Interviewer identities as valid instruments for selective panel survey attrition - an evaluation with matched survey-register data

Gerard J. van den Berg*
Maarten Lindeboom†
Marta López‡

March 15, 2007

Abstract

Instrumental variable methods to correct for panel attrition driven by unobservable characteristics rely on untestable exclusion restrictions. We use register information on attriters and non-attriters to assess whether characteristics of the survey interviewer are valid and informative instruments for attrition. The data concern unemployed workers, and the outcome of interest is exit to work. The analysis consists of the estimation of a range of parametric and semi-nonparametric binary outcome models and selection models, on the survey data and on the register data. The results show that there is attrition bias and that the interviewer identity is a valid and effective instrument to correct for this. This provides a justification of the use of this variable as an instrument for attrition.

Keywords: longitudinal data, interviewer effects, unemployment, sample selection, nonresponse, instrumental variables, exclusion restrictions, semi-nonparametric density estimation.

*Free University Amsterdam, Princeton University, IFAU-Uppsala, Netspar, CEPR, IZA, and IFS. Address: Department of Economics, Free University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. Email: gjvdberg@xs4all.nl

†Free University Amsterdam, Netspar, University of Bergen, and Tinbergen Institute. Email: mlindeboom@feweb.vu.nl

‡Free University Amsterdam. Email: mlopez@feweb.vu.nl

1 Introduction

Panel surveys often need to deal with the presence of attrition. If we estimate a statistical model on the sample of respondents, and the unobserved determinants of attrition are related to the endogenous variable of interest, then the estimates will be biased (see e.g. Heckman, 1979, and Vella, 1998). In the present paper we study transitions from unemployment to employment in survey data from the UK with possibly endogenous attrition. We combine the survey information with administrative records of the same workers (see Albæk and Holm Larsen, 1993, for a study with unemployment data that also uses register and survey information). The individual records in the survey data and the administrative data are linked. The administrative data contain information on actual labour market behaviour of all individuals in the original sampling frame (i.e., respondents and non-respondents). In particular, they supply the data at which the individual leaves unemployment. Basically, the administrative data provide us with a unique insight into the behaviour of the sample drop-outs and, in particular, allows us to see to what extent it differs from the behaviour of those who remain in the sample. A random sample of unemployed workers was taken from administrative records, and among these a survey about labor market outcomes and personal characteristics was conducted six months later, as well as follow-up surveys at regular time intervals. We focus on participation on the second wave of the survey (six months later) conditional on participation on the first wave. We are concerned only about unit non-response in the survey, not with item non-response. These data have been previously used by Dolton and O'Neill (1995, 1996a,b), O'Neill and Dolton (2002) and Van den Berg et al. (2006). We refer to this last paper for an analysis of the non-response to the first wave of the survey. The data were originally collected to evaluate the impact of the "Restart" policy program on unemployed individuals, for which the original sample was randomly divided into a treatment and a control group.

There are two main reasons why the unobserved determinants of non-response and the finding of a job are related. First of all, job search behaviour and the behaviour towards survey participation may be affected by the same underlying unobserved individual-specific characteristics. An individual with a relative dislike for social contacts may refuse to cooperate with the survey interview and may also be reluctant to apply for a job and/or to be exposed

to job search counselling by his case worker. An individual who spends a lot of his time searching for a job may not want to spend time with a survey interview. Badly motivated people may have difficulties finding a job and may be less inclined to participate in a survey, especially when this survey is about job search behaviour and labour market prospects. In sum, the unobserved determinants of job search behaviour and non-response may be related, and this gives rise to a selection effect. The second reason for a relation between non-response and the finding of a job is that the acceptance of a job makes it more difficult for the agency to contact the individual. Job acceptance may entail a movement of the individual to another geographical location - which could easily be out of the scope of the survey. Also, the individual may be away from home more often. These concern a causal effect of a job exit on non-response. The second relation is fundamentally different from the first relation, as the causal effect runs directly from job acceptance to non-response, and this effect does not depend on the presence of unobserved characteristics. In the presence of a causal effect, if one of two identical individuals purely by chance finds a job before the survey date, that individual has a higher probability of non-response, and the survey estimates will be biased.

We estimate a selection model to account for endogenous attrition. However, identification of the coefficients in this model is troublesome if the explanatory variables in the selection equation are the same as in the outcome equation, since identification through nonlinearity of the inverse Mills ratio is often weak. As a solution it is common to look for an instrument or an excluded variable (see, for example, Bhattacharya et al., 2005), that is, a variable that affects attrition behaviour but doesn't affect the endogenous variable of interest (in our case a binary outcome variable that indicates the finding of a job between the two first waves of the survey). Our complete data set allows us to check the validity of a candidate instrument. If an instrument is valid we can then use it in the set-up of a sample selection model and compare the outcomes of this model with the outcomes of the model that is estimated on the full sample. We can also evaluate whether the instrument provides a good solution to the selection problem by comparing the results for the selectivity model with those of the model estimated on the sample of respondents.

As candidates for instruments we examine the duration of the first interview, the number of interviews assigned to the interviewer in the first wave

and the identity of the interviewer who carried out this same survey. The first candidate seems a priori reasonable, since the interview duration may act as a proxy for the disutility the previous experience caused the respondent and hence influence their likelihood of responding at the second wave. On the other hand, it is plausible to think that the best interviewers (those who get the highest response rates in past experiences) are assigned more interviews, which would justify the choice of the second candidate. Besides, there's nothing that indicates that they correlate with the relevant chosen outcome, so they could then act as exclusion restrictions. However, neither of these two candidates turn out to be valid.

Interviewer effects—which are typically measured in terms of interviewer variance—on non-response have been extensively studied in the literature. Campanelli et al., 1997 investigate survey refusals and non-contacts and O'Muircheartaigh and Campanelli, 1999 use a multilevel approach that suggests that interviewers who are good at reducing whole household refusals are also good at reducing whole household non-contacts. In panel surveys, the identity of the interviewer in the first wave is useful, as individual differences in interviewer style and personality may have a bearing on the experience for the respondent and hence influence the likelihood of future response (see Pickery et al., 2001). Since in addition there is no reason to think that they are correlated with the outcome of interest, when there is endogenous attrition the characteristics of the interviewers and the interviewing process seem good candidates for instruments (see Fitzgerald et al., 1998 and Nicoletti and Peracchi, 2005). The data at our disposal contain interviewer identifiers, but no personal characteristics are available. We find that the interviewer identity information is a valid instrument and we use it to correct for selection bias. The structural equation in our selection model has a binary outcome that concerns the finding a job between the first and the second survey. We apply parametric and semi-nonparametric methods.

The paper is organized as follows. Section 2 presents the data and gives descriptive statistics. Section 3 describes the selection problem and studies the appropriateness of different instruments. Section 4 extends the parametric selection model to the semiparametric case, carrying out a simulation experiment and applying the method to the data. Section 5 concludes.

2 Data

The Restart Program provides counselling interviews for people in the UK who have been unemployed for more than six months. During these meetings the counsellor offers advice on job search, and he may place workers in contact with employers or training agencies. It also provides training courses for people who have been unemployed for more than two years. To avoid confusion, it must be stressed from the outset that the Restart interviews are not survey interviews. For the purposes of the present paper, the main relevance of the Restart interviews is that the planned date of the first Restart interview (6 months after entry into unemployment) affects the sampling design. To evaluate the program a random sample of 8925 unemployed workers was selected around March/April 1989 who would approach their 6th month of unemployment around May/June 1989. The median of the distribution of the Restart interview date is at the end of May 1989. Individuals were retained in the sample even if they subsequently did not attend a scheduled Restart interview. Every Employment Office throughout Britain was contacted while constructing the sample, in order to eliminate regional biases. Individuals were selected for the sample from the inflow lists, on the basis of their National Insurance (NI) numbers. For this sample, administrative information on some personal characteristics (sex, age, travel-to-work area) was collected from the Employment Services. The information on an individual's travel-to-work area was linked to the National Online Manpower Information System (NOMIS) data, in order to obtain data on local labour market conditions. In addition, the data are linked to the Joint Unemployment and Vacancies Operating System (JUVOS) Cohort database collected by the Employment Service. The JUVOS data provide accurate administrative records on the claimant's unemployment history from 1982 up to January 1995. In the present study we focus on the unemployment spell that has led to the invitation to the Restart interview. Unfortunately, the administrative data do not record the destination state upon exit out of unemployment. This could be employment, a training programme or simply signing off the claiming of unemployment benefit (to obtain benefits, one needs to register at the Employment Service). However, by comparing the administrative data to the survey data for respondents, O'Neill and Dolton (2002) show that most exits out of unemployment amount to a transition into employment.

After excluding individuals who lacked JUVOS data or travel-to-work area information (180 and 736 individuals, respectively), or whose age was below 16 or above 65 (141 individuals), or whose unemployment duration was substantially longer or shorter than 6 months in May 1989 (37 individuals), or whose elapsed unemployment duration was very large we are left with a sample of 8004. Of these, 509 are in the experimental control group. Members of the control group, although eligible for a Restart interview, were deliberately not offered a Restart interview after the first 6 months of unemployment. The existence of a random control group allows for the evaluation of the impact of the program without having to deal with the issue of self-selection. Only 221 out of the 8004 unemployment spells (approximately 3%) are right-censored at the end of the observation window.

