Empirical Monte Carlo Evidence on Estimation of Timing-of-Events Models*

Stefano Lombardi† Gerard J. van den Berg‡ Johan Vikström§

May 9, 2019

Abstract

This paper uses an Empirical Monte Carlo simulation approach to study estimation of Timing-of-Events (ToE) models. We exploit rich Swedish data of jobseekers with information on participation in a training program to simulate placebo treatment durations. We first use these simulations to examine which covariates are major confounding variables to be included in selection models. We then show that the joint inclusion of specific types of short-term employment history variables (notably, the share of time spent in employment), together with baseline socio-economic characteristics, regional and inflow timing information, is able to remove selection bias. Next, we omit sets of covariates and estimate ToE models with discrete distributions for the ensuing systematic unobserved heterogeneity. In many cases the ToE approach provides accurate effect estimates, especially if calendar-time variation is taken into account. However, assuming too many or too few support points for unobserved heterogeneity may lead to large biases. Information criteria, in particular those penalizing parameter abundance, are useful to select the number of support points.

JEL-codes: C14; C15; C41; J64

*We thank Paul Muller, Oskar Nordström Skans, seminar participants at IFAU, and conference participants at EEA and EALE for useful suggestions. Estimations were performed on supercomputing resources provided by the Swedish National Infrastructure for Computing (SNIC) at the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). Vikström and Lombardi acknowledge support from FORTE.

†Uppsala University, IFAU-Uppsala, IZA, Uppsala Center for Labor Studies.

‡University of Bristol, IFAU-Uppsala, IZA, ZEW, CEPR and J–PAL.

§IFAU-Uppsala, Uppsala Center for Labor Studies and Uppsala University.

1 Introduction

The Timing-of-Events (ToE) approach focuses on the effect of a treatment given during a spell in some state on the rate of leaving that state, if systematic unobserved confounders cannot be ruled out. To this purpose, Abbring and van den Berg (2003) specify a bivariate Mixed Proportional Hazard (MPH) model and establish conditions under which all parts of the model, including the treatment effect, are non-parametrically identified. The fact that this approach allows for unobserved confounders is one reason for why it has been applied in many settings. An early example is Abbring et al. (2005), who study the effect of benefit sanctions on the transition rate out of unemployment, with unobserved factors such as personal motivation potentially affecting both the time to a benefit sanction (treatment) and time in unemployment (outcome). Recent examples include Crépon et al. (2018), Richardson and van den Berg (2013), Caliendo et al. (2016), Busk (2016), Lindeboom et al. (2016), Holm et al. (2017), Bergemann et al. (2017) on labor market policies; Van Ours and Williams (2009, 2012), and McVicar et al. (2018) on cannabis use; van Ours et al. (2013), van den Berg and Gupta (2015), Palali and van Ours (2017) in health settings; Bijwaard et al. (2014) on migration; Jahn and Rosholm (2013) on temporary work; and Baert et al. (2013) on overeducation.

Several factors must be taken into account when estimating the ToE model. First, the model is often specified by approximating the unknown bivariate unobserved heterogeneity distribution by means of a discrete distribution (Lindsay, 1983; Heckman and Singer, 1984). In practice this can be implemented in several ways. One is to pre-specify a (relatively low) number of support points and increase their number until computational problems arise. Alternatively, one could use an information criterion to select the number of support points. Sample size may be a relevant factor, since estimation of (non-linear) MPH models with many parameters may be problematic with small samples. Also, different sources of variation, such as variation from time-varying covariates, may improve identification and the estimation.

In this paper, we use a new simulation design based on actual data to evaluate these and related specification issues for the implementation of the ToE model in practice. To this end, we modify the novel Empirical Monte Carlo design (EMC) proposed by Huber et al. (2013). In their study, they compare different methods to estimate treatment effects under unconfoundedness.[1] The key idea is to use actual data on treated units

[1]Other studies using the EMC simulation design include Huber et al. (2016) on the performance of parametric and semiparametric estimators commonly used in mediation analysis; Frölich et al. (2017) study the performance of a broad set of semi and nonparametric estimators for evaluation under conditional independence; Lechner and Strittmatter (2017) compare different procedures to deal with

to simulate placebo treatments for non-treated units and then base the simulations on these placebo treatments. This ensures that the true effect is zero, that the selection model is known, and that the unconfoundedness assumption holds by construction. The fact that real data is used instead of a data generating process chosen by the researcher makes the simulation exercise arguably more relevant for real applications.

Previous EMC implementations study estimators based on conditional independence assumptions. Here, we propose and implement a variant of the basic EMC approach, which allows us to study the ToE model. In our simulation design, we use rich administrative data on Swedish jobseekers, with information on participation in a training program (the treatment). For each jobseeker, we create detailed background information. This is used to estimate a duration model for the time to treatment using data on both treated and non-treated units. We then use the estimated model to simulate placebo treatment durations for each non-treated unit. By construction, the effect of these placebo treatments is zero and the treatment assignment process is known. With the simulated data we estimate various ToE models. Here, the key aspect is that we leave out some of the covariates that were used to simulate the placebo treatments. Since the excluded covariates were used to generate the placebo treatments, and since they also affect the outcome duration (re-employment rate), we obtain a bivariate duration model with correlated unobserved determinants, i.e. the ToE setting. This new simulation design allows us to use real data to examine a number of ToE-specification issues.[2]

An important question that has been studied for a long time is how to best specify the distribution of unobserved heterogeneity. Initial simulation evidence was provided by Heckman and Singer (1984), Ridder (1987), and Huh and Sickles (1994). More recently, Baker and Melino (2000) study a univariate duration model with unobserved heterogeneity and duration dependence. One conclusion is that model specifications with too many support points over-correct for unobserved heterogeneity (through an overly-dispersed unobserved heterogeneity distribution). This in turns leads to bias in all model components. Gaure et al. (2007) also use simulated data and examine a similar bivariate duration model as the one analyzed in this paper. One finding is that a discrete support points approach is generally reliable if the sample is large and there

common support problems; and Bodory et al. (2016) consider several inference methods for matching and weighting methods.

[2]Recently, Advani et al. (2018) use the LaLonde (1986) data to provide a critical assessment of the internal validity of the EMC simulation design. This critique is rebutted by Huber et al. (2016), who, among other things, stress that the LaLonde data is small in size (hence, the Monte Carlo samples are not drawn from an infinite population) and also contains only a few covariates (hence, the selection process is not well-captured).

is some exogenous variation, such as variation due to time-varying variables. On the other hand, unjustified restrictions – such as pre-specifying an extremely low number of support points for the unobserved heterogeneity – or deviations from the model assumptions, may cause substantial bias.

Our study adds to this evidence by using a simulation design based on actual data. This leads to several conclusions. In the main analyses, we leave out a large number of covariates from the model, so that the estimated effect of the placebo treatment is far from the true zero effect, i.e. there is substantial bias. However, two support points are already able to eliminate a large share of the bias. We also find a substantial risk of over-correcting for unobserved heterogeneity. With many support points, the average bias is more than twice as large as with a few support points, and the variance increases in the number of support points. The over-correction problem occurs because too many support points lead to an overly-dispersed distribution of unobserved heterogeneity, and to fit the data this is compensated in the model by bias in the treatment effect and the duration dependence.

