panel in the next wave. Of all sample persons, 7.9% dropped temporarily out of the panel. For simplicity, they were excluded from the analysis. After this exclusion we were left with 10,720 sample persons. As unemployment is asked in a retrospective manner for the previous year, the period of observation was chosen as 1 January 1995 to 31 December 1999. Spells by sample persons beginning during this period were chosen for the analysis. Table 2 shows how

the spells are distributed among the 10,720 sample persons. The majority of the sample persons had no unemployment spells at all during the five-year period. Among those having one spell or more, the mean number of spells was 3.7.

4 Determinants of initial nonresponse and attrition

We estimated models for nonresponse to find out the determinants of and to test the type of the missingness mechanism. The mechanisms leading to initial nonresponse and attrition are likely to be different from each other (Lepkowski and Couper 2002). Therefore, we estimated separate models for initial nonresponse and attrition. Although in attrition analysis there is the possibility to use the survey responses obtained from waves prior to attrition, we used in both the initial nonresponse model and the attrition model only explanatory variables obtained from registers in order to maintain comparability of results between the models.

We restricted the analysis to sample persons having at least one spell of unemployment during the observation period because our aim was to study how nonresponse affects an analysis of unemployment duration. There were 2956 sample persons eligible for the analysis. Three sets of covariates were used in the analyses. The first covariate set, the *spell covariates*, were constructed from the unemployment spell information. The spell covariates were used to test the assumption of ignorability of the initial nonresponse and attrition mechanisms. The second covariate set consists of the *spell covariates plus covariates from the model of interest*, which is the unemployment duration model. Our aim was to find out whether the covariates from the unemployment duration model help to protect against nonignorable nonresponse. The third covariate set enlarges the second set by including covariates that were found to explain nonresponse in preliminary analyses (Pyy-Martikainen et al. 2004). This *full covariate set* was used to get a better picture of the processes leading to initial nonresponse and attrition. We also maintained the spell covariates in the third set in order to see whether their effect is attenuated when a rich set of other covariates is controlled for.

The spell covariates consist of the number of days spent in unemployment as well as the number of unemployment spells. For the initial nonresponse analysis, the number of unemployment days and the number of spells were calculated both before and after the time of interview (or time of contact, if an interview was not obtained) in 1996. In the attrition analysis, the number of unemployment days and the number of spells were calculated for each year t before and after the last obtained interview, i.e. the interview at year $t-1$. If the initial nonresponse mechanism is MAR, none of the spell covariates should explain the probability of nonresponse. In the attrition model, a MAR nonresponse mechanism implies that the spell covariates measured after the last obtained interview should not affect the probability of nonresponse. We were thus able to test the type of the missingness mechanisms by looking at the statistical significance of the spell covariates. This would, of course, normally not be possible. However, the availability of register data provides in this case a unique opportunity to directly test the type of the missingness mechanisms.

The covariates from the model of interest are described in Sect. 6.2. We used a subset of the covariates of the unemployment duration model which are not directly

related to the spells being modeled. Thus, for example, the starting year of the spell was not used in the initial nonresponse and attrition models. The following covariates from the unemployment duration model were used: sex, age, level of education, residential area at NUTS2 level, and degree of urbanisation of the municipality. In the initial nonresponse analysis, the variables were measured at the end of 1995. In the attrition analysis, the variable values refer to the end of the years 1996–1999. For example, the hazard of attrition in 2000 is explained by values of the covariates at the end of 1999.

The additional covariates used in the full covariate set were: marital status; main activity status with three classes: employed, unemployed, out of labour force (OLF); household disposable income quartiles; family size; household socio-economic status with six classes: wage-earners, entrepreneurs, farmers, pensioners, and other (e.g. students). Household socioeconomic status was derived from the sample stratum information (for the sampling design of FI ECHP, see Pyy-Martikainen et al. (2004)) and it refers to year 1995. The reference time for other covariates is defined in the same way as for the covariates from the model of interest. Household disposable income at a specific year is the amount obtained during the whole year. The covariate means (covariate values evaluated at the end of 1995) calculated separately for total respondents, attriters and initial nonrespondents are shown in Appendix A.

The initial nonresponse analysis was conducted by estimating logit models for the probability of being a nonrespondent at the first wave of the panel. The initial nonresponse analysis includes all the 2956 sample persons having at least one unemployment spell during the observation period. The attrition process was modeled by a discrete-time hazard model (Cox 1972) where the conditional probability of attrition at a specific year, given that the person has remained in the survey until the year in question, is explained by a set of time-varying covariates. Excluding initial nonrespondents leaves 2085 sample persons for the attrition analysis.

The results of the initial nonresponse analysis are shown in Table 3. In the first two columns, estimates from the model with spell covariates only are reported. We tried several transformations of the number of unemployment days variable to detect a possible nonlinear effect: the squared number of days; unemployment time as a proportion of follow-up time before/after interview; the squared proportion; number of unemployment days quartile indicators as well as an indicator for belonging to the 10. decile. For the number of spells, the following variable transformations were tested: an indicator whether a person had at least one spell before/after the interview and an indicator of belonging to the 10. decile. With the only exception being the indicator of unemployment time after time of interview belonging to the 10. decile, the transformed variables did not have more explanatory power (measured by p-values of coefficients) than the untransformed ones. Looking at the spell covariates in the first two columns of Table 3, we see that there is evidence of nonignorable nonresponse. Belonging to the 10. decile with respect to number of unemployment days after time of interview raises the odds of nonresponse by 27.6%.[2] The explanatory power of the model is very low: pseudo-$R^2 = 0.002$.[3]

[2]Calculated as $(\exp(0.244) - 1)$.

[3]Pseudo-$R^2 = (D_0 - D_M)/D_0$, where D_0 is -2 times the log-likelihood of a model with intercept only and D_M is the corresponding measure for the model of interest.

Table 3 Initial nonresponse analysis. Estimates of logit models

Variable	Spell covariates		Spell covariates plus covariates from the model of interest		Full covariate set	
	$\hat{\beta}$	(*s.e.*)	$\hat{\beta}$	(*s.e.*)	$\hat{\beta}$	(*s.e.*)
Intercept	**−0.897**	(0.056)	−0.184	(0.357)	**−2.884**	(0.493)
No. spells before	−0.013	(0.019)	−0.009	(0.018)	0.005	(0.017)
No. spells after	−0.009	(0.008)	−0.004	(0.008)	0.000	(0.008)
UE[a] time before[b]	0.044	(0.037)	−0.050	(0.038)	−0.066	(0.161)
UE[a] time after in top 10%	*0.244*	(0.134)	*0.266*	(0.140)	0.197	(0.151)
Woman			−0.133	(0.082)	−0.115	(0.089)
Age			−0.012	(0.023)	**0.116**	(0.028)
Age squared			0.000	(0.000)	**−0.001**	(0.000)
Upper secondary education			−0.024	(0.094)	−0.095	(0.101)
Higher education			**−0.421**	(0.171)	**−0.650**	(0.183)
Semi-urban municipality			**−0.284**	(0.117)	*−0.249*	(0.126)
Rural municipality			**−0.431**	(0.105)	**−0.489**	(0.116)
Southern Finland			−0.184	(0.112)	−0.136	(0.120)
Eastern Finland			**−0.511**	(0.147)	**−0.422**	(0.157)
Central Finland			−0.208	(0.146)	−0.082	(0.156)
Northern Finland			−0.127	(0.151)	−0.027	(0.163)
Married					**−0.445**	(0.125)
Unemployed					*0.244*	(0.128)
Out of labour force					**0.361**	(0.131)
HH disposable income in Q2					0.219	(0.141)
HH disposable income in Q3					**0.803**	(0.146)
HH disposable income in Q4					**2.136**	(0.150)
Entrepreneur					**0.750**	(0.160)
Farmer					0.153	(0.180)
Pensioner					**0.770**	(0.221)
Other					**0.392**	(0.107)
Family size					**−0.212**	(0.040)
$-2\log L$	3,576.7		3,520.5		3,182.7	
Pseudo-R^2	0.002		0.018		0.112	
Number of persons: 2956						
Estimates significant at 5% (10%) risk level are displayed in **boldface** (*italics*)						

[a]UE unemployment

[b]Coefficients and standard errors multiplied by 100

The effect of unemployment time after time of interview remains significant at the 10% level even after inclusion of covariates from the unemployment duration model. Persons with a high level of education and living in semi-urban or rural municipalities in Eastern Finland are more likely to respond than other persons.

The effect of being in the 10. decile with respect to unemployment time loses its statistical significance when additional covariates are included in the model. Age becomes a statistically significant predictor of nonresponse. The probability of nonresponse rises along with age until the age of approximately 40 years and then starts to decrease. The effects of education, degree of urbanisation and area of living remain qualitatively the same as before. The effect of having a high level of education grows in absolute value. The new covariates raise the pseudo-R^2 to 0.112, which is almost ten times larger than in the model including only covariates from the unemployment duration model. Married persons are more likely to respond than the unmarried. Having been unemployed or outside the labour force at the end of 1995 increases the probability of nonresponse relative to those employed at the end of 1995. The higher the level of household disposable income, the more probable nonresponse is. Entrepreneurs, pensioners and households with other socio-economic statuses have a higher probability of nonresponse than wage-earners. Finally, a larger family size is related to a higher probability of response.

The results from the discrete-time attrition hazard model are reported in Table 4. We also estimated separate logit models for each year in order to see whether the covariate effects differ from year to year. The estimates from the separate logit models are not shown, but they are discussed in the following text when relevant.

The year dummies show the dependence of the attrition hazard on time. The attrition hazard is roughly constant during 1997–1999 and rises sharply in 2000. The data collection of the ECHP was joined with the Income Distribution Survey (IDS) during 1996–1997. As the IDS is a two-year panel survey, the ECHP was continued from 1998 as a stand-alone survey. This may be reflected as a small rise in the attrition hazard in 1998. The sharp rise in 2000 reflects the fact that the fieldwork of year 2000 was particularly difficult due to previous uncertainty about the continuation of the panel (see Pyy-Martikainen et al. 2004).

As for the number of unemployment days and number of unemployment spells variables, the following transformations were tested in order to detect a possible nonlinear effect: the squared number before/after last obtained interview and an indicator of belonging to the 10. decile. For the number of unemployment spells variables, an indicator whether a person had at least one spell before/after the last obtained interview was also tested. It turned out, however, that none of the transformed covariates had explanatory power. The amount of unemployment time after the last obtained interview is statistically significant at 10% level. However, the effect of unemployment time is small: an increase of 100 unemployment days after last obtained interview increases the odds of attrition hazard only by 3%.

The effect of unemployment time after the last obtained interview maintains its statistical significance also when covariates from the unemployment duration model are controlled for. This suggests that attrition is nonignorable with respect to the analysis of unemployment spells. However, the resulting bias is likely to be small due to the small magnitude of the effect.

The older the person, the more probably he or she stays in the panel. We did not include squared age in the attrition model as the squared term interfered with the estimation of the baseline hazard parameters. Having a high level of education decreases the odds of attrition hazard by roughly 30%. This effect is mainly due

Table 4 Attrition analysis. Estimates of discrete-time hazard models

Variable	Spell covariates		Spell covariates plus covariates from the model of interest		Full covariate set	
	$\hat{\beta}$	(*s.e.*)	$\hat{\beta}$	(*s.e.*)	$\hat{\beta}$	(*s.e.*)
Year 1997	**−2.430**	(0.098)	**−2.078**	(0.171)	**−1.945**	(0.209)
Year 1998	**−2.132**	(0.086)	**−1.757**	(0.169)	**−1.627**	(0.204)
Year 1999	**−2.440**	(0.099)	**−2.046**	(0.179)	**−1.931**	(0.211)
Year 2000	**−1.153**	(0.081)	**−0.733**	(0.173)	**−0.617**	(0.205)
N spells before	−0.001	(0.006)	−0.002	(0.006)	−0.002	(0.006)
N spells after	0.001	(0.008)	−0.001	(0.008)	−0.001	(0.008)
UE[a] time before[b]	0.011	(0.013)	0.020	(0.013)	0.016	(0.014)
UE[a] time after[b]	*0.030*	(0.016)	**0.040**	(0.017)	0.029	(0.019)
Woman			0.015	(0.073)	0.009	(0.074)
Age			**−0.010**	(0.003)	**−0.010**	(0.004)
Upper secondary education			−0.128	(0.081)	−0.130	(0.082)
Higher education			**−0.360**	(0.133)	**−0.340**	(0.136)
Semi-urban municipality			−0.105	(0.103)	−0.098	(0.105)
Rural municipality			−0.110	(0.091)	−0.145	(0.095)
Southern Finland			0.095	(0.108)	0.080	(0.109)
Eastern Finland			0.030	(0.129)	0.008	(0.131)
Central Finland			0.001	(0.137)	−0.030	(0.138)
Northern Finland			**0.573**	(0.133)	**0.575**	(0.135)
Married					0.050	(0.095)
Unemployed					0.074	(0.103)
Out of labour force					−0.065	(0.098)
HH disposable income in Q2					**−0.258**	(0.106)
HH disposable income in Q3					**−0.436**	(0.116)
HH disposable income in Q4					*−0.201*	(0.114)
Entrepreneur					0.171	(0.138)
Farmer					0.191	(0.135)
Pensioner					0.128	(0.209)
Other					−0.063	(0.086)
Family size					0.025	(0.029)
$-2\log L$	5,282.6		5,236.9		5,215.1	
Pseudo-R^2	0.041		0.050		0.053	
Number of persons: 2085						
Estimates significant at 5% (10%) risk level are displayed in **boldface** (*italics*)						

[a]UE unemployment

[b]Coefficients and standard errors multiplied by 100

to the strong effect of education in year 2000, which is revealed by looking at the separate models for each year. The effect of residential area is also different in 2000 from the other years. Compared to persons living in Uusimaa, the capital region, persons living in southern Finland or in eastern Finland have a higher probability of staying in the survey during 1997–1999, whereas in 2000, the effect is strongly of the opposite sign. It seems that in these regions, a rapid negative change in the attitudes towards the survey occurred in 2000. Persons living in northern Finland were the least willing to continue in the survey in 2000: their odds of attrition hazard are roughly three times higher compared to persons living in Uusimaa (estimate from a model not reported here). The estimates of the discrete-time hazard model average the effect of residential area over the years. Thus, only the dummy for northern Finland maintains its statistical significance in the dynamic model.

The last two columns of Table 4 show the results using the full set of covariates. Unemployment time after the last obtained interview is no longer statistically significant, as none of the other spell covariates. The estimates of age, education and residential area remain roughly the same as in the previous model. Persons with household disposable income in the second, third or fourth quartile are *less* likely to attrite compared to persons in the first quartile. Persons with highest household income levels are not likely to participate at all in the survey, which means that respondents at wave one are already a selected sample with respect to level of income.

It is possible that major *changes* in life affect the attrition hazard. For example, it may well be that a change in marital status, not the status per se, affects the hazard of attrition. We tested the effect of a divorce, a move from employment to unemployment and a change of residential area in the hazard of attrition. Somewhat surprisingly, however, none of these covariates had any explanatory power.

The explanatory power of the attrition models remain low irrespective of the covariates included. There seems to be more randomness in the attrition process than in the initial nonresponse process.

5 A taxonomy of unemployment spells

For each sample person having one or more spells during the observation period, the participation behaviour in the survey is known. It is assumed that unemployment spells are observed until the time of the last interview or until the end of the observation period, whichever comes first. This creates a number of different cases:

(a) Spells that end before the last interview (or before 31 December 1999, whichever comes first) are regarded as *fully observed.*
(b) Spells ongoing at the time of the last interview, which is followed by attrition, are regarded as *right censored by attrition* at the time of the last interview.
(c) Spells that start after the last interview, which is followed by attrition, are *not observed by attrition.*
(d) Spells by persons without any interviews are *not observed by initial nonresponse.*

Unless censored by attrition, spells ongoing at 31 December 1999 are censored by the end of the follow-up period. Table 5 shows the distribution of the type of

Table 5 Taxonomy of unemployment spells

Taxonomy	Frequency	Percent
(a) Fully observed	6,243	58.2
(b) Censored by attrition	253	2.4
(c) Not observed by attrition	1,216	11.3
(d) Not observed by initial nonresponse	3,022	28.2
All	10,734	100.0

unemployment spells. Spells lasting at most two days were excluded from the analysis as they were not considered as "true" unemployment spells but just registrations into the records of the employment office for some legislative reason. This way we got 10,734 spells from 2,930 persons. At the level of unemployment spells, the most important pattern of nonresponse was initial nonresponse: 28.2% of all spells were not observed for this reason. Further, 11.3% of spells were not observed by attrition whereas only 2.4% of spells were right-censored for the same reason. The small percentage of spells right-censored by attrition is a consequence of the high frequency of spells with short duration. In the context of event history analysis, discussion of missing data has concentrated mostly on the right censoring of event times. However, as Table 5 indicates, spells being unobserved by attrition or by initial nonresponse may be a far more important issue in event history analysis based on survey data.

The boxplots in Appendix B show the distribution of spell length according to the taxonomy of spells. The widths of the boxes reflect the relative sample sizes in different categories. For spells right-censored by attrition, the whole duration is used.

6 Size of bias due to initial nonresponse and attrition

On the basis of the taxonomy developed in the previous section, three different sets of unemployment spells were constructed:

The *full information* set of spells uses the entire register information without restrictions by initial nonresponse or attrition.
The *partial information* set of spells is a subset of the full information set of spells, obtained by excluding spells unobserved by initial nonresponse.
The *observed information* set of spells is a subset of the partial information set of spells, obtained by excluding spells unobserved by attrition and the remaining length of the spells censored by attrition.

The full information set of spells consists of 10,734 spells, whereas the partial information set of spells and the observed information set of spells contain 7,712 and 6,496 spells, respectively. The size of bias due to initial nonresponse and attrition was evaluated by comparing Kaplan–Meier estimates of survival function and estimates of regression coefficients from a Cox shared frailty model based on the three sets of unemployment spells. The difference between full information estimates and partial

information estimates shows the size of bias due to initial nonresponse, whereas the difference between partial information estimates and observed information estimates shows the size of bias due to attrition. The size of bias due to both initial nonresponse and attrition can be evaluated by comparing full information estimates and observed information estimates. Note that the observed information set of spells corresponds to the set of spells normally available for a survey data analyst.

6.1 Kaplan–Meier estimates

The Kaplan–Meier estimator is a nonparametric estimator of the survival function. Let $t_1 < t_2 < \cdots < t_k$ be the distinct observed unemployment durations. For each $l = 1, \ldots, k$, define the risk set $R(t_l)$ as the set of spells ongoing just before time t_l. Let r_l be the size of the risk set and d_l^c be the number of spells with outcome c at time t_l. The Kaplan–Meier estimator for outcome c is $S_c(t_l) = \prod_{j=1}^{l}(1 - d_j^c/r_j)$. This is an estimator of the cause-specific survival function $S_c(t) = P\ (T \geq t, C = c)$ that gives the probability that an event of type c occurs later than time t (Allison 1995). The value of $S_c(t)$ changes each time a spell is ended by outcome c. Spells with length t_l but with other outcomes, as well as spells censored at t_l affect only the risk sets $R_1, \ldots, R_{t_l}$. The overall survival function that ignores the reason for the ending of a spell can be obtained as a product of the cause-specific survival functions. The outcome of interest in our analysis is getting employed. Of all unemployment spells exceeding two days, 47.6% ended because of getting employed. Other spells ended because of transition out of labour force, subsidised work or for other reasons.

Figure 1 shows the estimates of cause-specific survival functions for the full information, partial information and observed information sets of spells along with 95% confidence intervals for the full information survival function. At most time points, the partial information and observed information estimates are below the lower confidence bound of the full information estimate. The partial information and observed information estimates are virtually identical. Thus, the downward bias in the survival curves is caused by initial nonresponse and not by attrition. This is confirmed by plotting the survival curves by response status separately for initial nonrespondents, attriters and total respondents (figure not shown here). Initial nonrespondents tend to get employed much more slowly than attriters and total respondents. Estimates of survival functions along with their standard errors at time points $t = 3, 100, 200, 300, \ldots$ are shown in Appendix C.

6.2 Estimates from Cox shared frailty models

In order to detect the possible bias in the estimates of covariate effects, we estimated Cox shared frailty models (see e.g. Therneau and Grambsch 2000 or Therneau et al. 2003). The shared frailty model allows us to take into account the possible correlation

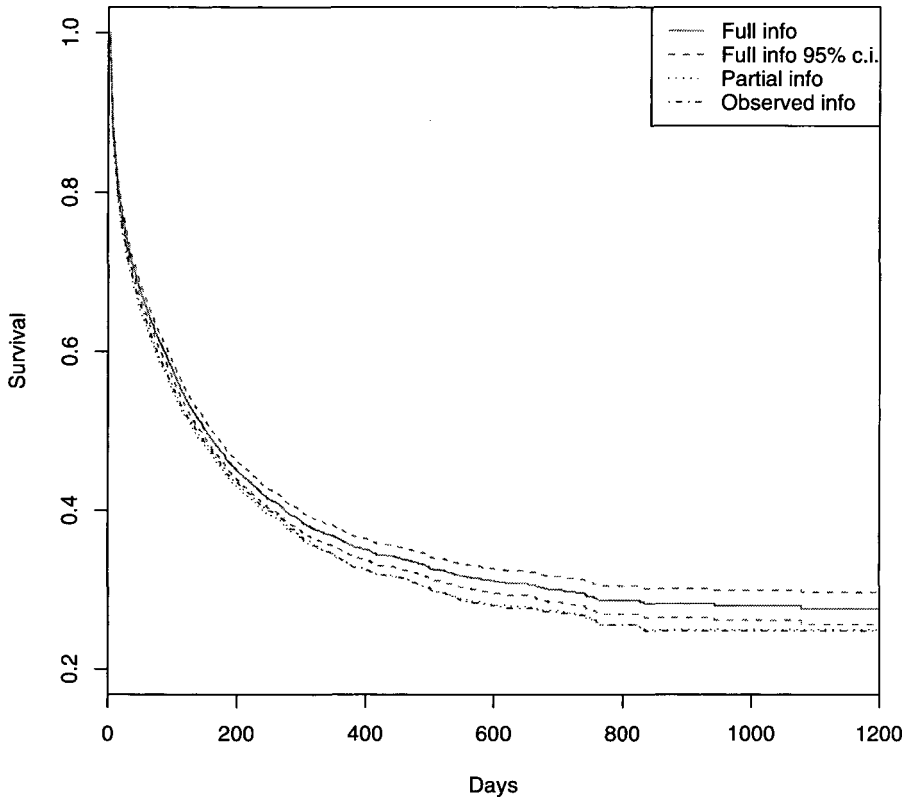


Fig. 1 Bias due to initial nonresponse and attrition in estimates of survival function

between spells from the same person.[4] We have q persons generating altogether n spells. The shared frailty model specifies the cause-specific hazard function of spell j and outcome c in the following way:

$$\lambda_{jc}(t) = \lambda_{0c}(t)\exp(x_j\beta_c + z_j\omega_c), \quad j = 1, \ldots, n,$$

where $\lambda_{0c}(t)$ is an unspecified function of time, x_j is a $(1 \times p)$ vector of covariates related to spell j and β_c is a $(p \times 1)$ vector of regression parameters. The $(q \times 1)$ vector ω_c contains random variables or frailties that measure the effect of unobserved covariates on the hazard and z_j is a $(1 \times q)$ vector of indicator variables such that $z_{ij} = 1$ when spell j belongs to person i and 0 otherwise. The ω_{ic}'s are assumed to be common across spells from the same person. They could measure, for example, persons' motivation to search for a job. Because of this common unknown factor, spells from the same person are positively correlated. The ω_{ic}'s are distributed as

[4]The shared frailty model is also in line with economic theories of job search where it is the *individual* hazard of finding a job that is of interest, as opposed to the population averaged hazard estimated by nonfrailty models (van den Berg 2001).

the logarithms of iid gamma random variables with mean one and variance σ.[5] The model is estimated by maximising a penalised partial log-likelihood.

The set of explanatory variables used is similar to those used in econometric analyses of unemployment duration (see e.g. Meyer 1990; Carling et al. 1996; Abbring et al. 2005). The variables are spell-specific and they are usually measured at the end of the year preceding the start of the unemployment spell. Age is measured in years. Level of education divides persons into three classes. Basic education corresponds to the completion of comprehensive school. Upper secondary education comprises matriculation examination and upper secondary vocational education. Higher education comprises, for example, vocational college education and university education. The possible state dependency in unemployment durations is measured by the proportion of time (since 1 January 1995) spent in unemployment before the spell in question. Variation in local labour market conditions is taken into account by information on residential area and statistical grouping of municipalities. The residential area dummies are based on the NUTS2 classification of regions. The statistical grouping of municipalities divides municipalities into urban, semi-urban and rural by the proportion of the population living in urban settlements and by the population of the largest urban settlement. Earnings-related unemployment benefit indicates whether a person has received this kind of benefit at the starting year of the unemployment spell. This variable, or variants of it, is often the variable of main interest in an econometric unemployment duration analysis. Year dummies indicating the starting year of the unemployment spell aim at capturing the effect of economic fluctuations over time.

The estimation results of Cox shared frailty models are shown in Table 6. Efron's (1977) method was used to handle tied event times. The standard error estimates are based on the inverse of the second derivative matrix for the penalised log-likelihood. In order to compare the estimates from the full and the restricted data sets, we use a variant of the Hausman test. Hausman (1978) used the asymptotic result that the variance of the difference of an efficient and a consistent estimator can be computed by the difference of the variances of the single estimates. In our application of this result we take the estimator on the full data set as the efficient estimate. Under the null-hypothesis that nonresponse is ignorable for the estimation of the model, i.e. conditional on the covariates of the model nonresponse is purely random, the estimation on the basis of the observed information is still consistent.[6] The Hausman test statistic is $(\hat{\beta}_{\text{res}} - \hat{\beta}_{\text{full}})'(\hat{\Sigma}_{\text{res}} - \hat{\Sigma}_{\text{full}})^{-1}(\hat{\beta}_{\text{res}} - \hat{\beta}_{\text{full}})$, where $\hat{\beta}_{\text{res}}$ and $\hat{\beta}_{\text{full}}$ are the $(p \times 1)$ vectors of parameter estimates and $\hat{\Sigma}_{\text{res}}$ and $\hat{\Sigma}_{\text{full}}$ the $(p \times p)$ covariance matrices of parameter estimates from the restricted model and full model. Under the null hypothesis of no bias, the Hausman test statistics is asymptotically χ^2_p distributed.

---

[5]The distribution of ω_{ic}'s is asymmetric, implying that some individuals have a very low exit rate from unemployment to employment. This seems to be a plausible assumption given the distribution of unemployment spells that is heavily skewed towards long durations.