In September/October 1989 (6 months after the identification of the full sample) a survey organization (Social and Community Planning Research, or SCPR) conducted a survey (SCPR1) of these individuals, with the purpose to provide additional information on background variables and job search behavior. After another 6 months, a second wave of the SCPR survey (SCPR2) took place. In these interviews detailed information was obtained on subsequent work history, personal characteristics, the Restart interview, previous employment history, search behaviour and benefit income.

At the individual level, the survey is carried out as follows. First, the Employment Office provides the information necessary to locate the sample member. The address is the address given by the sample member for official unemployment related business. Next, the interviewer attempts to establish contact with the sample member him- or herself, to make an appointment for the face-to-face interview. If the attempt does not result in a contact then the interviewer makes another attempt, up to at least four times in total. Different attempts are always made at different days of the week and at different times of the day. The interviewer's earnings depend on the number of actual interviews. There is anecdotal evidence that interviewers often continue to try to establish contact if all four attempts were unsuccessful. After the interview, the interviewer returns the completed response forms by mail to the survey agency.

Of the original sample of 8004 individuals, a total of 4706 individuals completed the first survey. We are interested in those who were unemployed when the first survey took place (our selection date), which leads us to discard

a total of 2200 individuals of this 4706. Furthermore, of the remaining ones, a total of 486 were actually unemployed, but had left unemployment between the first selection date (March/April 1989) and the SCPR1 date (Sept/Oct 1989) and had become unemployed again before the first wave. These special cases are subject to different behaviour patterns and we exclude them from the analysis, which then leads us to the final number of 2004. A total of 1396 out of them responded to SCPR2. A diagram with the spells valid for our study is presented in Figure 1. Similarly, those cases discarded are shown in Figure 2. We are interested in attrition between SCPR1 and SCPR2. We refer to Van den Berg et al. (2006) for an analysis of the non-response to the first wave of the survey.

For the purposes of our analysis we need to create a variable that concerns whether an individual found a job between SCPR1 and SCPR2. To do so, and given that the planned date for the second survey is not available for its non-respondents, we have estimated it given the available SCPR2 dates. We have computed the mean and the median of the second survey dates for the respondents. These dates range from February 1990 to July 1990, and the median is 1.5 weeks earlier than the mean. In order to reduce this range we have attempted to evaluate item non-response instead of unit non-response. In this sense, we have looked at respondents to a specific question in SCPR1 with a high level of non-response, instead of the whole survey, in the hope of reducing the size of the sample, and with it the range of dates. Those concerning the reception of benefits in the family registered the highest level of non-response, though not large enough for our purposes (3%). As an alternative, we studied only those individuals who responded within the date initially fixed for SCPR1, expecting that those among them who responded to SCPR2 would also do it in the expected time initially assigned for it. Unfortunately, the range and variance didn't experiment a significant reduction. As a third option we considered item non-response in SCPR2 among its respondents, which would imply that the interview date for the second wave would be available for all individuals considered. Again the lack of variables that register a high non-response is an obstacle.

In table 1 we present some descriptive statistics for the most relevant characteristics of the whole initial sample as well as for respondents to SCPR1 and SCPR2. It is clear from this table that relatively minor effects of attrition are found on the mean of the variables. More specifically, the average age of

Table 1: Descriptive statistics of variable among SCPR1 respondents

Variable	Resp. SCPR1 ($n = 2004$) Mean (std. dev.)	Resp. SCPR2 ($n = 1396$) Mean (std. dev.)
age	34.83 (12.99)	35.73 (13.29)
female	0.29	0.30
local unempl. rate decline	0.34 (0.05)	0.34 (0.05)
icity	0.22	0.21
control	0.08	0.09
married	0.45	0.49
number dep. kids	0.53 (1.00)	0.56 (1.01)
educational or technical qual.	0.51	0.51
driver license	0.47	0.48
find job between SCPR1-SCPR2	0.53	0.53
censored	.065	.069
unempl. dur. beyond SCPR1 date	362.32	371.57

the respondents to SCPR2 is higher than that of the respondents to the first wave. Similarly, the residual unemployment duration, measured in days (measured from the date at which the first survey took place) is higher for respondents to the second wave, which indicates that, on average, individuals with lower exit rates remain in the sample.

3 Valid instruments to correct for sample attrition

Considering the sample of respondents to the first survey as a startpoint, we are interested in studying the effect of certain socioeconomic characteristics on the probability of finding a job in the time between this survey and the second wave 6 months later. If the finding of a job were only observed for respondents to SCPR2 we would face a selection problem, as respondents in the second wave might not be a random sample of the respondents to SCPR1, but instead there exist unobserved characteristics that determine the finding