Another result is that information criteria are useful for selecting the number of support points. In particular, the Akaike information criterion (AIC), Bayesian information criterion (BIC), and Hannan-Quinn information criterion (HQIC) all perform well. They protect against over-correction by penalizing parameter abundance. They also guard against under-correction by rejecting models without or with only weak correction for unobserved heterogeneity. However, information criteria with little penalty for parameter abundance, such as those solely based on the maximum likelihood (ML) criterion, should be avoided altogether. This is because they tend to favor models with too many support points, and this leads to over-correction problems.

We mainly focus on the above-mentioned specification choices, but the simulation results also indicate that the ToE model is generally able to adjust for a large share of the bias due to unobserved heterogeneity. This holds in our baseline model, where the only source of variation is across cross-sectional units through time-fixed covariates. When we introduce more exogenous variation in the form of time-varying covariates through the unemployment rate in the local labor market (measured at monthly intervals), the bias is further reduced. This holds even when the setting is characterized by substantial heterogeneity induced by omitting a large set of covariates, including a wide range of short- and long-term labor market history variables. The importance of time-varying variables echoes the results in Gaure et al. (2007).

The results on how to specify the distribution of unobserved heterogeneity are not only relevant for ToE models, but also for all kinds of selection models with random

effects, including univariate duration models, general competing risks models,[3] non-parametric maximum likelihood estimators for non-duration outcomes and structural models with unobserved heterogeneity. Univariate duration models with unobserved heterogeneity are used in studies of factors behind duration dependence in aggregate re-employment rates. The latter may be explained by individual-level duration dependence or dynamic sorting of unemployed with low exit probabilities into long-term unemployment (see e.g., Abbring et al., 2001). In labor economics, competing risks models are used in studies of unemployment durations with competing exits to employment and non-employment (e.g. Narendranathan and Stewart, 1993) as well as exits to different types of jobs (Baert et al., 2013; Jahn and Rosholm, 2013). In health economics and epidemiology, two often studied competing risks are disease relapse and death (see e.g., Gooley et al., 1999). Non-parametric maximum likelihood estimators have also been extensively used when modelling non-duration outcomes. One example is consumer choice analysis (Briesch et al. 2010). Other random effects models are multinomial choice models with unobserved determinants (Ichimura and Thompson, 1998; Fox et al., 2012; Gautier and Kitamura, 2013).

Another important contribution of our paper is that we evaluate the relevance of different set of covariates when measuring causal effects of active labor market programs. Not only is this important for evaluations based on conditional independence assumptions (CIA), but also for identification strategies that allow for unobserved heterogeneity, since we also characterize the unobserved heterogeneity that needs to be taken into account. Starting from the seminal work by LaLonde (1986), who mainly focused on the performance of different non-experimental estimators against an experimental benchmark, the relevance of different covariates has been an increasingly studied research topic. In a series of influential papers on the reliability of alternative non-experimental estimators for the evaluation of social programs, Heckman et al. (1996, 1997, 1998) show that local labor market indicators, baseline socio-economic characteristics, county- and metropolitan-level unemployment rates, employment status and earnings in the four quarters prior to participation, and short-term labor force transitions are important predictors of program participation.[4] Evidence from Mueser et al. (2007) is consistent with these findings, and Smith and Todd (2005) additionally

[3]The ToE model is a type of competing risks model where one duration (treatment duration) is assumed to have a causal impact on the other duration (outcome duration). In general, there are many other competing risks models with related unobserved heterogeneity.

[4]They use experimental data from the U.S. National Job Training Partnership Act, as compared to LaLonde (1986), who uses data from the Current Population Survey and the Panel Study of Income Dynamics. See e.g. Smith and Todd (2005) for a details on the differences between the data sources.

stress that information on socio-economic characteristics and past earnings alone is in general not sufficient to eliminate selection bias.

More recently, Lechner and Wunsch (2013) use administrative data on jobseekers in Germany and EMC methods to assess the importance of covariates when analyzing job search assistance and training programs. Their idea is to consider a much more detailed set of covariates as compared to the previous literature, using essentially all variables that are important for the selection process in various CIA-based evaluations of active labor market programs (ALMP). With this data, they perform simulations as in Huber et al. (2013), i.e. they simulate placebo treatments for the non-treated using the full data. Then, to assess the relative importance of alternative covariates, they leave out different blocks of observables, re-estimate the model, and compare the size of the bias (non-zero treatment effect) across specifications.[5] Their findings are in line with the previous literature that stresses the importance of controlling for baseline socio-economic characteristics, timing of the program, region, UI benefits payments, pre-treatment outcomes and short-term labor market history. However, they also show that it is not sufficient to control only for certain aspects of the short-term labor market history. Instead, detailed information is needed on past employment, unemployment, out-of-labor-force condition, earnings, and characteristics of the last job.

We use our EMC-simulated data in a similar way. As a first step, we use the Swedish data to construct detailed covariates as in the German setting of Lechner and Wunsch (2013). This allows us to examine to what extent the results in their setting carry over to other countries and programs. However, we also include additional variables not used by Lechner and Wunsch (2013). First, since we model treatment durations and not binary treatment indicators, we also include previous employment and unemployment durations in the set of covariates. Previous durations may capture aspects related to how long one stays unemployed in a better way than non-duration history variables such as the employment rate over a certain time period. Second, the covariates in Lechner and Wunsch (2013) reflect important aspects of labor market attachment, skills and benefit variables, but more general unobserved skills may also be relevant. To this end, we use parental income, which is a commonly used proxy for such general unobserved skills. Third, since we model the treatment duration, time-varying variables, such as local business cycle conditions, may play a role, especially during longer unemployment spells. Overall, our analyses are complementary to those in Lechner and Wunsch (2013) since we consider a duration outcome framework.

[5]Other studies additionally assess the importance of usually unobserved variables, such as personality traits and attitudes (Caliendo et al., 2017).

When analyzing the relevance of different covariates, we find that short-term labor market history variables are particularly important to adjust for. Moreover, adjusting for employment history is relatively more important than adjusting for unemployment, earnings and welfare history (out-of-labor-force status). We also find that adding information about long-term labor market history (last ten years) on top of controlling for short-term history (last two years) is unimportant. When comparing different short-term employment characteristics, we see that short-term employment history (in particular, the employment rate) is a relevant determinant of program participation, whereas short-term unemployment history is relatively less important.

Our work is also related to two papers by Kastoryano and van der Klaauw (2011) and Muller and van der Klaauw (2017), who use different quasi-experimental and non-experimental dynamic treatment evaluation approaches, including the ToE model, to evaluate job search assistance programs in The Netherlands. Our paper is complementary to theirs in different ways. First, here we study in detail what covariates are relevant predictors of program participation, which is not the focus of Kastoryano and van der Klaauw (2011) and Muller and van der Klaauw (2017). Second, in our simulation exercises, we use an EMC design and exclusively focus on inference based on the ToE model and on how to specify its components. The two above mentioned papers, instead, focus on comparing estimates from different approaches, which, as mentioned, include the ToE model as well as alternatives to it. In doing so, they do not simulate data, nor do they consider the average bias corresponding to a certain degree of generated unobserved heterogeneity, as we do through our EMC design.

The paper proceeds as follows. Section 2 presents the Timing-of-Events model proposed by Abbring and van den Berg (2003). Section 3 describes the simulation design and the data used in the simulations. Section 4 describes the estimated selection model that is used to simulate the placebo treatments and compares the bias when different sets of covariates are included in the model. In Section 5, we present the EMC simulation results, and Section 6 concludes.