[6]The behaviour of the efficient and the consistent estimator under the alternative is exchanged here. In econometric application the efficient estimator becomes inconsistent while the consistent estimator remains consistent under the alternative. As the Hausman test is evaluated under the null hypothesis this change is irrelevant here.

Table 6 Analysis of unemployment duration. Estimates of Cox shared frailty models

Variable	Full information		Partial information		Hausman test	Observed information		Hausman test
	$\hat{\beta}$	*s.e.*	$\hat{\beta}$	*s.e.*	*p*-value	$\hat{\beta}$	*s.e.*	*p*-value
Female	**0.131**	(0.062)	0.070	(0.074)	0.131	0.066	(0.081)	0.202
Age	**0.105**	(0.018)	**0.127**	(0.021)	0.051	**0.147**	(0.023)	0.006
Age squared	**−0.002**	(0.000)	**−0.002**	(0.000)	0.074	**−0.002**	(0.000)	0.009
Upper secondary educ.	*0.118*	(0.064)	0.052	(0.076)	0.105	0.088	(0.085)	0.595
Higher education	**0.504**	(0.105)	**0.373**	(0.125)	0.051	**0.452**	(0.134)	0.531
Proportion of UE[a] time	**0.011**	(0.001)	**0.012**	(0.001)	0.043	**0.014**	(0.001)	0.001
Semi urban municipality	**0.198**	(0.076)	**0.184**	(0.091)	0.794	0.126	(0.100)	0.276
Rural municipality	0.088	(0.071)	0.072	(0.082)	0.698	0.030	(0.091)	0.311
Southern Finland	**0.363**	(0.090)	**0.330**	(0.111)	0.607	**0.381**	(0.121)	0.818
Eastern Finland	**0.252**	(0.109)	0.146	(0.130)	0.135	0.196	(0.143)	0.539
Central Finland	**0.290**	(0.112)	*0.252*	(0.135)	0.611	*0.257*	(0.149)	0.736
Northern Finland	**0.452**	(0.113)	**0.393**	(0.136)	0.434	**0.461**	(0.151)	0.921
Earnings-rel. UE[a] benefit	0.012	(0.053)	−0.068	(0.062)	0.014	**−0.222**	(0.071)	0.000
Year 1996	*−0.100*	(0.052)	**−0.121**	(0.060)	0.488	**−0.142**	(0.063)	0.237
Year 1997	0.010	(0.054)	−0.057	(0.065)	0.034	−0.096	(0.071)	0.012
Year 1998	**0.145**	(0.057)	*0.109*	(0.064)	0.238	0.080	(0.071)	0.129
Year 1999	**0.306**	(0.061)	**0.262**	(0.070)	0.188	**0.252**	(0.079)	0.289
$\hat{\sigma}$	**1.61**		**1.63**			**1.78**		
Joint Hausman test *p*-value			0.080			0.001		
Number of spells	10,734		7,712			6,496		

Estimates significant at 5% (10%) risk level are displayed in **boldface** (*italics*)

[a]UE Unemployment

Looking at the full information estimates we see that females have a higher exit rate from unemployment to employment. The effect of age on the exit rate has an inverted u-shape. Higher levels of education increase the exit hazard. Somewhat unexpectedly, a larger proportion of time spent in unemployment before the current spell is related to a higher hazard of exit. There is also variation among the exit hazard with respect to local labour market conditions and economic fluctuations over time. According to the LR test,[7] the variance of frailty terms is highly significant. There is thus heterogeneity among the individuals that is not captured by observed explanatory variables.

According to Hausman tests for single covariates, the partial information model coefficient estimates of age, higher education, proportion of unemployment time, earnings-related unemployment benefit and year 1997 are statistically significantly (at 10% risk level) different from the full information estimates. Looking at the observed information estimates, we see that by and large the same differences are statistically significant, the significance levels being somewhat higher. Thus, both initial nonresponse and attrition cause bias in the coefficient estimates. Joint Hausman tests over all covariates comparing partial information and observed information estimates with full information estimates show statistically significant attrition biases (at 8% and 0.1% risk levels). The variance of frailty terms remains highly significant both in the partial information model and in the observed information model.

The largest bias due to nonresponse is caused to the effect of getting earnings-related unemployment benefits. This is remarkable as the effect of this variable, or variants of it, is often the main focus of an unemployment duration analysis. A higher benefit level is usually thought to encourage the unemployed to search longer or less intensively for new employment, leading to longer unemployment spells (Atkinson and Micklewright 1991). Estimation of separate models for initial nonrespondents, attriters and total respondents reveals that the effect of getting earnings-related unemployment benefit is totally different in the three groups (results not shown here). In the group of total respondents, the effect is negative whereas in the group of initial nonrespondents, the effect is positive. Both effects are statistically significant at 5% risk level. In the group of attriters, the coefficient is not statistically significant from zero.

We also estimated the shared frailty models with Gaussian random terms in order to see whether the results are sensitive with respect to the choice of the frailty distribution. This was not the case, however, as the results were very similar to those from a model with gamma random terms. As discussed by van den Berg (2001), in the case of multiple spell data, the estimates of frailty models are robust with respect to the functional form specification of the frailty terms.

[7] As noted by Therneau and Grambsch (2000), the LR test involves the boundary of the parameter space, but it has been shown (Nielsen et al. 1992) that the one degree of freedom chi-square approximation is valid.

7 Conclusions

We demonstrated a novel way to conduct a nonresponse analysis of longitudinal survey data. The linking of register data at person-level to survey data enabled us to (1) analyse and compare the processes leading to initial nonresponse and attrition, (2) test the type of the missingness mechanisms and (3) estimate the size of bias due to initial nonresponse and attrition. We used the FI ECHP data combined at person-level by longitudinal register data. Spells of unemployment were used as study variables of interest. By taking data on unemployment spells and covariates from the register we got directly comparable data both for survey respondents and nonrespondents and were able to detect a pure nonresponse effect free from measurement errors.

Our nonresponse analysis shows that initial nonresponse and attrition are different processes driven by different background variables. Having a low education level, being approximately 40 years of age, living in an urban municipality in the capital region, being not married, being unemployed or outside the labour force, having high household disposable income, being a pensioner and having a small family are all associated with a high probability of initial nonresponse. There are far fewer strong predictors of attrition, which suggests that attrition is less selective than initial nonresponse. Being young, having a low education level, living in northern Finland and having low household disposable income are associated with high attrition probability. The difficulties in the fieldwork in year 2000 are evident both in the shape of the attrition hazard and in some covariate effects.

The normally unobserved values of study variables have a statistically significant effect on the probability of initial nonresponse and attrition which indicates that both missing data mechanisms are nonignorable with respect to analysis of unemployment duration. The size of bias due to initial nonresponse and attrition was estimated by comparing data sets restricted by nonresponse to a benchmark data set without any restrictions by nonresponse. Initial nonresponse causes downward bias in the estimated survival function whereas attrition does not have a biasing effect. Both initial nonresponse and attrition cause bias in the coefficient estimates of a Cox shared frailty model. It is remarkable that the largest bias is caused to the effect of getting earnings-related unemployment benefit, as this is often the main focus of an econometric unemployment duration analysis.

Our results suggest that initial nonresponse may be at least as important a source of bias as attrition in panel surveys. Other recent studies have reached similar conclusions. The studies by Fitzgerald et al. (1998) and by Sisto (2003) even suggest that a bias in estimates caused by initial nonresponse may fade away over the life of the panel. These results challenge the common view of attrition being the main threat to the value of panel data. A practical recommendation for the survey organisation running a panel is to draw attention to nonresponse at the initial wave of the survey. For the survey data analyst, we do not have a recipe to heal the nonresponse bias. The use of weights aimed at correcting for nonresponse bias may not always be helpful (unpublished paper by Pyy-Martikainen 2006). The bias-reducing power of joint models for the nonresponse mechanism and unemployment duration is a subject for future research.

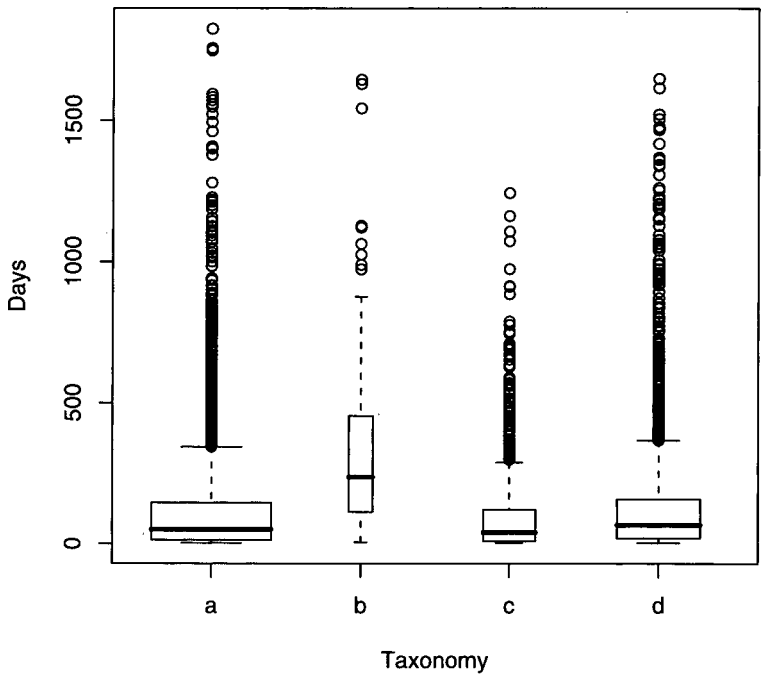
Appendix A: Means of explanatory variables

Variable	Total respondents	Attriters	Initial nonrespondents	F test *p*-value
No. spells before	1.102	0.979	0.971	0.489
No. spells after	3.027	3.317	2.852	0.287
UE[a] time before	82.56	86.58	90.79	0.285
UE[a] time after	342.5	376.4	381.1	0.024
Woman	0.533	0.536	0.497	0.180
Age	33.71	32.85	32.94	0.231
Basic education	0.356	0.413	0.403	0.017
Upper secondary education	0.534	0.502	0.528	0.304
Higher education	0.109	0.086	0.069	0.006
Urban municipality	0.493	0.497	0.602	0.000
Semi-urban municipality	0.178	0.172	0.158	0.498
Rural municipality	0.329	0.330	0.240	0.000
Uusimaa	0.196	0.157	0.239	0.000
Southern Finland	0.351	0.354	0.367	0.736
Eastern Finland	0.183	0.187	0.126	0.001
Central Finland	0.176	0.142	0.144	0.055
Northern Finland	0.094	0.159	0.124	0.000
Married	0.454	0.453	0.378	0.001
Employed	0.366	0.362	0.319	0.065
Unemployed	0.342	0.347	0.366	0.495
Out of labour force	0.293	0.291	0.315	0.476
HH disposable income	143,400	153,600	199,700	0.000
Wage earner	0.518	0.497	0.475	0.161
Entrepreneur	0.070	0.086	0.101	0.041
Farmer	0.080	0.094	0.071	0.196
Pensioner	0.029	0.034	0.053	0.016
Other	0.303	0.289	0.300	0.786
Family size	3.209	3.395	3.206	0.006
Number of persons	1162	923	871	

Basis of analysis: sample persons having 1 + UE spells during 1.1.1995-31.12.1999
For the spell variables, the time of reference is 1996 interview/contact time
Other variables are measured at the end of 1995

[a]UE Unemployment

Appendix B: Boxplots of spell length by type of spell



Appendix C: Kaplan–Meier estimates of survival function

t	Full information		Partial information		Observed information	
	$\hat{S}(t)$	*s.e.*	$\hat{S}(t)$	*s.e.*	$\hat{S}(t)$	*s.e.*
3	0.966	(0.002)	0.964	(0.002)	0.965	(0.002)
100	0.575	(0.005)	0.550	(0.006)	0.555	(0.007)
200	0.450	(0.006)	0.430	(0.007)	0.435	(0.008)
300	0.387	(0.006)	0.365	(0.007)	0.368	(0.008)
400	0.351	(0.007)	0.326	(0.008)	0.326	(0.009)
500	0.328	(0.007)	0.301	(0.008)	0.302	(0.009)
600	0.311	(0.008)	0.282	(0.009)	0.280	(0.010)
700	0.300	(0.008)	0.274	(0.009)	0.273	(0.011)
800	0.287	(0.009)	0.256	(0.011)	0.256	(0.012)
900	0.283	(0.009)	0.250	(0.011)	0.248	(0.013)
1000	0.280	(0.010)	0.250	(0.011)	0.248	(0.013)
1100	0.276	(0.010)	0.250	(0.011)	0.248	(0.013)
1200	0.276	(0.010)	0.250	(0.011)	0.248	(0.013)
1300	0.269	(0.012)	0.238	(0.016)	0.234	(0.019)
1400	0.254	(0.016)				

References

Abbring, J., van den Berg, G., van Ours, J.: The effect of unemployment insurance sanctions on the transition rate from unemployment to employment. Econ. J. **115**, 602–630 (2005)

Allison, P.: Survival Analysis Using the SAS System: A Practical Guide. SAS Institute, Cary (1995)

Atkinson, A., Micklewright, J.: Unemployment compensation and labor market transitions: a critical review. J. Econ. Lit. **XXIX**, 1679–1727 (1991)

Bring , J., Carling, K.: Attrition and misclassification of drop-outs in the analysis of unemployment duration. J. Off. Stat. **16**, 321–330 (2000)

Brown, J., Duncan, G., Rodgers, W.: Measurement error in cross-sectional and longitudinal labour market surveys: validation study evidence. In: Hartog, J., Ridder, G., Theeuwes, J. (eds.) Panel Data and Labour Market Studies. North-Holland, Amsterdam (1990)

Carling, K., Edin, P., Harkman, A., Holmlund, B.: Unemployment duration, unemployment benefits and labor market programs in Sweden. J. Pub. Econ. **59**, 313–334 (1996)

Cox, D.: Regression models and life tables. J. R. Stat. Soc. B **34**, 187–220 (1972)

Duncan, G.: Using panel studies to understand household behaviour and well-being. In: Rose, D. (ed.) Researching Social and Econ. Change. The Uses of Household Panel Surveys. Routledge, London (2000)

Efron, B.: The efficiency of Cox's likelihood function for censored data. J. Am. Stat. Assoc. **72**, 557–565 (1977)

Fitzgerald, J., Gottschalk, P., Moffitt, R.: An analysis of sample attrition in panel data. The Michigan panel study of income dynamics. J. Hum. Resour. **2**, 251–299 (1998)

Hausman, J.: Specification tests in econometrics. Econometrica **46**, 1251–1271 (1978)

Hovi, M., Nordberg, L., Penttilä, I.: Interview and register data in income distribution analysis. Experiences from the finnish European community panel survey in 1996. Stat. Finl. Rev. **2000**, 9 (2000)

Lepkowski, J., Couper, M.: Non-response in the second wave of longitudinal household surveys. In: Groves, R., Dillman, D., Eltinge, J., Little, R. (eds.) Survey Non-response. Wiley, New York (2002)

Little, R., Rubin, D.: Statistical Analysis with Missing Data. Wiley, New Jersey (2002)

Meyer, B.: Unemployment insurance and unemployment spells. Econometrica **58**, 757–782 (1990)

Nielsen, G., Gill, R., Andersen, P., Sørensen, T.: A counting process approach to maximum likelihood estimation of frailty models. Scand. J. Stat. **19**, 25–43 (1992)

Peracchi, F.: The European community household panel: a review. Empir. Econ. **27**, 63–90 (2002)

Pyy-Martikainen, M.: Survey nonresponse, attrition and unemployment duration. Unpublished Manuscript (2006)

Pyy-Martikainen, M., Rendtel, U.: The effects of panel attrition on the analysis of unemployment spells. CHINTEX Working Paper No. 10 (2003)

Pyy-Martikainen, M., Sisto, J., Reijo, M.: The ECHP study in Finland. Quality report. Stat. Finl. Living Cond. **2004**, 1 (2004)

Rubin, D.: Inference and missing data. Biometrika **63**, 581–592 (1976)

Sisto, J.: Attrition effects on the design based estimates of disposable household income. CHINTEX Working Paper No. 9 (2003)

Steiner, V., Kraus, F.: Modelling the heaping effect in unemployment duration models. With an application to retrospective event data in the German socio-economic panel. ZEW Discussion Paper No. 95-09, Mannheim (1995)

Therneau, T., Grambsch, P.: Modeling Survival Data. Extending the Cox Model. Springer, New York (2000)

Therneau, T., Grambsch, P., Pankratz, V.: Penalised survival models and frailty. J. Comput. Graph. Stat. **12**, 156–175 (2003)

Trivellato, U.: Issues in the design and analysis of panel studies: a cursory review. Qual. Quant. **33**, 339–352 (1999)

van den Berg, G.: Duration models: specification, identification and multiple durations. In: Heckman, J., Leamer, E. (eds.) Handbook of Econometrics, vol. 5. Elsevier, Amsterdam (2001)

van den Berg, G., Lindeboom, M., Dolton, P.: Survey non-response and the duration of unemployment. J. R. Stat. Soc. A **169**, 585–604 (2006)

Article II

***Pyy-Martikainen, M. & Rendtel, U.**, Measurement errors in retrospective reports of event histories. A validation study with Finnish register data. Survey Research Methods 3, 3, 139–155, 2009.*

Measurement Errors in Retrospective Reports of Event Histories A Validation Study with Finnish Register Data

Marjo Pyy-Martikainen
Åbo Akademi University and Statistics Finland

Ulrich Rendtel
Freie Universitat Berlin

It is well known that retrospective survey reports of event histories are affected by measurement errors. Yet little is known about the determinants of measurement errors in event history data or their effects on event history analysis. Making use of longitudinal register data linked at person-level with longitudinal survey data, we provide novel evidence about 1) type and magnitude of measurement errors in survey reports of event histories, 2) validity of classical assumptions about measurement errors, 3) measurement error bias and 4) effect of measurement accuracy in event history analysis. The classical assumptions about measurement errors are not supported by our measurement error models. Measurement error in both spell durations and spell outcomes are shown to be important causes of bias in an event history analysis. The effects of education and earnings-related unemployment benefit are estimated with sizeable bias. The magnitude of bias in estimated covariate effects does not depend on model type whereas the Cox model produces clearly less biased estimates of baseline hazard compared to the Weibull model. The large bias in the Weibull baseline hazard is shown to be almost entirely due to low measurement accuracy in survey data.
Keywords: measurement error bias, validation study, event history data, unemployment spells

1 Introduction

Event history data are frequently used to analyze person-specific processes such as fertility, poverty and labour market transitions. Event history data typically consist of information about durations of spells in a state of interest (such as poverty, unemployment, having no children), the outcome or terminal event of the spell (transition to non-poverty, to employment or out of labour force, birth of first child), as well as a set of covariates explaining the durations and outcomes.

Event history data can be collected retrospectively by using either a multi-state or an event occurrence framework (see Lawless 2003). In the multi-state framework the reference period of interest is split into shorter time intervals and for each interval, the state occupied by the person is determined. The event occurrence framework asks for dates of specific events such as transitions between the states of interest. The Survey of Labour and Income Dynamics (SLID) uses the event occurrence framework for information on job and jobless spells during the year preceding the interview. The Survey of Income and Program Participation (SIPP) collects information about spells on food stamps program and spells without health insurance by using a multi-state framework where the 4-month reference period is split into time intervals of one month. The European Community Statistics on Income and Living Conditions (EU-SILC) uses a multi-state framework very similar to that used in the European Community Household Panel (ECHP) to collect month-level labour market state information for the year preceding the interview.

It is well-known that retrospective survey reports of event histories are affected by measurement errors (Eisenhower, Mathiowetz and Morganstein 1991; Bound, Brown and Mathiowetz 2001). A measurement error is the discrepancy between the observed value of a variable provided by the survey respondent and its underlying true value. Measurement errors in event histories are manifested as failure to report a spell (omission), reporting a spell that did not occur (overreporting) and misreporting the duration of a spell (misdating) (Mathiowetz 1986; Holt, McDonald and Skinner 1991).[1] In longitudinal surveys, misdating is typically manifested as the heaping of spell starts and ends at the seam between two reference periods, a phenomenon called the seam effect.[2] Even though spell outcomes may also be misreported (e.g. misclassification of a transition out of labour force as a transition to employment), this topic has received little attention in the literature.

Bound, Brown and Mathiowetz (2001) discuss the causes of measurement errors in survey reports. The respondents' ability to report accurately is believed to depend on the cognitive processes related to answering a survey ques-

Contact information: Marjo Pyy-Martikainen, Department of Economics and Statistics, Åbo Akademi University and Statistics Finland, FIN-00022 Statistics Finland, e-mail: marjo.pyy-martikainen@stat.fi

[1] These definitions of measurement error types are somewhat different from those used in a recent study by Jäckle (2008a). She uses definitions that are based on single events and not, as in our case, on spells which consist of two events (initial and terminal) and the time in between.

[2] As pointed out by Jäckle (2008a), seam effects can also arise as a consequence of chopping of long spells spanning three or more reference periods. Chopping may occur e.g. due to misclassification of the state at the middle waves.

tion, the social desirability of the event being reported and on various features of the survey design. The longer the recall period, the more difficult the reporting task and the less salient the event, the more difficult it is to retrieve the information requested. Socially undesirable events tend to go unreported while the opposite is true for socially desirable events. Survey design features, such as mode and method of data collection, interviewer characteristics, frequency and time interval between interviews of a longitudinal survey are likely to affect survey data quality (Groves 1989). However, as noted by Bound, Brown and Mathiowetz (2001), there are no decisive results with respect to the direction and magnitude of measurement errors attributable to these survey design features.

Despite the recognition of the existence of measurement errors in survey-based data on event histories, little is known about their effects on an event history analysis. Skinner and Humphreys (1999) studied spells generated from a Weibull distribution under the assumption of no censoring. They showed both analytically and by a simulation study that the standard estimators of regression coefficients of a Weibull model are approximately unbiased when measurement errors in spells are independent of each other, spell durations and covariates. The estimator of the shape parameter that determines the duration dependence of the hazard is, however, biased. Empirical evidence of measurement error bias in event history analysis concerns residence histories (Courgeau 1992), occupational spells (Hill 1994), time to benefit receipt or to nonemployment (Pierret 2001) and spells of benefit receipt (Jäckle 2008b). The findings from these studies are mixed: both attenuation and strengthening of covariate effects as well as both weakening and strengthening of duration dependence of baseline hazard were detected. Moreover, the studies by Hill (1994) and Pierret (2001) are not able to provide precise information about measurement error bias as they are based on the comparison of two survey data sets having different data collection methods or recall periods. Both data sets are thus subject to measurement errors as well as possibly different non-response patterns.

The studies by Skinner and Humphreys (1999) and Augustin (1999) are the only studies we are aware of that propose methods to adjust for measurement errors in spells. A common feature of the methods proposed is that they rely on rather restrictive assumptions: that spells are generated from certain parametric duration models, there is no censoring and measurement errors are independent of each other, spell durations and covariates.

Our study provides novel evidence of measurement errors in event history data by using longitudinal register data linked at person-level with longitudinal survey data. The combined survey-register longitudinal data enables us to 1) provide information on the type and magnitude of measurement errors in survey reports of event histories, 2) test the plausibility of common assumptions about measurement errors and 3) study measurement error bias in event history analysis. The survey data used in our study is collected by a multi-state framework with a reference period of one year split into one-month intervals. Comparisons of the survey data with register data measured at day level are affected by differences in the measurement accuracy. A fourth aim of our study is to evaluate the separate biasing effects of measurement accuracy and measurement error. This is done by discretizing the day-level register data into month-level data and by comparing results from the three data sets.

The next section discusses the details of the data and the research design. Section 3 studies the magnitude and type of measurement errors in survey reports of event histories. Section 4 specifies models for the process of reporting event histories in order to assess the validity of common assumptions about measurement errors. Section 5 shows how measurement errors affect standard event history analyses. Section 6 evaluates the separate biasing effects of measurement accuracy and measurement error. The findings and implications of our study are discussed in Section 7.

2 The data

Unemployment spells were used as the study variables of interest. We conducted a complete record-check validation study of reports of unemployment spells in the Finnish subset of European Community Household Panel (FI ECHP) data by making use of longitudinal register data linked at person-level with FI ECHP survey data. The register data were assumed to contain true, error-free information about unemployment spells. This is, of course, a simplifying assumption. However, as unemployed persons need to register into the records of an employment office in order to receive unemployment benefits, the register data can be claimed to be more accurate than the survey data.

The ECHP is an input-harmonised sample survey conducted in 15 EU member states between 1994 and 2001 and co-ordinated by Eurostat. The ECHP covers a wide range of topics concerning living conditions, the core topics being income and employment, see Peracchi (2002) for a review of the ECHP. The Finnish ECHP started in 1996. The FI ECHP is documented in Pyy-Martikainen et al. (2004). We used the first five waves of FI ECHP covering the years 1996-2000.

In the ECHP, retrospective labour market state data were collected by a multi-state framework in the form of a month-by-month main activity state calendar obtained for the year preceding the interview. The respondent was first asked whether there were changes in his/her main activity state during the preceding year. If not, the respondent was asked to choose a main activity state from a showcard with 10 options. If there were changes, the respondent was asked to choose a main activity state from the showcard for each month of the year beginning from January:

> "Were there any changes in your main activity in <year>?" [yes/no]
>
> if no:
>
> "What was your main activity state in <year> according to this list?"
>
> if yes:
>
> "What was your main activity state in <month>?"

Interviewers were given the following instructions: if a person's weekly working hours are 15 or more, an option related to employment should be chosen. If a person has had various activity states during a month, employment should be preferred over other states. Thus, in principle, having worked for 15 hours during one week in a specific month is enough to be defined as having been employed in that month. In FI ECHP, a person is defined as unemployed if he/she is without a job, available for work and looking for work through the employment office or newspaper advertisements or some other way. Persons dismissed temporarily are also regarded as unemployed.[3]

Our analysis was based on the FI ECHP sample persons aged 16 or over and thus eligible for a personal interview at the beginning of 1996 (11,641 persons altogether). The sample persons were defined as all members of the initial sample of households. Initial non-respondents (3,146 persons, 27.0%) were excluded because no survey information was available for them (for missingness patterns in the FI ECHP, see Pyy-Martikainen and Rendtel 2008). Temporary dropouts (921 persons, 7.9%) were also excluded because their inclusion would have posed the problem of left-censored spells. Left-censored spells are not only a source of bias in an event history analysis but they would have also artificially increased the heaping of spell starts in January. These restrictions left us with 7,574 (65.1%) sample persons, of whom 4,364 responded in each of the five interviews and 3210 attrited during years 1997 to 2000. For the total respondents, information about unemployment spells was obtained for the five-year period covering the years 1995-1999. For the attriters, information was obtained up to the end of the year that precedes the last interview. Unemployment spells ongoing at the end of the relevant reference period were right-censored. Spells ongoing in January 1995 were dropped because their starting date was unknown. The resulting survey data contain 2719 unemployment spells of 1,482 persons.