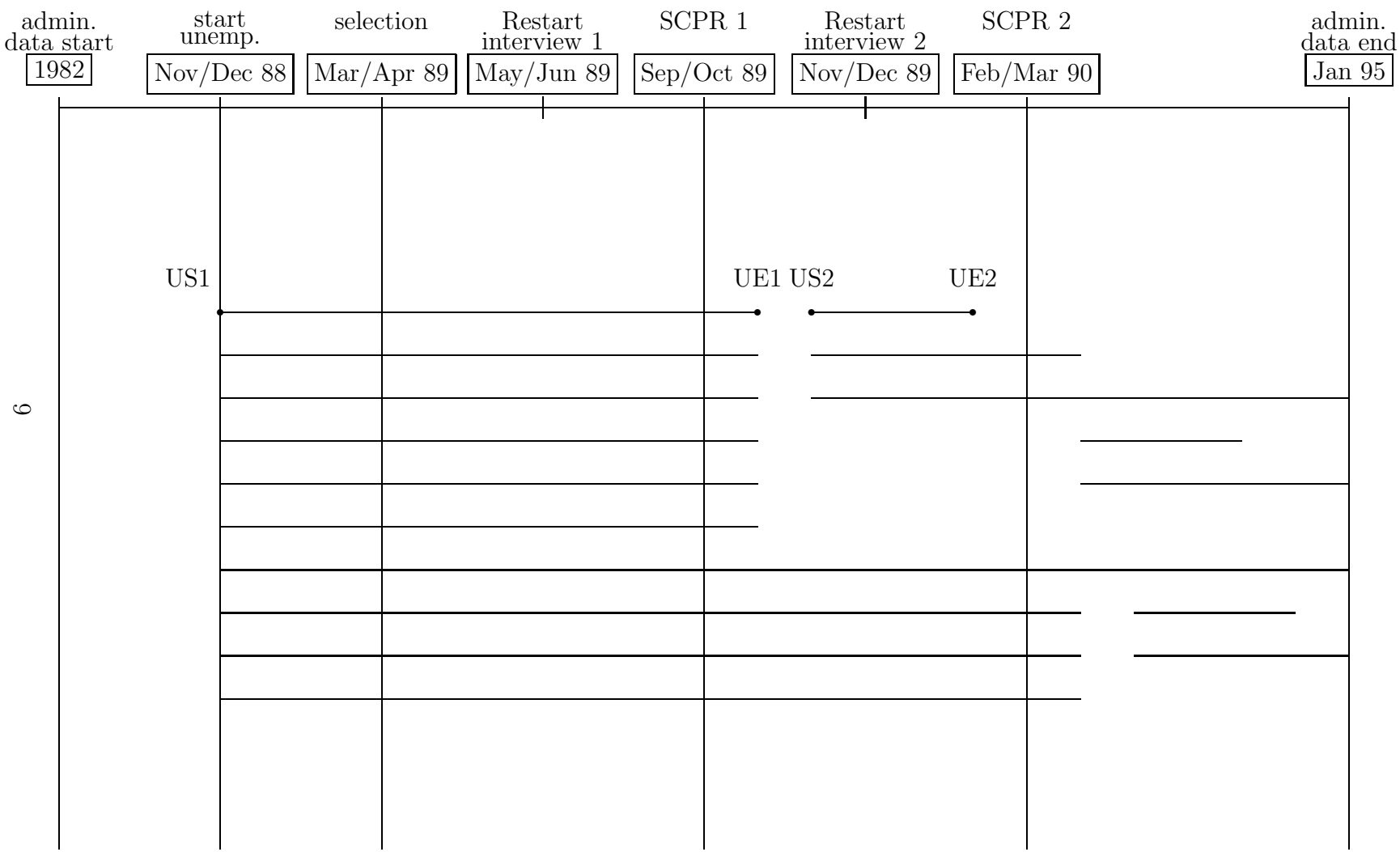

Figure 1: Unemployment spells in analysis

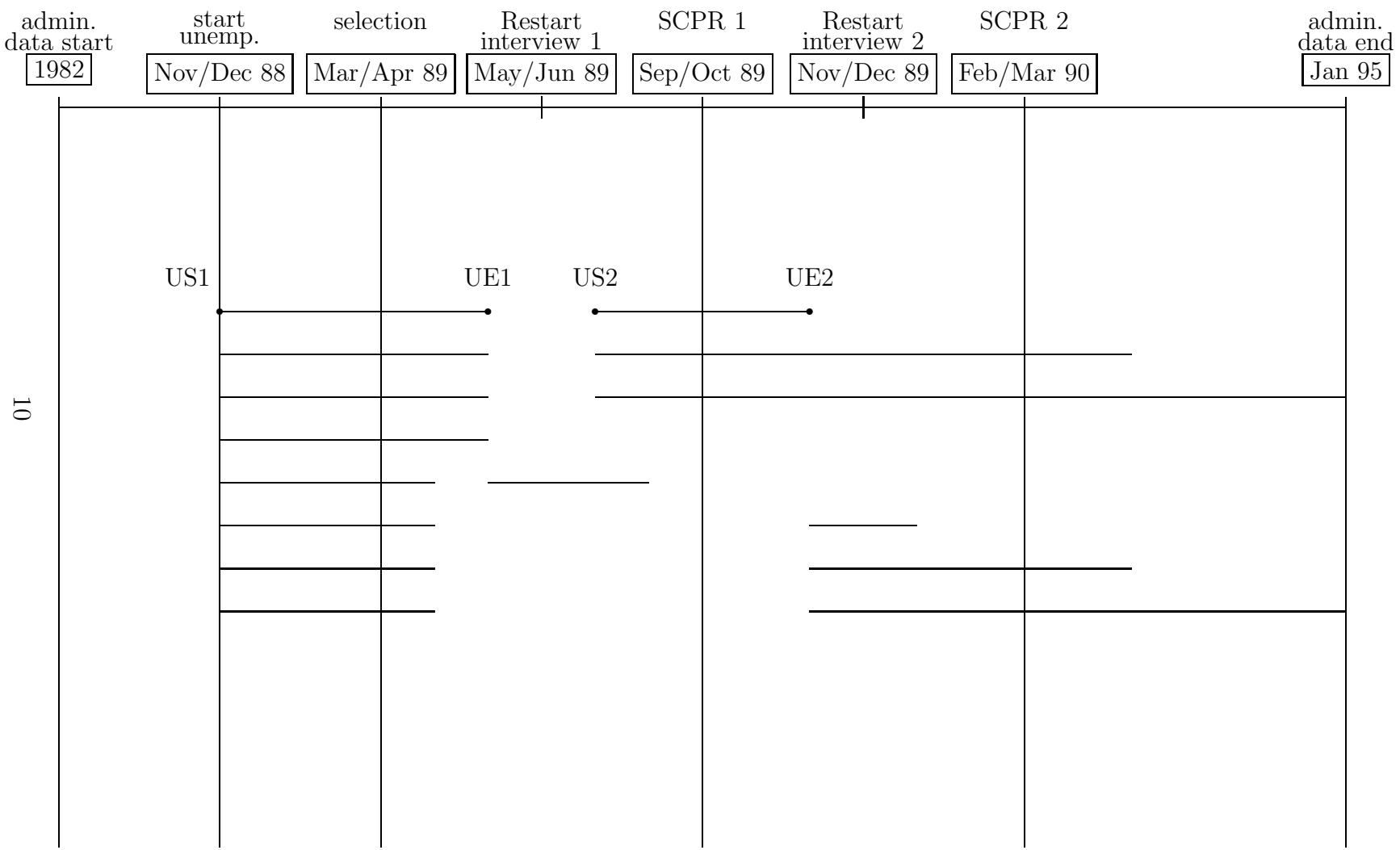

Figure 2: Special unemployment spells

of a job that seem to be related to unobserved characteristics that lead to attrition. This problem has been presented in the introduction and it is the basis of the current section. To express this in mathematical terms, let y_i^* be a latent endogenous variable with associated indicator function y_i (finding a job between SCPR1 and SCPR2), that is observed only when $d_i = 1$ (respondent to SCPR2), where:

$$\begin{aligned} & y_i^* = x_i'\beta + \varepsilon_{1i}; i = 1, \cdots, N \\ & d_i^* = x_i'\gamma_1 + z_i'\gamma_2 + \varepsilon_{2i}; i = 1, \cdots, N \\ & y_i = 1 \quad \text{iff} \quad y_i^* > 0; y_i = 0 \quad \text{otherwise} \\ & d_i = 1 \quad \text{iff} \quad d_i^* > 0; d_i = 0 \quad \text{otherwise} \\ & y_i \quad \text{observed iff} \quad d_i = 1 \end{aligned} \tag{1}$$

If there is a selection problem, then $\varepsilon_1 \perp \varepsilon_2$. We assume that $z \perp \varepsilon_2$ and $x \perp \varepsilon_1, \varepsilon_2$. For correction of the selection bias generated by attrition, it is essential to have instruments that affect attrition behaviour but that do not affect the distribution of the variable of interest. In equation (1) z represents the excluded variable.

A plausible candidate as an instrument would be information on the interviewer who performed the interview in the first wave of the survey. The most flexible way to incorporate interviewer characteristics is to use interviewer fixed effects (i.e., interviewer dummies). We evaluate their validity and performance parametrically, namely by probit analyses and Heckman's selection model, with an extension to the semiparametric case for the selection model in the next section. It must be stressed out that all the analyses concern the case when the second interview date for its non-respondents is estimated by the mean of the interview dates for respondents. If the median is taken instead of the mean the results are very close to those presented below, so they are not given in the present paper.

3.1 Validity

3.1.1 *Informativeness*

We perform a likelihood ratio test to check the joint significance of the coefficients of the interviewer dummies in the selection equation (i.e., $\gamma_2 = 0$ in

equation (1)). It must be stressed out that those interviewer dummies that predict non-response perfectly (only attriters or only respondents) are excluded, giving rise to a final sample of 1895. They will be considered both for the reduced and the full model, so that a likelihood ratio test can be carried out. This test yields that the set of interviewer dummies is jointly significant, with $2|\log LR| = 248.93 > \chi^2_{163,0.95} = 193.791$.

3.1.2 *Exclusion restriction*

To establish that an instrument is truly valid one must also verify that the variable in question does not correlate with the relevant chosen outcome measure in order that the variable can act as an exclusion restriction. Hence it can be tested by establishing if this set of interviewer dummies also adds to the model for the finding of a job between SCPR1 and SCPR2 estimated on the full sample (i.e., respondents to the second wave and attriters between the two waves). A likelihood ratio test is carried out. In our notation, we test whether $\gamma_3 = 0$ in $y_i^* = x_i'\beta + z_i'\gamma_3 + \varepsilon_{1i}, i = 1, \cdots, N$, with $y_i = 1$ iff $y_i^* > 0$ and $y_i = 0$ otherwise. Similarly to the previous point, we delete those interviewers that predict a transition from unemployment to employment perfectly, that is, who are assigned only to people who found a job between SCPR1 and SCPR2 or the oppossite, resulting after their deletion a total of 1945 individuals. The test doesn't reject the hypothesis that the coefficients of the interviewer dummies are jointly significantly different from zero ($2|\log LR| = 164.71 \not> \chi^2_{170,0.95} = 201.423$).

3.2 Performance of the instrument

3.2.1 *The selection problem*

We estimate a bivariate probit model using survey and administrative data (where the finding of a job is known both for respondents and non-respondents to SCPR2) and test whether the correlation coefficient ρ is significantly different from zero. The interviewer dummies are included in the selection equation. The hypothesis that $\rho = 0$ is rejected at a 95% confidence level (with test statistic following a χ^2_1 and taking value 4.57564, p-value=0.0324), which indicates there is a selection problem. The results are displayed in Table 2.

Table 2: Bivariate probit with administrative data

	Coef.	Std. Err.
find job between SCPR1-SCPR2		
intercept	-0.511	0.218*
age	-0.035	0.009*
age2	0.001	0.000*
female	0.407	0.067*
loc. unemp. rate decline	1.251	0.598*
control	-0.090	0.106
living in city area	-0.118	0.072
tot.dep. kids	-0.047	0.039
driver lic.	0.165	0.062*
married	0.475	0.081*
respondent in SCPR2		
intercept	-0.050	0.606
age	-0.002	0.010*
age2	0.001	0.001
female	0.199	0.076*
loc. unemp. rate decline	-0.118	1.055
control	0.181	0.121
living in city area	0.045	0.102
tot.dep. kids	0.027	0.046
driver lic.	0.010	0.071
married	0.230	0.094*
interviewer dummies	...	...
-log likelihood = 2267.7342		
observations = 1895		

Explanatory note: An asterisk denotes significance at the 5% level.

3.2.2 *Relevance of the instrument*

We estimate a parametric sample selection model for a binary outcome including the interviewer dummies in the selection equation (using survey data). We also estimate a probit model for the outcome equation using register and survey data and check their similarity. The results are shown in tables 3 and 4. Note that the number of cases in each table are different, since for the second one the interviewer dummies are not included and so no deletion of perfect predictors takes place. The hypothesis that $\rho = 0$ is rejected at a 90% confidence level (with test statistic following a χ^2_1 and taking value 2.89, p-value=0.0889).

3.2.3 *Is the selection problem remedied well?*

We compare the estimates in the outcome equation in a Heckman selection model to the corresponding estimates when we carry out a probit regression for the outcome equation using only survey data (with a total of 1396 respondents to SCPR2). Tables 3 and 4 display the results.

3.3 Checking the validity of the number of interviews and duration of first survey as instruments

We evaluate the validity as instruments of other variables. We consider the time spent in the face-to-face interview in the first wave of the survey. The argument for using such a variable is that a previous, time consuming experience of being surveyed may make the individual less likely to agree to respond in the next survey. Since a linear specification for this variable is too restrictive (and it is insignificant in the selection equation, with test statistic -0.29 and p-value=0.278) we also consider a quadratic and logarithmic specification. Though the conditions to be an exclusion restriction are verified, it is not informative, as it is insignificant in the selection equation (with $2|LR| = 0.81 \not> \chi^2_{2,0.95} = 5.99146$ in the likelihood ratio test for the joint significance of the parameters in the quadratic case and with test statistic -0.06 and p-value 0.950 in the logarithmic specification).