2 The Timing-of-Events model

This section presents the ToE approach as introduced by Abbring and van den Berg (2003). They specify a bivariate duration model for the duration in an initial state and the duration until the treatment of interest: T_e and T_p, with t_e and t_p being their realizations The model includes individual characteristics, X, and unobserved individual characteristics V_e and V_p, with realizations (x, v_e, v_p). Abbring and van den

Berg (2003) assume that the exit rate from the initial state, $\theta_e(t|D(t), x, V_e)$, and the treatment rate, $\theta_p(t|x, V_p)$, follow the Mixed Proportional Hazard (MPH) form:[6]

$$\begin{aligned} \ln \theta_e(t|x, D, V_e, t_p) &= \ln \lambda_e(t) + x'\beta_e + \delta D(t) + V_e, \\ \ln \theta_p(t|x, V_p) &= \ln \lambda_p(t) + x'\beta_p + V_p, \end{aligned} \tag{1}$$

where t is the elapsed duration, $D(t)$ is an indicator function taking the value one if the treatment has been imposed before t, δ represents the treatment effect, and $\lambda_e(t)$, $\lambda_p(t)$ capture duration dependence in the exit duration and the treatment duration, respectively. Also, let G denote the joint distribution of $V_e, V_p|x$ in the inflow into unemployment.

Abbring and Van den Berg (2003) show that all components of this model, including the treatment effect, δ, and the unobserved heterogeneity distribution, G, are identified under the following assumptions. The first assumption is no-anticipation, which means that future treatments are not allowed to affect current outcomes. This holds if the units do not know the exact time of the treatment or if they do not react on such information.[7] A second assumption is that X and V should be independently distributed, implying that the observed characteristics are uncorrelated with the unobserved characteristics. A third assumption is the proportional hazard structure (MPH model). We discuss these assumptions in more detail when we describe our simulation design. Abbring and van den Berg (2003) also impose several regularity conditions.

Identification is semi-parametric, in the sense that given the MPH structure, the ToE model does not rely on any other parametric assumptions. Moreover, unlike many other approaches, the ToE method does not require any exclusion restrictions. Instead, identification of the treatment effect follows from the variation in the moment of the treatment and the moment of the exit from the initial state. If treatment is closely followed by an exit from the initial state, regardless of the time since the treatment, then this is evidence of a causal effect, while any selection effects due to dependence of V_p and V_e do not give rise to the same type of quick succession of events. However, this requires some exogenous variation in the hazard rates. The most basic exogenous variation is generated through the time-invariant covariates, x, which create variation in the hazard rates across units. Strictly speaking, this is the only variation that is

[6]This is the most basic ToE model with time-constant and homogeneous treatment effect, but note that Abbring and Van den Berg (2003) also allow for time-varying treatment effects as well as other extensions of this basic model.

[7]The no-anticipation assumption also implies that any anticipation of the actual time of the exit from the initial state does not affect the current treatment rate.

needed for identification.

Previous studies suggest that time-varying variation, i.e. variables that change with the elapsed duration, for instance due to business cycle variation or seasonal variation, is a useful and more robust source of additional exogenous variation (Eberwein et al., 1997; Gaure et al., 2007). The intuition is that such time-varying covariates shift the hazard rates, and this helps to identify the influences of the unobserved heterogeneity. More specifically, current factors have an immediate impact on the exit rate, whereas past factors affect the current transition probabilities only through the selection process (for a more detailed discussion, see van den Berg and van Ours, 1994, and van den Berg and van Ours, 1996). We therefore examine both ToE models with only time-invariant variables and models with time-varying variables.

3 Simulation approach

3.1 The basic idea

The idea behind EMC is to simulate using real data instead of using a data generating process that is entirely specified by the researcher, such as in a typical Monte Carlo study. The argument is that real data is more closely linked to real applications with real outcomes and real covariates, and thus provides arguably more convincing simulation evidence. As a background to our simulation design, consider the EMC design adopted by Huber et al. (2013). They use real data on jobseekers in Germany to compare the performance of alternative estimators of treatment effects under conditional independence. They proceed in the following way. They first use the real data on both treated and non-treated units to capture the treatment selection process. The estimated selection model is then used to simulate placebo treatments for all non-treated units in the sample, effectively partitioning the sample of non-treated into placebo treated and placebo controls. This ensures that the selection process used for the simulations is known and that the conditional independence assumption holds by construction, even if real data is used in the simulations. Moreover, by construction, the true effect of the placebo treatments is zero. Then, Huber et al. (2013) use the resulting simulated data to analyze the performance of various CIA-based estimators.

We tweak this simulation design in some key dimensions with the aim of using the EMC approach to study the ToE model. Our simulations are also based on real data. We use rich Swedish register and survey data of jobseekers, with information on participation in a labor market training program. The outcome duration, T_e, is the time

in unemployment, while the treatment duration, T_p, is time to the training program. The data (described below) is also used to create detailed background information for each unit. Then, we use this data to generate placebo treatments, but we do this in a slightly different way than Huber et al. (2013). In particular, instead of simulating binary treatment indicators as they do, we use a hazard model for the treatment duration, and use this to simulate placebo treatment durations. As for the standard EMC approach, the effect of these placebo treatments is zero by construction. Unobserved heterogeneity is then generated by leaving out blocks of the covariates used in the true selection model. That is, we leave out some covariates that were used when generating the placebo treatment durations. This leads to a bivariate duration model with correlated unobserved determinants, since the excluded covariates affect both the time in unemployment (the outcome) and, by construction, the treatment duration.

The simulated data is used for various simulation exercises. We mainly focus on the estimation of the treatment effect. By construction, the true effect of the placebo treatments is zero, but since we leave out variables and generate correlated unobserved determinants, we introduce bias (estimated non-zero treatment effect). We evaluate important specification issues related to ToE estimation, and primarily focus on their impact on the bias and the variance of the treatment effects estimates. In addition, we also study their impact on other parts of the model. Some of the issues that we study, such as the specification of the unobserved heterogeneity distribution and of the baseline hazard, were raised by previous Monte Carlo simulations studies (Gaure et al., 2007; Baker and Melino, 2000). However, we also study specification aspects that have not been studied before. One example is that we exclude different blocks of covariates to study how the ToE approach performs with different types of unobserved heterogeneity.

One important reason to use the Swedish unemployment spell data is that there are many examples of evaluations that estimate ToE models using this type of data.[8] The use of unemployment spells also affects how we design our simulation study. Unemployment durations and labor market program entries are typically measured at the daily level, which is also the case in our setting. We treat the daily spell data as if it were continuous, and generate the placebo treatment durations measured at the daily level by using a continuous-time selection model. Accordingly, we estimate continuous-time ToE models.[9]

[8]Examples include Abbring et al. (2005), Lalive et al. (2005), Røed and Raaum (2006), Lalive et al. (2008), Kyyrä (2010), Richardson and van den Berg (2013), Kyyrä et al. (2013), Arni et al. (2013), and Van den Berg and Vikström (2014).