Validation data were obtained from the Ministry of Labour's Job-seekers Register. The register contains day-level information about unemployment spell starts and ends. For each spell, the outcome is also registered. In the register, an unemployed job seeker is defined as being without a job and seeking a new job. Registering with the employment office is considered as evidence of seeking a job. Persons dismissed temporarily are regarded as unemployed. Register spells ongoing between 1 January 1995 and 31 December 1999 were linked at person-level to the survey data by personal identification codes.[4] This time period corresponds to the main activity state reference periods of the first five years of the FI ECHP. We constructed register spell data covering, for each person, the same time span as his/her follow-up time in the survey data. For the total respondents, this means using register spells ongoing between 1 January 1995 and 31 December 1999. For the attriters, register spells ongoing between 1 January 1995 and the end of the year preceding the last interview were used. Spells ongoing at the end of the relevant reference period were right-censored. Left-censored spells (ongoing at 1 January 1995) were dropped. Spells lasting at most two days were also dropped as they were not regarded as true unemployment spells but registrations into the records of the employment office for some legislative reason. The register data contain 6,050 spells of 1,854 persons. Apart from covariates related to the fieldwork, covariates used in subsequent analyses were also taken from various administrative registers.

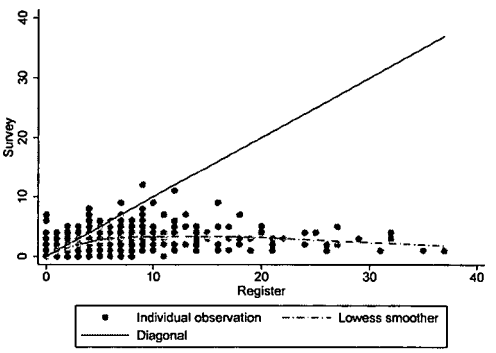


Figure 1. Number of unemployment spells in register and survey over the 5-year follow-up period

The magnitude and type of measurement errors were evaluated by person-level comparisons of survey reports and register data. The effects of measurement errors on event history analysis were assessed by comparing estimates based on the two data sources. No survey weights were used in the analysis. Likewise, no attempts were made to correct for the non-response bias. Although estimates based on both survey and register data are affected by non-response, the differences in the estimates cannot be attributed to non-response bias as both the survey and the register data contain the same persons. This was also the main reason why we neglected the use of survey weights in this study.

3 Magnitude and type of measurement errors

Figure 1 shows for each person the number of unemployment spells calculated both from the register and survey data over the 5-year follow-up period. For clarity, the x-axis

[3] The implementation of FI ECHP differs here from Eurostat recommendations, according to which main activity states apart from those related to employment be determined according to self-declaration on the basis of most time spent.

[4] All Finnish citizens are registered in the Finnish Population Information System (FPIS), which is a national register that contains basic information such as name, date of birth and address. As part of the registration process, citizens are issued with a personal identity code (PIC) that is used as a means of identification of persons. The FPIS is used throughout Finnish society's information services and management, including the production of statistics and research.

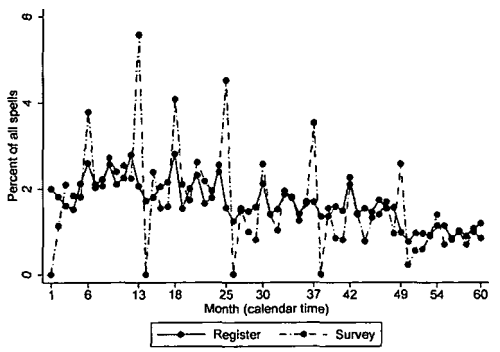


Figure 2. Spell starts in register and survey data

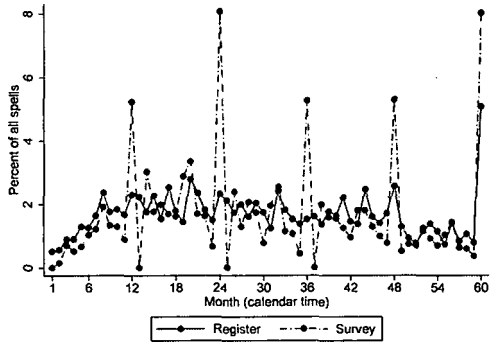


Figure 3. Spell ends in register and survey data

is truncated at 40.[5] A Lowess scatterplot smoother and a diagonal line are also shown.[6] If the number of survey spells and register spells were approximately equal, the points in Figure 1 would lie in the vicinity of the diagonal line. This is not the case, instead, the Lowess line is almost flat implying there is no association between the number of survey and register spells. There is both omitting and overreporting of unemployment spells, omitting being much more important. The omitting of unemployment spells is largely due to the differences in measurement accuracy in survey and register data.

There is a strong heaping effect of unemployment spell starts and ends at the seams between the reference periods of consecutive panel waves (Figures 2 and 3). Unemployment spells tend to start in January and end in December. There is also heaping of spell starts in June. Moreover, there is evidence of backward telescoping of spell starts: following the peaking of spell starts in January there is a lack of spells starting in February. This is likely a consequence of memory decay: events occurring early in the reference period are more difficult to recall.

Table 1: Spell outcomes in register and survey data

	Register		Survey	
Outcome	spells	%	spells	%
Employment	3,238	53.5	1,638	60.2
Subsidised work	720	11.9	58	2.1
OLF[a], Other	1,544	25.5	592	21.8
Attrition	274	4.5	213	7.8
End of follow-up	274	4.5	218	8.0
All	6,050	100.0	2,719	100.0

[a] OLF Out of Labour Force

An often ignored issue is that there may be measurement error in reported spell outcomes as well. In the analysis of unemployment duration, the outcome of interest is often becoming employed. In the survey data, 60.2% of spells ended in becoming employed, whereas in the register data only 53.5% of spells ended for this reason (Table 1). A person-level comparison of register and survey data shows that getting subsidised work is often misclassified by survey respondents as normal employment (Table 2). The higher percentage of survey spells that end because of attrition or end of follow-up reflects the fact that the survey spells are, on average, longer than register spells. The comparison in Table 2 was restricted to persons having one unemployment spell according to both survey and register data during the entire follow-up period. This restriction was done in order to make sure that the spells being compared are the same. The linking of multiple spells per person would have been too unreliable for measurement accuracy and measurement error reasons.

4 Determinants of measurement errors

Because of measurement errors, the true durations T^* are not observed in the survey. The reported durations T can be thought of as consisting of the true duration and a measurement error: $T = T^* + \epsilon$.[7] According to the classical assumptions (see e.g. Bound, Brown and Mathiowetz 2001, Skrondal and Rabe-Hesketh 2004) the measurement errors ϵ have zero mean and are independent of each other, true durations T^* and any covariates explaining T^*. We aimed at testing the validity of these assumptions by modelling $\epsilon = T - T^*$ as a

[5] Only three persons had more than 40 register spells during the follow-up period.

[6] For an introduction to the Lowess procedure see, for example, Fan and Gijbels (1996).

[7] An alternative for the additive measurement error model is the multiplicative model $T = T^* \times \epsilon$, see e.g. Skinner and Humphreys (1999) and Augustin (1999). According to the multiplicative model, the longer the spell lasts the larger the measurement error tends to be. Because of the way unemployment data was collected in the ECHP, there is substantial error in the measurement of short spells also -the reason why we chose to work with the additive model.

Table 2: Misclassification of spell outcomes (sample n: 351)

	Outcome in survey						
Outcome in register	(a)	(b)	(c)	(d)	(e)	(f)	All
(a) Employed	93.2	0.0	2.9	1.0	0.0	2.9	100.0
(b) Subsidised work	85.0	2.5	10.0	0.0	2.5	0.0	100.0
(c) OLF[a]	13.5	1.1	80.9	0.0	2.3	2.3	100.0
(d) Other	50.0	0.0	36.4	4.6	0.0	9.1	100.0
(e) End of follow-up	0.0	0.0	7.9	0.0	92.1	0.0	100.0
(f) Attrition	1.7	0.0	6.8	0.0	0.0	91.5	100.0

[a]OLF Out of Labour Force

function of the true duration and covariates x. We included in our models also some fieldwork-related covariates that are believed to affect measurement errors. Because the survey and register data can be reliably linked only at person-level (and not at spell-level), we defined our measurement error variable as the difference between the sum of unemployment durations from the survey and the sum of unemployment durations from the register, calculated separately for each person $i = 1, \ldots, n$ and for each panel wave $j = 1, \ldots, K_i$ in which the person was unemployed according to both survey and register:

$$\epsilon_{ij} = \sum_{s=1}^{S_{ij}} T_{sij} - \sum_{r=1}^{R_{ij}} T^*_{rij}.$$

S_{ij} and R_{ij} are the numbers of survey and register spells for person i and wave j. ϵ_{ij}'s can be thought of as estimates of cumulated measurement errors in the unemployment spells reported by person i in the wave j interview. To calculate ϵ_{ij}'s, unemployment spells extending over two or more waves were cut at the seams between the waves. We modelled measurement errors in two phases: in the first phase, we modelled the probability of reporting no unemployment spells in a specific wave, given that at least one unemployment spell was found in the register.[8] In the second phase, we modelled the magnitude of cumulated measurement error in the reported unemployment spells, given that at least one unemployment spell was both reported and found in the register.

For the first phase model, assume there are latent variables y^*_{ij} describing the propensity of person i to omit reporting unemployment spells occurring in wave j. The latent variables are assumed to follow the model

$$y^*_{ij} = x_{ij}\beta + \zeta_i + \varepsilon_{ij},$$

where x_{ij} is a $(1 \times p)$ vector of covariates (including a constant) possibly varying with time and person, β is a $(p \times 1)$ vector of the parameters to be estimated and $\zeta_i \sim N(0, \sigma^2_\zeta)$ are person-specific random effects. The random effects ζ_i were incorporated in the model in order to allow for the possibility of correlation of responses by the same person. Error terms ε_{ij} are assumed to be independent and to follow a logistic distribution with mean zero and variance $\sigma^2_\varepsilon = \pi^2/3$. [9] It is assumed that ε_{ij} and ζ_i are uncorrelated. The model can be alternatively expressed as

$$\text{logit}[P(y_{ij} = 1 \mid x_{ij}, \zeta_i)] = x_{ij}\beta + \zeta_i,$$

where

$$y_{ij} = \begin{cases} 1 & \text{if } y^*_{ij} > 0 \\ 0 & \text{if } y^*_{ij} \leq 0. \end{cases}$$

Variables y_{ij} are thus binary variables telling whether person i omits reporting unemployment spells occurring in wave j or not. The intracluster correlation i.e. the correlation among the latent responses by the same person is $\rho = \sigma^2_\zeta/(\sigma^2_\zeta + \frac{\pi^2}{3})$. The model is estimated by maximum likelihood, using a Gauss-Hermite quadrature to approximate the integral over the random terms ζ_i in the log-likelihood function (see e.g. Skrondal and Rabe-Hesketh 2004). In the empirical application, a 12-point quadrature was used.

The second phase model was specified as a random effects linear model:

$$\epsilon_{ij} = x_{ij}\gamma + \nu_i + \delta_{ij},$$

where ϵ_{ij} are the estimates of cumulated measurement errors defined earlier, x_{ij} is a $(1 \times p)$ vector of covariates (including a constant) possibly varying with time and person and γ is a $(p \times 1)$ vector of parameters to be estimated. The assumptions about the random terms ν_i and δ_{ij} are: $\nu_i \sim N(0, \sigma^2_\nu), \delta_{ij} \sim N(0, \sigma^2_\delta)$ and $\text{cov}(\nu_i, \delta_{ij}) = 0$. The intracluster correlation is $\rho = \sigma^2_\nu/(\sigma^2_\nu + \sigma^2_\delta)$. The model was estimated by maximum likelihood.

The distribution of the ϵ_{ij}'s is shown in Figure 4. Compared to a normal distribution (solid line), the empirical distribution (kernel density estimate shown by dashed line) has more mass in the vicinity of zero.

The model estimates are reported in Table 3. The covariates were arranged into three groups: 1) covariates related to

[8] We did not model the probability of overreporting spells given that the register data show none since such a reporting error was found in less than 1 % of person-years.

[9] Variance $\sigma^2_\varepsilon = \pi^2/3$ results from setting the scale parameter of logistic distribution equal to one.

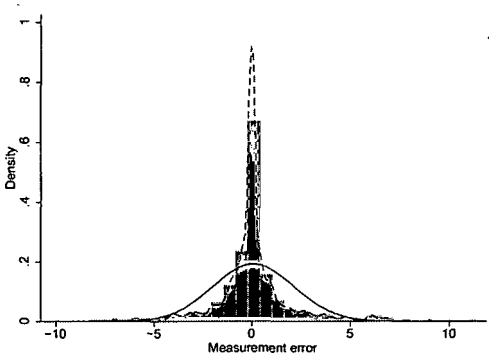


Figure 4. Distribution of cumulated measurement errors

the study variable of interest; 2) covariates used in the event history model (whose estimation is assumed to be the main target of analysis) and 3) covariates related to fieldwork. Covariates in groups 1) and 2) were used to test the classical assumptions about measurement errors. All the covariates are measured at the same year as the dependent variables. The covariates of the event history model are described in section 5. The covariates related to fieldwork include mode of interview (face-to-face vs. telephone), nature of the respondent (self vs. proxy) and the year of interview. Even though the mode of interview and the nature of the respondent are likely to influence the quality of survey reports, it is not clear from theory or empirical evidence how these survey design features affect the direction and magnitude of reporting errors (Bound, Brown and Mathiowetz 2001). Studies which do not control the assignment of respondents to self/proxy or face-to-face/telephone groups are subject to potential self-selection bias (Moore 1988). For example, it may well be that persons with more complex unemployment histories (and, therefore, more prone to reporting errors) are more difficult to reach and, therefore, less likely to give a personal face-to-face interview. However, this problem should be alleviated by the use of covariates related to unemployment history in the measurement error model. During 1996-1997, the fieldwork of the FI ECHP was conducted during February-May, whereas from 1998 onwards the fieldwork period was shifted to autumn. This caused a lengthening of the recall period by several months. Because of memory decay, this was expected to lead to a higher probability of omission and increased magnitude of measurement errors.

Having less than one month of cumulated unemployment time increases the odds of omission by a factor of almost five[10], a consequence of the lower accuracy of measurement and the preference given to activities related to employment on the survey questionnaire (Table 3, Model 1). Each additional month of unemployment decreases the odds of omission by 23.7%. [11] Being a female increases the odds of omission by 27.6%. Age has a u-shaped effect on the probability of omission. The probability decreases until the age of 37 and starts to increase thereafter. A higher probability of omission among the older is likely a consequence of decreasing cognitive ability along with age whereas the young tend to have shorter spells which are both more difficult to recall and more likely too short to be reported in the monthly main activity state scheme. Persons living in Eastern Finland and receiving earnings-related unemployment benefit are more likely than other persons to report unemployment spells. Conducting a proxy interview instead of an interview with the person of interest increases the odds of omission by 72.8%. During the years 1998-2000, the odds of omission are more than double compared to the year 1995. The estimated correlation between the latent responses by the same person is 0.281 and highly significant according to likelihood ratio test.

Both the amount of cumulated unemployment time and the number of unemployment spells affect the magnitude of cumulated measurement errors (Table 3, Model 2). Respondents with cumulative unemployment time less than one month are more likely to overreport which is expected since the reported unemployment time cannot be less than one month. Respondents with longer cumulative unemployment time and more unemployment spells are more likely to underreport. Females are more likely to overreport while persons with an upper secondary or higher education, living outside the capital region and receiving earnings-related unemployment benefit tend to underreport. The estimated correlation between the cumulated measurement errors by the same person is 0.123, again highly significant according to the likelihood ratio test.

5 Effects of measurement errors in event history analysis

Previous sections showed that measurement errors in spell durations are not only of nonnegligible magnitude but also do not conform to the classical independence assumptions. The spell outcomes were also shown to be misclassified. What is the impact of measurement errors in event history analysis based on survey data? This was evaluated by comparing Kaplan-Meier estimates of survival function and estimates from Cox and Weibull proportional hazards models based on register and survey data. The estimates based on register data were used as benchmarks against which the bias due to measurement errors in the survey-based estimates was evaluated.

The study design is described in Table 4. In the first phase, we assessed the impact of measurement errors in spell durations only. Measurement errors in spell durations include not only the effect of misdating of spells but also the effect of omissions and overreporting. Phase 1 analyses ignore spell outcome i.e. study the rate of exit from unemployment regardless of the reason for the exit. The Phase 1 survey data consist of survey spell durations and register

[10] Calculated as exp(1.604)

[11] Calculated as 1 – exp(–0.270)

Table 3: Determinants of measurement errors. Model 1: model for the probability of omission. Model 2: model for the magnitude of measurement error

	Model 1		Model 2	
	coef.	se	coef.	se
Constant	**1.551**	(0.528)	*0.724*	(0.420)
Covariates related to the study variable				
Sum of reg UE[a] months	**-0.270**	(0.016)	**-0.125**	(0.011)
Sum of reg UE[a] months lt 1	**1.604**	(0.154)	**1.138**	(0.193)
Number of reg UE[a] spells	0.019	(0.024)	**-0.075**	(0.022)
Covariates of the EH[b] model				
Female	**0.244**	(0.101)	**0.164**	(0.075)
Age	**-0.119**	(0.028)	0.025	(0.022)
Age squared	**0.002**	(0.000)	-0.000	(0.000)
Upper secondary education	0.051	(0.118)	**-0.183**	(0.085)
Higher education	0.227	(0.176)	**-0.524**	(0.135)
Semi urban municipality	0.074	(0.139)	0.000	(0.105)
Rural municipality	-0.050	(0.123)	0.131	(0.091)
Southern Finland	-0.210	(0.142)	**-0.331**	(0.109)
Eastern Finland	**-0.488**	(0.173)	**-0.273**	(0.128)
Central Finland	-0.105	(0.179)	**-0.295**	(0.138)
Northern Finland	-0.109	(0.194)	**-0.374**	(0.146)
Earnings-rel. UE[a] benefit	**-0.381**	(0.105)	**-0.268**	(0.079)
Covariates related to fieldwork				
Telephone interview	*0.199*	(0.115)	0.045	(0.092)
Proxy interview	**0.547**	(0.166)	0.139	(0.136)
Interview in 1997	0.112	(0.123)	-0.115	(0.084)
Interview in 1998	**0.926**	(0.129)	0.096	(0.097)
Interview in 1999	**0.939**	(0.139)	**-0.327**	(0.104)
Interview in 2000	**0.748**	(0.159)	-0.151	(0.121)
Intracluster correlation	**0.281**	(0.034)	**0.123**	(0.020)
-2 log likelihood	4,591		15,456	
number of persons	2,028		1,626	
number of person-years	5,103		3,673	
number of person-years with no reported spells	1,430		-	

Estimates significant at 5% (10%) risk level are displayed in **boldface** (*italics*).

[a] UE unemployment

[b] EH Event History

Table 4: Effects of measurement errors in event history analysis: study design

Phase	Measurement error in	Type of data	Benchmark data from	Survey data from
1	spell duration	spell duration	Register	Survey
		covariates	Register	Register
2	spell duration	spell duration	Register	Survey
	spell outcome	spell outcome	Register	Survey
		covariates	Register	Register

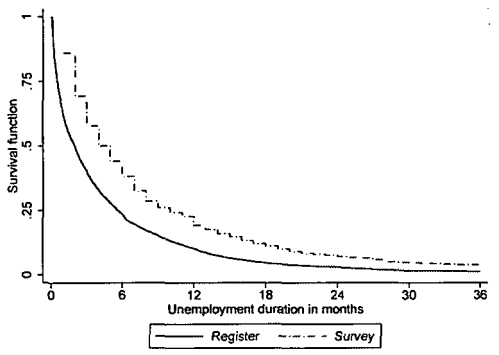


Figure 5. Phase 1. Kaplan-Meier survival function estimates for register and survey data.

covariates. By using the same source of covariates in the two data sets, the differences in estimates could only be attributed to differences in register and survey spells.

In the second phase, measurement errors in survey spell outcomes were taken into account by conducting a cause-specific analysis. In this analysis, the outcome of interest was becoming employed. Phase 2 survey data analyses were conducted using survey spell durations and outcomes, and register covariates.

Results from Phase 1 analyses are shown in the following whereas results from Phase 2 analyses are shown in the Appendix. Figure 5 shows Phase 1 Kaplan-Meier estimates for register and survey data. The Kaplan-Meier estimator is defined as $\hat{S}(t_l) = \prod_{j=1}^{l}(1 - d_j/r_j)$, where t_l is the duration of the lth ordered spell, r_j is the size of the risk set and d_j is the number of spells ending at time t_l. $\hat{S}(t)$ is an estimator of the survival function $S(t) = P(T \geq t)$ that describes the probability of a spell ending later than at time t.[12] In Figure 5 and in all subsequent figures describing the distribution of unemployment spells, the x-axis is truncated at 36 months because very few spells were longer than this. Survey spells end at a lower rate than register spells at all durations. The median duration of a spell is 2 months in the register and 5 months in the survey data. According to the cause-specific Kaplan-Meier estimates (Figure A.6 in Appendix), survey spells end in employment at a lower rate than register spells at durations less than 14 months. Thereafter, the situation is reversed. The crossing of the curves is due to the misclassification of subsidised work as normal employment by survey respondents. If in register data subsidised work is classified as normal employment, the register-based Kaplan-Meier curve lies below the survey-based curve at all durations (results not shown here).

We estimated both Cox and Weibull proportional hazards models in order to assess the measurement error bias in the estimates of the covariate effects and the baseline hazard. A proportional hazards model specifies the hazard function as a product of two terms: $\lambda(t \mid x) = \lambda_0(t)g(x)$. The hazard function $\lambda(t \mid x)$ describes the conditional probability of exit from unemployment, given the covariates and given that the spell has not ended before time t. Function $\lambda_0(t)$ is a baseline hazard specifying the dependency of the hazard function on the duration of interest. The covariates have a multiplicative effect on the hazard function via $g(x)$. Usually $g(x) = \exp(x\beta)$, where x is a $(1 \times p)$ vector of (possibly time-varying) covariates and β is a $(p \times 1)$ vector of parameters.[13] For the Weibull model, the baseline hazard is specified as $\lambda_0(t) = pt^{p-1}$. The shape parameter p determines whether the hazard function is monotonically decreasing ($p < 1$), increasing ($p > 1$) or constant ($p = 1$). The Cox model is estimated by a partial likelihood function that does not involve the $\lambda_0(t)$ terms. The shape of the hazard function is therefore completely unrestricted, which makes the model flexible when compared to fully parameterized models. Both belonging to the class of proportional hazards models, the parameter estimates of Cox and Weibull models are directly comparable. The parameter estimates of proportional hazards models are reported as hazard ratios. The hazard ratio of the i^{th} coefficient is calculated as $\exp(\beta_i)$ and it is interpreted as the ratio of the hazards for a 1-unit increase in the i^{th} covariate.

We hypothesize that estimates of the covariate effects of duration models with a flexible baseline hazard, such as the Cox proportional hazards model, are less biased by measurement errors than estimates from fully parameterized models. For example, it may well be that the effect of heaping of spell starts and ends is absorbed by a flexible baseline hazard. Van den Berg et al. (2004) found that covariate estimates of a Cox proportional hazards model were less biased by non-response than estimates of an exponential or a Weibull model. In order to assess our hypothesis, we compared the size of bias of the survey estimates of the Cox and Weibull proportional hazards models.

Sometimes dummies for heaping months are included as covariates in an attempt to correct for the heaping effect (e.g. Hujer and Schneider 1989, Hunt 1995, Kraus and Steiner 1998). We estimated models both with and without dummies for January and December in order to see whether such heaping dummies protect against measurement error bias in covariate effects or in the baseline hazard.

A set of covariates similar to those used in econometric analyses of unemployment duration was used (see e.g. Meyer 1990, Carling et al. 1996, Abbring et al. 2005). The covariates are spell-specific and they are usually measured at the end of the year preceding the start of the unemployment spell. Age is measured in years. Level of education divides persons into three classes. Basic education corresponds to the completion of comprehensive school. Upper secondary education comprises matriculation examina-

[12] The cause-specific Kaplan-Meier estimator $\hat{S}_c(t)$ describes the probability of an event of type c occurring later than at time t.

[13] A cause-specific proportional hazards model describes the conditional probability of exit due to the event of interest at time t, given that the spell has not ended before t.

tion and upper secondary vocational education. Higher education comprises, for example, tertiary vocational college education and university education. The possible state dependency in unemployment durations is measured by the proportion of time (since 1 January 1995) spent in unemployment before the spell in question. Variation in local labor market conditions is taken into account by information on residential area and statistical grouping of municipalities. The residential area dummies are based on the NUTS2 classification of regions. The statistical grouping of municipalities divides municipalities into urban, semi-urban and rural ones by the proportion of the population living in urban settlements and by the population of the largest urban settlement. Earnings-related unemployment benefit indicates whether a person has received this kind of benefit at the starting year of the unemployment spell. This variable, or variants of it, is often the variable of main interest in an unemployment duration analysis. Other covariates were directly determined by the spell itself and were therefore always taken from the same data source as the spell information. Indicators for the starting year of the unemployment spell aim at capturing the effect of economic fluctuations over time. The January dummy indicates whether the spell started in January (January 1995 excluded). The December dummy is specified as a time-varying indicator variable that gets value 1 in December and zero otherwise.[14]

The estimates from Phase 1 regression analyses are shown in Table 5.[15] The estimates from Phase 2 analyses are shown in Table A.1 in the Appendix. For each model, covariate hazard ratios and their standard errors are reported. Robust estimates of standard errors were calculated in order to take into account the clustering of spells within persons (Lin 1994).

Except for the year dummies, the magnitude and direction of measurement error bias in estimated covariate effects are similar in all estimated models (Table 5). The survey estimates of the year dummies are very much affected by the inclusion of heaping dummies, see footnote 16. The estimated effects of sex, level of education and the dummy for living in Northern Finland have all large biases, the absolute values exceeding 10 percentage points. The effect of education is larger, i.e. further from 1, in the survey-based models, whereas the opposite is true for the effects of sex and living in Northern Finland. Having high education has a markedly stronger effect in the survey-based models: the bias ranging from 18 to 30 percentage points. The shape parameters of the Weibull models are badly biased, which is clearly illustrated in Figure A.2. Both the Cox and the Weibull models show similar effects of January and December dummies. The register spells are less likely to end in December than in other months. This seasonal variation effect in spell ends is masked in the survey estimate by the heaping of spell ends in December. Survey spells beginning in January have a lower hazard of exit, implying longer spell durations while the January dummy has no effect in the register data. This is an indication of backward telescoping of survey spell starts. The effect of January and December dummies in other estimated covariate effects is negligible except for the year dummies of the survey models.[16] The results in Table 5 do not give support to our hypothesis about the Cox model coefficient estimates having smaller bias.