Another candidate for instrument would be the number of interviews assigned to each interviewer, since it makes sense to think that better interviewers are assigned more interviews based on previous response rates.

Table 3: Probit with sample selection

	Coef.	Std. Err.
Outcome equation		
intercept	-0.679	0.289*
age	-0.032	0.011*
age2	0.001	0.000*
female	0.434	0.077*
loc. unemp. rate decline	0.646	0.697
control	-0.106	0.123
living in city area	-0.114	0.084
tot.dep. kids	-0.062	0.046
driver lic.	0.137	0.074
married	0.575	0.093*
Selection equation		
intercept	-0.179	0.591
age	-0.006	0.010
age2	0.000	0.001
female	0.198	0.075*
loc. unemp. rate decline	0.248	1.061
control	0.180	0.120
living in city area	0.026	0.101
tot.dep. kids	0.028	0.045
driver lic.	-0.004	0.071
married	0.237	0.093*
interviewer dummies	...	...
-log likelihood = 1871.842		
observations = 1895		

Explanatory note: An asterisk denotes significance at the 5% level.

Table 4: Probit on outcome equation

	Full sample		Survey data	
	Coef.	Std. Err.	Coef.	Std. Err.
find job between SCPR1-SCPR2				
intercept	-0.511	0.212*	-0.379	0.255
age	-0.035	0.009*	-0.032	0.011*
age2	0.001	0.000*	0.001	0.000*
female	0.380	0.065*	0.387	0.076*
loc. unemp. rate decline	1.230	0.579*	0.618	0.696
control	-0.042	0.103	-0.069	0.121
living in city area	-0.098	0.070	-0.086	0.085
tot.dep. kids	-0.050	0.038	-0.081	0.045
driver lic.	0.158	0.061*	0.141	0.073*
married	0.464	0.079*	0.537	0.093*
-log likelihood	1312.9961		910.3783	
observations	2004		1396	

Explanatory note: An asterisk denotes significance at the 5% level.

However, this variable turns out not to be valid, since it is not significant in the selection equation when we run a probit analysis. We reject the hypothesis that this variable is zero with test statistic 1.52 and p-value 0.128 in the linear specification, as well as when taking its logarithm (test statistic 1.10 and p-value 0.273). The same occurs when a quadratic specification is considered ($2|LR| = 3.14 \not> \chi^2_{2,0.95} = 5.99146$ in the likelihood ratio test).

The fact that none of these unidimensional variables is valid as an instrument but the multidimensional interviewer identities are, implies that we cannot bound the effect of changes in the explanatory variables (see Bhattacharya et al., 2005), since these bounds are based on a monotonous instrumental variable assumption. Besides, these unidimensional candidates would mean a computational advantage with respect to the interviewer dummies when we turn to semi-nonparametric estimation methods. This would entitle us to consider alternatives that are computationally cumbersome with the case of the interviewer identities (like Klein and Spady, 1993, in combination with the series approach of Newey, 1988). Since these candidates are not valid we have attempted to reduce the number of interviewer dummies by

deleting those cases with interviewers that are assigned less than a certain number of interviews (5 or 6). We don't obtain a significant reduction (less than 3%).

4 Extension: semiparametric estimation of binary-outcome selection models

4.1 Model

So far we have discussed a fully parametric approach to estimate the selection model that assumes bivariate normality of the error terms in the outcome and selection equation. This might be too restrictive. Besides, the sensitivity of the parameter estimates to the distributional assumption has been discussed in the literature (see Manski, 1989). A review of parametric and semi-nonparametric methods to estimate models with sample selection bias can be found in Vella (1998). In this paper we adapt the semi-nonparametric maximum likelihood estimation method of Gallant and Nychka (1987), that approximates the true distribution of the error terms by a Hermite series, to our binary-outcome selection model.

As in section 3, consider a latent endogenous variable y_i^* with associated indicator function y_i, that is observed only when $d_i = 1$, where:

$$\begin{aligned} & y_i^* = x_i'\beta + \varepsilon_{1i}; i = 1, \cdots, N \\ & d_i^* = z_i'\gamma + \varepsilon_{2i}; i = 1, \cdots, N \\ & y_i = 1 \quad \text{iff} \quad y_i^* > 0; y_i = 0 \quad \text{otherwise} \\ & d_i = 1 \quad \text{iff} \quad d_i^* > 0; d_i = 0 \quad \text{otherwise} \\ & y_i \quad \text{observed iff} \quad d_i = 1 \end{aligned} \tag{2}$$

Focusing on the distribution of the error terms $(\varepsilon_{1i}, \varepsilon_{2i})$, the loglikelihood

function for this model is:

$$\begin{aligned}\log L = &\sum_{\substack{d_t=1\\y_t=1}} \log\left[\int_{-z_t'\gamma}^{+\infty}\int_{-x_t'\beta}^{+\infty} h(\varepsilon_1,\varepsilon_2)d\varepsilon_1 d\varepsilon_2\right]\\ &+\sum_{\substack{d_t=1\\y_t=0}} \log\left[\int_{-z_t'\gamma}^{+\infty}\int_{-\infty}^{-x_t'\beta} h(\varepsilon_1,\varepsilon_2)d\varepsilon_1 d\varepsilon_2\right]\\ &+\sum_{d_t=0} \log\left[\int_{-\infty}^{-z_t'\gamma}\int_{-\infty}^{+\infty} h(\varepsilon_1,\varepsilon_2)d\varepsilon_1 d\varepsilon_2\right].\end{aligned} \tag{3}$$

Gallant and Nychka propose h to be of the form:

$$\begin{aligned}h(\varepsilon_1,\varepsilon_2) &= \left[\sum_{i=0}^{K}\sum_{j=0}^{K}\alpha_{ij}\varepsilon_1^i\varepsilon_2^j\right]^2 e^{\frac{-\varepsilon_1^2}{\delta_1^2}} e^{\frac{-\varepsilon_2^2}{\delta_2^2}}\\ &= \sum_{i,j,k,l=0}^{K}\alpha_{ij}\alpha_{kl}\varepsilon_1^{i+k}\varepsilon_2^{j+l} e^{-\frac{\varepsilon_1^2}{\delta_1^2}} e^{-\frac{\varepsilon_2^2}{\delta_2^2}}.\end{aligned} \tag{4}$$

This combination of linear univariate standard normal densities will be used to approximate the true density of the error terms. To ensure integration to 1 the former expression must be divided by:

$$S = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h(\varepsilon_1,\varepsilon_2)d\varepsilon_1 d\varepsilon_2,$$

Since the α's are identified up to a scale only, they can be normalized by, e.g., setting $\alpha_{00}=1$. Besides, for the means of ε_1 and ε_2 to be equal to zero some restrictions on the α_{ij} can be imposed. For the case $K=1$ these restrictions are $\alpha_{01}=\alpha_{10}=0$, but for $K\geq 2$ they become more complicated. In this case, we could proceed as suggested by Melenberg and van Soest (1993): restricting the intercepts in both the selection and the outcome equation. On the other hand, for the identification of the scales of equations in (2) we can set $\delta_1=\sqrt{2}$ and $\delta_2=\sqrt{2}$ (that is, we normalize by setting $\sigma_1^2=\delta_1^2/2=1$ and $\sigma_2^2=\delta_2^2/2=1$). Note that if $K=0$ then h boils down to a bivariate normal distribution with zero correlation between ε_1 and ε_2.

Gallant and Nychka's choice of h as a Hermite form is quite advantageous

computationally, since with it expression (3) becomes:

$$\begin{aligned}\log L = &\sum_{\substack{d_t=1\\y_t=1}} \log\left[\sum_{i,j,k,l=0}^{K} \alpha_{ij}\alpha_{kl} I_{j+l}(-z_t'\gamma, +\infty; \sqrt{2}) \cdot I_{i+k}(-x_t'\beta, +\infty; \sqrt{2})\right] \\ &+ \sum_{\substack{d_t=1\\y_t=0}} \log\left[\sum_{i,j,k,l=0}^{K} \alpha_{ij}\alpha_{kl} I_{j+l}(-z_t'\gamma, +\infty; \sqrt{2}) \cdot I_{i+k}(-\infty, -x_t'\beta; \sqrt{2})\right] \\ &+ \sum_{d_t=0} \log\left[\sum_{i,j,k,l=0}^{K} \alpha_{ij}\alpha_{kl} I_{j+l}(-\infty, -z_t'\gamma; \sqrt{2}) \cdot I_{i+k}(-\infty, +\infty; \sqrt{2})\right] \\ &- \sum_{d_t, y_t} \log\left[\sum_{i,j,k,l=0}^{K} \alpha_{ij}\alpha_{kl} I_{j+l}(-\infty, +\infty; \sqrt{2}) \cdot I_{i+k}(-\infty, +\infty; \sqrt{2})\right],\end{aligned}$$

where:

$$I_k(a, b; \delta) = \int_a^b u^k e^{\frac{-u^2}{\delta^2}} du.$$

By partial integration it can be obtained (see Van der Klaauw and Koning, 2003) that:

$$I_k(a, b; \delta) = \begin{cases} \delta\sqrt{\pi}\left(\Phi\left(\frac{\sqrt{2}b}{\delta}\right) - \Phi\left(\frac{\sqrt{2}a}{\delta}\right)\right) & , k = 0 \\ \frac{\delta^2}{2}\left(\exp(-a^2/\delta^2) - \exp(-b^2/\delta^2)\right) & , k = 1 \\ \frac{\delta^2}{2}\left(a^{k-1}\exp(-a^2/\delta^2) - b^{k-1}\exp(-b^2/\delta^2)\right) + \frac{(k-1)\delta^2}{2} I_{k-2}(a, b; \delta) & , k \geq 2. \end{cases}$$

For a particular K we can estimate this selection model by maximum likelihood like a parametric model. Gallant and Nychka show that the estimates of β and γ are consistent providing K tends to infinity as the sample size increases.