[9]Continuous-time models are often estimated in the literature, even when using discrete data (daily,

Next, let us relate our simulated data to the assumptions made in the ToE approach. By construction, the no-anticipation assumption holds, because the units cannot anticipate and react to placebo treatments. However, there are other ToE assumptions that may not hold in this simulation design. First, the assumption requiring independence between X and V (random effects assumption) may not hold in our simulations, since the excluded covariates representing unobserved heterogeneity may be correlated with the covariates that were actually used in the ToE estimation. To explore this, in extended simulations we estimate ToE models when leaving out blocks of covariates that are alternatively highly or mildly correlated with the observables. It turns out that the degree of correlation between the observed and unobserved factors is relatively unimportant. Second, since the outcome duration is not modeled, the outcome hazard (re-employment rate) may not follow the MPH structure. Third, a duration model without embedded unobserved heterogeneity is used to model the treatment selection process. This means that although we use an extremely rich set of covariates to estimate the selection process, if there are some omitted characteristics, the model will be misspecified.

All these three potential violations of the ToE assumption arise because we use a simulation design based on real data, which most likely does not follow a MPH structure. However, one may argue that this is the benefit of our approach, because we explore estimation of the ToE model using arguably more realistic data.

3.2 The relevance of different covariates

The analysis of the ToE model specification is the main contribution of our paper. However, by leaving out different blocks of covariates, we can also evaluate the relevance of different observables when measuring causal effects of active labor market programs. To this end, we use the simulated data with placebo treated and non-treated units, for which the "true" treatment effect is known to be zero. Then, to assess the relative importance of different covariates, we leave out alternative blocks of observables and compare the bias size across the resulting specifications.

This analysis benefits from the rich Swedish data that we use. In creating the variables, we follow Lechner and Wunsch (2013), who claim that they consider all relevant determinants of program participation by focusing on information used in a range of CIA-based evaluations of ALMPs. They use German data, while here we

weekly, monthly or yearly). For instance, this is the case for Palali and van Ours (2017), Tatsiramos (2010), Jahn and Rosholm (2013), Kyyrä et al. (2013), McVicar et AL (2018), Muller et al. (2017), van Ours and Williams (2009), van Ours and Williams (2012), and van Ours et AL (2013).

use Swedish databases to re-construct similar variables. However, we also include additional information not used by Lechner and Wunsch (2013). First, since we model treatment durations and not binary treatment indicators, we also include variables that capture the duration element of employment and unemployment histories. In principle, previous durations may capture aspects of the ongoing unemployment spell in a better way than non-duration history variables. By comparing their inclusion with that of other unemployment and employment history variables, such as the employment rate, we can see if previous durations matter relatively more for the current duration outcomes. Second, the covariates in Lechner and Wunsch (2013) reflect important features of labor market attachment, skills and benefit variables, but more general unobserved skills may also be relevant. To this end, we use parental income, which is a commonly used proxy for such general unobserved skills. Third, since we model the treatment duration, time-varying variables may play an important role. In particular, business cycle conditions change over time, especially during longer unemployment spells. Another difference compared Lechner and Wunsch (2013) is that we consider a duration outcome framework, and use duration models to study the relevance of different blocks of covariates.

Note that this procedure holds under the assumption of CIA with the full set of covariates. Lechner and Wunsch (2013) provide good arguments as to why CIA should be valid in their German setting when they use their full set of covariates, and Vikström (2017) provides similar arguments for Sweden. This can of course always be questioned, for instance because treatment selection is based on unobserved motivation and skills. Thus, we study the relevance of the different observed covariates, keeping in mind that there may also be important information that is not included in our data.

3.3 The training program

One often-studied treatment for jobseekers is labor market training. This motivates our use of data on a Swedish vocational training program called AMU (Arbetsmarknadsutbildning). The program and the type of administrative data that we use resemble those of other countries. The main purpose of the program, which typically lasts for around 6 months, is to improve the skills of the jobseekers so as to enhance their chances of finding a job. Training courses include manufacturing, machine operator, office/warehouse work, health care, and computer skills. The basic eligibility criterion is to be at least 25 years old. During the training, participants receive a grant. Those who are entitled to unemployment insurance (UI) receive a grant equal to their UI benefits level, while

for those not entitled to UI the grant is smaller. In all cases, training is free of charge.

Previous evaluations of the AMU training program include Harkman and Johansson (1999), de Luna et al. (2008), Richardson and van den Berg (2013), and Vikström and van den Berg (2017). These papers describe the program in more detail.

3.4 Data sources and sampling

We combine data from several administrative registers and surveys. The Swedish Public Employment Service provides daily unemployment and labor market program records of all unemployed in Sweden. We use this information to construct spell data on the treatment duration (time to training program) and the outcome duration (time to employment), both measured in days. We sample all unemployment spells starting during the period of 2002–2011.[10] The analyses are restricted to the prime-age population (age 25–55) since younger workers are subject to different labor market programs and to avoid patterns due to early retirement decisions of older workers. We also exclude disabled workers. In total, the sampled spells are 2.6 million, of which 3% involve training participation. The mean unemployment duration in the sample is 370 days. In case a jobseeker enters into training multiple times, only the first instance is considered.

For each spell, we construct detailed information on individual-level characteristics. We start by constructing similar variables as in the German data in Lechner and Wunsch (2013).[11] The population register LOUISE provides basic socio-economic information, such as country of origin, civil status, regional indicators and level of education. Matched employer-employee data (RAMS) and wage statistics from Statistics Sweden are used to construct information on the characteristics of the last job (wages, type of occupation, skill-level), and to retrieve information on the characteristics of the last firm (firm size, industry and average worker characteristics). From Unemployment Insurance (UI) records we obtain information on UI eligibility.

Data from the Public Employment Service is used to construct unemployment history variables. It is also used to construct information on the regional unemployment rate. Earnings records and information on welfare participation are used to construct employment, out-of-labor force and earnings histories. For the history variables, we construct both short-run history (last two years) and long-run history (last ten years). Altogether, these variables capture detailed aspects of the workers' employment and

[10] Any ongoing spells are right-censored on December 31, 2013.

[11] There are some differences between the Swedish and German data. The classification of occupations differs, we lack some firm-level characteristics, and we have less information on UI claims. We also use welfare benefits transfers to construct measures of out-of-labor-force status.

earnings history in the last two and ten years.

As already mentioned, we also include additional variables not used by Lechner and Wunsch (2013). These include previous unemployment and employment durations, the idea being that previous durations may capture the current ones in a better way than the above mentioned employment history variables. To this aim, we construct time spent in the last employment spell, time in the last unemployment spell as well as indicators for no previous unemployment/employment spell. We also study the relevance of controlling for the mother's and father's income, under the assumption that parental income may capture general unobserved skills. We exploit the Swedish multi-generational register (linking children to parents) together with income registers to create information on parental income (father and mother income, averaged over age 35-55 of the parent). Finally, we also explore time-varying variables, and include the local unemployment rate in the region during each month as a time-varying covariate (Sweden has 21 regions).

Finally, the outcome considered in this paper is the re-employment rate (job exit rate). We consider as an exit to employment a transition to a part-time or full-time job that is maintained for at least 30 days.

All variables that are used in the analyses are summarized in Table 1. The statistics in the table show that immigrants from outside Europe, males, married and the less educated jobseekers are overrepresented among the training participants. Training participants also also more likely to be employed in firms with lower wages, and there are fewer previous managers and more mechanical workers among the treated workers. All labor market history measures point in the same direction: training participants have worse unemployment and welfare characteristics in the last two and ten years.