The competing risks analysis with becoming employed as the outcome of interest (Table A.1) shows similar biases in the effects of sex and level of education as before (Table 5). The survey-based models underestimate the effect of receiving earnings-related unemployment benefit by over 28 percentage points. [17] Compared to the analysis that ignores the outcome of interest, the biases in the year dummies and in the shape parameters of the Weibull models have become more pronounced. Moreover, most of the area dummies have now large biases. Introducing an additional source of measurement error, error in spell outcome, has apparently increased the measurement error bias. The effect of the heaping dummies as well as their effect on other estimated covariate effects is similar to before. Again, there is no indication of the Cox model coefficient estimates being more robust with respect to measurement error bias.

Figures A.1 and A.2 show the estimated baseline hazard functions for the Cox model and for the Weibull model without the heaping dummies.[18] For the estimated baseline hazard contributions of the Cox model (see Kalbfleisch and Prentice 2002), a kernel smoother with the Epanechnikov kernel function and a bandwidth of two months was applied (see e.g. Klein and Moeschberger 2003). The hazard func-

[14] Note that the December dummy is defined in a different time scale than the analysis time. The analysis time is specified as time from the beginning of each unemployment spell, whereas the December dummy is specified in calendar time.

[15] For the survey data, we estimated also complementary log-log (cloglog) models corresponding to Cox proportional hazard and Weibull models. The cloglog model is suitable for survival times that are grouped into discrete intervals of time but that are intrinsically continuous. Estimates from the cloglog models were very close to the results from ordinary continuous time Cox and Weibull models.

[16] The effect of the year dummies is weaker in the survey models without the January dummy. This is because the effect of a spell beginning in January is confounded with the effect of the starting year of the spell. Compared to the year 1995, spells beginning during the years 1996-1999 have a higher hazard of exit. The fact that spells beginning in January 1995 are excluded (because they are left-censored) attenuates this effect as spells beginning in January have also a lower hazard of exit.

[17] In the register-based models, the effect of receiving earnings-related unemployment benefit instead of basic unemployment allowance is to increase the exit rate into employment, which is contrary to expectations. A similar effect was found by Hujer and Schneider (1989) and, as noted by Hunt (1995), is likely a result of positive unobserved qualities of receivers of earnings-related unemployment benefit. In a study making use of the same data set, Pyy-Martikainen and Rendtel (2008) estimated a shared frailty Cox hazard model that controls for person-specific unobserved heterogeneity. The effect of receiving earnings-related unemployment benefit was to lower the hazard of exit, which is in accordance with the results from search theory.

[18] The estimated baseline hazards from the models including heaping dummies are almost identical and, therefore, not reported.

Table 5: Phase 1. Proportional hazards models. Measurement error in spell duration only

	1. Cox			2. Cox, heaping dummies			3. Weibull			4. Weibull, heaping dummies		
	Register	Survey		Register	Survey		Register	Survey		Register	Survey	
Variable	hr[a] (se)	hr (se)	bias[b]	hr (se)	hr (se)	bias	hr (se)	hr (se)	bias	hr (se)	hr (se)	bias
Female	**1.128** (0.067)	0.925 (0.045)	-20.3	*1.121* (0.667)	0.944 (0.047)	-17.7	**1.129** (0.069)	0.921 (0.052)	-20.8	*1.124* (0.069)	0.936 (0.053)	-18.8
Age	1.014 (0.017)	*1.024* (0.014)	1.0	1.015 (0.017)	*1.026* (0.015)	1.1	1.014 (0.018)	1.026 (0.016)	1.2	1.015 (0.018)	*1.029* (0.017)	1.4
Age squared	*1.000* (0.000)	**0.999** (0.000)	-0.1	*1.000* (0.000)	**0.999** (0.000)	-0.1	*1.000* (0.000)	**0.999** (0.000)	-0.1	*1.000* (0.000)	**0.999** (0.000)	-0.1
Upper secondary educ.	1.052 (0.074)	**1.128** (0.065)	7.6	1.052 (0.074)	**1.120** (0.065)	6.8	1.051 (0.076)	**1.153** (0.076)	10.2	1.051 (0.076)	**1.145** (0.076)	9.4
Higher education	**1.373** (0.134)	**1.555** (0.124)	18.2	**1.365** (0.134)	**1.584** (0.127)	21.9	**1.390** (0.140)	**1.662** (0.152)	27.2	**1.383** (0.139)	**1.684** (0.155)	30.1
Proportion of UE tabfinm[c] time	1.001 (0.001)	*0.998* (0.001)	-0.3	1.001 (0.001)	**0.997** (0.001)	-0.4	1.001 (0.001)	**0.998** (0.001)	-0.3	1.001 (0.001)	**0.997** (0.001)	-0.4
Semi urban municipality	1.135 (0.106)	1.071 (0.075)	-6.4	1.137 (0.107)	1.066 (0.077)	-7.1	1.140 (0.107)	1.067 (0.088)	-7.3	1.141 (0.108)	1.054 (0.089)	-8.7
Rural municipality	1.029 (0.070)	0.970 (0.058)	-5.9	1.028 (0.070)	0.979 (0.060)	-4.9	1.032 (0.072)	0.961 (0.066)	-7.1	1.031 (0.072)	0.972 (0.067)	-5.9
Southern Finland	1.189 (0.164)	**1.213** (0.085)	2.4	1.194 (0.165)	**1.162** (0.081)	-3.2	1.200 (0.165)	**1.224** (0.097)	2.4	1.204 (0.166)	*1.158* (0.091)	-4.6
Eastern Finland	*1.151* (0.174)	1.134 (0.095)	-1.7	1.157 (0.175)	1.091 (0.092)	-6.6	1.160 (0.175)	1.161 (0.109)	0.1	1.164 (0.176)	1.110 (0.104)	-5.4
Central Finland	1.069 (0.146)	1.145 (0.106)	7.6	1.074 (0.146)	1.105 (0.104)	3.1	1.079 (0.148)	1.127 (0.120)	4.8	1.082 (0.148)	1.078 (0.116)	-0.4
Northern Finland	*1.331* (0.200)	1.156 (0.110)	-17.5	*1.334* (0.201)	1.136 (0.108)	-19.8	**1.356** (0.205)	1.181 (0.125)	-17.5	**1.358** (0.205)	1.159 (0.121)	-19.9
Earnings-rel. UE[d] benefit	1.068 (0.067)	1.010 (0.053)	-5.8	1.063 (0.067)	0.996 (0.053)	-6.7	1.070 (0.069)	1.022 (0.062)	-4.8	1.067 (0.068)	1.003 (0.062)	-6.4
Year 1996	1.073 (0.049)	**1.137** (0.063)	6.4	1.069 (0.050)	**1.228** (0.071)	15.9	1.079 (0.052)	**1.163** (0.079)	8.4	1.076 (0.052)	**1.268** (0.090)	19.2
Year 1997	**1.151** (0.062)	**1.156** (0.072)	0.5	**1.145** (0.063)	**1.275** (0.084)	13	**1.168** (0.065)	*1.207* (0.088)	3.9	**1.164** (0.066)	**1.361** (0.104)	19.7
Year 1998	**1.326** (0.088)	**1.342** (0.090)	1.6	**1.314** (0.089)	**1.421** (0.102)	10.7	**1.365** (0.093)	**1.486** (0.111)	12.1	**1.357** (0.094)	**1.639** (0.129)	28.2
Year 1999	**1.363** (0.118)	1.003 (0.096)	-36.0	**1.297** (0.115)	0.943 (0.111)	-35.4	**1.427** (0.125)	**1.225** (0.122)	-20.2	**1.363** (0.122)	**1.320** (0.148)	-4.3
Begin in January				1.003 (0.057)	**0.716** (0.041)	-28.7				0.994 (0.059)	**0.678** (0.044)	-31.6
December				**0.532** (0.037)	**1.341** (0.103)	80.9				**0.538** (0.037)	1.119 (0.079)	58.1
Weibull shape							**0.760** (0.020)	**1.081** (0.017)	32.1	**0.759** (0.020)	**1.087** (0.018)	32.8
-2 log pseudolikelihood	85,037	31,867		84,922	31,804		21,407	7,493		21,296	7,442	

number of register spells: 6,050 of which number of events: 5,502
number of survey spells: 2,717 of which number of events: 2,287
Estimates significant at 5% (10 %) risk level are displayed in **boldface** (*italics*).
Standard errors adjusted for clustering of spells within persons.

[a] hr hazard ratio. $H_0 : hr = 1$

[b] bias= 100 × (Survey – Register)

[c] UE Unemployment

tion estimates were calculated setting the continuous variables at their mean values and the dummy variables to zero.[19] The survey baseline hazard from the Cox model is close to the register baseline hazard, although it displays a tendency towards underestimation. Due to the lower accuracy of measurement of spells in the survey, the survey baseline hazard is not able to reach the spike displayed by the register at the shortest durations. The survey baseline hazard from the Weibull model is nearly constant while the register baseline hazard shows negative duration dependence. The survey-based Weibull baseline hazard thus leads to erroneous conclusions about the duration dependence while the Cox baseline hazards from survey and register both display negative duration dependence. With respect to the estimation of the baseline hazard, the flexibility of the Cox model is clearly an advantage. As will be shown in section 6, the shape of the Weibull hazard is completely determined by spells shorter than one month.

Taking spell outcome into account markedly increases the measurement error bias in the estimated baseline hazard. The survey-based cause-specific hazard from the Cox model severely overestimates the true baseline hazard (Figure A.7 in Appendix). Moreover, the survey-based hazard is more kinked than the corresponding register-based hazard. If in the register data subsidised work is classified as employment, the register baseline hazard shifts somewhat upwards and exhibits similar kinks (results not shown here). The cause-specific Weibull baseline hazards from survey and register data lead again to different conclusions about the duration dependence (Figure A.8 in Appendix).

6 Effect of measurement accuracy

The previous section showed that the survey-based estimates of both the distribution of spells and of covariate effects were biased. This is a consequence of not only measurement errors but also of the way event history data were collected in the survey. In ECHP, information on main activity state is collected at the accuracy of one month. Moreover, as employment is preferred over unemployment, it is difficult to obtain information on unemployment spells shorter than one month. We aimed at separating the biases due to measurement error and measurement accuracy by discretizing the register spells and repeating the analyses with discretized data. Discrepancies between estimates based on survey data and discretized register data could then be taken as estimates of bias due to measurement error. Respectively, bias due to measurement accuracy could be evaluated by comparing results from original and discretized register data.

Register data were discretized in the following way: for each month, the number of unemployment days was calculated. If the number of days was at least 28, the register-based state of that month was defined as unemployed. The unemployment spell duration was then calculated by using these monthly indicators of unemployment state. Obtaining spell outcome information was not possible as this would have necessitated register information about other spells than unemployment. This information was not available in our data. Spells ongoing at December were censored if the person in question attrited from the survey the following year or if the spell was ongoing at the end of the reference period (December 1999).

Figure A.3 shows that the upward bias in the survey-based survival curve is to a large extent due to the lack of short spells. The survey and register curves are now for all practical purposes equal at durations less than approximately 8 months. The median spell duration in the discretized register data is 4 months, which is only one month shorter than in the survey data.

The estimates from the proportional hazard models based on the discretized register data, as well as the estimated biases due to measurement error and measurement accuracy, are shown in Table 6. Coarsening the measurement accuracy in register data diminishes the effect of being a female. This is due to the fact that females have a shorter median unemployment duration and thus, dropping out short spells affects more females than males. Measurement error operates in the same direction as measurement accuracy, attenuating the effect of being a female. Both measurement accuracy and measurement error cause a positive bias in the effect of higher education, the bias due to measurement error being markedly larger. This suggests that persons with higher education tend to underreport spell durations, a result supported by the model for the magnitude of measurement error (see Table 3). The area dummies have large biases due to both measurement accuracy and measurement error, but the biases tend to work in opposite directions. As for the time dummies, the biases due to measurement error and measurement accuracy are largest for year 1999, but they mostly work in opposite directions.[20] The biases in January and December dummies show that the heaping of spell starts and ends really is a measurement error and not a measurement accuracy problem. By contrast, the bias in the shape parameters of the Weibull models is for the most part due to measurement accuracy and, more specifically, the lack of short spells.

The estimated baseline hazard functions from the Cox proportional hazard models without time dummies are shown in Figure A.4. The lack of short spells in discretized register data and in survey data leads to underestimation of the baseline hazard for durations shorter than six months. For longer durations, the biases due to measurement accuracy and mea-

[19] This corresponds to a 36-year-old male with a basic level of education living in an urban municipality in the capital region, receiving basic unemployment allowance and having been unemployed 34 percent of the follow-up time before the spell in question. His unemployment spell started in 1995 the models including time dummies, the spell did not start in January and did not include December.

[20] The effect of the dummy for year 1999 increases markedly when the heaping dummies are included (models 2 and 4). In the discretized register data, an unemployment spell is ongoing in December 1999 is always censored because it is the last month of the follow-up period. this attenuates the effect of year dummies and, especially the effect of year 1999, in models not containing heaping dummies.

Table 6: Phase 1. Proportional hazards models. Measurement error in spell duration only

	1. Cox			2. Cox, heaping dummies			3. Weibull			4. Weibull, heaping dummies		
	Register2	MA	ME	Register2	MA	ME	Register2	MA	ME	Register2	MA	ME
Variable	hr[a] (se)	bias[b]	bias[c]	hr (se)	bias	bias	hr (se)	bias	bias	hr (se)	bias	bias
Female	0.995 (0.043)	-13.3	-7.0	0.985 (0.042)	-13.6	-4.1	1.001 (0.048)	-12.8	-8.0	0.989 (0.047)	-13.5	-5.3
Age	1.017 (0.013)	0.3	0.7	1.017 (0.129)	0.2	0.9	1.023 (0.015)	0.9	0.3	1.020 (0.014)	0.5	0.9
Age squared	**0.999** (0.000)	-0.1	0.0	**0.999** (0.000)	-0.1	0.0	**0.999** (0.000)	-0.1	0.0	**0.999** (0.000)	-0.1	0.0
Upper secondary educ.	1.061 (0.054)	0.9	6.7	1.076 (0.054)	2.4	4.4	1.074 (0.061)	2.3	7.9	1.087 (0.061)	3.6	5.8
Higher education	**1.420** (0.109)	4.7	13.5	**1.402** (0.106)	3.7	18.2	**1.461** (0.125)	7.1	20.1	**1.443** (0.122)	6.0	24.1
Proportion of UE[d] time	**0.997** (0.001)	-0.4	0.1	**0.998** (0.001)	-0.3	-0.1	**0.997** (0.001)	-0.4	0.1	**0.997** (0.001)	-0.4	0.0
Semi urban municipality	**1.200** (0.067)	6.5	-12.9	**1.190** (0.065)	5.3	-12.4	**1.248** (0.076)	10.8	-18.1	**1.229** (0.073)	8.8	-17.5
Rural municipality	**1.119** (0.060)	9.0	-14.0	*1.106* (0.058)	7.8	-12.7	**1.137** (0.068)	10.5	-17.6	*1.113* (0.065)	8.2	-14.1
Southern Finland	1.075 (0.065)	-11.4	13.8	1.081 (0.064)	-11.3	8.1	1.076 (0.072)	-12.4	14.8	1.083 (0.070)	-12.1	7.5
Eastern Finland	1.052 (0.076)	-9.9	8.2	1.069 (0.075)	-8.8	2.2	1.047 (0.084)	-11.3	11.4	1.068 (0.083)	-9.6	4.2
Central Finland	1.034 (0.080)	-3.5	11.1	1.045 (0.080)	-2.9	6.0	1.019 (0.088)	-6.0	10.8	1.035 (0.087)	-4.7	4.3
Northern Finland	1.134 (0.095)	-19.7	2.2	1.130 (0.094)	-20.4	0.6	1.139 (0.105)	-21.7	4.2	1.135 (0.103)	-22.3	2.4
Earnings-rel. UE[f] benefit	1.039 (0.046)	-2.9	-2.9	1.026 (0.045)	-3.7	-3.0	1.046 (0.053)	-2.4	-2.4	1.036 (0.051)	-3.1	-3.3
Year 1996	**1.157** (0.056)	8.4	-2.0	**1.168** (0.058)	9.9	6.0	**1.196** (0.067)	11.7	-3.3	**1.214** (0.070)	13.8	5.4
Year 1997	**1.259** (0.067)	10.8	-10.3	**1.288** (0.070)	14.3	-1.3	**1.326** (0.080)	15.8	-11.9	**1.362** (0.083)	19.8	-0.1
Year 1998	**1.366** (0.079)	4.0	-2.4	**1.465** (0.083)	15.1	-4.4	**1.493** (0.096)	12.8	-0.7	**1.624** (0.102)	26.7	1.5
Year 1999	**1.238** (0.100)	-12.5	-23.5	**1.857** (0.141)	56.0	-91.4	**1.492** (0.126)	6.5	-26.7	**2.209** (0.171)	84.6	-88.9
Begin in January				0.929 (0.059)	-7.4	-21.3				0.900 (0.064)	-9.4	-22.2
December				**0.364** (0.033)	-16.8	97.7				**0.358** (0.031)	-18.0	76.1
Weibull shape							**1.139** (0.015)	37.9	-5.8	**1.163** (0.015)	40.4	-7.6
-2 log pseudolikelihood	42,824			42,610			9,261			9,031		

number of discretized register spells: 3,448 of which number of events: 2,998
Estimates significant at 5% (10 %) risk level are displayed in **boldface** (*italics*).
Standard errors adjusted for clustering of spells into persons.

[a] hr hazard ratio. $H_0 : hr = 1$

[b] bias due to measurement accuracy,

calculated as = 100 × (Register2 − Register)

[c] bias due to measurement error,

calculated as = 100 × (Survey − Register2)

[d] UE unemployment

surement error work in opposite directions. Measurement accuracy creates a small positive bias leading to overestimation of the baseline hazard. The hazard spikes are however correctly placed in time. As measurement error creates a large negative bias, the joint effect of these two sources of bias leads to the underestimation of the baseline hazard. The effect of measurement error is, moreover, to flatten the shape of the baseline hazard. Figure A.5 shows the estimated Weibull hazard functions. Measurement accuracy has a dominating effect here: the exclusion of short spells leads to a badly biased shape of the baseline hazard, while measurement error only leads to slight underestimation of the level of the hazard.

7 Conclusion

Our study provided novel evidence on the existence, determinants and effects of measurement errors in event history analysis. Using longitudinal register data linked at person-level with longitudinal survey data, we were able to 1) provide information on the type and magnitude of measurement errors in retrospective survey reports of event histories, 2) assess the plausibility of classical assumptions about measurement errors, 3) study measurement error bias and 4) study the effect of measurement accuracy on event history analysis based on survey data.

Unemployment spells obtained from the FI ECHP data were used as the study variables of interest. Register data on unemployed jobseekers were used as the validation data. Available for all sample persons, having a definition of unemployment similar to that in the survey and giving precise information not only about the beginning and ending dates of each spell but also about spell outcomes, the validation data used in this study can be considered as being of outstanding quality.

According to our analysis, unemployment spells were subject to both omissions and, to a lesser extent, overreporting. Spell starts and ends were strongly heaped at the seams between the reference periods of consecutive panel waves. These findings are consistent with earlier studies on measurement errors in unemployment spells (Mathiowetz 1986, Mathiowetz and Duncan 1988, Kraus and Steiner 1998). A usually unnoticed issue is the classification error in reported spell outcomes. There was an excess of exits into employment in the survey data due to the fact that exits into subsidised work were often misclassified by respondents as becoming employed.

The model for the magnitude of measurement errors showed that the classical assumptions about measurement errors are not valid: cumulated measurement errors were correlated across survey waves, with variables related to true spells and with covariates used to explain the duration of spells. The model for the probability of omission of spells exhibited similar dependencies. Conducting a proxy interview instead of an interview with the person of interest and the lengthening of the recall period increased sharply the probability of omission while these survey design features had no effect on the magnitude of the cumulated measurement error.

The measurement error bias in an event history analysis was shown to result from both erroneously measured spell durations and misclassified spell outcomes. The survey data overestimated both the median duration of unemployment and the median time to becoming employed. There was no evidence of an overall attenuation effect of measurement errors on the estimated covariate effects, a result consistent with earlier empirical studies. The effect of education and in the competing risks analysis also the effect of receiving earnings-related unemployment benefit were estimated with sizeable bias. As for the estimated covariate effects, neither dummies for the heaping months nor the more flexible Cox model did protect against measurement error bias whereas the baseline hazard was much more accurately estimated by the Cox model. The survey-based estimates of the Weibull baseline hazard led to erroneous conclusions about the duration dependency of the hazard. The misclassification of spell outcomes was shown to be an important cause of bias in the estimate of cause-specific baseline hazard from the Cox model.

The survey data used in our study is collected by a multi-state framework with an accuracy of one month. Comparisons of the survey data with register data measured at day level are affected by differences in the measurement accuracy. Our attempts to separate the bias due to measurement accuracy and measurement error showed that measurement accuracy is an important source of bias in both the estimates of the distribution of spells and of covariate effects. Most notably, the bias in the Weibull baseline hazard was shown to be almost entirely due to lower measurement accuracy.

It is well-known that retrospective survey reports of event histories are affected by measurement errors. A few recent studies – including ours – suggest that measurement errors in survey spells have a non-negligible effect on an event history analysis. This has implications both for the survey organization collecting event history data and for the data analyst. In the light of our study, the use of proxy interviews should be kept to a minimum as they tend to lead to spell omissions. For the same reason, the time interval between the survey interview and the end of the reference period of the event history questions should be kept as short as possible. Paying attention to a careful definition of states in a multi-state data collection framework is important in order to avoid misclassification errors. Information about the spell distributions should be taken into account already in the questionnaire design phase in order to find an appropriate level of measurement accuracy. Our results suggest that attempts to control for heaping effects in the analysis phase by the inclusion of dummies for the heaping months are not helpful. As for the estimated covariate effects, the Cox model did not turn out to be more robust with respect to measurement error bias than the Weibull model. This contradicts earlier empirical findings concerning the robustness of Cox model with respect to non-response bias (van den Berg et al. 2004). However, the flexibility of the Cox model was clearly advantageous in the estimation of the baseline hazard. There have been only few attempts to develop methods to adjust for bias due to measurement error in spells in event history

analysis. Moreover, the proposed methods assume that measurement errors are independent of each other, of the true durations and of the covariates used to explain the durations. Our study suggests that methods making more realistic assumptions need to be developed in order to effectively adjust for measurement errors.

References

Abbring, J., van den Berg, G., & van Ours, J. (2005). The Effect of unemployment insurance sanctions on the transition rate from unemployment to employment. *The Economic Journal, 115*, 602-630.

Augustin, T. (1999). *Correcting for measurement error in parametric duration models by quasi-likelihood.* Collaborative Research Center 386, Discussion Paper 157.

Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. In J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (Vol. 5, p. 3705-3833). Amsterdam: Elsevier.

Carling, K., Edin, P., Harkman, A., & Holmlund, B. (1996). Unemployment duration, unemployment benefits and labor market programs in Sweden. *Journal of Public Economics, 59*, 313-334.

Courgeau, D. (1992). Impact of response errors on event history analysis. *Population: an English Selection, 4*, 97-110.

Eisenhower, D., Mathiowetz, N., & Morganstein, D. (1991). Recall error: Sources and bias reduction techniques. In P. Biemer, R. G. L. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (p. 127-144). New York: Wiley.

Fan, J., & Gijbels, I. (1996). *Local polynomial modeling and its applications.* London: Chapman and Hall.

Groves, R. (1989). *Survey errors and survey costs.* New York: Wiley.

Hill, D. (1994). The relative empirical validity of dependent and independent data collection in a panel survey. *Journal of Official Statistics, 10*, 359-380.

Holt, D., McDonald, J., & Skinner, C. (1991). The Effect of measurement errors on event history analysis. In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (p. 127-144). New York: Wiley.

Hujer, R., & Schneider, H. (1989). The Analysis of labor market mobility using panel data. *European Economic Review, 33*, 530-536.

Hunt, J. (1995). The Effect of unemployment compensation on unemployment duration in Germany. *Journal of Labor Economics, 13*, 88-120.

Jäckle, A. (2008a). *Measurement error and data collection methods: Effects on estimates from event history data.* ISER Working Paper Series 13.

Jäckle, A. (2008b). *The causes of seam effects in panel surveys.* ISER Working Paper Series 14.

Kalbfleisch, J., & Prentice, R. (2002). *The statistical analysis of failure time data* (2nd ed.). New York: Wiley.

Klein, J., & Moeschberger, M. (2003). *Survival analysis. Techniques for censored and truncated data* (2nd ed.). New York: Springer.

Kraus, F., & Steiner, V. (1998). *Modelling heaping effects in unemployment duration models – With an application to retrospective event data in the German Socio-Economic Panel.* Stuttgart: In Jahrbücher für Nationalökonomie und Statistik. Lucius & Lucius.

Lawless, J. (2003). Event History Analysis and Longitudinal Surveys. In R. Chambers & C. Skinner (Eds.), *Analysis of Survey Data.* Chichester: Wiley.

Lin, D. (1994). Cox regression analysis of multivariate failure time data: The marginal approach. *Statistics in Medicine, 13*, 2233-2247.

Mathiowetz, N. (1986). *The problem of omissions and telescoping error: New evidence from a study of unemployment.* Proceedings of the section on Survey Research Methods, American Statistical Association.

Mathiowetz, N., & Duncan, G. (1988). Out of work, out of mind: Response errors in retrospective reports of unemployment. *Journal of Business and Economic Statistics, 6*, 221-229.

Meyer, B. (1990). Unemployment insurance and unemployment spells. *Econometrica, 58*, 757-782.

Moore, J. (1988). Self/proxy response status and survey response quality. *Journal of Official Statistics, 4*, 155-172.

Peracchi, F. (2002). The European Community Household Panel: A review. *Empirical Economics, 27*, 63-90.

Pierret, C. (2001). Event history data and survey recall. *Journal of Human Resources, 36*, 439-466.

Pyy-Martikainen, M., & Rendtel, U. (2008). Assessing the impact of initial nonresponse and attrition in the analysis of unemployment duration with panel surveys. *Advances in Statistical Analysis, 92*, 297-318.

Pyy-Martikainen, M., Sisto, J., & Reijo, M. (2004). *The ECHP study in Finland. Quality report.* Helsinki: Statistics Finland.

Skinner, C., & Humphreys, K. (1999). Weibull regression for lifetimes measured with error. *Lifetime Data Analysis, 5*, 23-37.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling.* Boca Raton, FL: Chapman and Hall.

van den Berg, G., Lindeboom, M., & Dolton, P. (2004). *Survey non-response and unemployment duration.* IFAU working paper 2004:12.

Appendix

Figure A.1. Phase 1. Estimated baseline hazard function from the Cox model. No heaping dummies.