As an extension, a bivariate normal density can be considered instead of the univariate densities product. This generalized semi-nonparametric approach enables us to develop a normality test in line with Van der Klaauw and Koning (2003), for our particular case with a binary outcome, in the Appendix. The power of this test is expected to be lower than in their case, given the limited information contained in the outcome variable. The joint

density of the error terms is now:

$$h(\varepsilon_1, \varepsilon_2) = \sum_{i,j,k,l=0}^{K} \alpha_{ij}\alpha_{kl}\varepsilon_1^{i+k}\varepsilon_2^{j+l} \exp\left(-\varepsilon'\Sigma^{-1}\varepsilon\right). \tag{5}$$

Again the scale normalizations and conditions for identification for the more restrictive density above apply to this case [1] , and the density must be divided by $S = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h(\varepsilon_1, \varepsilon_2)d\varepsilon_1 d\varepsilon_2$. The following integrals will have to be solved:

$$\int_a^b \int_c^d h(\varepsilon_1, \varepsilon_2)d\varepsilon_1 d\varepsilon_2 = \sum_{i,j,k,l=0}^{K} \alpha_{ij}\alpha_{kl} \int_a^b \int_c^d \varepsilon_1^{i+k}\varepsilon_2^{j+l} \exp\left(-\varepsilon'\Sigma^{-1}\varepsilon\right)d\varepsilon_1 d\varepsilon_2.$$

Expressing this in terms of the bivariate normal pdf, the conditional normal pdf and the univariate normal pdf, those integrals are:[2]

$$\int_a^b \int_c^d \varepsilon_1^{i+k}\varepsilon_2^{j+l}\phi(\varepsilon_1, \varepsilon_2)d\varepsilon_1 d\varepsilon_2 = \int_a^b \varepsilon_2^{j+l}\phi(\varepsilon_2) \int_c^d \varepsilon_1^{i+k}\phi(\varepsilon_1|\varepsilon_2)d\varepsilon_1 d\varepsilon_2. \tag{6}$$

As a first step, we compute $\int_c^d \varepsilon_1^{i+k}\phi(\varepsilon_1|\varepsilon_2)d\varepsilon_1$, which leads to:

$$\int_c^d \varepsilon_1^{i+k} \exp\left[-\frac{(\varepsilon_1 - \rho\sigma_1\varepsilon_2/\sigma_2)^2}{2\sigma_1^2(1-\rho^2)}\right] d\varepsilon_1.$$

With the change of variable $u = (\varepsilon_1 - \rho\sigma_1\varepsilon_2/\sigma_2)/(\sigma_1\sqrt{1-\rho^2})$ we have:

$$\int_m^n (\sigma_1\sqrt{1-\rho^2}u + \rho\sigma_1\varepsilon_2/\sigma_2)^{i+k} \exp(-u^2/2)du,$$

where:

$$m = \frac{c - \rho\sigma_1\varepsilon_2/\sigma_2}{\sigma_1\sqrt{(1-\rho^2)}}$$
$$n = \frac{d - \rho\sigma_1\varepsilon_2/\sigma_2}{\sigma_1\sqrt{(1-\rho^2)}}.$$

For the case when $K = 1$, we would have to solve these integrals for $i + k = 0, 1, 2$. Again, using the formulae in Van der Klaauw and Koning (2003):

[1] In this case, we fix elements $\Sigma_{11} = \Sigma_{22} = \sqrt{2}$, which is equivalent to fixing $\sigma_{11} = \sigma_{22} = 1$ if we express (5) in terms of the bivariate normal density.

[2] Since constants that are common for any i, j, k, l cancel out when substracting S, they are not considered. From now on, multiplying constants verifying this will be omitted.

$$\begin{cases} I_0(m,n;2) & ,i+k=0 \\ \\ \sigma_1\sqrt{1-\rho^2}I_1(m,n;2)+\rho\sigma_1\varepsilon_2/\sigma_2 I_0(m,n;2) & ,i+k=1 \\ \\ \sigma_1^2(1-\rho^2)I_2(m,n;2)+2\sigma_1^2\rho\sqrt{1-\rho^2}\varepsilon_2/\sigma_2 I_1(m,n;2) & \\ +(\rho\sigma_1\varepsilon_2/\sigma_2)^2 I_0(m,n;2) & ,i+k=2. \end{cases}$$

Then the second step will in equation (6) will be to compute the following integrals by quadrature methods:

$$\begin{cases} \int_a^b \varepsilon_2^{j+l}\phi(\varepsilon_2)I_0(m,n;\sqrt{2})d\varepsilon_2 & ,i+k=0 \\ \\ \sigma_1\sqrt{1-\rho^2}\int_a^b \varepsilon_2^{j+l}\phi(\varepsilon_2)I_1(m,n;\sqrt{2})d\varepsilon_2 & \\ +\rho\sigma_1/\sigma_2\int_a^b \varepsilon_2^{j+l+1}\phi(\varepsilon_2)I_0(m,n;\sqrt{2})d\varepsilon_2 & ,i+k=1 \\ \\ \sigma_1^2(1-\rho^2)\int_a^b \varepsilon_2^{j+l}\phi(\varepsilon_2)I_2(m,n;\sqrt{2})d\varepsilon_2 & \\ +2\sigma_1^2\rho\sqrt{1-\rho^2}/\sigma_2\int_a^b \varepsilon_2^{j+l+1}\phi(\varepsilon_2)I_1(m,n;\sqrt{2})d\varepsilon_2 & \\ +(\rho\sigma_1/\sigma_2)^2\int_a^b \varepsilon_2^{j+l+2}\phi(\varepsilon_2)I_0(m,n;\sqrt{2})d\varepsilon_2 & ,i+k=2. \end{cases}$$

If $K=0$ this reduces to the parametric case when the joint distribution of the error terms is assumed to be bivariate normal.

Finally, it must be stressed out that for both methods described in this section K grows to infinity as the sample size increases, and it must grow at a fast enough rate to achieve consistency (though its asymptotic distribution hasn't been derived yet). In the literature Gallant and Nychka's method is applied to different values of K out of which the most appealing is selected (see Melenberg and van Soest, 1993, or Gabler et al., 1993).

4.2 Monte Carlo simulations

The following experiment is considered:

$$\begin{aligned} y_i^* &= \beta_0+\beta_1 x_i+\beta_2 w_i+\varepsilon_{1i} \\ d_i^* &= \gamma_0+\gamma_1 v_i+\gamma_2 w_i+\varepsilon_{2i}, \qquad i=1,\cdots,N, \end{aligned}$$

with true parameters $\beta_1=1,\beta_2=.5,\beta_3=-.5,\gamma_1=1,\gamma_2=-1,\gamma_3=1$. The exogenous variables x_i and v_i are independently $N(0,3)$ distributed and w_i

is distributed uniformly on [-3,3]. For all experiments it is imposed that:

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

The errors $(\varepsilon_1, \varepsilon_2)$ are drawn, from a bivariate normal distribution with mean 0, a bivariate t distribution and a centered chi-squared distribution. For the first case we take $(\varepsilon_1, \varepsilon_2)' = C \cdot (u_1, u_2)'$, where u_1 and u_2 are independent draws of a standard normal distribution and:

$$C = \begin{pmatrix} 1 & 0 \\ 0.5 & 0.86603 \end{pmatrix}$$

is the Cholesky decomposition such that $\Sigma = CC'$. In the bivariate t distribution we take $(\varepsilon_1, \varepsilon_2)' = C \cdot (u_1/\sqrt{3}, u_2/\sqrt{3})'$, where u_1 and u_2 are independent draws from a t_3 distribution. Similarly, for the chi-squared we take $(\varepsilon_1, \varepsilon_2)' = C \cdot ((u_1 - 2)/2, (u_2 - 3)/\sqrt{6})'$, with u_1 and u_2 being independent draws of independent chi-squared distributions with 2 and 3 degrees of freedom, respectively. A total of 100 simulations are performed for each case, using samples of size $n = 500$. It would be convenient to extend this to bigger samples, with e.g. $n = 1000$ (though the time for computations would be then disproportionally long, especially for the generalized semi-nonparametric approach, that uses quadrature methods).