3.5 Simulation details

Selection model. The first step of the EMC design is to estimate the treatment selection model. We use a continuous-time parametric proportional hazard model for the treatment hazard, $\theta_p(t|x)$, at time, t, conditional on a set of covariates, x, which includes time-fixed regressors and time-varying monthly regional unemployment rate:

$$\theta_p(t|x) = \lambda_p(t) \cdot \exp(x\beta_p) \tag{2}$$

The baseline hazard, $\lambda_p(t)$, is taken as piecewise constant, with $\ln \lambda_p(t) = \alpha_m$ for $t \in [t_{m-1}, t_m)$, where m is an indicator for the m^{th} time interval. We use eight time intervals, with splits after 31, 61, 122, 183, 244, 365 and 548 days. The observed

variables, X, include all covariates described in Table 1. The model estimates reported in Table 1 show that the daily treatment rate peaks after roughly 300 days. They also confirm the same patterns found for the sample statistics: immigrants, younger workers, males, high-school graduates, and UI recipients are more likely to be treated. Short- and long-term unemployment and employment history variables are also important determinants of treatment assignment.

After estimating the selection model by using the full population of actual treated and controls (i.e. the never treated), the treated units are discarded and play no further role in the simulations. Next, we use (2) to simulate the placebo times to treatment for each non-treated, T_s, which is generated according to (dropping x to simplify notation):

$$\exp\left(-\int_0^{T_p}\theta_p(\tau)d\tau\right)=U, \tag{3}$$

where $U\sim\mathcal{U}[0,1]$. Since $\theta_p(t)>0\ \forall t$, the integrated hazard $\int_0^{T_p}\theta_p(\tau)d\tau$ is strictly increasing in T_p. By first randomly selecting U for each unit and then finding the unique solution to (3), we can retrieve T_p for each observation.[12]

Simulated treatments that occur after the actual exit from unemployment are ignored. Thus, the placebo treated units are those with a placebo treatment realized before the exit to job. During this procedure, $\hat{\theta}_p(t|x_i)$ is multiplied by a constant γ, which selected such that the share of placebo treated is around 20%. This ensures that there is a fairly large number of treated units in each sample. A similar approach is adopted by Huber et al. (2013).

Simulations. The placebo treatments are simulated for all non-treated units. Next, we draw random samples of size N from this full sample (independent draws with replacement). We set N = 10,000, 40,000 and 160,000 because ToE models are rarely estimated with small sample sizes. If the estimator is N-convergent, increasing the sample size by a factor of 4 (by going from 10,000 to 40,000, or from 40,000

[12]The actual distribution for the integrated hazard will depend on the specification of the selection model (2). In the simple case where all covariates are time-fixed and the placebo treatments are generated by using a proportional hazard model that has two piecewise constant parts, with θ_s^0 for $t\in[0,t_1)$ and θ_s^1 for $t>t_1$:

$$\exp\left(-\int_0^{T_s}\theta_s(\tau)d\tau\right)=\begin{cases}\exp\left(-\int_0^{T_s}\theta_s^0 d\tau\right) & \text{if } U>\exp\left(-\int_0^{t_1}\theta_s^0 d\tau\right)\\ \exp\left(-\int_0^{t_1}\theta_s^0 d\tau-\int_{t_1}^{T_s}\theta_s^1 d\tau\right) & \text{otherwise}\end{cases}$$

This can be easily extended to the case where the baseline hazard has more than two locally constant pieces and where X contains time-varying regressors (in both cases, the integrated hazard shifts in correspondence of changes in such regressors over calendar- or duration-time).

to 160,000) should reduce the standard error by 50%. For each ToE specification we perform 500 replications.

3.6 Implementation of the bivariate duration model

We estimate a continuous-time ToE model for the treatment and outcome hazards as defined in Equation (1). The unknown distribution of the unobserved heterogeneity is approximated by a discrete mass points distribution (Lindsay, 1983; Heckman and Singer, 1984; Gaure et al., 2007).

Likelihood function. For each unit $i = 1, \ldots, N$ we formulate the conditional likelihood contribution, $L_i(v_i)$, conditional on the vector of unobserved variables $v_i = (v_{ei}, v_{pi})$. Then, the individual likelihood contribution, L_i, is obtained by integrating $L_i(v_i)$ over the distribution of the unobserved heterogeneity, G. For the duration dependence ($\lambda_e(t)$, $\lambda_p(t)$), we use a piecewise constant specification with $\lambda_s(t) = \exp(\alpha_{sm})$ where the spell-duration indicators are $\alpha_{sm} = \mathbb{1}\left[t \in [t_{m-1}, t_m)\right]$, for $m = 1, \ldots, M$ cut-offs. In the baseline setting we fix the cut-offs to $31, 61, 122, 183, 244, 365, 548, 2160$. The actual covariates X used in the model are explained in the next section.

To set up $L_i(v_i)$, we split the spells into parts where all right-hand side variables in (1) are constant. Splits occur at each new spell-duration indicator and when the treatment status changes. In all baseline ToE specifications the variables specified are calendar-time constant. In additional specifications where the time-varying local unemployment rate is included, calendar-time variation leads to additional (monthly) splits. Spell part j for unit i is denoted by c_{ij}, and has length l_{ij}. Let C_i be the set of spell parts for unit i. Each part, c_{ij}, is fully described in terms of l_{ij}, α_{sm}, x_i and the outcome indicator, y_{sij}, which equals one if the spell part ends with a transition to state s and zero otherwise. There are two such possible states (job exit and treatment start). Then, with approximately continuous durations, $L_i(v_i)$ is:

$$L_i(v_i) = \prod_{c_{ij} \in C_i} \left[\exp\left(-l_{ij} \sum_{s \in S_{it}} \theta_s(t, x_i, D_{it}, v_{si}|\cdot) \right) \times \prod_{s \in S_{it}} \theta_s(t|\cdot)^{y_{sij}} \right], \tag{4}$$

with

$$\theta_s(t|\cdot) = \begin{cases} \lambda_e(t)\ \exp(x_i'\beta_e)\ \exp(\delta D_{it})\ v_{ei} \\ \lambda_p(t)\ \exp(x_i'\beta_p)\ v_{pi}. \end{cases}$$

L_i is obtained by integrating $L_i(v_i)$ over $G(V)$. Let p_w be the probability associated with support point, w, with $w = 1, \ldots, W$, such that $\sum_{w=1^W} p_w = 1$. Then, the

log-likelihood function is:

$$\mathcal{L} = \sum_{i=1}^{N} \left(\sum_{w=1}^{W} p_w \ln L_i(v_w) \right) \equiv \sum_{i=1}^{N} L_i. \tag{5}$$

Search algorithm. In order to estimate the discrete support points, we use the iterative search algorithm in Gaure et al. (2007). For each replication we estimate models with up to $\overline{W}$ support points. We can then select the appropriate model using alternative information criteria (see below). Let $\hat{\vartheta}_W$ be the maximum likelihood (ML) estimate with W support points. The search algorithm is:

Step 1: Set $W = 1$ and compute the ML estimate $\hat{\vartheta}_W$.

Step 2: Increment W by 1. Fix all ϑ_W elements but (v_W, p_W) to $\hat{\vartheta}_{W-1}$. Use the simulated annealing method (Goffe et al., 1994) to search for an additional support point, and return the $(\tilde{v}_W, \tilde{p}_W)$ values for the new support point.

Step 3: Perform ML maximization with respect to the full parameters vector $\vartheta_W = (\beta, v, p)$ by using $\hat{\vartheta}_{W-1}$ and $(\tilde{v}_W, \tilde{p}_W)$ as initial values. Return $\hat{\vartheta}_W$.