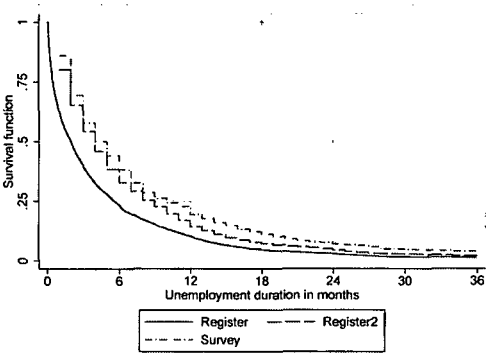


Figure A.3. Phase 1. Kaplan-Meier survival function estimates. Register 2: discretized register data.

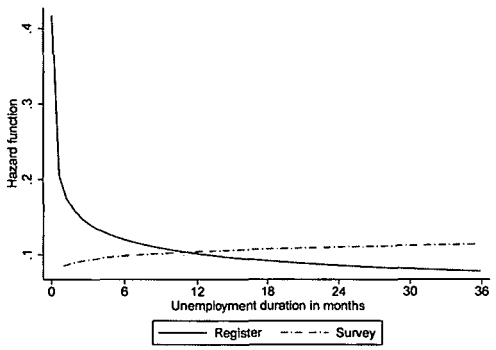


Figure A.2. Phase 1. Estimated baseline hazard function from the Weibull model. No heaping dummies.

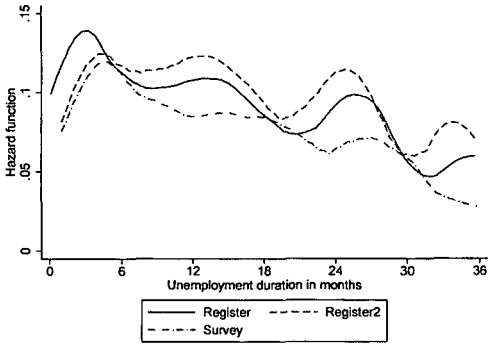


Figure A.4. Phase 1. Estimated baseline hazard function from the Cox model. No time dummies. Register 2: discretized register data.

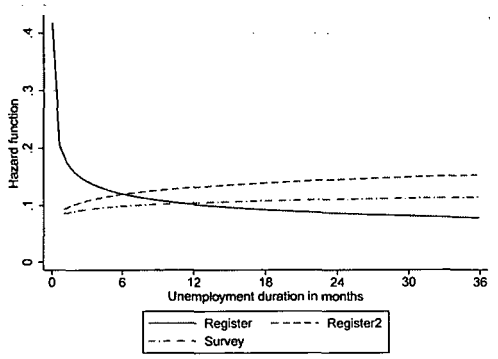


Figure A.5. Phase 1. Estimated baseline hazard function from the Weibull model. No time dummies. Register 2: discretized register data.

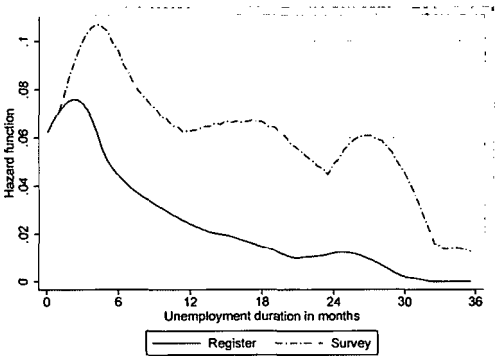


Figure A.7. Phase 2. Estimated baseline hazard function from the Cox proportional hazards model. No time dummies. Outcome of interest: becoming employed.

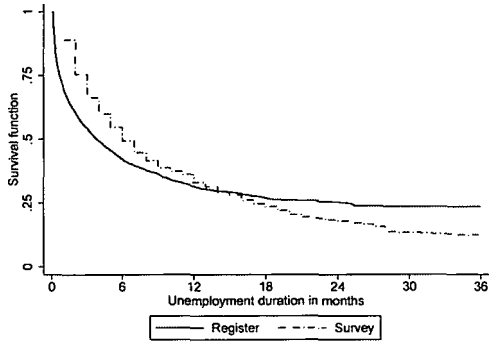


Figure A.6. Phase 2. Kaplan-Meier survival function estimates for the register and survey data. Outcome of interest: becoming employed.

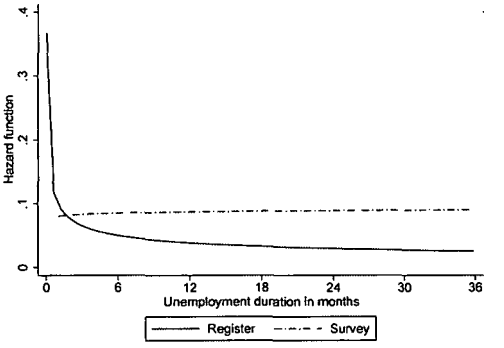


Figure A.8. Phase 2. Estimated baseline hazard function from the Weibull model. No time dummies. Outcome of interest: becoming employed.

Table A.1: Phase 2. Proportional hazards models. Outcome of interest: becoming employed

	1. Cox			2. Cox, heaping dummies			3. Weibull			4. Weibull, heaping dummies		
	Register	Survey		Register	Survey		Register	Survey		Register	Survey	
Variable	hr[a] (se)	hr (se)	bias[b]	hr (se)	hr (se)	bias	hr (se)	hr (se)	bias	hr (se)	hr (se)	bias
Female	1.039 (0.099)	**0.798** (0.047)	-24.2	1.033 (0.099)	**0.807** (0.048)	-22.6	1.033 (0.102)	**0.794** (0.052)	-23.9	1.028 (0.101)	**0.803** (0.053)	-22.4
Age	**1.098** (0.032)	**1.129** (0.020)	3.2	**1.099** (0.032)	**1.134** (0.020)	3.5	**1.105** (0.033)	**1.131** (0.022)	2.6	**1.105** (0.033)	**1.135** (0.023)	3.0
Age squared	**0.999** (0.000)	**0.998** (0.000)	-0.1	**0.999** (0.000)	**0.998** (0.000)	-0.1	**0.999** (0.000)	**0.998** (0.000)	-0.1	**0.999** (0.000)	**0.998** (0.000)	-0.1
Upper secondary educ.	1.034 (0.118)	1.106 (0.077)	7.2	1.033 (0.119)	1.097 (0.077)	6.4	1.047 (0.124)	1.137 (0.089)	9.0	1.046 (0.124)	1.126 (0.088)	7.9
Higher education	**1.377** (0.204)	**1.558** (0.151)	18.1	**1.366** (0.203)	**1.565** (0.152)	19.9	**1.433** (0.219)	**1.690** (0.179)	25.8	**1.423** (0.217)	**1.687** (0.179)	26.4
Proportion of UE[c] time	1.002 (0.002)	**0.997** (0.001)	-0.5	*1.003* (0.002)	**0.997** (0.001)	-0.6	1.002 (0.002)	**0.997** (0.001)	-0.5	1.002 (0.002)	**0.997** (0.001)	-0.5
Semi urban municipality	1.133 (0.171)	1.114 (0.092)	-1.9	1.135 (0.172)	1.107 (0.093)	-2.8	1.158 (0.175)	1.114 (0.104)	-4.5	1.160 (0.175)	1.098 (0.104)	-6.2
Rural municipality	1.016 (0.110)	0.972 (0.074)	-4.4	1.013 (0.110)	0.981 (0.075)	-3.2	1.031 (0.116)	0.957 (0.080)	-7.4	1.030 (0.116)	0.967 (0.081)	-6.3
Southern Finland	1.345 (0.339)	1.141 (0.099)	-20.4	1.354 (0.341)	1.091 (0.094)	-26.3	1.380 (0.348)	*1.173* (0.112)	-20.7	1.387 (0.349)	1.104 (0.104)	-28.2
Eastern Finland	1.283 (0.348)	1.135 (0.118)	-14.8	1.296 (0.351)	1.095 (0.113)	-20.0	1.307 (0.354)	1.178 (0.134)	-12.9	1.315 (0.356)	1.124 (0.126)	-19.2
Central Finland	1.171 (0.291)	1.186 (0.133)	1.5	1.178 (0.292)	1.147 (0.130)	-3.0	1.193 (0.297)	1.179 (0.148)	-1.4	1.198 (0.298)	1.132 (0.141)	-6.7
Northern Finland	*1.568* (0.414)	1.175 (0.137)	-39.3	*1.574* (0.415)	1.154 (0.134)	-42.0	*1.628* (0.430)	1.212 (0.154)	-41.6	*1.633* (0.430)	1.186 (0.148)	-44.7
Earnings-rel. UE[c] benefit	**1.319** (0.140)	1.035 (0.065)	-28.4	**1.310** (0.139)	1.016 (0.063)	-29.4	**1.331** (0.143)	1.046 (0.073)	-28.5	**1.325** (0.143)	1.018 (0.070)	-30.7
Year 1996	1.087 (0.077)	**1.151** (0.077)	6.4	1.078 (0.076)	**1.250** (0.088)	17.2	1.118 (0.083)	**1.192** (0.093)	7.4	1.110 (0.083)	**1.315** (0.108)	20.5
Year 1997	1.091 (0.091)	*1.158* (0.089)	6.7	1.081 (0.091)	**1.292** (0.105)	21.1	1.141 (0.101)	**1.219** (0.105)	7.8	1.133 (0.101)	**1.399** (0.128)	26.6
Year 1998	**1.323** (0.135)	**1.383** (0.111)	6.0	**1.304** (0.136)	**1.526** (0.129)	22.2	**1.440** (0.152)	**1.558** (0.137)	11.8	**1.423** (0.154)	**1.796** (0.164)	37.3
Year 1999	**1.321** (0.168)	1.004 (0.110)	-31.8	*1.250* (0.161)	1.105 (0.147)	-14.5	**1.508** (0.193)	*1.246* (0.142)	-26.2	**1.434** (0.187)	**1.539** (0.199)	10.5
Begin in January				1.050 (0.079)	**0.693** (0.047)	-35.7				1.036 (0.083)	**0.646** (0.049)	-39.0
December				**0.536** (0.050)	1.025 (0.095)	48.9				**0.545** (0.049)	0.873 (0.077)	32.8
Weibull shape							**0.616** (0.019)	*1.032* (0.018)	41.6	**0.615** (0.018)	**1.050** (0.019)	43.5
-2 log pseudo likelihood	51,838	23,084		51,770	23,055		17.973	6,679		17,910	6,637	

#(register spells): 6050 of which #(events): 3238
#(survey spells): 2717 of which #(events): 1637
Estimates significant at 5% (10%) risk level are displayed in **boldface** (*italics*).
Standard errors adjusted for clustering of spells within persons.

[a] hr hazard ratio. $H_0 : hr = 1$

[b] bias= 100 × (Survey – Register)

[c] UE Unemployment

Article III

Pyy-Martikainen, M. & Nordberg, L., Inverse probability of censoring weighting method in survival analysis based on survey data. Statistics in Transition, 8, 3, 487–501, 2007..

INVERSE PROBABILITY OF CENSORING WEIGHTING METHOD IN SURVIVAL ANALYSIS BASED ON SURVEY DATA

Marjo Pyy-Martikainen[1], Leif Nordberg[2]

ABSTRACT

In survival analysis based on survey data, attrition implies that a part of the event times are right-censored: it is only known that the true time exceeds that observed. To simplify the analysis, it is usually assumed that the process generating right-censoring is independent of the remaining event time. In practice, the assumption of independent censoring may not always hold. Dependent censoring may cause a bias in survival analysis. An inverse probability of censoring weighting (IPCW) method has been proposed to adjust for bias in survival analysis due to dependent censoring. To our knowledge, however, there are no empirical applications of the method in a complex survey data context. We use simulation methods to study the statistical properties of IPCW method in an artificial 2-wave panel survey. Our simulation study shows that the IPCW method is able to reduce bias in survival estimation also when there is only little information about the determinants of the right-censoring mechanism.

Key words: Inverse probability of censoring weights; survival analysis; complex surveys; simulation study.

1. Introduction

In panel surveys, data on durations spent in various states is often collected. Examples include duration of unemployment or employment, duration of poverty, duration of social assistance benefit receipt etc. In the analysis of durations or spells based on survey data, attrition implies that the ending date of some of the spells is unknown. For these *right-censored* spells, it is only known that the length

[1] Department of Economics and Statistics, Åbo Akademi University and Statistics Finland, e-mail: marjo.pyy-martikainen@stat.fi

[2] Department of Economics and Statistics, Åbo Akademi University, e-mail: lnordber@abo.fi

of the spell was at least that observed. Right-censoring may also occur because of end of follow-up time.

It is usually assumed, in order to make the analysis easier, that the right-censoring mechanism is *independent* of the remaining event time. This means that the right-censoring mechanism does not remove individuals from the survey because of particularly long or short durations. Under an independent right-censoring mechanism censoring does not cause bias and can thus be ignored in the analysis. In social surveys, the probability of attrition may be related for example to social exclusion which may be manifested by a long duration of unemployment or poverty. In such a situation, analyses of unemployment or poverty duration that ignore the censoring mechanism will lead to biased estimates.

Robins (1993) introduced an *inverse probability of censoring weighting* (IPCW) method that aims to correct for bias due to dependent right-censoring utilizing auxiliary variables related to both censoring and the duration of interest. He showed that if right-censoring is conditionally independent given the auxiliary variables, then using IPCW versions of Kaplan-Meier and Cox partial likelihood estimators result in consistent estimation. In simulation studies by van der Laan and Hubbard (1997) and van der Laan and Robins (1998) it has been shown that IPCW-based estimators perform remarkably well in a non- or semiparametric setting and in situations where the information about survival times is very limited.

In social surveys it is likely that all auxiliary variables needed to achieve conditional independence of censoring and event times are not observed. The censoring mechanism may thus contain some information on the event time of interest even after the IPCW correction. We conduct a simulation study in order to investigate the performance of the IPCW method in such a less-than-perfect situation. Our aim is to find out how strong the auxiliary information has to be for the IPCW method to be a useful tool. We take a design-based approach in the analysis. Consequently, our target parameters are the finite population regression coefficient B and survival function $S(t)$ that would be obtained from the estimation procedure if all data values in the finite population were available instead of having a sample only. The use of the IPCW method in a complex survey data context has previously been discussed by Lawless (2003a). However, we are unaware of any empirical applications of the method in the analysis of complex survey data. Our simulation study indicates that the method may be very useful even when the censoring mechanism is only partially known.

The paper is organized as follows. Section 2 gives a short introduction to some basic concepts of survival analysis. Section 3 discusses model-based and design-based approaches to survival analysis and introduces the design-based versions of Kaplan-Meier estimator and the partial likelihood function used to estimate the parameters of the Cox proportional hazards model. Section 4 introduces the concept of independent censoring and the IPCW method aimed to adjust for bias caused by violation of this assumption. The performance of the IPCW method is

studied by simulation methods in section 5. Section 6 concludes by discussing the findings from the simulation study.

2. Basic concepts of survival analysis

We are interested in making inferences about a duration or spell variable T. Because of censoring, only $t = \min(T, C)$ and $\delta = I(T \leq C)$ are observed, where C is a censoring time and δ is an event indicator. If $\delta = 1$, T is observed, and if $\delta = 0$, then we know only that the event time is longer than the censoring time. In longitudinal surveys, C depends on the length of the follow-up time, on the time at which the duration began and on the time of attrition (Lawless 2003a). The survival function and the hazard function are the two most important ways to express the distribution of a duration variable T. The value of the survival function

$$S(t) = P(T \geq t)$$

at time t is the probability that the spell is at least as long as t. The value of the hazard function

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt \mid T \geq t)}{dt}$$

at time t describes the conditional probability of spell completion at time t, given that the spell has lasted until t. Models for the duration variables are usually constructed by defining the way covariates affect the hazard function.

3. Model-based and design-based approaches to survival analysis

A classical model-based analysis assumes that observations for different units are independent and that the sampling design is noninformative. Under a noninformative sampling design, the sample inclusion probabilities are not related to the values of outcome variables (for two alternative more formal definitions of noninformative sampling, see Chambers and Skinner, 2003, p.4 and Pfeffermann and Sverchkov, 2003, p.176). Longitudinal social surveys often have a complex sampling design with unequal probabilities of selection, stratification and clustering of observations. As a consequence, both the assumptions of independence and noninformativeness may be violated. In such a case it is necessary to take the impact of the sampling design into account in the analysis. A common approach to the design-based inference about a model parameter θ is to specify a finite population parameter θ_U that would be obtained from the model estimation procedure if all data values in the finite population U were available instead of

having a sample only. An estimate of θ_U is then obtained using sample data values and sample weights. This approach is also followed in our study. We are interested in estimating the finite population regression coefficient B from a Cox proportional hazards model. The design-based theory of Cox proportional hazards model was developed by Binder (1992). Similarly, we are interested in estimating the empirical survival function $S(t) = \frac{1}{N}\sum_{i=1}^{N} I(T_i \geq t)$ based on all units in the finite population (Lawless, 2003a). The following two subsections introduce the design-based versions of Kaplan-Meier estimator and the partial likelihood function used to estimate the parameters of the Cox proportional hazards model. In both cases we assume the population is constant over time and that we have at most one spell per unit.

Without going deeper in the discussion concerning the relative merits of design-based versus model-based analysis (see e.g. Pfeffermann, 1993), we see it for many reasons as both interesting and worthwhile to try to derive optimal estimators for the results one would get if one made the same analysis in an "ideal" situation, i.e. in the case where all data values in the finite population U were available.

3.1. Kaplan-Meier estimator

The Kaplan-Meier estimator (Kaplan and Meier, 1958) is a nonparametric estimator of the survival function $S(t)$. Folsom, Lavange and Williams (1989) developed an estimator that is appropriate when survival data is obtained from a complex survey. A lucid discussion of this estimator can be found in Lawless (2003b). Let $t_i, i = 1,\ldots,n$ be the observed event and censoring times in the sample of size n. Let $t_{(1)},\ldots,t_{(h)},\ldots,t_{(r)}$ be the ordered event times. The weighted number of observations undergoing an event at $t_{(h)}$ is $D_{(h)} = \sum_{i=1}^{n} I(t_i = t_{(h)})\delta_i w_i$, where w_i is the weight attached to observation i and δ_i is the event indicator defined earlier. The weighted number of observations with event or censoring times exceeding $t_{(h)}$ is $N_{(h)} = \sum_{i=1}^{n} I(t_{(h)} \leq t_i) w_i$. The weighted Kaplan-Meier estimator of the survival function is defined as

$$\hat{S}(t) = \prod_{h=1}^{r}\left(1 - \frac{D_{(h)}}{N_{(h)}}\right)^{I(t_{(h)} \leq t)}. \qquad (1)$$

Note that $D_{(h)}$ estimates the number of population units that undergo an event at time $t_{(h)}$ and $N_{(h)}$ estimates the population size of the risk set at time $t_{(h)}$. $\hat{S}(t)$ is thus an estimator of a population survival function that would be obtained if all the units of the finite population of interest were available for analysis.

3.2. Cox proportional hazard model

It is often of interest to find out how certain covariates $x=(x_1,\ldots,x_p)$ are related to the event time T. One of the most popular tools to study the association between T and x is the Cox proportional hazards model (Cox, 1972). The model specifies the hazard function as a product of two terms:

$$\lambda(t \mid x) = \lambda_0(t)\exp(x\beta),$$

where $\lambda_0(t)$ is a baseline hazard function that depends only on the event time and $\exp(x\beta)$ defines the way covariates x affect the hazard function. One reason for the popularity of the Cox proportional hazards model is the fact that the model parameters β can be estimated without assuming any parametric distribution for the event time variable T.

For survival data obtained from a complex survey, Binder (1992) used a pseudo-likelihood method to estimate the parameters and their variances for a Cox proportional hazards model. The unequal selection probabilities are taken into account by using sample weights. The dependence between observations is not modelled explicitly but is taken into account in variance estimation. The model is estimated by maximising a partial likelihood function. For a population of N units, the partial likelihood function is defined as

$$PL = \prod_{i=1}^{N}\left[\frac{\lambda(t_i \mid x_i)}{\sum_{j=1}^{N} I(t_i \le t_j)\lambda(t_i \mid x_j)}\right]^{\delta_i},$$

where x_i is the covariate vector, t_i is the spell length, and δ_i is the event indicator related to unit i. $I(t \le t_j)$ indicates whether the spell of unit j is still going on at time t. The sum $\sum_{j=1}^{N} I(t \le t_j)$ defines the size of the risk set, i.e. the number of spells still going on at time t. Note that the part of the hazard function that

depends on event time only is common to each unit and cancels from the expression. The partial likelihood function can thus be expressed as

$$PL = \prod_{i=1}^{N} \left[\frac{\exp(x_i B)}{\sum_{j=1}^{N} I(t_i \le t_j) \exp(x_j B)} \right]^{\delta_i},$$

where B is the vector of population regression coefficients. B is determined as the solution to the score equations:

$$\frac{\partial \log PL}{\partial B} = \sum_{i=1}^{N} \delta_i \left[x_i - \frac{\sum_{j=1}^{N} I(t_i \le t_j) x_j \exp(x_j B)}{\sum_{j=1}^{N} I(t_i \le t_j) \exp(x_j B)} \right] = 0.$$

As noted by Roberts and Kovacevic (2007), if all units of the finite population do not experience spells, then N is the size of the subpopulation that experiences spells. To estimate the population regression coefficient B from a sample of n observations, Binder (1992) proposed the following pseudo-score estimating equations:

$$\sum_{i=1}^{n} w_i \delta_i \left[x_i - \frac{\sum_{j=1}^{n} w_j I(t_i \le t_j) x_j \exp(x_j \hat{B})}{\sum_{j=1}^{n} w_j I(t_i \le t_j) \exp(x_j \hat{B})} \right] = 0, \qquad (2)$$

where $w_j, j = 1, \ldots, n$ are the sample weights attached to the sample observations. The estimator $\hat{B}$ that solves equation (2) is the pseudo-maximum likelihood estimator of B (Binder, 1992). Binder (1992) and Roberts and Kovacevic (2007) discuss the design-based estimation of variance of $\hat{B}$.

4. Dependent censoring and the IPCW method

Censoring of spells occurs because of the shortness of the follow-up period or because of attrition. Censoring is thus related to the data collection and not to the phenomenon under study. Therefore, censoring should not affect the analysis of spells. To make analysis easier, it is usually assumed that the censoring mechanism is *independent* of the remaining event time (see e.g. Kalbfleisch and Prentice,

1980, pp. 119-121). For independent censoring mechanisms, the cause-specific hazard of T equals the marginal hazard:

$$\begin{aligned} \lambda_T(t \mid x, C \geq t) &= \lim_{dt \to 0} \frac{P(t \leq T < t+dt \mid T, C \geq t, x)}{dt} \\ &= \lim_{dt \to 0} \frac{P(t \leq T < t+dt \mid T \geq t, x)}{dt} = \lambda(t \mid x). \end{aligned} \tag{3}$$

Assumption (3) means that censoring and failure mechanisms are conditionally independent, given x. A (conditionally) independent right-censoring mechanism does not remove units from the survey because of particularly long or short spells. Under a (conditionally) independent censoring mechanism censoring does not cause bias and can thus be ignored in the analysis. In reality the assumption of independent censoring may not always hold. In social surveys, attrition may be related to particularly long or short spells of unemployment, poverty, supplementary benefit receipt etc. If this is not properly taken into account, the corresponding estimates from spell analyses are biased. It should be noted that bias is an outcome-specific issue. Attrition that is selective with respect to duration of benefit receipt may not cause bias in the analysis of unemployment duration.

4.1. Inverse probability of censoring weights

The inverse probability of censoring weighting (IPCW) method was introduced by Robins in 1993. The method aims to correct for bias due to dependent censoring. The assumption of independent censoring (equation (3)), can be shown to be equivalent to

$$\lambda_C(t \mid x, T, T > t) = \lambda_C(t \mid x, T > t), \tag{4}$$

which means that the cause-specific hazard of censoring does not depend on the (possibly unobserved) event time T. Dependent censoring mechanisms thus violate equation (4). The fundamental assumption underlying the IPCW method is that, given a vector of auxiliary variables z,

$$\lambda_C(t \mid x, z, T, T > t) = \lambda_C(t \mid x, z, T > t). \tag{5}$$

Given assumption (5), equation (4) is true if z does not predict censoring, ie. if

$$\lambda_C(t \mid x, z, T > t) = \lambda_C(t \mid x, T > t). \tag{6}$$

The assumption of independent censoring can thus be tested by modelling the cause-specific hazard of censoring using e.g. the Cox proportional hazard model. If the auxiliary variables z explain the cause-specific hazard of censoring, then

censoring is dependent, which has to be taken into account in the analysis. The auxiliary variables z are variables which are not of interest as such, but which are used to correct for bias due to dependent censoring. In order to be effective in this respect, the auxiliary variables should be associated with both censoring and event times. Robins (1993) showed that if the assumption (5) holds, then using in equations (1) and (2) weights defined by

$$w_i(t) = \frac{1}{S_C(t \mid x_i, z_i)},$$

where $S_C(t \mid x_i, z_i)$ is the cause-specific survival function for unit i, results in consistent estimation under dependent censoring. The estimate of $S_C(t \mid x_i, z_i)$ can be based on a fit of a Cox proportional hazard model with censoring as the event of interest. The weights $w_i(t)$ are time-dependent and inversely proportional to the conditional probability of having remained uncensored until time t, given x_i and z_i.

The approach of Robins is purely model-based. The sampling design has no role in the analysis and the weights are used only to correct for dependent censoring. Lawless (2003a) discussed the use of IPC weights in the estimation of survival function based on complex survey data. He showed that, in general, weights related to both the sampling design and censoring mechanism are needed for consistent estimation. He proposed the use of weights defined by

$$w_i^*(t) = \frac{1}{\pi_i \times S_C(t \mid x_i, z_i)}, \qquad (7)$$

where $\pi_i = P(i \in s)$ is the sample inclusion probability for unit $i, i = 1, \ldots, N$.

5. Simulation study

If the censoring model is correctly specified, then the IPC weighted Kaplan-Meier and Cox partial likelihood estimators can fully correct the bias due to dependent censoring. In longitudinal surveys, attrition depends on many variables, some of which may not be observed. Thus, the estimated censoring model should be considered as an approximation of the true model and, consequently, there is likely to be some residual dependency between T and C even after conditioning on z. Our aim is to study the bias-reducing power of the IPCW method in the presence of dependent censoring under the following scenarios:

1. The variable determining the censoring mechanism, z, is known.

2. We observe a variable that is either a) strongly or b) weakly associated with the variable z.
3. The variable z is unknown.