Results for the semi-nonparametric (SNP) approach are given, as well as for the generalized semi-nonparametric approach (GSNP). We show the results for $K = 1$ with restrictions on the α_{ij}'s for SNP and GSNP. The corresponding estimates for $K = 1, 2, 3$ using Melenberg and van Soest's approach for SNP are also presented. Note that the GSNP case for the latter has only been estimated for $K = 1$, since it becomes computationally costly. The fully parametric method performs very well for all three different distribution of the disturbances. Under Melenberg and van Soest's approach, and because the true values of the intercepts are not zero, we expect some α's to differ from zero to allow for a nonzero mean of the joint distribution of the disturbances. In the cases of the experiment, it seems that in general increasing the value of K originates imprecise estimates, as well as a significant increase in computational time.

Table 5: Gallant and Nychka estimates with restrictions on α_{ij}'s with bivariate normal distributed disturbances (Standard errors between parentheses)

N=500, K=1	Parametric		SNP		GSNP	
β_0	0,99	(0,158)	1,14	(0,137)	0,98	(0,224)
β_1	0,50	(0,070)	0,56	(0,085)	0,49	(0,122)
β_2	-0,50	(0,084)	-0,57	(0,080)	-0,50	(0,123)
γ_0	1,02	(0,114)	1,11	(0,144)	0,98	(0,195)
γ_1	-1,00	(0,102)	-1,09	(0,133)	-0,97	(0,189)
γ_2	1,01	(0,095)	1,10	(0,132)	0,97	(0,199)
ρ	0,55	(0,222)	0		0,28	(0,465)
α_{01}	0		0		0	
α_{10}	0		0		0	
α_{11}	0		0,30	(0,177)	0,12	(0,267)
-log-likl	251,42		251,76		251,22	

4.3 Results for the data

We apply the semi-nonparametric approach of Gallant and Nychka for $K = 1$ with restrictions on the α's, as well as for $K = 1$ with restrictions on the intercepts to our data (see Melenberg and van Soest, 1993). Table 11 compares these results to those using a fully parametric assumption for the distribution of the error terms. The results when fixing the α's are very close to the those of the parametric model, whereas when restricting the intercepts some differences are observed.

Table 6: Gallant and Nychka estimates with Melenberg and van Soest approach with bivariate normal distributed disturbances (Standard errors between parentheses)

N=500	GSNP		SNP					
			K=1		K=2		K=3	
β_0	0		0		0		0	
β_1	0,46	(0,060)	0,42	(0,054)	0,51	(0,078)	0,61	(0,109)
β_2	-0,46	(0,070)	-0,45	(0,050)	-0,51	(0,074)	-0,62	(0,128)
γ_0	0		0		0		0	
γ_1	-0,91	(0,099)	-0,87	(0,081)	-1,01	(0,127)	-1,20	(0,182)
γ_2	0,91	(0,095)	0,86	(0,074)	1,01	(0,120)	1,19	(0,171)
ρ	0,61	(0,233)	0		0		0	
α_{01}	0,52	(0,287)	0,53	(0,097)	0,28	(0,143)	0,41	(0,317)
α_{02}					-0,09	(0,064)	-0,17	(0,191)
α_{03}							-0,05	(0,066)
α_{10}	0,24	(0,380)	0,60	(0,128)	0,23	(0,209)	0,45	(0,396)
α_{11}	0,17	(0,166)	0,45	(0,100)	0,47	(0,108)	0,44	(0,409)
α_{12}					0,16	(0,069)	-0,05	(0,214)
α_{13}							-0,06	(0,085)
α_{20}					-0,14	(0,126)	-0,23	(0,225)
α_{21}					0,17	(0,073)	-0,04	(0,211)
α_{22}					0,10	(0,047)	0,05	(0,175)
α_{23}							0,01	(0,056)
α_{30}							-0,07	(0,103)
α_{31}							-0,07	(0,095)
α_{32}							0,01	(0,056)
α_{33}							0,01	(0,018)
-log-likl	251,09		253,78		250,92		249,24	

Table 7: Gallant and Nychka estimates with restrictions on α_{ij}'s with bivariate t distributed disturbances (Standard errors between parentheses)

N=500, K=1	Parametric		SNP		GSNP	
β_0	1,30	(0,202)	1,43	(0,199)	1,34	(0,236)
β_1	0,65	(0,086)	0,69	(0,094)	0,62	(0,105)
β_2	-0,65	(0,099)	-0,70	(0,099)	-0,65	(0,119)
γ_0	1,11	(0,144)	1,14	(0,154)	1,06	(0,163)
γ_1	-1,12	(0,148)	-1,16	(0,155)	-1,05	(0,167)
γ_2	1,12	(0,131)	1,16	(0,138)	1,05	(0,147)
ρ	0,47	(0,312)	0		-0,31	(0,470)
α_{01}	0		0		0	
α_{10}	0		0		0	
α_{11}	0		0,18	(0,120)	0,28	(0,195)
-log-likl	220,01		220,41		218,26	

Table 8: Gallant and Nychka estimates with Melenberg and van Soest approach with bivariate t distributed disturbances (Standard errors between parentheses)

N=500	GSNP		SNP					
			K=1		K=2		K=3	
β_0	0		0		0		0	
β_1	0,60	(0,084)	0,51	(0,055)	0,63	(0,072)	0,69	(0,142)
β_2	-0,59	(0,079)	-0,50	(0,045)	-0,64	(0,079)	-0,71	(0,151)
γ_0	0		0		0		0	
γ_1	-1,10	(0,144)	-0,99	(0,107)	-1,10	(0,170)	-1,18	(0,286)
γ_2	1,10	(0,126)	0,99	(0,093)	1,09	(0,144)	1,16	(0,256)
ρ	0,51	(0,247)	0		0		0	
α_{01}	0,55	(0,457)	0,73	(0,131)	0,39	(0,171)	0,54	(0,405)
α_{02}					-0,12	(0,085)	-0,21	(0,168)
α_{03}							-0,09	(0,075)
α_{10}	0,71	(0,644)	0,82	(0,112)	0,48	(0,276)	0,62	(0,641)
α_{11}	0,55	(0,296)	0,67	(0,126)	0,69	(0,200)	0,63	(0,781)
α_{12}					0,19	(0,122)	-0,10	(0,269)
α_{13}							-0,10	(0,118)
α_{20}					0,03	(0,174)	-0,15	(0,318)
α_{21}					0,25	(0,116)	-0,01	(0,437)
α_{22}					0,12	(0,054)	0,05	(0,165)
α_{23}							0,01	(0,076)
α_{30}							-0,08	(0,129)
α_{31}							-0,08	(0,137)
α_{32}							0,02	(0,050)
α_{33}							0,01	(0,023)
-log-likl	217,99		221,54		217,97		216,58	

Table 9: Gallant and Nychka estimates with restrictions on α_{ij}'s with bivariate Chi-squared distributed disturbances (Standard errors between parentheses)

N=500, K=1	Parametric		SNP		GSNP	
β_0	1,19	(0,166)	1,30	(0,188)	1,10	(0,245)
β_1	0,65	(0,070)	0,68	(0,097)	0,59	(0,122)
β_2	-0,67	(0,083)	-0,72	(0,102)	-0,62	(0,130)
γ_0	1,06	(0,117)	1,10	(0,171)	0,96	(0,195)
γ_1	-1,06	(0,119)	-1,11	(0,149)	-0,97	(0,165)
γ_2	1,07	(0,102)	1,11	(0,135)	0,97	(0,156)
ρ	0,43	(0,287)	0		0,23	(0,539)
α_{01}	0		0		0	
α_{10}	0		0		0	
α_{11}	0		0,18	(0,217)	0,04	(0,272)
-log-likl	229,43		229,79		229,04	

Table 10: Gallant and Nychka estimates with Melenberg and van Soest approach with bivariate Chi-squared distributed disturbances (Standard errors between parentheses)

N=500	GSNP		SNP					
			K=1		K=2		K=3	
β_0	0		0		0		0	
β_1	0,56	(0,057)	0,51	(0,051)	0,56	(0,061)	0,63	(0,092)
β_2	-0,57	(0,057)	-0,53	(0,046)	-0,59	(0,068)	-0,63	(0,092)
γ_0	0		0		0		0	
γ_1	-0,94	(0,094)	-0,90	(0,083)	-0,95	(0,106)	-1,07	(0,198)
γ_2	0,95	(0,081)	0,90	(0,073)	0,96	(0,102)	1,08	(0,194)
ρ	0,47	(0,285)	0		0		0	
α_{01}	0,43	(0,113)	0,57	(0,058)	0,40	(0,091)	0,65	(0,171)
α_{02}					-0,07	(0,044)	-0,21	(0,151)
α_{03}							-0,11	(0,043)
α_{10}	0,53	(0,177)	0,70	(0,054)	0,56	(0,124)	0,88	(0,140)
α_{11}	0,21	(0,127)	0,42	(0,052)	0,44	(0,071)	0,62	(0,173)
α_{12}					0,07	(0,033)	-0,19	(0,125)
α_{13}							-0,11	(0,041)
α_{20}					-0,03	(0,065)	-0,15	(0,180)
α_{21}					0,11	(0,032)	-0,09	(0,147)
α_{22}					0,06	(0,022)	0,01	(0,132)
α_{23}							0,01	(0,043)
α_{30}							-0,13	(0,063)
α_{31}							-0,10	(0,055)
α_{32}							0,02	(0,048)
α_{33}							0,01	(0,014)
-log-likl	228,39		230,58		227,65		221,88	

Table 11: Parametric estimates and Gallant and Nychka estimates for $K = 1$ with restrictions on the α's (SNPa) and restrictions on the intercepts (SNPb). (Estimates corresponding to interviewer dummies, in the selection equation, not included)