Step 4: Store $\{\hat{\vartheta}_W,\ \mathcal{L}(\hat{\vartheta}_W)\}$. If $W < \overline{W}$ return to Step 2, else stop.

Step 1 corresponds to a model without unobserved heterogeneity, since $\hat{v}$ cannot be distinguished from the intercept in X. In *Step 2* the algorithm searches for a new support point in the $[-3, 3]$ interval.[13] In this step, all other parameters of the model are fixed. This explains why in *Step 3* we perform a ML maximization over all parameters, including the new support point. At the end of the procedure we obtain $\overline{W}$ maximum likelihood estimates: $\{\hat{\vartheta}_W,\ \mathcal{L}(\hat{\vartheta}_W)\}_{W=1}^{\overline{W}}$.

Information criteria. We use different approaches to choose between the $\overline{W}$ estimates. First, we report results where we pre-specify the number of support points (up to six points). An alternative approach is to increase the number of support points until there is no further improvement in the likelihood (ML criterion).

We also use information criteria that penalize parameter abundance. Specifically, the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the Hannan-Quinn information criterion (HQIC). The latter two are more restrictive since they impose a larger penalty on parameter abundance. Formally, $AIC = \mathcal{L}(\hat{\vartheta}_W) -$

[13] As starting values we set $v_W = 0.5$ and $p_W = \exp(-4)$. The simulated annealing is stopped once it finds a support point with a likelihood improvement of at least 0.01. In most cases, the algorithm finds a likelihood improvement within the first 200 iterations.

k, $BIC = \mathcal{L}(\hat{\vartheta}_W) - 0.5k \cdot \ln N$ and $HQIC = \mathcal{L}(\hat{\vartheta}_W) - k \cdot \ln(\ln N)$, where $k \equiv k(W)$ is the number of estimated model parameters and N is the total number of spell parts used for the estimation.[14] The ML criterion is defined as $ML = \mathcal{L}(\hat{\vartheta}_W)$, where only likelihood increases greater than 0.01 are considered. The criteria are calculated for each replication, so that the selected number of support points may vary both across replications and criteria. This allows us to compute the average bias and the mean square error for all information criteria. Throughout our analyses we report the absolute level of bias (i.e. the placebo effect estimates).

4 Available covariates and evaluations of ALMPs

We now evaluate the relevance of different types of covariates. Specifically, we leave out various blocks of covariates and compare the size of the bias – the difference between the estimated treatment effect and the true zero effect of the placebo treatments – across specifications. All covariates are a subset of those used to generate the placebo treatments. For each specification, the full sample of placebo treated and placebo non-treated units is used to estimate a parametric proportional hazard (PH) model. Here, the baseline hazard is specified in the same way as for the model used to simulate the placebo treatments.[15] Table 1 lists all covariates in each block.

The main results are given in Table 2. In each panel of the table, we start with the covariates from the proceeding panels and add additional information to the covariates already in the model, so that the model is extended sequentially by adding blocks of covariates one by one. We add the covariates in a similar order as Lechner and Wunsch (2013), who argue that the order resembles the ease, likelihood and cost of obtaining the respective information. This will, for instance, reveal the relevance of adding information on long-term labor market history on top of more basic variables such as short-term history and baseline socio-economic characteristics.

In Panel A, we start with a baseline model with a set of baseline socio-economic characteristics, which returns a positive and sizable bias of around 6.9%. That is, the estimated treatment effect is 0.069 when the true effect of these placebo treatments is equal to zero. Additionally controlling for calendar time (inflow year and month dummies) and regional information (regional dummies and local unemployment rate

[14] We follow Gaure et al. (2007) and use the grand total number of spell parts. N can be alternatively used, but our simulations indicate that this is of minor importance in practice.

[15] We have also estimated the bias using other duration models, including a Cox-model, leading to similar results.

at inflow) reduces the bias from 6.9% to 6.2%.[16] Since the corresponding excluded covariates include short- and long-term labor market history, the positive bias means that training participants tend to have more favorable labor market histories.

Panel B compares the relevance of short-term employment, unemployment, earnings and welfare benefit histories. Here, we compare the relevance of entire blocks of covariates, while later we do so for individual variables, such as previous employment rates against employment durations. All blocks of short-term history variables reduce the bias. However, adjusting for short-term employment history is relatively more important than adjusting for unemployment, earnings and welfare history (out-of-labor-force status). If we adjust for unemployment history and earnings history, the bias drops to 5.0% and 4.0%, respectively, whereas if the model includes employment history the bias is much closer to zero. In fact, the sign of the bias is even reversed (slightly negative, -1.4%) when adjusting for short-term employment history. These results indicate that participants in labor market training are to a large extent selected based on their previous employment records. One explanation may be that caseworkers aim to select jobseekers with an occupational history aligned with the vocational training program.

Table 3 examines individual short-term employment and unemployment variables. They are added in addition to the baseline covariates in a similar way as in Panel B of Table 2. The aim is to understand what specific aspects of employment and unemployment are the most important to adjust for. In our comparisons we control for either past employment duration, different measures of the share of time spent in employment (employment rate), employment status at a given point in time, or other history variables. A reason for this exercise is that we model treatment durations and not a binary treatment status. Accordingly, it may be the case that previous durations capture aspects of the ongoing unemployment spell in a better way than previous employment rates and employment status at a given point in time. We also compare the relevance of similarly constructed short-term unemployment history variables.

The results show that information on previous employment duration reduces the bias considerably: from 6.2% in the baseline specification to 3.9% (Panel A). However, adding information on past employment rates or other short-term employment history variables reduces the bias even more, leading to biases of -0.04% and 0.2%, respectively (Panel B and C). In particular, Panel B shows that all variables measuring past employment rate single-handedly capture a large part of the bias. For instance, by only adjusting for months employed in the last six months before the unemployment spell,

[16]For completeness, we also report estimates when using these time and regional variables only, without including the baseline socio-economic characteristics. This leads to larger bias.

the bias reduces from 6.2% to 1.7%. Panel C also shows that employment status one year before the unemployment spell has a similar impact on the bias. On the other hand, employment status two years before the spell and other short-term employment variables appear to be less important. Interestingly, the bias is positive or close to zero in all cases, so that the reversal of the bias sign that was observed in Panel B of Table 2 occurs only once all short-term employment variables are included together. That is, even if some short-term history variables are more relevant, they all capture different aspects of the selection process, so that adjusting for both previous employment durations and rates is important.

Panels D to F of Table 3 report estimates from a similar exercise where we control for the short-term unemployment history and duration variables one at a time. This confirms that unemployment history variables have a modest impact on the estimated bias compared to the employment history variables. For instance, while adjusting for previous employment duration reduced the bias from 6.2% to 3.9%, now including previous unemployment duration only reduces the bias from 6.2% to 5.5%. All in all, this suggests that for training programs with emphasis on human capital accumulation, the most important characteristics to control for are those related to employment history.