We assume that we are interested in the survival function and how a single covariate affects the hazard function. The parameters of interest are thus the values of the finite population survival function $S(t)$ at certain time points and the finite population regression coefficient B of the covariate. The statistical properties of IPC weighted estimators $\hat{B}$ and $\hat{S}(t)$ using weights defined in equation (7) were studied by simulation methods. Four different artificial populations corresponding to the above scenarios were generated and from each population, $K = 500$ independent samples were drawn. For each sample $s_j, j = 1..., K$, drawn, estimates $\hat{B}_j$ and $\hat{S}_j(t)$ were calculated. The distribution of the K estimates was used as an approximation of the sampling distribution and the following estimators

$$\bar{\hat{B}} = \frac{1}{K}\sum_{j=1}^{K}\hat{B}_j$$

$$S^2_{\hat{B}} = \frac{1}{K-1}\sum_{j=1}^{K}(\hat{B}_j - \bar{\hat{B}})^2$$

were used to estimate the mean and variance of the sampling distribution of IPC weighted $\hat{B}$. For the IPC weighted Kaplan-Meier estimator, the mean and variance of the sampling distribution were estimated at time points $t = 1,25,50,75,100,125,150$.

5.1. Generation of the populations

The longitudinal population of interest is defined as individuals belonging to the target population at the beginning of the survey. The population is assumed to remain constant over time. Thus, there are no exits from or entrances to the target population. Each of the 4 populations are of size $N = 10000$ individuals and consist of $D = 8$ subgroups defined by binary variables sex, level of education and social exclusion. We are interested in the overall survival function of unemployment spells and the effect of education on the spell length.

Social exclusion is an unobserved variable that determines the probability of attrition. Sex is used as a stratification variable in the sampling design and as an auxiliary variable in the censoring model.

Table 1. The association between variables sex and social exclusion in populations 1, 2a, 2b and 3.

Population	Degree of association	ϕ	% of men excluded	% of women excluded
1	perfect	1	100	0
2a	strong	0.6	80	20
2b	weak	0.2	60	40
3	none	0	50	50

The 4 populations differ by the degree of association between variables sex and social exclusion according to Table 1. The phi coefficient ϕ measures association between 2 binary variables. The binary variables are considered positively associated if most of the data falls along the diagonal cells of a 2-by-2 frequency table. Value $\phi = 1$ corresponds to perfect positive association, value $\phi = -1$ to perfect negative association and value $\phi = 0$ to no association.

We consider a single spell analysis with one unemployment spell per individual (in an empirical analysis it could be e.g. the first spell beginning during the observation period). The unemployment spells were generated from the Weibull distribution, whose hazard function is

$$\lambda(t \mid a,b) = \frac{a}{b}(\frac{x}{b})^{a-1}. \qquad (8)$$

The shape parameter a was set equal to 0.8 in each subpopulation. This corresponds to a decreasing hazard rate. The scale parameter b_1 of subpopulation 1, socially excluded men with low education, was set equal to 100, which corresponds a median duration of 63 days. The other scale parameters $b_d, d = 2,\ldots,8$, were chosen according to the hazard rates of Table 2.

The median duration of the unemployment spells as well as the effect of education on the hazard of spell completion are different among the excluded and among the non-excluded. The ratio of hazard rates (the hazard ratio or briefly hr) for the variable education is 3 among the non-excluded and 1.5 among the excluded. Censoring that depends on exclusion status thus biases both the estimates of survival function $S(t)$ and the estimate of the regression coefficient B.

Table 2. The hazard rates among the subpopulations. Median unemployment durations in parentheses.

	Excluded		Non-excluded	
	Low education	High education	Low education	High education
Men	h_1 (63)	$h_2 = 1.5h_1$ (38)	$h_3 = 2h_1$ (27)	$h_4 = 6h_1$ (7)
Women	$h_5 = h_1$ (63)	$h_6 = 1.5h_1$ (38)	$h_7 = 2h_1$ (27)	$h_8 = 6h_1$ (7)

5.2. Sampling design

We drew stratified simple random samples without replacement using sex as a stratification variable. Each sample is selected at a single time point, say day 0, at which there are $N = 10000$ individuals in the finite population. For each unit in the population is generated an unemployment spell according to relevant Weibull distribution. Unemployment spells start randomly during days $1,\ldots,30$. Each sample has 350 men and 250 women corresponding to inclusion probabilities of 0.07 and 0.05. This is a relatively simple sampling design where sample weights are however needed to produce design-unbiased population-level estimates. The sample weights are defined as the inverses of the inclusion probabilities. For each sample, an artificial 2-wave panel survey with attrition was conducted. The first wave interview is assumed to occur at day 30. It is assumed that there is no nonresponse at wave 1. Consequently, each unemployment spell is observed at least until day 30. We generated selective attrition by stratifying the samples according to exclusion status and drawing 80% samples among the non-excluded and 20% samples among the excluded. This corresponds to an attrition rate of 20% among the non-excluded and an attrition rate of 80% among the excluded. For the non-attriters, the observed duration is determined as $t = \min(T,600)$. For each sample, the IPC weights were constructed using sex as an auxiliary variable in the censoring model and weighted estimates of $S(t)$ and B were calculated. As the weights are time-varying the data had to be transformed into a counting process form (see e.g. Therneau and Grambsch, 2000, p. 68), where each unemployment spell is split into several intervals, the splitting points being defined by the times at which censoring occurs in the sample.

5.3. Results

The results from our simulation study are shown in Table 3. The number of replicate samples was 500. The true population parameters that are being estimated are shown in the first column. For both the design-weighted and IPC weighted

estimators, the mean, standard deviation and percent bias are reported. The percent bias of the design-weighted estimators shows how much selective attrition distorts the results. In population 1, the association between sex and social exclusion is perfect. This corresponds to a situation where the censoring mechanism is known and is, thus, an ideal situation for the IPC correction. Looking at the last column of Table 3, we see that the bias due to selective attrition has indeed almost vanished. A small positive bias remains in both $\hat{B}$ and $\hat{S}(t)$. As noted by Binder (1992), $\hat{B}$ is a design-consistent, but not a design-unbiased estimator of B. We are not aware of results concerning the design-based properties of the weighted Kaplan-Meier estimator.

In general, the bias of IPC weighted estimators grows as the association of sex and social exclusion gets weaker but is always less than the bias of design-weighted estimators. When sex and social exclusion are independent, the bias of IPC weighted Kaplan-Meier estimators is equal to that of design-weighted estimators. In that case there is thus no gain from using IPC weights in survival curve estimation.

Interestingly, the IPC weighted estimators of the hazard ratio $\exp(B)$ perform quite well relative to design-weighted estimators even when the association between sex and social exclusion is weak or when the variables are independent. This may be explained in the following way. Because of censoring, the IPC weights grow over time. This means that persons who remain a long time in the risk set and, therefore, are more likely to be excluded, get larger values of weights. This corrects the estimates in the right direction.

6. Discussion

We conducted a simulation study to investigate the performance of IPCW method in survival analysis based on complex survey data. If the censoring and event times are conditionally independent, given a set of auxiliary variables, then using this information in the construction of IPC weights can remove bias due to dependent censoring. However, in real-world situations, what is often observed are not the variables determining the censoring mechanism but some correlates of them. As a consequence, there may be residual dependency between censoring and event times even after the IPCW correction. Our simulation study shows that the IPCW method may be useful in survival analysis based on complex survey data even in such less-than-perfect real-world situations. Remarkably, there are gains from using the IPCW method in the estimation of the population regression coefficient even when the censoring mechanism is completely unknown. The development of design-based variance estimation methodology for IPCW Kaplan-Meier and Cox partial likelihood estimators remains an area where further research is needed (Lawless, 2003a).

Table 3. Results of a simulation study with 500 replications. $hr = \exp(B)$.

	Population parameters	Design weighted estimates mean	s.d.	% bias	IPC weighted estimates mean	s.d.	% bias
	$hr = 1.687$	1.977	0.213	17.2	1.710	0.235	1.4
	$S(1) = 0.940$	0.940	0.009	0.0	0.940	0.009	0.0
	$S(25) = 0.493$	0.477	0.022	-3.3	0.502	0.022	1.8
1	$S(50) = 0.333$	0.294	0.025	-11.8	0.342	0.027	2.6
	$S(75) = 0.248$	0.203	0.023	-18.2	0.253	0.027	2.3
	$S(100) = 0.191$	0.149	0.021	-22.3	0.197	0.027	2.8
	$S(125) = 0.147$	0.109	0.019	-26.1	0.151	0.026	2.6
	$S(150) = 0.116$	0.083	0.017	-28.3	0.119	0.024	2.2
	$hr = 1.663$	1.928	0.197	15.9	1.748	0.212	5.1
	$S(1) = 0.936$	0.935	0.010	0.0	0.935	0.010	0.0
	$S(25) = 0.492$	0.476	0.023	-3.2	0.484	0.023	-1.6
2a	$S(50) = 0.336$	0.296	0.026	-11.9	0.313	0.026	-7.0
	$S(75) = 0.242$	0.200	0.024	-17.2	0.215	0.024	-10.9
	$S(100) = 0.185$	0.145	0.021	-21.7	0.159	0.022	-13.9
	$S(125) = 0.142$	0.106	0.019	-25.4	0.119	0.021	-16.5
	$S(150) = 0.113$	0.082	0.017	-27.4	0.093	0.019	-17.7
	$hr = 1.655$	1.960	0.231	18.4	1.807	0.233	9.2
	$S(1) = 0.938$	0.939	0.009	0.1	0.939	0.009	0.1
	$S(25) = 0.504$	0.494	0.023	-2.0	0.495	0.023	-1.7
2b	$S(50) = 0.340$	0.300	0.025	-11.9	0.302	0.024	-11.1
	$S(75) = 0.246$	0.206	0.023	-16.3	0.207	0.023	-15.7
	$S(100) = 0.181$	0.143	0.021	-20.9	0.144	0.021	-20.4
	$S(125) = 0.138$	0.103	0.019	-25.4	0.104	0.019	-24.6
	$S(150) = 0.109$	0.079	0.017	-27.3	0.080	0.017	-26.4
	$hr = 1.720$	2.003	0.230	16.5	1.853	0.245	7.7
	$S(1) = 0.937$	0.938	0.009	0.1	0.938	0.009	0.1
	$S(25) = 0.495$	0.475	0.023	-3.9	0.476	0.023	-3.9
3	$S(50) = 0.332$	0.286	0.024	-13.9	0.286	0.024	-13.9
	$S(75) = 0.243$	0.195	0.024	-19.6	0.195	0.024	-19.6
	$S(100) = 0.182$	0.136	0.021	-24.7	0.136	0.021	-24.6
	$S(125) = 0.139$	0.099	0.018	-28.4	0.099	0.018	-28.3
	$S(150) = 0.110$	0.074	0.016	-32.8	0.074	0.016	-32.7

Acknowledgements

The authors would like to thank an anonymous referee for comments that greatly improved the readability of the paper.

REFERENCES

BINDER, D. (1992). Fitting Cox's Proportional Hazards Models from Survey Data. Biometrika 79, 1, 139—147.

CHAMBERS, R. and SKINNER, C. (2003). Analysis of Survey Data. Wiley.

COX, D. (1972). Regression Models and Life Tables. Journal of Royal Statistical Society B, 34, 187—220.

FOLSOM, R., LAVANGE, L. and WILLIAMS, R. (1989). A Probability Sampling Perspective on Panel Data Analysis. In Panel Surveys (D. Kasprzyk, G. Duncan, G. Kalton and M. Singh, eds), pp. 108—38. Wiley.

KALBFLEISCH, J. and PRENTICE, R. (1980). The Statistical Analysis of Failure Time Data. Wiley.

KAPLAN, E. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. Journal of American Statistical Association 53, 457—481.

LAWLESS, J. (2003a). Censoring and Weighting in Survival Estimation from Survey Data. SSC Annual Meeting, June 2003, Proceedings of the Survey Methods Section.

LAWLESS, J. (2003b). Event History Analysis and Longitudinal Surveys. In Analysis of Survey Data (R. Chambers and C. Skinner, eds), 221—243. Wiley.

PFEFFERMANN, D. (1993). The Role of Sampling Weights when Modeling Survey Data. International Statistical Review, 61, 317—337.

PFEFFERMANN, D. and SVERCHKOV, M. (2003). Fitting Generalized Linear Models under Informative Sampling. In Analysis of Survey Data (R. Chambers and C. Skinner, eds), 175—195. Wiley.

ROBERTS, G. and KOVACEVIC, M. (2007). Modelling Durations of Multiple Spells From Longitudinal Surveys. Survey Methodology, 33, 13—22.

ROBINS, J. (1993). Information Recovery and Bias Adjustment in Proportional Hazards Regression Analysis of Randomized Trials Using Surrogate Markers.

In Proceedings of the Biopharmaceutical Section, American Statistical Association, 24—33. Alexandria, Virginia: American Statistical Association.

THERNEAU, T. and GRAMBSCH, P. (2000). Modeling Survival Data. Extending the Cox Model. Springer-Verlag.

VAN DER LAAN, M. and HUBBARD, A. (1997). Estimation with Interval Censored Data and Covariates. Lifetime Data Analysis 3, 77—91.

VAN DER LAAN, M. and ROBINS, J. (1998). Locally Efficient Estimation with Current Status Data and Time-Dependent Covariates. JASA, 93, 693—701.

WILLIAMS, R. (1995). Product-Limit Survival Functions with Correlated Survival Times. Lifetime Data Analysis, 1, 171—186.

Article IV

***Pyy-Martikainen, M.**, Approaches for event history analysis based on complex longitudinal survey data. Advances in Statistical Analysis, 97, 297–315, 2013..*

Approaches for event history analysis based on complex longitudinal survey data

Marjo Pyy-Martikainen

Received: 16 June 2011 / Accepted: 19 November 2012 / Published online: 14 December 2012

Abstract A researcher using complex longitudinal survey data for event history analysis has to make several choices that affect the analysis results. These choices include the following: whether a design-based or a model-based approach for the analysis is taken, which subset of data to use and, if a design-based approach is chosen, which weights to use. We discuss different choices and illustrate their effects using longitudinal register data linked at person-level with the Finnish subset of the European Community Household Panel data. The use of register data enables us to construct an event history data set without nonresponse and attrition. Design-based estimates from these data are used as benchmarks against design-based and model-based estimates from subsets of data usually available for a survey data analyst. Our illustration suggests that the often recommended way to use panel data for longitudinal analyses, data from total respondents and weights from the last wave analysed may not be the best way to go. Instead, using all available data and weights from the first survey wave appears to be a safe choice for longitudinal analyses based on multipurpose survey data.

Keywords Combined survey-register data · Complex longitudinal survey data · Event history analysis · Multiple spells

1 Introduction

Event history data consist of information about durations of spells in a state of interest (such as poverty, unemployment, having no children), the outcome of the spell (transition to non-poverty, to employment, birth of first child), as well as a set of covariates

explaining the durations and outcomes. Multiple spells from the same person may be involved. We consider the analysis of event history data based on a complex longitudinal survey. Data based on a complex longitudinal survey may involve stratification, clustering and unequal selection probabilities of units. Nonresponse and attrition bring an additional stage of complexity to longitudinal survey data. How should these complexities of survey data be taken into account at the analysis stage?

Pfeffermann (1993, 1996), Chambers and Skinner (2003), Lohr (1999) and Korn and Graubard (1999) discuss two approaches for the analysis of survey data. The model-based approach is the traditional approach for analytic inference. In the model-based approach, the target parameters of interest are parameters β of a superpopulation model that is assumed to have generated the variable values in the finite population. The standard model-based approach ignores the probability distribution $P(S)$ induced by the sampling design in the estimation procedure. The only source of random variation in the superpopulation model parameter estimator $\hat{\beta}$ is due to the random component in the model. An analyst taking the model-based approach would ignore the survey weights. Information on the sample design variables might be incorporated as covariates of the model. The model-based standard errors of parameter estimates reflect the uncertainty due to the model.

The design-based approach is traditionally used for descriptive inference. In this approach, the only source of random variation in the estimation procedure is the probability distribution induced by the sampling design. The ideas of design-based inference can be applied to analytic inference as well. In this approach, the target parameter of interest is a finite population parameter B that would be obtained from the model estimation procedure if all data values in the finite population were available. An analyst taking the design-based approach would conduct a weighted analysis. The design information would be used to calculate the standard errors of parameter estimates. These standard errors reflect the uncertainty due to making inferences on the basis of only a sample instead of the whole population. Pfeffermann (1993) shows that design-consistent estimators $\hat{B}$ of the finite population parameters B are consistent estimators for the model parameters β assuming that the finite population values are generated by the model. Thus, if the model is correctly specified, both $\hat{B}$ and $\hat{\beta}$ should be close to β.

Kovačević and Roberts (2007) discuss and demonstrate the model-based and the design-based approaches in the context of event history data. They favour the design-based approach for variance estimation reasons. Boudreau (2003) and Boudreau and Lawless (2006) recommend the use of weights when the sample design is non-ignorable. However, they note that the use of weights is sometimes controversial and may even render questionable the results if the weights are not properly defined.

An additional choice the analyst of event history data has to make is which subset of data to use in the analysis. Longitudinal analyses often use only respondents to each wave of interest (Eurostat 2003; Kalton and Brick 2000). Even though the available data until the time of attrition could be used, attriters are often discarded from the analysis. In an analysis using weights this can be motivated by the fact that weights are usually adjusted for nonresponse and attrition. However, the general purpose weights usually included in a survey data set may not fully correct for

nonresponse and attrition that is selective with respect to the particular response variable of interest. The inclusion of the available data from the attriters might in this case help to reduce the bias due to attrition. This is a topic briefly discussed by Roberts and Kovačević (2001). To our knowledge, no empirical research exists on this issue.

We discuss the design-based and the model-based approaches for event history analysis and the choice between them. The choice of the subset of data to be used in a longitudinal analysis is also discussed. By using the Finnish subset of the European Community Household Panel (FI ECHP) data linked at person-level with longitudinal register data, we illustrate how these choices affect the results from event history analysis. Unemployment spells are used as the study variables of interest. Register data are used as a source of information on unemployment spells and covariates. The data involve multiple unemployment spells from the same persons. Register data are available for all sample persons irrespective of their response status. Survey data are used only to obtain information of the occurrence and timing of nonresponse. This way, unemployment is defined in an identical way for both respondents and nonrespondents. On the basis of combined survey-register data, different sets of unemployment spells were constructed. The full information set of spells uses all register information available for the FI ECHP sample persons, without restrictions by nonresponse or attrition. The design-based estimates from the full information set of spells are taken to be the best available estimates of B, the finite population regression parameters, which, if the model postulated is correct, in turn estimate the model parameters β, see Pfeffermann (1993). These estimates are used as benchmarks against model-based and design-based estimates based on subsets of data normally available for the analyst.

The Cox proportional hazards model is used as an example of event history analysis. We model the marginal distributions of the multiple spells and take the correlation among the spells by the same person into account in estimating the variances of the model coefficients. A marginal model-based Cox proportional hazards model was also used in an analysis of measurement errors in the FI ECHP data (Pyy-Martikainen and Rendtel 2009). The correlation among the spells can be alternatively taken into account by allowing cluster-specific random effects. The latter approach is not used in this paper.

The next two sections discuss the model-based and design-based approaches for event history analysis. Section 4 discusses the concept of ignorability of sample design and how to test for the ignorability. Survey data with ignorable sample design can be analysed using model-based procedures. Section 5 discusses the choice of the subset of data used in the analysis. The effects of these choices are illustrated by empirical analyses in Sect. 6. Section 7 concludes by discussing the findings and implications of the study.

2 The model-based approach for event history analysis

2.1 Single-spell analysis

We first consider the case where a single spell is observed for each person. The event times t_i^*, $i = 1, \ldots, N$ in the finite population of N persons are assumed to be

realizations of random variables T_i^* following a superpopulation model with a hazard function $\lambda(t \mid x)$. The T_i^* are usually subject to right-censoring. Therefore, only $T_i = \min(T_i^*, C_i)$ is observed, where C_i is the censoring time related to spell i. If censoring times are independent of the event times (Kalbfleisch and Prentice 1980, pp. 119–121), one does not need to model the distribution of censoring times to draw inferences about the event times. In longitudinal surveys, C_i depend on the length of the follow-up time, on the time at which the spell began and on the time of attrition, see Lawless (2003). The finite population values consist of (t_i, δ_i, x_i), $i = 1, \ldots, N$, where t_i is the realized event or censoring time, the event times following the superpopulation model $\lambda(t \mid x)$. The realized value of the event indicator Δ_i tells whether t_i is an event ($\delta_i = 1$) or a censoring time ($\delta_i = 0$). Hereafter, t_i is called the duration of spell i.

Under the proportional hazards model, the hazard function is specified as a product of two terms: $\lambda(t \mid x) = \lambda_0(t)\exp(x\beta)$. The function $\lambda_0(t)$ is a baseline hazard describing the dependency of the hazard function on the spell duration. The $(1 \times p)$ vector of covariates x affects the hazard via $\exp(x\beta)$, where β is a $(p \times 1)$ vector of regression coefficients, the superpopulation model parameters of interest. A popular proportional hazards model, the Cox (1972) model, uses a partial likelihood function to estimate the parameters β. The partial likelihood function for a sample of n independent observations (t_i, δ_i, x_i) is

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(x_i\beta)}{\sum_{j=1}^{n} I(t_i \leq t_j)\exp(x_j\beta)} \right]^{\delta_i}. \tag{1}$$

This formula assumes there are no ties in the data. Ties arise if two or more events occur at the same time. In this case the temporal ordering of the events is not clear. This can be taken into account by modifying the partial likelihood, see, e.g. Therneau and Grambsch (2000). The maximum partial likelihood estimator $\hat{\beta}$ is the solution to the score equations:

$$U(\beta) = \frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} u_i(\beta) = 0, \tag{2}$$

where

$$u_i(\beta) = \delta_i \left[x_i - \frac{\sum_{j=1}^{n} I(t_i \leq t_j) x_j \exp(x_j\beta)}{\sum_{j=1}^{n} I(t_i \leq t_j)\exp(x_j\beta)} \right].$$

The solution $\hat{\beta}$ is consistent under the superpopulation model distribution. It is also asymptotically normally distributed with mean β, the true superpopulation parameter vector and covariance equal to the inverse of the expected information matrix. In practice, the inverse of the observed information matrix evaluated at $\hat{\beta}$ is used as a covariance estimator:

$$\hat{V}(\hat{\beta}) = \left\{ \frac{\partial U(\beta)}{\partial \beta} \right\}_{\hat{\beta}}^{-1}. \tag{3}$$

2.2 Multiple-spell analysis

Multiple spells arise when more than one spell may be observed for each person. The observations consist of $(t_{ij}, \delta_{ij}, x_{ij})$, where $i = 1, \ldots, N$ identifies the persons and $j = 1, \ldots, n_j$ the spells by the same person. Spells by the same person are likely to be correlated, thus violating the assumption of independency. We take a marginal modeling approach for multiple-spell analysis. Accordingly, the marginal distributions of the multiple spells are modeled, leaving the dependence structure of the spells by the same person unspecified. This permits the use of (2) for estimating the model parameters. In case of cluster-correlated data, the estimator (3) does not provide correct covariance estimates for $\hat{\beta}$. Lin (1994) proposed the following covariance estimator of $\hat{\beta}$ for cluster-correlated data:

$$\hat{V}(\hat{\beta}) = \left\{ \frac{\partial U(\beta)}{\partial \beta} \right\}_{\hat{\beta}}^{-1} B(\hat{\beta}) \left\{ \frac{\partial U(\beta)}{\partial \beta} \right\}_{\hat{\beta}}^{-1}, \tag{4}$$

where $B(\hat{\beta}) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} W_{ij}(\hat{\beta})' W_{ik}(\hat{\beta})$, the term W_{ij} being the score residual related to spell j by person i, see Lin (2000). This estimator takes the correlation of spells by the same person into account but assumes that spells from different persons are independent.

3 The design-based approach for event history analysis

3.1 Single-spell analysis

For event history data obtained from a complex survey, Binder (1992) used a pseudo-likelihood method, see, e.g. Skinner (1989), to estimate the parameters and their variances of a Cox proportional hazards model. For spells of the finite population of N persons, the partial likelihood function is defined as

$$L(B) = \prod_{i=1}^{N} \left[\frac{\exp(x_i B)}{\sum_{j=1}^{N} I(t_i \leq t_j) \exp(x_j B)} \right]^{\delta_i}. \tag{5}$$

The finite population parameter vector B is determined as the solution to the score equations

$$U(B) = \sum_{i=1}^{N} u_i(B) = 0,$$

where

$$u_i(B) = \delta_i \left[x_i - \frac{\sum_{j=1}^{N} I(t_i \leq t_j) x_j \exp(x_j B)}{\sum_{j=1}^{N} I(t_i \leq t_j) \exp(x_j B)} \right].$$

To estimate the population regression coefficient B from a sample of n observations, Binder (1992) proposed the following pseudo-score equations:

$$\hat{U}(\hat{B}) = \sum_{i=1}^{n} w_i \hat{u}_i(\hat{B}) = 0, \tag{6}$$

where

$$\hat{u}_i(\hat{B}) = w_i \delta_i \left[x_i - \frac{\sum_{j=1}^{n} w_j I(t_i \leq t_j) x_j \exp(x_j \hat{B})}{\sum_{j=1}^{n} w_j I(t_i \leq t_j) \exp(x_j \hat{B})} \right],$$

and $w_i, i = 1, \ldots, n$ are the sample weights attached to the sample observations. The estimator $\hat{B}$ that solves Eq. (6) is the pseudo-maximum likelihood estimator of B Binder (1992). Binder (1992) notes that $\hat{B}$ is a design-consistent but slightly biased estimator of B. Binder (1983) proposed a method for deriving the variance of parameter estimators which satisfy equations of the form (6). A design-consistent estimator of the variance of $\hat{B}$ is

$$\hat{V}(\hat{B}) = \left\{ \frac{\partial \hat{U}(B)}{\partial B} \right\}_{\hat{B}}^{-1} \hat{V}(\hat{U}(\hat{B})) \left\{ \frac{\partial \hat{U}(B)}{\partial B} \right\}_{\hat{B}}^{-1}. \tag{7}$$

As $\hat{U}(\hat{B})$ is an estimator of a population total, $\hat{V}(\hat{U}(\hat{B}))$ can be calculated using standard variance estimation formulas. However, to use Eq. (7) to derive a design-consistent variance estimator, $\hat{U}(\hat{B})$ must be expressed as a sum of independent random vectors (Binder 1992). Equation (6) does not satisfy this condition as each $\hat{u}_i(\hat{B})$ involves sums over all n observations. By using Taylor linearization, Binder (1992) derived an alternative expression for $\hat{U}(\hat{B})$ that satisfies the independency condition. Denote this alternative expression by $\hat{U}^*(\hat{B})$. The variance of $\hat{B}$ can be estimated by estimating $\hat{V}(\hat{U}^*(\hat{B}))$ by standard design-based methods and plugging the formula into Eq. (7). For a stratified clustered design where clusters can be assumed independent, $\hat{V}(\hat{U}^*(\hat{B}))$ would be estimated by the between-cluster variance estimator (Kovačević and Roberts 2007; Williams 2000):

$$\hat{V}(\hat{U}^*(\hat{B})) = \sum_{h=1}^{H} \frac{c_h}{c_h - 1} \sum_{c=1}^{c_h} (t_{hc} - \bar{t}_h)(t_{hc} - \bar{t}_h)', \tag{8}$$

where $h = 1, \ldots, H$ identifies the stratum, $c = 1, \ldots, c_h$ identifies the cluster or primary sampling unit (PSU) within the stratum, and $i = 1, \ldots, n_{hc}$ identifies the observation within the cluster, $t_{hc} = \sum_{i=1}^{n_{hc}} w_{hci} \hat{u}_{hci*}$ and $\bar{t}_h = \sum_{c=1}^{c_h} \frac{t_{hc}}{c_h}$.