	Parametric		**SNPa**		**SNPb**	
	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.
Outcome equation						
intercept	-0.679	0.289*	-0.655	0.322*		
age	-0.032	0.011*	-0.034	0.011*	-0.070	0.019*
age2	0.001	0.000*	0.001	0.000*	0.002	0.001*
female	0.434	0.077*	0.458	0.088*	1.196	0.133*
loc. unemp. rate decline	0.646	0.697	0.715	0.753	0.837	0.487
control	-0.106	0.123	-0.121	0.131	-0.278	0.222
living in city area	-0.114	0.084	-0.124	0.090	-0.295	0.148
tot.dep. kids	-0.062	0.046	-0.073	0.048	-0.008	0.086
driver lic.	0.137	0.074	0.150	0.078	0.206	0.134
married	0.575	0.093*	0.610	0.110*	1.446	0.198*
Selection equation						
intercept	-0.179	0.591	0.254	0.495		
age	-0.006	0.010	-0.005	0.011	0.001	0.021
age2	0.000	0.001	0.000	0.001	0.001	0.000
female	0.198	0.075*	0.206	0.080*	0.243	0.141
loc. unemp. rate decline	0.248	1.061	0.171	1.109	1.550	1.259
control	0.180	0.120	0.186	0.127	0.347	0.206
living in city area	0.026	0.101	0.033	0.106	0.083	0.213
tot.dep. kids	0.028	0.045	0.028	0.048	0.058	0.076
driver lic.	-0.004	0.071	-0.001	0.075	0.111	0.134
married	0.237	0.093*	0.248	0.100*	0.538	0.181*
interviewer dummies	...	...	...	...	...	...
α_{01}					-0.846	0.220*
α_{10}					0.759	0.280*
α_{11}			0.227	0.152	5.044	2.027*
observations=1895	-log likl=1871.842		-log likl=1872.28		-log likl=1846.95	

5 Conclusions

Most longitudinal surveys suffer from attrition at least some of which may not occur at random from the sample. Attrition may cause a bias in estimates based on data from respondents, since unobserved determinants of non-response behaviour might be related to the endogenous variable of interest.

In this paper we study the finding of a job between two waves of a survey as the binary outcome of interest and attrition between these two surveys. Our complete dataset combines survey and administrative records. The administrative records provide information on individuals labour market behaviour and personal characteristics for the complete sample of participants and non-respondents. We look for appropriate instruments or excluded variables to guarantee the identification of the parameters in the selection model. We find that there is attrition bias and that information on the interviewer that carries out an interview on the first wave of a panel survey can act as a valid and effective instrument to correct for this. Other candidates, like the number of interviews assigned to each interview in the first wave or the duration of the first interview do not verify the conditions to be valid. Our results are of interest for agencies that run surveys as well as for researchers who are not so well endowed with data as in the present paper. We also estimate the selection model semi-nonparametrically, adapting Gallant and Nychka's (1987) method to our particular case with a binary outcome, running a simulation experiment to check its performance. We then apply this to our data, finding some differences in the estimates depending on the restrictions on the parameters imposed to ensure a zero mean.

References

Albæk, K. and A. Holm Larsen (1993). Unemployment data, from surveys and administrative registers. In H. Bunzel, P. Jensen, and N. Westergård-Nielsen (Eds.), *Panel Data and Labour Market Dynamics*, pp. 123–147. Amsterdam: North Holland.

Bhattacharya, J., A. Shaikh, and E. Vytlacil (2005). Treatment effect bounds: an application to swan-ganz catheterization. Working paper, NBER No. W11263.

Campanelli, P., P. Sturgis, and S. Purdon (1997). *Can you hear me knocking: an investigation into the impact of interviewers on survey response rates*. London: Social and Community Planning Research.

Dolton, P. and D. O'Neill (1995). The impact of restart on reservation wages and long-term unemployment. *Oxford Bulletin of Economics and Statistics 57*, 451–470.

Dolton, P. and D. O'Neill (1996a). The restart effect and the return to full-time stable employment. *Journal of the Royal Statistical Society, Series A 159*, 275–288.

Dolton, P. and D. O'Neill (1996b). Unemployment duration and the restart effect: some experimental evidence. *Economic Journal 106*, 387–400.

Fitzgerald, J., P. Gottschalk, and R. Moffitt (1998). An analysis of sample attrition in panel data: the michigan panel study of income dynamics. *Journal of Human Resources 33*(2), 251–299.

Gabler, S., F. Laisney, and M. Lechner (1993). Semi-nonparametric maximum likelihood estimation of binary choice models with and application to labour force participation. *Journal of Business and Economic Statistics 11*, 61–80.

Gallant, A. and D. Nychka (1987). Seminonparametric maximum likelihood estimation. *Econometrica 55*, 363–390.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica 47*, 153–161.

Klein, R. and R. Spady (1993). An efficient semiparametric estimator of the binary response model. *Econometrica 61*, 387–423.

Manski, C. (1989). Anatomy of the selection problem. *Journal of Human Resources 24*(3), 343–360.

Melenberg, B. and A. van Soest (1993). Semi-parametric estimation of the sample selection model. Working paper, CentER, Tilburg.

Newey, W. (1988). Two step series estimation of sample selection models. Unpublished manuscript.

Nicoletti, C. and F. Peracchi (2005). Survey response and survey characteristics: microlevel evidence from the european community household panel. *Journal of the Royal Statistical Society, Series A 168*(4), 763–781.

O'Muircheartaigh, C. and P. Campanelli (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society, Series A 162*, 437–446.

O'Neill, D. and P. Dolton (2002). The long-run effects of unemployed monitoring and work search programmes. *Journal of Labor Economics 20*, 381–403.

Pickery, J., G. Loosveldt, and A. Carton (2001). The effects of interviewer and respondent characteristics on response behavior in panel surveys: a multilevel approach. *Sociological Methods & Research 29*(4), 509–523.

Van den Berg, G., M. Lindeboom, and P. Dolton (2006). Survey non-response and the duration of unemployment. *Journal of the Royal Statistical Society, Series A 169*, 585–604.

Van der Klaauw, B. and R. Koning (2003). Testing the normality assumption in the sample selection model with an application to travel demand. *Journal of Business and Economic Statistics 21*(1), 31–42.

Vella, F. (1998). Estimating models with sample selection bias: a survey. *Journal of Human Resources 33*(1), 127–169.

A Appendix

We assume that the true density of $(\varepsilon_1, \varepsilon_2)$ is a member of the flexible class of density functions $\mathcal{H}_{\mathcal{K}}$. We test the null hypothesis that $(\varepsilon_1, \varepsilon_2)$ has a normal distribution against the alternative hypothesis that it has some other mean zero bivariate distribution function in the class $\mathcal{H}_{\mathcal{K}}$ for any fixed K. This is equivalent to testing for the joint significance of the $\alpha_{ij}, i + j \geq 1$. Since the α's are normalized by setting $\alpha_{00} = 1$ and there are two restrictions on these parameters to fix the location (which for $K = 1$ are $\alpha_{01} = \alpha_{10} = 0$), the null hypothesis will be rejected with $100(1 - \alpha)\%$ confidence when the likelihood ratio (LR) verifies $2|\log LR| > \chi_{(K+1)^2-3,1-\alpha}$. The results are not very encouraging: even though there is a 0% rejections in the case where the true distribution of the error terms is bivariate normal, it is 32% for the t and to an extremely low 7% for the Chi-squared. Different pictures (figures 3 to 5) with the distribution of the coefficients together with their distribution when fixing $\alpha_{ij} = 0, \quad i + j \geq 1$ are also given as a graphical help.