Next, we return to Table 2. Here, Panel C shows that adding information on long-term labor market history (last 10 years) on top of short-term history (last two years) has a minor impact on the bias of the estimated treatment effect. The same holds when in Panel D we adjust for various characteristics of the last job (e.g., previous wage and occupation) as well as for detailed information about the last firm (e.g., industry and composition of worker). Lechner and Wunsch (2013) and before them Heckman et al. (1996, 1997, 1998) also find that, after controlling for calendar time, local labor market conditions and short-term labor market history, adding additional covariates such as long-term labor market history is relatively unimportant. One difference compared to Lechner and Wunsch (2013) is that in our setting, adjusting for short-term employment history is enough to obtain small bias, whereas they find that it is important to also adjust for all aspects of the short-term history (employment, unemployment, out-of-labor-force status, earnings, and non-firm characteristics of the last job) in order to obtain a low bias. Our results are also in line with those of Heckman et al. (1996, 1997, 1998), who stress the importance of controlling for socio-economic characteristics, timing of the spell start, local labor market information, and short-term labor market history. However, here we have access to more extensive information to study how alternative types of labor market history variables impact on the bias.

Finally, Panel D examines the relevance of parental income, the idea being that

father's and mother's income proxy for more general unobserved skills. This may be important if unobserved skills are not captured by the variables discussed so far, which are mainly related to labor market attachment. However, parents' income turns out to have limited impact on the bias, indicating that labor market histories are also able to capture more general unobserved skills.[17]

5 Specification of ToE models

This section presents the main simulation results. The focus here is on the estimation of (placebo) treatment effects. We study to what extent the ToE model is able to adjust for the bias observed in the previous section, and which specification of the model leads to the best results. Results are presented in the form of average bias, variance of the placebo estimates, and mean squared error (MSE). In the main analyses, the only source of variation used for identification is provided by time-fixed covariates. In Section 5.7, we introduce additional exogenous variation deriving from time-varying local unemployment rate.

5.1 Baseline results

Table 4 reports results from the baseline simulations where we compare different specifications of the discrete unobserved heterogeneity distribution. In these simulations we adjust for baseline socio-economic characteristics, inflow time dummies, regional indicators and unemployment rate (the covariates in Panels A–B, Table 1).[18] First, consider the results for a sample size of 10,000 in Columns 1–3. In Panel A, we fix the number of support points to a pre-specified number in all replications. The first row shows that the baseline model without unobserved heterogeneity (one support point) leads to large bias (6.0%).[19] This confirms that under-correcting for unobserved heterogeneity may lead to substantial bias.

However, two support points already reduce the bias from 6.0% to 2.7%.[20] For three

[17]This confirms the results in Caliendo et al. (2017), who find that once one controls for rich observables of the type that we include here, additional (usually unobserved) variables measuring personality traits and preferences become redundant.

[18]Here, we control for time-fixed regional unemployment rate (measured as the month of inflow into unemployment). In Table 8, we estimate ToE models where this covariate varies on a monthly basis.

[19]This is roughly the same bias as in the corresponding model estimated with the full sample in Panel A of Table 2. The minor difference is due to sampling variation since here we report the average bias from random drawings, whereas estimates in Table 2 are obtained from the full set of placebo treated and non-treated observations.

[20]Here, we focus on the bias of the treatment effect, but previous simulation studies using simulated

or more support points, the average bias is even larger and keeps increasing in the same direction when adding additional support points. In fact, with six support points the average bias (6.4%) is more than twice as large as the average bias with two support points (2.7%). Moreover, both the variance and the MSE increase in the number of support points (Columns 2–3). The increased bias due to too many support points is in line with the results from Baker and Melino (2000), who argue that specifications with too many (spurious) mass points tend to over-correct for unobserved heterogeneity. This happens because too many support points lead to an overly-dispersed distribution of unobserved heterogeneity. Thus, in order to fit the data, the model compensates this with changes (bias) in the treatment effect, and presumably also in the duration dependence. This pattern contradicts the general intuition that one should adjust for unobserved heterogeneity in the most flexible way in order to avoid bias due to unaccounted unobserved heterogeneity.

To better understand the over-correction pattern due to too many spurious support points, Figure 1 shows the distribution of the treatment effect estimates for one, two and six support points. With one support point, the estimates are centered around a bias of around 6% and the variance of the estimates is rather low. With two support points the entire distribution shifts towards zero (although the average bias is non-zero), but the variance gets larger than for one support point. With six support points, there is a further increase in the variance. Perhaps more importantly, the entire distribution of the estimates shifts to the right (larger positive bias). This shows that the increased bias is not explained by a few extreme estimates. Instead, the overly-dispersed distribution of the unobserved heterogeneity has a more general effect for almost all replications.

Interestingly, the problem with over-correcting for unobserved heterogeneity does not occur to the same extent in the simulated data used by Gaure et al. (2007). They highlight that the main problem is under-correction with too few support points.[21] Our simulation results that are based on real data, instead, suggest that both under- and over-correction are important problems when estimating ToE models. Thus, finding a way to select the appropriate number of support points appears to be important.

data show that failing to account for unobserved heterogeneity also leads to biased spell-duration and covariate effects (see e.g. Gaure et al., 2007).

[21] In their main simulations, Gaure et al. (2007) find no evidence that too many support points over-correct for unobserved heterogeneity. However, when they reduce the sample size they also find evidence of over-correction. Here, the sample size is 10,000 observations, but we will show that over-correction is a problem also with larger sample sizes.

5.2 Information criteria

Panel B of Table 4 provides simulation results when the distribution of the unobserved heterogeneity (number of support points) is specified by using alternative information criteria. Panel C reports the average number of support points that are selected according to each criterion. The ML criterion, where the number of support points is increased as long as the likelihood is improved, leads to 4.11 support points on average. The bias and variance are large compared to simply pre-specifying a low number of support points. Hence, the ML criterion tends to select too many support points, leading to an over-correction problem (too many spurious support points are included). As a result, criteria with little penalty for parameter abundance, such as the ML criterion, should be avoided altogether.

The results for AIC, BIC and HQIC are much more encouraging. All three criteria produce models with rather few unobserved heterogeneity support points (often two support points). These specifications corresponds to those with the lowest bias achieved when pre-specifying a low number of support points. We conclude that these more restrictive information criteria protect against over-correction problems due to too many support points. They do so by penalizing the number of parameters in the discrete heterogeneity distribution. They also guard against under-correction problems (too few support points) by favoring models with unobserved heterogeneity over models without unobserved heterogeneity (one support point).

A comparison between the AIC, BIC and HQIC criteria reveals rather small differences. As expected, the two more restrictive information criteria (BIC and HQIC) lead to models with fewer support points, and the average bias is slightly lower than for the less restrictive AIC criterion. The variance is also slightly lower for BIC and HQIC than for AIC. This is because these more restrictive criteria tend to select fewer support points and the variance of the estimated treatment effects is increasing in the number of support points. However, later we will see that none of the three criteria is superior in all settings. All three penalize parameter abundance, and this protects against problems of over-correction due to spurious support points. In some cases, the risk of under-correcting is relatively more important, and this favors the less restrictive AIC criterion. In other cases, the opposite holds, and this favors the more restrictive BIC and HQIC criteria. Thus, using all three criteria and reporting several estimates as robustness check appears to be a reasonable approach.

The main interest here is in providing background information on the alternative specification choices. However, Table 4 also provides some insights on the overall idea of using ToE models to adjust for unobserved heterogeneity. In general, the table shows

that the ToE approach corrects for a large share of the bias, which is reduced from 6.0% for the model without unobserved heterogeneity to around 2.7% when information criteria are used to select the number of support points (see Column 1 of Table 4). This holds even though the only source of exogenous variation derives from time-fixed covariates in X. In subsequent analyses, we explore if additional sources of exogenous variation in the form of calendar-time variation are able to eliminate the bias entirely.