3.2 Multiple-spell analysis

As noted by Kovačević and Roberts (2007), a longitudinal survey with a multistage design implies clustering of spells within multiple levels. The spells are clustered

within persons. Persons may be clustered, e.g. within households and households in turn within geographical areas. Assuming independency of primary sampling units, i.e. the sampling units for the first stage in a multistage sampling design (geographical areas in our example), an estimate of between-PSU variability captures the variation among the spells within the PSUs regardless of the type of dependency among spells within the same PSU. In this case, the between-cluster variance estimator can be applied without any modifications to multiple-spell data by defining the PSU's as clusters (Kovačević and Roberts 2007).

4 How to choose the approach?

Longitudinal surveys have often complex sample designs involving stratification, clustering and unequal selection probabilities of sampling units. If the selection probabilities depend on values of strata or cluster variables or some size measure that is related to the response variable, the sample distribution of the response variable may be very different from the corresponding finite population distribution. In this case the failure to take the sample design into account may lead to bias in the inference. Korn and Graubard (1999) give examples of studies where a model-based analysis that ignores the design information leads to misleading results. In these studies, the selection probabilities either depend directly on the response variable or on a design variable that is strongly related to the response variable.

Having a complex sample design does not necessarily mean that a design-based approach for analysis needs to be taken, however. By considering sampling from a finite population as a special case of a mechanism generating missing data (the nonsampled units being missing), Rubin (1976) developed conditions under which a sample design can be ignored in the analysis. In the following, we summarize the ignorability conditions along the lines of Pfeffermann (1993, 1996). Denote by Z the $(k \times N)$ matrix of values of the design variables that are assumed to be known for each unit $i = 1, \ldots, N$ in the population and by Y the $(p \times N)$ matrix of response variable values for the population units. The dimension of p depends on the number of survey waves and on the amount of information collected at each wave. The matrix Y is partitioned into $[Y_S, Y_{\bar{S}}]$, where $Y_s = \{Y_i, i \in S\}$ are the response variable values for the sampled units and $Y_{\bar{s}} = \{Y_i, i \notin S\}$ the response variable values for the nonsampled units. $I = [I_i, \ldots, I_N]$ is the vector of sample inclusion indicator variables so that $I_i = 1$ if $i \in S$ and $I_i = 0$ otherwise. The probability of drawing a sample depends on the design variables and possibly also on the response variables: $P(S) = P(I \mid Y, Z; \phi)$.

The parameters of interest in the analysis (the model coefficients from a Cox proportional hazards model in our example) are the parameters θ of a superpopulation model $f(Y; \theta)$ that is assumed to have generated the values Y. The model may include covariates x but they are omitted from the notation for convenience. Assuming that the design variables are known for each unit in the population, the observed data consist of (Y_s, I, Z). The joint distribution of Y_s and I, given Z, is

$$\int P(I \mid Y_s, Y_{\bar{s}}, Z; \phi) f(Y_s, Y_{\bar{s}} \mid Z; \theta_1) \mathrm{d}Y_{\bar{s}}. \tag{9}$$

Now if the selection to the sample depends only on the values of the design variables, i.e. if

$$P(I \mid Y, Z; \phi) = P(I \mid Z; \phi), \tag{10}$$

then the first term in the integrand in (9) can be ignored and inference about Y can be based on

$$f(Y_s \mid Z; \theta_1) = \int f(Y_s, Y_{\bar{s}} \mid Z; \theta_1) \mathrm{d}Y_{\bar{s}}. \tag{11}$$

When inference based on (11) is equivalent to inference based on (9), the sample design is ignorable. Otherwise, it is nonignorable. As noted by Pfeffermann (1993), ignorability of the design refers to the information provided by the sample design beyond what is already provided by the design variables.

Samples with ignorable designs can be analysed using standard model-based procedures ignoring weights. This requires, however, that the sample design does not depend on the response variables, all the values of the design variables are known for each population unit and inference is based on the conditional distribution of Y_s given Z. In practice, the data analyst rarely has access to all the relevant design information.

Sugden and Smith (1984) studied the conditions under which sample designs that satisfy condition (10) are ignorable, given only partial information on the design. Denoting by $D_s(Z)$ the available design information, the key condition for the ignorability under partial information is $P(I \mid Z) = P(I \mid D_s(Z))$. Thus, incorporating only partial design information may be sufficient for a model-based analysis to produce unbiased estimates of the parameters of interest.

Pfeffermann and Sverchkov (2009) contains a review on different tests for the ignorability of sample design proposed in literature. By using the results of Hausman (1978), Pfeffermann (1993) proposed the following test statistic for likelihood-based inference on superpopulation model parameters β:

$$\lambda = (\hat{B} - \hat{\beta})' \left(\hat{V}(\hat{B}) - \hat{V}(\hat{\beta}) \right)^{-1} (\hat{B} - \hat{\beta}), \tag{12}$$

where $\hat{B}$ is the pseudo-maximum likelihood estimator of B and $\hat{\beta}$ is the maximum likelihood estimator of β, conditioning on the available design information, $\hat{V}(\hat{B})$ and $\hat{V}(\hat{\beta})$ being estimated by design-based methods. Under the ignorability assumption, the asymptotic distribution of λ is χ^2_p, p being the dimension of the vector β.

Pfeffermann (1993) discusses the rationale of λ. The design-based estimator $\hat{B}$ is design-consistent (i.e. consistent under the randomization distribution $P(S)$ induced by the sampling design) for the finite population parameter B. The finite population parameter B is not affected by the sample design and is therefore model-consistent (i.e. consistent under the superpopulation model distribution) for β.[1] Consequently,

[1] The concept of design-consistency requires that both the sample size n and the population size N are assumed to increase to infinity. The concept of model-consistency requires the latter assumption. See Särndal et al. (1992) for a discussion on consistency in the model-based and design-based frameworks.

$\hat{B}$ is consistent for β even when the sample design is not ignorable. For a more formal discussion on the consistency of $\hat{B}$ as an estimator of β, see Pfeffermann (1993) and Roberts and Kovačević (2003). When the sample design is ignorable, the model-based estimator $\hat{\beta}$ is also consistent for β. If the sampling design is not ignorable, $\hat{\beta}$ is no longer consistent for β and the two estimators tend to differ from each other.

A large value of λ can also indicate misspecification of the superpopulation model. Korn and Graubard (1999) give an example where the aim is to model the association between gestational age and birthweight. The model is misspecified for the population as a linear regression while the true relationship between the variables is curvilinear. Due to oversampling of babies with low birthweight, the sample and population distributions of the independent variable birthweight differ. As a consequence, the weighted and unweighted regressions are attempting to fit a straight line to different parts of the curvilinear relationship. Under a misspecified population model, the model-based parameter β and its estimator may no longer have a meaningful interpretation and $\hat{\beta}$ will change depending upon the sample design while $\hat{B}$ is estimating a well-defined quantity: the finite population regression coefficient. In this sense, design-based analysis can be claimed to be robust with respect to misspecification of the population model.

The test statistic λ can thus be used to identify misspecification of the population model or omission of important design variables which, when included, make the design ignorable. However, as Chambers and Skinner (2003) note, the need to condition on design variables means changing the focus from the model of interest $f(Y;\theta)$ to the conditional model $f(Y \mid Z;\theta_1)$. This means that the sample design is driving the specification of the model and target parameters of interest. Chambers and Skinner (2003) argue that it is more appropriate to define the target parameters first on the basis of the scientific questions of interest before considering how the sample design influences inference about the target parameters. This is a strong argument in favor of the design-based approach. Moreover, a design-based approach provides protection against nonignorable sample designs and against population model misspecification—although the latter in a rather limited sense. However, as discussed by Pfeffermann (1993, 1996), the benefits of a design-based approach come with a price: if the model postulated is at least approximately correct, i.e. if the population values can be considered as realizations from the model, the use of design-based estimators is less efficient compared with model-based estimators. The smaller the sample size and the larger the variability of the weights, the larger is the loss in efficiency. Binder (1992) showed that estimators defined by Eq. (6) are asymptotically normally distributed. This permits the use of normal-based confidence intervals and test statistics for the model parameters. Asymptotic design-based arguments assume a sequence of finite populations and samples, the size of both increasing to infinity. In practice, the sample size in the domain of interest may be small. The small sample properties of design-based estimators of model parameters are usually unknown. Finally, as discussed by Pfeffermann (1996), the protection against misspecification offered by design-based estimators is limited to inferences concerning populations with a similar structure to that of the population under study.

Both the design-based and model-based approaches have their pros and cons. Fortunately, the model-based and design-based estimates of model parameters may not

be dramatically different in practice. Korn and Graubard (1999) note that for the design-based and model-based estimates of association to differ greatly, the type of association must be very misspecified, an omitted design variable must have a strong interaction with the independent variables, or the inclusion probabilities must depend directly upon the response variable.

Korn and Graubard (1999) recommend comparing design-based and model-based estimates of model parameters. The test statistics (12) can be used for likelihood-based inferences such as parameter estimates from event history models. If the difference between the estimates is not statistically significant, the sample design is ignorable and a model-based approach can be taken. If the differences remain statistically significant even after conditioning on available design information and attempts to postulate a model that correctly represents the population, a design-based approach should be taken.

5 Which subset of data and which weights to use?

Longitudinal analyses for the time interval $[1, t]$ are usually concerned only with persons who exist in the population throughout this interval, excluding both entrants and leavers (Kalton and Brick 2000). This population is defined by the intersection of the cross-sectional populations of the time interval $[1, t]$: $P_1 \cap \ldots \cap P_t$ and is represented by data involving respondents to all waves of interest (Smith et al. 2009). The weighting of these data involves adjusting the design weights for nonresponse at wave 1 and attrition at all other waves (assuming no temporary nonresponse patterns are allowed). Weights from wave t, possibly calibrated to conform to the auxiliary population totals from the population $P_1 \cap \ldots \cap P_t$ would be used in design-based analyses. Such longitudinal population totals may be difficult to construct in practice (Smith et al. 2009). If leavers from the survey population are also included, the relevant population is P_1, the population at wave 1 (Smith et al. 2009). This population is represented by respondents to all waves of interest as well as persons with incomplete data due to exit from the survey population. Weights from wave t, adjusted for nonresponse and attrition, and calibrated to conform to the P_1 population totals, would be used in a design-based analysis.

Restricting the analysis to respondents to all waves of interest (possibly including also leavers from the survey population) means that all persons with incomplete data for the set of waves of interest are discarded. This may be a sizeable part of the data. If nonresponse and attrition are driven by processes related to the response variable and the weighting procedure fails to adjust for that, the estimates based on total respondents only may be badly biased. As Smith et al. (2009) point out, the main objective of household panel surveys is to provide multitopic data for a broad set of purposes. Therefore, the weights attached to the data must also serve multiple purposes and may not fully adjust for nonresponse and attrition selective with respect to the response variable of a particular analysis. In order to reduce the bias due to attrition, it might be useful to include the available data from attriters in the analysis (Roberts and Kovačević 2001). In an event history analysis this would involve the inclusion of spells by attriters until the time of attrition. Event history methods are able to make

full use of the incomplete information contained in the spells censored by attrition. It is not clear, however, which weights should be used in an analysis including attriters. The wave t weights are not available for the attriters. The wave 1 weights do not adjust for attrition. Roberts and Kovačević (2001) suggest using the wave 1 weights or weights from the starting or ending time of the spell. The Survey of Income and Program Participation Users' Guide (Westat 2001) contains a special section on spell analysis which recommends the use of the weight from the starting time of the spell.

6 Illustration with multiple unemployment spell data

The effects of different choices discussed in earlier sections are illustrated by an analysis of multiple unemployment spell data. We use the Finnish subset of the European Community Household Panel (FI ECHP) data combined at person-level with longitudinal register data. For a description of the FI ECHP data, see Pyy-Martikainen et al. (2004).

In the ECHP, information on unemployment spells is obtained from answers to a month-by-month main activity state calendar relating to the year preceding the interview. For the following illustration, information on unemployment spells was, however, taken solely from administrative registers. This was done to have directly comparable spell information for all the FI ECHP sample persons, both respondents and nonrespondents. The survey data were used only to obtain information on the occurrence and timing of nonresponse. The same approach was used by Pyy-Martikainen and Rendtel (2008) in a study of survey nonresponse and attrition.

Three different sets of unemployment spells were constructed on the basis of combined survey-register data: the *full information* set of spells uses all register information available for the FI ECHP sample persons, without restrictions by nonresponse or attrition. Sample persons are defined as all members of the initial sample of households. The *observed information* set of spells is a subset of the full information set of spells, obtained by excluding spells unobserved by nonresponse and attrition and the remaining length of the spells censored due to attrition. The observed information set of spells corresponds to the data normally available for a survey data analyst. The *total respondents* set of spells is a subset of the observed information set of spells, obtained by excluding all spells by attriters. This subset of data is often used for longitudinal analyses.

Estimates based on the full information set of spells were taken as benchmark estimates $\hat{B}_{\text{bm}}$ against which estimates based on the observed information and the total respondents sets of spells were evaluated. These benchmark estimates are free from the effects of nonresponse and attrition. The benchmark estimates were taken to be the best available estimates of B, the finite population regression parameters, which, if the model postulated is correct, in turn estimate the model parameters β (Pfeffermann 1993). The population of interest is P_1.

6.1 The data

The construction of combined survey-register data is described in Pyy-Martikainen and Rendtel (2008). Unemployment spells beginning during 1 Jan 1995–31 Dec 1999, and

a set of covariates, all retrieved from administrative registers, were linked by personal identification numbers for all FI ECHP sample persons aged 16 or over at the beginning of 1996. The observation period corresponds to the first five waves of the FI ECHP data. There were 11,641 sample persons of whom 4,364 responded at each wave, 3,210 attrited during waves 2–5, and 3,146 did not respond in any wave. For simplicity, the 921 temporary drop-outs were excluded from the analysis. Of the remaining 10,720 sample persons, 2,930 persons had altogether 10,734 unemployment spells during the observation period. These spells form the full information set of spells. The observed information set of spells has 6,496 spells and the total respondents set of spells has 4,066 spells. In the following, we discuss briefly the target population, the sampling design and the construction of weights in the FI ECHP.

The target population of the FI ECHP consists of members of private households permanently resident in Finland. Persons living abroad or in institutions and persons without a permanent place of residence do not belong to the target population. As most household panel surveys, the FI ECHP aims to remain cross-sectionally representative of the household population over time. This is achieved by the appropriate follow-up rules of the sample persons, see Pyy-Martikainen et al. (2004).

The FI ECHP sample is a two-phase stratified network sample. The population information system of the Population Register Centre was used as the frame. The frame population consists of persons permanently living in Finland aged 15 and over. In the first phase, a master sample of target persons was drawn from the frame. Dwelling units were constructed by adding to the master sample all persons sharing the same domicile code as the target persons. The master sample was merged with the most recent taxation records from which information was used to form a socio-economic group for each target person. The socio-economic groups were formed by dividing wage earners, entrepreneurs, farmers, pensioners and other non-active persons into subgroups defined by aggregate taxable income. The second phase consisted of drawing the final sample from the master sample using stratification according to socio-economic group; farmers, entrepreneurs and high-income wage earners having the largest sampling fractions. The selection probabilities depend on the size of the dwelling unit and its socio-economic group. It is plausible that socio-economic group is related to both the length and outcome of an unemployment spell. However, it is not clear in advance whether and how this affects an unemployment spell analysis.

Three types of weights are provided in the ECHP User Data Base: base weights to be used in longitudinal analyses, and personal weights and household weights to be used in cross-sectional analyses (Eurostat 2003). Base weights are defined for the sample persons only and they represent the basis for adjusting weights from one wave to the next. Personal weights and household weights are derived from the base weights by weight sharing. The weighting of the first wave of the FI ECHP was conducted at Statistics Finland (Pyy-Martikainen et al. 2004). For the subsequent waves, the common weighting procedure developed at Eurostat was used (Eurostat 2000, 2002a,b; Peracchi 2002). For longitudinal analyses involving waves 1 to t, the wave t base weights are recommended in Eurostat (2003). The wave t weights are, however, calibrated using population totals from the wave t, which does not conform with the population of interest, P_1.

6.2 Results

Our empirical analysis consists of model-based and design-based estimates of Cox proportional hazards models for the total respondents and observed information data sets. Cause-specific analyses were conducted, the outcome of interest being the transition from unemployment to employment. A set of covariates similar to those used in econometric analyses of unemployment duration was used, see, e.g. Meyer (1990); Carling et al. (1996); Abbring et al. (2005). The construction of the covariates is described in Pyy-Martikainen and Rendtel (2008).

The model-based analyses were conducted both without any design information and with design information related to the household selection probabilities. This information consists of socio-economic strata and the number of dwelling unit members aged at least 15.[2] The design information used in the calibration of weights was ignored to restrict the number of parameters to be estimated. The inclusion of this partial design information should, if not eliminate, at least diminish the bias in parameter estimators (Pfeffermann 1993; Sugden and Smith 1984).

The design-based total respondent analyses were weighted by the last wave base weights. This is an often recommended way to use panel data for longitudinal analyses, e.g. Kalton and Brick (2000), Eurostat (2003). Leavers from the survey population were excluded as they do not have the appropriate weigths in the FI ECHP.[3]

The design-based observed information estimates were calculated using both first wave base weights and base weights from the starting wave of the spell, as suggested by Roberts and Kovačević (2001). The ignorability of the sample design was tested by Hausman tests. It was expected that the inclusion of the design variables would bring the model-based estimates of coefficients closer to the design-based estimates, this being manifested by a smaller value of the Hausman test statistics.

Design-based estimates from the full information set of spells were used as benchmark estimates. Weights for this set of spells were generated by calibrating the wave 1 design weights of all sampled households (including nonrespondents) to the same auxiliary population totals as the wave 1 base weights, see Pyy-Martikainen et al. (2004). To facilitate comparison of estimates from the different models with benchmark estimates, Mahalanobis type distances of estimated regression coefficient vectors were computed using the following formula:

$$d(\hat{\beta}, \hat{B}_{\text{bm}}) = \sqrt{(\hat{\beta} - \hat{B}_{\text{bm}})' \hat{V}(\hat{B}_{\text{bm}})^{-1} (\hat{\beta} - \hat{B}_{\text{bm}})} \tag{13}$$

where $\hat{\beta}$ is the vector of coefficient estimates from the model being evaluated, $\hat{B}_{\text{bm}}$ is the vector of benchmark estimates, and $\hat{V}(\hat{B}_{\text{bm}})^{-1}$ is the inverse of the estimated covariance matrix of $\hat{B}_{\text{bm}}$.

[2] An alternative way to take the stratum information into account is to use stratified Cox proportional hazards models where the baseline hazard function is estimated separately for each sample stratum, see Boudreau (2003).

[3] This implied the exclusion of only 44 spells.

Table 1 Analysis of multiple unemployment spells

	Design based	Model based		Model based		Design based	Design based	
	Full information	Total respondents		Observed information		Total respondents	Observed information	
						lw	fw	sw
Covariates	est. (se)	est. (se)	est. (se)	est.(se)	est. (se)	est. (se)	est. (se)	est. (se)
Covariates of main interest								
Female	**0.225** (0.085)	0.193 (0.120)	0.195 (0.119)	0.124 (0.088)	0.126 (0.087)	0.098 (0.146)	0.163 (0.115)	0.143 (0.113)
Age	**0.075** (0.025)	**0.095** (0.035)	**0.094** (0.038)	**0.088** (0.026)	**0.089** (0.027)	**0.096** (0.039)	**0.095** (0.032)	**0.109** (0.032)
Age squared	**−0.001** (0.000)	**−0.001** (0.001)	**−0.001** (0.001)	**−0.001** (0.000)	**−0.001** (0.000)	**−0.001** (0.001)	**−0.001** (0.000)	**−0.002** (0.000)
Upper secondary educ.	0.044 (0.116)	0.019 (0.154)	0.051 (0.154)	0.058 (0.116)	0.073 (0.116)	0.090 (0.187)	0.024 (0.159)	0.045 (0.152)
Higher education	**0.531** (0.148)	*0.327* (0.190)	*0.405* (0.189)	**0.411** (0.148)	**0.450** (0.148)	**0.473** (0.213)	**0.564** (0.186)	**0.558** (0.177)
Proportion of ue time	**0.006** (0.002)	**0.010** (0.003)	**0.009** (0.003)	**0.008** (0.002)	**0.008** (0.002)	**0.011** (0.003)	**0.009** (0.002)	**0.009** (0.002)
Semi urban municipality	0.194 (0.136)	0.185 (0.197)	0.171 (0.194)	0.149 (0.149)	0.139 (0.152)	0.081 (0.204)	0.176 (0.188)	0.139 (0.183)
Rural municipality	**0.216** (0.105)	0.033 (0.151)	0.005 (0.161)	0.002 (0.109)	−0.018 (0.116)	−0.144 (0.167)	0.049 (0.141)	−0.013 (0.140)
Southern Finland	0.276 (0.204)	0.116 (0.320)	0.109 (0.320)	0.258 (0.258)	0.273 (0.262)	−0.079 (0.339)	0.191 (0.284)	0.159 (0.276)
Eastern Finland	0.083 (0.221)	0.137 (0.358)	0.149 (0.349)	0.207 (0.275)	0.233(0.278)	0.002 (0.364)	0.076 (0.302)	0.081 (0.289)
Central Finland	0.206 (0.205)	0.002 (0.316)	−0.049 (0.330)	0.147 (0.256)	0.137 (0.267)	−0.040 (0.326)	0.194 (0.282)	0.126 (0.270)
Northern Finland	*0.414* (0.218)	0.273 (0.333)	0.204 (0.347)	*0.489* (0.274)	*0.473* (0.282)	0.391 (0.345)	it 0.512 (0.298)	*0.503* (0.296)
Earnings-rel. ue benefit	**0.348** (0.105)	*0.223* (0.136)	*0.206* (0.132)	*0.191* (0.102)	*0.167* (0.101)	0.228 (0.158)	0.220 (0.138)	0.202 (0.136)
Year 1996	0.036 (0.078)	−0.143 (0.093)	−0.136 (0.095)	−0.017 (0.077)	−0.017 (0.078)	−0.164 (0.111)	0.003 (0.092)	0.032 (0.092)
Year 1997	0.041 (0.103)	−0.174 (0.119)	−0.155 (0.121)	−0.052 (0.091)	−0.043 (0.093)	*−0.295* (0.156)	−0.050 (0.122)	−0.041 (0.123)
Year 1998	**0.273** (0.127)	0.007 (0.154)	0.024 (0.152)	0.171 (0.115)	0.183 (0.116)	−0.146 (0.195)	0.206 (0.159)	0.186 (0.159)
Year 1999	**0.310** (0.141)	0.026 (0.167)	0.037 (0.164)	*0.251* (0.128)	**0.256** (0.130)	−0.230 (0.216)	0.153 (0.174)	0.009 (0.185)
Distance from Benchmark estimates		4.54	4.85	3.88	4.14	5.85	3.31	4.39

Table 1 continued

	Design based	Model based		Model based		Design based	Design based	
	Full information	Total respondents		Observed information		Total respondents	Observed information	
						lw	fw	sw
Covariates	est. (se)	est. (se)	est. (se)	est. (se)	est. (se)	est. (se)	est. (se)	est. (se)
Covariates related to sample design								
#(hh members 15+)			0.034 (0.053)		0.030 (0.037)			
Wage and salary earners 2			−0.284 (0.228)		−0.225 (0.152)			
Wage and salary earners 3			**−0.559** (0.257)		**−0.538** (0.203)			
Wage and salary earners 4			−0.413 (0.371)		−0.193 (0.366)			
Entrepreneurs 1			−0.104 (0.315)		−0.266 (0.213)			
Entrepreneurs 2			*−0.567* (0.297)		**−0.522** (0.238)			
Farmers 1			−0.195 (0.226)		−0.168 (0.159)			
Farmers 2			−0.412 (0.292)		**−0.644** (0.250)			
Pensioners 1			−0.484 (0.339)		*−0.421* (0.251)			
Pensioners 2			**−1.662** (0.576)		**−1.813** (0.486)			
Other 1			−0.165 (0.213)		−0.204 (0.167)			
Other 2			−0.112 (0.220)		−0.1114 (0.163)			
No tax record			−0.227 (0.345)		−0.165 (0.275)			

Full information: 10,734 spells, 5,111 events
Obs information: 6,496 spells, 3,212 events
Total respondents: 4,066 spells, 2,005 events
Estimates significant at 5 % (10 %) risk level are displayed in boldface (italics)
Standard errors adjusted for clustering of spells within persons. Results from Cox proportional hazard models. *lw* last wave weights, *fw* first wave weights, *sw* weights from starting wave of the panel, *ue* unemployment

Table 2 Hausman test statistics

Model-based	Total respondents LW	Design-based	
		Observed information FW	Observed information SW
Total respondents, no design variables	14.088 (0.661)		
Total respondents, design variables	17.000 (0.454)		
Observed information, no design variables		15.692 (0.546)	17.504 (0.421)
Observed information, design variables		18.287 (0.371)	20.096 (0.269)

p values in parentheses
LW last wave weights, *FW* first wave weights, *SW* weights from the starting wave of the panel

The design-based estimates were calculated using Sudaan's Survival procedure. The sampling design was approximated by Sudaan's design = WR option which implies with replacement sampling of clusters within strata. The clustering was taken into account at the person level. Even though persons are clustered in households, the clustering at household level was not found to be important among persons experiencing unemployment spells. SAS Phreg procedure was used to calculate the model-based estimates. Robust standard errors were calculated by the method of Lin (1994). The Efron (1977) approximation of partial likelihood was used to handle tied event times in all analyses.