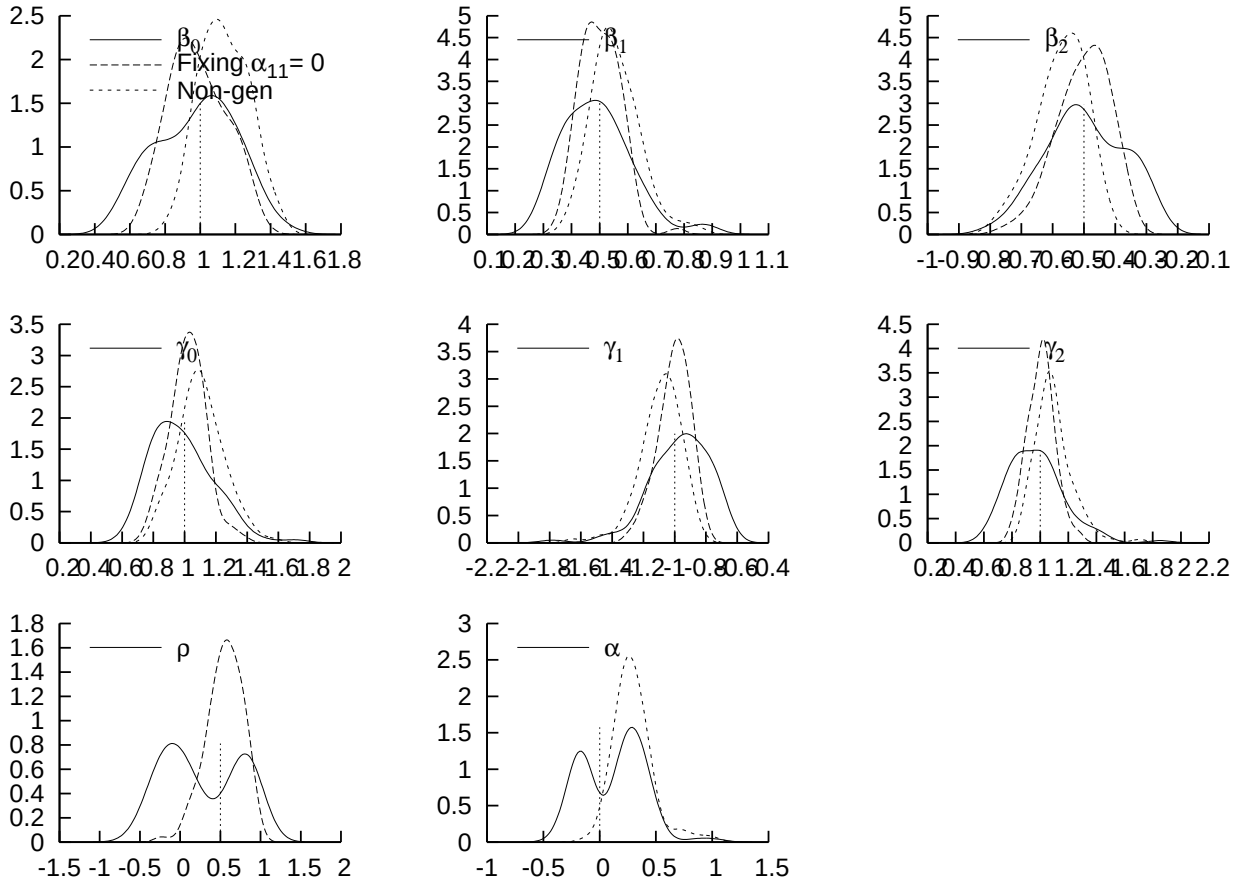

Figure 3: Bivariate normal with $\mathrm{Cov}(\varepsilon_1, \varepsilon_2) = 0.5$.

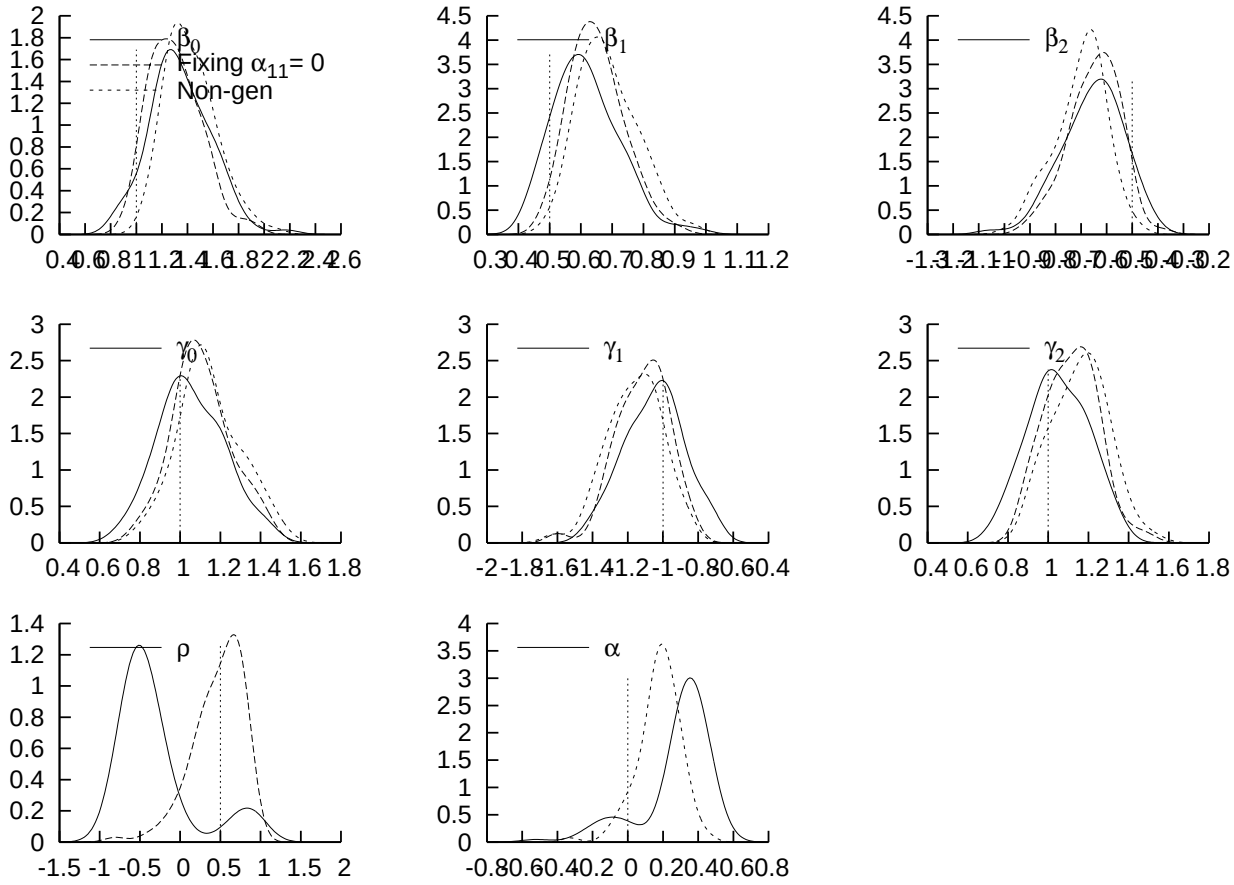

Figure 4: Bivariate t with $\mathrm{Cov}(\varepsilon_1, \varepsilon_2) = 0.5$.

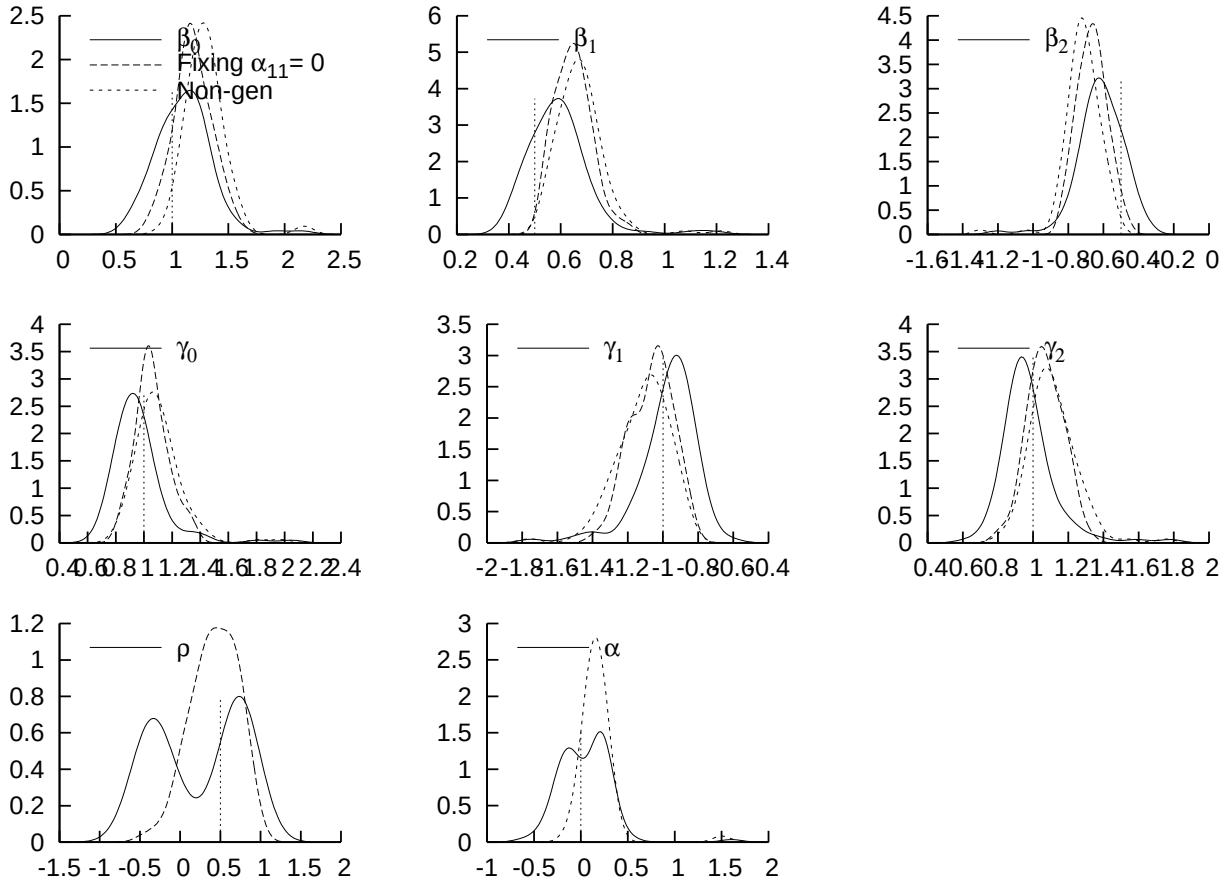

Figure 5: Bivariate Chi-squared with $\mathrm{Cov}(\varepsilon_1, \varepsilon_2) = 0.5$.