The estimation results are shown in Table 1 and the results from the Hausman tests in Table 2. The Hausman tests were calculated using the set of covariates of main interest and the data sets normally available to the analyst; the observed information and the total respondents data sets. None of the tests indicate nonignorability of sample design and thus, model-based analysis would be valid in this case. Contrary to expectations, the inclusion of design variables increases the value of the Hausman test statistics. Looking at the distance measures in Table 1 it is obvious that in our example, the observed information estimates are superior to the total respondents estimates. The design-based observed information estimates weighted by the first wave weights are closest to the benchmark estimates while the design-based estimates based on total respondents set of data and weighted by the last wave weights are furthest from the benchmark estimates. Neither the last wave weights nor the weights from the starting wave of the spell are helpful in correcting bias due to attrition in our example: estimates weighted by these weights are further from the benchmark estimates than the corresponding unweighted estimates. Including design variables in the analysis brings the estimated effect of having a higher education closer to the benchmark estimate. As regards other covariate effects, the inclusion of design variables does not make much difference. Contrary to expectations, the overall effect of the design variables is to move the estimates somewhat further from the benchmark estimates (diagnostic checks did not indicate a multicollinearity problem). Given their small effect on the parameter estimates of interest, the design variables are surprisingly powerful predictors of the

exit rate from unemployment to employment. Compared with the reference group of low-income wage earners, high-income wage earners, entrepreneurs, farmers, and pensioners have a lower exit rate.

7 Conclusions

We discussed different choices that a researcher using complex longitudinal survey data for event history analysis has to make. These choices include the following: whether a model-based or a design-based approach for the analysis is taken, which subset of data to use and, if a design-based approach is taken, which set of weights to use.

We illustrated the implication of these choices by using longitudinal register data linked at person-level with FI ECHP survey data. Unemployment spells retrieved from administrative registers were used as study variables of interest and the marginal Cox proportional hazards model as an example of event history analysis. The survey data were used only to obtain information of the occurrence and timing of nonresponse. On the basis of combined survey-register data, different sets of unemployment spells were constructed. The full information set of spells uses all register information available for the FI ECHP sample persons, without restrictions by nonresponse or attrition. Design-based estimates from this set of spells were used as benchmarks against which the model-based and design-based estimates from the other sets of spells subject to nonresponse and attrition were compared.

Measured by a Mahalanobis type of distance measure, the observed information estimates were closer to the benchmark estimates than the total respondents estimates. Thus using all the available data in the analysis, including the spells by attriters until the time of attrition, helped to reduce bias due to attrition in our illustration. Comparison of the model-based and the design-based estimates revealed that the weighting correction for attrition is not very helpful in our particular example. The weights from the last wave analysed and the weights from the starting wave of the spell produced estimates that were further from the benchmark than the corresponding unweighted estimates. The design-based estimates with total respondents data and the last wave weights were furthest from the benchmark estimates. This is remarkable as this is the often recommended way to use panel data for longitudinal analyses (Kalton and Brick 2000; Eurostat 2003). The design-based estimates from the observed information data and weighted by the first-wave weights were closest to the benchmark estimates, thus suggesting in this particular case that this may be the best way to go. However, the Hausman tests indicated ignorability of sample design and a model-based analysis would be valid in this case. Contrary to expectations, the inclusion of design variables moved estimates further from the benchmark estimates.

The general purpose weights of a survey gathering data on a wide range of topics may not be able to correct effectively enough for bias due to attrition in all analyses. Including available data from attriters in the analysis may then be an effective means to reduce bias due to attrition. The Hausman test may be used to decide whether a design-based or a model-based approach for the analysis should be taken. This test is based on the assumption that the design-based estimates are closer to the true

superpopulation parameters than the model-based estimates. In cases where this is not true the interpretation of the Hausman test becomes problematic. If a design-based approach is chosen, there is the additional choice of the appropriate weights to be used in a longitudinal analysis. Our illustration suggests that the first wave weights may be a safe choice in multipurpose surveys. Our results indicate that the choice of the data set and the weights to be used in the analysis are important stages in event history analysis based on survey data.

References

Abbring, J., et al.: The effect of unemployment insurance sanctions on the transition rate from unemployment to employment. Econ. J. **115**(505), 602–630 (2005)

Binder, D.: On the variances of asymptotically normal estimators from complex surveys. Intern. Stat. Rev. **51**, 279–292 (1983)

Binder, D.: Fitting Cox's proportional hazards models from survey data. Biometrika **79**(1), 139–147 (1992)

Boudreau, C.: Duration data analysis in longitudinal surveys. PhD thesis, University of Waterloo, Canada (2003)

Boudreau, C., Lawless, J.: Survival analysis based on the proportional hazards model and survey data. Can. J. Stat. **34**(2), 203–216 (2006)

Carling, K., et al.: Unemployment duration, unemployment benefits and labor market programs in sweden. J. Public Econ. **59**(3), 313–334 (1996)

Chambers, R., Skinner, C. (eds.): Analysis of Survey Data, chap 1. Wiley, New York (2003)

Cox, D.: Regression models and life tables. J. Royal Stat. Soc. B **34**(2), 187–220 (1972)

Efron, B.: The efficiency of Cox's likelihood function for censored data. J. Am. Stat. Assoc. **72**(359), 557–565 (1977)

Eurostat: Construction of weights in the ECHP. Tech. rep., Doc. Pan 165/00 (2000)

Eurostat: ECHP weighting—overview of procedures applied for the issue of the waves 1994–1999 udb. Tech. rep., Doc. Pan 189/02 (2002a)

Eurostat: ECHP weighting -review after wave 5. Tech. rep., Doc. Pan 183/02 (2002b)

Eurostat: ECHP UDB manual: European Community Household Panel Longitudinal User's Database, waves 1 to 7, survey years 1994 to 2000. Doc. Pan 168/2003-6 (2003)

Hausman, J.: Specification tests in econometrics. Econometrica **46**(6), 1251–1271 (1978)

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

Kalton, G., Brick, M.: Weighting in household panel surveys. In: Rose, D. (ed.) Researching Social and Economic Change. The Uses of Household Panel Studies, Routledge (2000)

Korn, E., Graubard, B.: Analysis of Health Surveys. Wiley, New York (1999)

Kovačević, M., Roberts, G.: Modelling durations of multiple spells from longitudinal survey data. Survey Methodol. **33**(1), 13–22 (2007)

Lawless, J.: Event history analysis and longitudinal surveys. In: Chambers, R., Skinner, C. (eds.) Analysis of Survey Data, pp. 221–243. Wiley, New York (2003)

Lin, D.: Cox regression analysis of multivariate failure time data: the marginal approach. Stat. Med. **13**, 2233–2247 (1994)

Lin, D.: On fitting Cox's proportional hazards models to survey data. Biometrika **87**(1), 37–47 (2000)

Lohr, S.: Sampling: Design and Analysis. Duxbury Press, USA (1999)

Meyer, B.: Unemployment insurance and unemployment spells. Econometrica **58**(4), 757–782 (1990)

Peracchi, F.: The European Community Household Panel: a review. Empir. Econ. **27**(1), 63–90 (2002)

Pfeffermann, D.: The role of sampling weights when modeling survey data. Intern. Stat. Rev. **61**(2), 317–337 (1993)

Pfeffermann, D.: The use of sampling weights for survey data analysis. Stat. Methods Med. Res. **5**, 239–261 (1996)

Pfeffermann, D., Sverchkov, M.: Inference under informative sampling. In:Pfeffermann D., Rao, C. (eds.) Sample Surveys: Inference and Analysis, vol. 29B, pp. 455–487. Elsevier, Amsterdam (2009)

Pyy-Martikainen, M., Rendtel, U.:Assessing the impact of initial nonresponse and attrition in the analysis of unemployment duration with panel surveys.Adv. Stat. Anal. 92:297–318. doi:10.1007/s10182-008-0069-y.
Pyy-Martikainen, M., Rendtel, U.: Measurement errors in retrospective reports of event histories. A validation study with Finnish register data. Survey Res. Methods **3**(3), 139–155 (2009)
Pyy-Martikainen, M., et al.: The ECHP study in Finland. Quality report. Living conditions 2004:1, Statistics Finland (2004)
Roberts, G., Kovačević, M.: New research problems in analysis of duration data arising from complexities of longitudinal surveys. In: Proceedings of the Survey Methods Section, SSC Annual Meeting, pp 111–116 (2001)
Roberts, G., Kovačević, M.: Design-based and model-based methods for estimating model parameters. In: Chambers, R., Skinner, C. (eds.) Analysis of Survey Data, pp. 29–48. Wiley, New York (2003)
Rubin, D.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)
Särndal, C.E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer, Berlin (1992)
Skinner, C.: Domain means, regression and multivariate analysis. In: Skinner, C., Holt, D., Smith, F. (eds.) Analysis of Complex Surveys, pp. 80–84. Wiley, New York (1989)
Smith, P., et al.: Sample design for longitudinal surveys. In: Lynn, P. (ed.) Methodology of Longitudinal Surveys, chap 2. Wiley, New york (2009)
Sugden, R., Smith, T.: Ignorable and informative designs in survey sampling inference. Biometrika **71**(3), 495–506 (1984)
Therneau, T., Grambsch, P.: Modeling Survival Data. Springer, Berlin (2000)
Westat: Survey of income and program participation users' guide. Tech.rep. (2001)
Williams, R.: A note on robust variance estimation for cluster-correlated data. Biometrics **56**, 645–646 (2000)

Tutkimuksia-sarja
Research Reports Series

Tilastokeskus on julkaissut Tutkimuksia v. 1966 alkaen, v. 1990 lähtien ovat ilmestyneet seuraavat:

164. **Henry Takala,** Kunnat ja kuntainliitot kansantalouden tilinpidossa. Tammikuu 1990. 60 s.

165. **Jarmo Hyrkkö,** Palkansaajien ansiotasoindeksi 1985=100. Tammikuu 1990. 66 s.

166. **Pekka Rytkönen,** Siivouspalvelu, ympäristöhuolto ja pesulapalvelu 1980-luvulla. Tammikuu 1990. 70 s.

167. **Jukka Muukkonen,** Luonnonvaratilinpito kestävän kehityksen kuvaajana. 1990. 119 s.

168. **Juha-Pekka Ollila,** Tieliikenteen tavarankuljetus 1980-luvulla. Helmikuu 1990. 45 s.

169. **Tuovi Allén – Seppo Laaksonen – Päivi Keinänen – Seija Ilmakunnas,** Palkkaa työstä ja sukupuolesta. Huhtikuu 1990. 90 s.

170. **Ari Tyrkkö,** Asuinolotiedot väestölaskennassa ja kotitaloustiedustelussa. Huhtikuu 1990. 63 s.

171. **Hannu Isoaho – Osmo Kivinen – Risto Rinne,** Nuorten koulutus ja kotitausta. Toukokuu 1990. 115 s.

171b. **Hannu Isoaho – Osmo Kivinen – Risto Rinne,** *Education and the family* background of the young in Finland. 1990. 115 pp.

172. **Tapani Valkonen – Tuija Martelin – Arja Rimpelä,** Eriarvoisuus kuoleman edessä. *Sosioekonomiset kuolleisuuserot* Suomessa 1971–85. Kesäkuu 1990. 145 s.

173. **Jukka Muukkonen,** Sustainable development and natural resource accounting. August 1990. 96 pp.

174. **Iiris Niemi – Hannu Pääkkönen,** Time use changes in Finland in the 1980s. August 1990. 118 pp.

175. **Väinö Kannisto,** Mortality of the elderly in late 19th and early 20th century Finland. August 1990. 50 pp.

176. **Tapani Valkonen – Tuija Martelin – Arja Rimpelä,** *Socio-economic* mortality differences in Finland 1971–85. December 1990. 108 pp.

177. **Jaana Lähteenmaa – Lasse Siurala,** Nuoret ja muutos. Tammikuu 1991. 211 s.

178. **Tuomo Martikainen – Risto Yrjönen,** Vaalit, puolueet ja yhteiskunnan muutos. Maaliskuu 1991. 120 s.

179. **Seppo Laaksonen,** Comparative Adjustments for Missingness in Short-term Panels. April 1991. 74 pp.

180. **Ágnes Babarczy – István Harcsa – Hannu Pääkkönen,** Time use trends in Finland and in Hungary. April 1991. 72 pp.

181. **Timo Matala,** Asumisen tuki 1988. Kesäkuu 1991. 64 s.

182. **Iiris Niemi – Parsla Eglite – Algimantas Mitrikas – V.D. Patrushev – Hannu Pääkkönen,** Time Use in Finland, Latvia, Lithuania and Russia. July 1991. 80 pp.

183. **Iiris Niemi – Hannu Pääkkönen,** Vuotuinen ajankäyttö. Joulukuu 1992. 83 s.

183b. **Iiris Niemi – Hannu Pääkkönen – Veli Rajaniemi – Seppo Laaksonen – Jarmo Lauri,** Vuotuinen ajankäyttö. Ajankäyttötutkimuksen 1987–88 taulukot. Elokuu 1991. 116 s.

184. **Ari Leppälahti – Mikael Åkerblom,** Industrial Innovation in Finland. August 1991. 82 pp.

185. **Maarit Säynevirta,** Indeksiteoria ja ansiotasoindeksi. Lokakuu 1991. 95 s.

186. **Ari Tyrkkö,** Ahtaasti asuvat. Syyskuu 1991. 134 s.

187. **Tuomo Martikainen – Risto Yrjönen,** Voting, parties and social change in Finland. October 1991. 108 pp.

188. **Timo Kolu,** Työelämän laatu 1977–1990. Työn ja hyvinvoinnin koettuja muutoksia. Tammikuu 1992. 194 s.

189. **Anna-Maija Lehto,** Työelämän laatu ja tasa-arvo. Tammikuu 1992. 196 s.

190. **Tuovi Allén – Päivi Keinänen – Seppo Laaksonen – Seija Ilmakunnas,** Wage from Work and Gender. A Study on Wage Differentials in Finland in 1985. 88 pp.

191. **Kirsti Ahlqvist,** Kodinomistajaksi velalla. Maaliskuu 1992. 98 s.

192. **Matti Simpanen – Irja Blomqvist,** Aikuiskoulutukseen osallistuminen. Aikuiskoulutustutkimus 1990. Toukokuu 1992. 135 s.

193. **Leena M. Kirjavainen – Bistra Anachkova – Seppo Laaksonen – Iiris Niemi – Hannu Pääkkönen – Zahari Staikov,** Housework Time in Bulgaria and Finland. June 1992. 131 pp.

194. **Pekka Haapala – Seppo Kouvonen,** Kuntasektorin työvoimakustannukset. Kesäkuu 1992. 70 s.

195. **Pirkko Aulin-Ahmavaara,** The Productivity of a Nation. November 1992. 72 pp.

196. **Tuula Melkas,** Valtion ja markkinoiden tuolla puolen. Kanssaihmisten apu Suomessa 1980-luvun lopulla. Joulukuu 1992. 150 s.

197. **Fjalar Finnäs,** Formation of unions and families in Finnish cohorts born 1938–67. April 1993. 58 pp.

198. **Antti Siikanen – Ari Tyrkkö,** Koti – Talous – Asuntomarkkinat. Kesäkuu 1993. 167 s.

199. **Timo Matala,** Asumisen tuki ja aravavuokralaiset. Kesäkuu 1993. 84 s.

200. **Arja Kinnunen,** Kuluttajahintaindeksi 1990=100. Menetelmät ja käytäntö. Elokuu 1993. 89 s.

201. **Matti Simpanen,** Aikuiskoulutus ja työelämä. Aikuiskoulutustutkimus 1990. Syyskuu 1993. 150 s.

202. **Martti Puohiniemi,** Suomalaisten arvot ja tulevaisuus. Lokakuu 1993. 100 s.

203. **Juha Kivinen – Ari Mäkinen,** Suomen elintarvike- ja metallituoteteollisuuden rakenteen, kannattavuuden ja suhdannevaihteluiden yhteys; ekonometrinen analyysi vuosilta 1974 – 1990. Marraskuu 1993. 92 s.

204. **Juha Nurmela,** Kotitalouksien energian kokonaiskulutus 1990. Marraskuu 1993. 108 s.

205a. **Georg Luther,** Suomen tilastotoimen historia vuoteen 1970. Joulukuu 1993. 382 s.

205b. **Georg Luther,** Statistikens historia i Finland till 1970. December 1993. 380 s.

206. **Riitta Harala – Eva Hänninen-Salmelin – Kaisa Kauppinen-Toropainen – Päivi Keinänen – Tuulikki Petäjäniemi – Sinikka Vanhala,** Naiset huipulla. Huhtikuu 1994. 64 s.

207. **Wangqiu Song,** Hedoninen regressioanalyysi kuluttajahintaindeksissä. Huhtikuu 1994. 100 s.

208. **Anne Koponen,** Työolot ja ammattillinen aikuiskoulutus 1990. Toukokuu 1994. 118 s.

209. **Fjalar Finnäs,** Language Shifts and Migration. May 1994. 37 pp.

210. **Erkki Pahkinen – Veijo Ritola,** Suhdannekäänne ja taloudelliset aikasarjat. Kesäkuu 1994. 200 s.

211. **Riitta Harala – Eva Hänninen-Salmelin – Kaisa Kauppinen-Toropainen – Päivi Keinänen – Tuulikki Petäjäniemi – Sinikka Vanhala,** Women at the Top. July 1994. 66 pp.

212. **Olavi Lehtoranta,** Teollisuuden tuottavuuskehityksen mittaminen toimialatasolla. Tammikuu 1995. 73 s.

213. **Kristiina Manderbacka,** Terveydentilan mittarit. Syyskuu 1995. 121 s.

214. **Andres Vikat,** Perheellistyminen Virossa ja Suomessa. Joulukuu 1995. 52 s.

215. **Mika Maliranta,** Suomen tehdasteollisuuden tuottavuus. Helmikuu 1996. 189 s.

216. **Juha Nurmela,** Kotitaloudet ja energia vuonna 2015. Huhtikuu 1996. 285 s.

217. **Rauno Sairinen,** Suomalaiset ja ympäristöpolitiikka. Elokuu 1996. 179 s.

218. **Johanna Moisander,** Attitudes and Ecologically Responsible Consumption. August 1996. 159 pp.

219. **Seppo Laaksonen** (ed.), International Perspectives on Nonresponse. Proceedings of the Sixth International Workshop on Household Survey Nonresponse. December 1996. 240 pp.

220. **Jukka Hoffrén,** Metsien ekologisen laadun mittaaminen. Elokuu 1996. 79 s.

221. **Jarmo Rusanen – Arvo Naukkarinen – Alfred Colpaert – Toivo Muilu,** Differences in the Spatial Structure of the Population Between Finland and Sweden in 1995 – a GIS viewpoint. March 1997. 46 pp.

222. **Anna-Maija Lehto,** Työolot tutkimuskohteena. Marraskuu 1996. 289 s.

223. **Seppo Laaksonen** (ed.), The Evolution of Firms and Industries. June 1997. 505 pp.

224. **Jukka Hoffrén,** Finnish Forest Resource Accounting and Ecological Sustainability. June 1997. 132 pp.

225. **Eero Tanskanen,** Suomalaiset ja ympäristö kansainvälisestä näkökulmasta. Elokuu 1997. 153 s.

226. **Jukka Hoffrén,** Talous hyvinvoinnin ja ympäristöhaittojen tuottajana – Suomen ekotehokkuuden mittaaminen. Toukokuu 1999. 154 s.

227. **Sirpa Kolehmainen,** Naisten ja miesten työt. Työmarkkinoiden segregoituminen Suomessa 1970–1990. Lokakuu 1999. 321 s.

228. **Seppo Paananen,** Suomalaisuuden armoilla. Ulkomaalaisten työnhakijoiden luokittelu. Lokakuu 1999. 152 s.

229. **Jukka Hoffrén,** Measuring the Eco-efficiency of the Finnish Economy. October 1999. 80 pp.

230. **Anna-Maija Lehto – Noora Järnefelt** (toim.), Jaksaen ja joustaen. Artikkeleita työolotutkimuksesta. Joulukuu 2000. 264 s.

231. **Kari Djerf,** Properties of some estimators under unit nonresponse. January 2001. 76 pp.

232. **Ismo Teikari,** Poisson mixture sampling in controlling the distribution of response burden in longitudinal and cross section business surveys. March 2001. 120 pp.

233. **Jukka Hoffrén,** Measuring the Eco-efficiency of Welfare Generation in a National Economy. The Case of Finland. November 2001. 199 pp.

234. **Pia Pulkkinen,** "Vähän enemmän arvoinen" Tutkimus tasa-arvokokemuksista työpaikoilla. Tammikuu 2002. 154 s.

235. **Noora Järnefelt – Anna-Maija Lehto,** Työhulluja vai hulluja töitä? Tutkimus kiirekokemuksista työpaikoilla. Huhtikuu 2002. 130 s.

236. **Markku Heiskanen,** Väkivalta, pelko, turvattomuus. Surveytutkimusten näkökulmia suomalaisten turvallisuuteen. Huhtikuu 2002. 323 s.

237. **Tuula Melkas,** Sosiaalisesta muodosta toiseen. Suomalaisten yksityiselämän sosiaalisuuden tarkastelua vuosilta 1986 ja 1994. Huhtikuu 2003. 195 s.

238. **Rune Höglund – Markus Jäntti – Gunnar Rosenqvist** (eds.), Statistics, econometrics and society: Essays in honour of Leif Nordberg. April 2003. 260 pp.

239. **Johanna Laiho – Tarja Nieminen** (toim.), Terveys 2000 -tutkimus. Aikuisväestön haastatteluaineiston tilastollinen laatu. Otanta-asetelma, tiedonkeruu, vastauskato ja estimointi- ja analyysiasetelma. Maaliskuu 2004. 95 s.

240. **Pauli Ollila,** A Theoretical Overview for Variance Estimation in Sampling Theory with Some New Techniques for Complex Estimators. September 2004. 151 pp.

241. **Minna Piispa.** Väkivalta ja parisuhde. Nuorten naisten kokeman parisuhdeväkivallan määrittely surveytutkimuksessa. Syyskuu 2004. 216 s.

242. **Eugen Koev.** Combining Classification and Hedonic Quality Adjustment in Constructing a House Price Index. Tammikuu 2013. 67 pp.

243. **Henna Isoniemi – Irmeli Penttilä** (toim.), Perheiden muuttuvat elinolot. Artikkeleita lapsiperheiden elämänmuutoksista. Syyskuu 2005. 168 s.

244. **Anna-Maija Lehto – Hanna Sutela – Arto Miettinen** (toim.), Kaikilla mausteilla. Artikkeleita työolotutkimuksesta. Toukokuu 2006. 385 s.

245. **Jukka Jalava – Jari Eloranta – Jari Ojala** (toim.) Muutoksen merkit – Kvantitatiivisia perspektiivejä Suomen taloushistoriaan. Tammikuu 2007. 373 s.

246. **Jari Kauppila.** The Structure and Short-Term Development of Finnish Industries in the 1920s and 1930s. An Input-output Approach. Elokuu 2007. 274 s.

247. **Mikko Myrskylä.** Generalised Regression Estimation for Domain Class Frequencies. Elokuu 2007. 137 pp.

248. **Jukka Jalava.** Essays on Finnish Economic Growth and Productivity, 1860–2005. Joulukuu 2007. 154 s.

249. **Yrjö Tala.** Kirkon vai valtion kirjat? Uskontokuntasidonnaisuuden ongelma Suomen väestökirjanpidossa 1839–1904. 317 s.

250. **Hanna-Kaisa Rättö.** Hyvinvointi ja hyvinvoinnin mittaamisen kehittäminen. Huhtikuu 2009. 82 s.

251. **Pertti Koistinen** (toim.), Työn hiipuvat rajat. Tutkielmia palkkatyön, hoivan ja vapaaehtoistyön muuttuvista suhteista. Huhtikuu 2009. 159 s.

252. **Kirsti Ahlqvist.** Kulutus, tieto, hallinta. Kulutuksen tilastoinnin muutokset 1900-luvun Suomessa. Maaliskuu 2010. 314 s.

253. **Jukka Hoffrén (editor) – Eeva-Lotta Apajalahti – Hanna Rättö.** Economy-wide MFA with Hidden Flows for Finland: 1945–2008. 89 pp.

254. **Hannu Pääkkönen.** Perheiden aika ja ajankäyttö. Tutkimuksia kokonaistyöajasta, vapaaehtoistyöstä, lapsista ja kiireestä. Toukokuu 2010. 260 s.

255. **Anna Pärnänen.** Organisaatioiden ikäpolitiikat:strategiat, instituutiot ja moraali. Helmikuu 2011. 291 s.

256. **Ilja Kristian Kavonius.** Kädestä suuhun – Makro- ja mikrotaloudellinen tarkastelu suomalaisten kotitalouksien säästämisestä ja sen mittaamisesta 1950-luvulla. Keskäkuu 2011. 193 s.

257. **Vesa Kuusela.** Paradigms in Statistical Inference for Finite Populations. Up to the 1950s. Elokuu 2011. 236 pp.

258. **Arto Kokkinen.** On Finland's Economic Growth and Convergence with Sweden and the EU15 in the 20th Century. Maaliskuu 2012. 253 pp.

259. **Hanna Sutela.** Määräaikainen työ ja perheellistyminen Suomessa 1984–2008. Joulukuu 2012. 208 s.

260. **Marjo Pyy-Martikainen.** Statistical Analysis of Survey-based Event History Data with Application to Modeling of Unemployment Duration. Lokakuu 2013. 114 pp.

The Research Reports series describes Finnish society in the light of up-to-date research results. Scientific studies that are carried out at Statistics Finland or are based on the datasets of Statistics Finland are published in the series.

Longitudinal surveys are increasingly used to collect event history data on person-specific processes such as transitions between labour market states. Survey-based event history data pose a number of challenges for statistical analysis. These challenges include survey errors due to sampling, non-response, attrition and measurement.

This study deals with non-response, attrition and measurement errors in event history data and the bias caused by them in event history analysis. The study also discusses some choices faced by a researcher using longitudinal survey data for event history analysis and demonstrates their effects. These choices include, whether a design-based or a model-based approach is taken, which subset of data to use and, if a design-based approach is taken, which weights to use.

The study takes advantage of the possibility to use combined longitudinal survey register data. The Finnish subset of European Community Household Panel (FI ECHP) survey for waves 1–5 were linked at person-level with longitudinal register data.

ISSN 0355-2071
= Research Reports
ISBN 978-952-244-456-1 (pdf)
ISBN 978-952-244-455-4 (print)

Tietopalvelu, Tilastokeskus
puh. 09 1734 2220
www.tilastokeskus.fi

Julkaisutilaukset, Edita Publishing Oy
puh. 020 450 05
asiakaspalvelu.publishing@edita.fi
www.editapublishing.fi

Informationstjänst, Statistikcentralen
tel. +358 9 1734 2220
www.stat.fi

Beställning av publikationer, Edita Publishing Oy
tel. +358 20 450 05
www.editapublishing.fi

Information Service, Statistics Finland
tel. +358 9 1734 2220
www.stat.fi

Publication orders, Edita Publishing Oy
tel. +358 20 450 05
www.editapublishing.fi