5.3 Sample size

In Columns 4–6 and 7–9 of Table 4, the sample size is increased to 40,000 and 160,000 observations, respectively. For both these sample sizes we see that two support points are associated with the lowest bias, but here the increase in the bias after three support points is smaller than for 10,000 observations. For instance, with 10,000 observations, going from two to six support points increases the bias from 2.7% to 6.4%, and with 40,000 observations, it increases from 2.2% to 3.7%. For the largest sample with 160,000 observations, the increase in the bias when going from 2 to 6 mass points is even smaller. This suggests that over-correction due to too many support points tends to be a problem with small samples. Note that what constitutes a small sample size most likely differs across applications. For instance, it might be related to the number of parameters in the model, the fraction of treated units, the number of exit states, and the variation in the observed covariates.

Another result is that for larger sample sizes there are smaller differences between the ML criterion and the three other information criteria. For instance, with a sample size of 160,000, there are virtually no differences in the average bias between the four information criteria.

5.4 Excluded covariates

We next vary the unobserved heterogeneity by excluding different sets of covariates when estimating the ToE models. In the baseline simulations, the ToE model includes baseline socio-economic characteristics, inflow time dummies and regional information. Here, we generate more unobserved heterogeneity by excluding additional covariates (all the socio-economic characteristics reported in Panel A of Table 1) and less heterogeneity by excluding fewer covariates (previous earnings). Table 3 shows that these models generate a bias of 9.5% and 4.0%, respectively, in the full sample of placebo treated and controls (Panels A and B). These values can be compared to the bias of 6.2% in the baseline setting.

Columns 1–3 of Table 5 report the results for the model with more extensive unobserved heterogeneity. Again, the ToE model adjusts for a large share of the bias due to unobserved heterogeneity. For instance, with a sample size of 10,000, the bias for the specification without unobserved heterogeneity is 9.4%, but it drops to 2–3% when we adjust for unobserved heterogeneity using the AIC, BIC or HQIC criteria (Panel A). As before, these more restrictive criteria return the lowest bias, whereas the ML criterion leads to a model with too many support points. We obtain similar results with 40,000 observations, but here the difference between the ML criterion and the other criteria is small. Again, this is consistent with previous results. It confirms that it is important to use an appropriate information criterion to select the number of support points, because this avoids problems with over-correction due to spurious support points.

Overall, the specification with less substantial unobserved heterogeneity, obtained by excluding fewer covariates, produces similar patterns (Columns 4–6 of Table 5). The main difference concerns the relative performance of the AIC, BIC and HQIC criteria. Consider the results for a sample size of 40,000. With more extensive unobserved heterogeneity (Columns 1–3), the bias for the AIC criterion is 0.9%, whereas it is 1.8% and 1.9% for the BIC and HQIC criteria. This suggests that the more restrictive information criteria (BIC and HQIC) may under-correct for unobserved heterogeneity by favoring models with too few support points, and this leads to larger bias. This pattern is reversed when we create less substantial unobserved heterogeneity by excluding fewer covariates (Columns 4–6). Here, the average bias is lower for the more restrictive BIC and HQIC criteria than for AIC. This is because for this specification, there likely is a larger risk of over-correcting for unobserved heterogeneity, which leads to better bias performance for the criteria with a larger penalty for parameter abundance. From this, we conclude that neither one of the information criteria is superior in all settings.

5.5 Degree of correlation between X and V

Since we use single-spell data, identification of the ToE model requires independence between the included regressors and the unobserved heterogeneity (random effects assumption). This may not hold in our setting, because we create unobserved heterogeneity by leaving out certain blocks of covariates, and these excluded covariates may be correlated with those that we include when we estimate the ToE model. We therefore perform additional simulation exercises leaving out different covariates from the model. We consider three settings with strongly positive, mildly positive and negative corre-

lation between the covariates used in the ToE model and the excluded covariates.[22] We select variables so that the starting bias, corresponding to the specifications with 1 mass point (no unobserved heterogeneity), is similar across the alternative degrees of correlation (between 4.4% and 4.8%).

Panel A of Table 6 shows the simulation results with samples of size 10,000. It shows that the information criteria perform similarly as before. The ML criterion selects a larger number of mass points which leads to larger bias, and the AIC, BIC and HQIC criteria select more parsimonious models characterized by lower bias than for the ML criterion. Importantly, this holds regardless of the degree of correlation between the observed and the unobserved variables. It holds with a strong positive correlation (Columns 1–3), mildly positive correlation (Columns 4–6) and negative correlation (Columns 7–9). This is reassuring: even when the variables left out from the model are largely related with those left in the ToE model, the relative performance of the information criteria does not appear to be affected. We obtain similar results when drawing samples of size 40,000 (Panel B of Table 6).

5.6 Approximation of the unobserved heterogeneity

So far we have focused on the estimation of the treatment effect. The overall performance of the ToE model can be also checked by inspecting to what extent the estimated discrete distributions for the unobserved heterogeneity approximates the true one. To examine this, we focus on the estimation of the unobserved heterogeneity for the treatment duration. For this duration, the true unobserved heterogeneity is known since we create it by leaving out certain blocks of covariates. However, since we do not simulate the outcome durations, the exact composition of V_e is unknown.

Specifically, for each actual treated and control unit, we use the coefficients of the estimated selection model reported in Table 1 to compute the linear predictor of the variables left out from the model. We compare the first two moments with the corresponding moments for the estimated unobserved heterogeneity from the ToE models (with samples of size 10,000). We include the estimated constant in the linear predictor, which leads to relatively small values of both true and approximated $\exp(V_p)$.

The results from this exercise are shown in Table 7. The table reports results for the

[22]To compute the correlation, we use the estimates from the selection model with all regressors described in Table 1. Then for each cross-sectional unit, the estimated parameters are used to compute the linear predictor of the excluded covariates. This linear predictor equals V in the simulation. Finally, we correlate this with the observed covariates used in the model (linear predictor of all included covariates). This produces one measure of the correlation between the observed and unobserved variables in the model.

true unobserved heterogeneity (Panel A) and the estimated unobserved heterogeneity (Panels B–C). Panel B shows that larger numbers of mass points tend to overestimate the dispersion of the unobserved heterogeneity. On the other hand, the mean of the unobserved heterogeneity distribution tends to be slightly underestimated, regardless of the number of mass points chosen. Panel C indicates that the ML criterion returns an unobserved heterogeneity with too large variance when compared to the true variance, whereas for the more restrictive information criteria (AIC, BIC and HQIC) the variance is too small. However, overall, the ToE model appears to approximate well the true underlying unobserved heterogeneity distribution of the selection model.[23]

5.7 Exogenous variation

Identification of the ToE model requires variation in the observed exogenous covariates, which is needed to produce exogenous changes in the hazard rates. This was the only source of exogenous variation exploited in the baseline simulations above. It resulted in several insights on how to specify the unobserved heterogeneity distribution when estimating ToE models. Overall, we found that the ToE model was able to adjust for a large part of the selection due to unobserved heterogeneity, but it did not eliminate the bias entirely. For this reason we now consider an additional source of identification in the form of time-varying covariates (local unemployment rate).[24] The idea is that time-varying covariates should be useful for identification since they generate shifts in the hazard rates that help to recover the distribution of the unobserved heterogeneity. The results from this exercise are presented in Table 8. The first row of Panel A shows that the bias without adjusting for unobserved heterogeneity (one support point) is 5.6%. As before, additional support points are then stepwise included (Panel A). The results confirm what was found in the baseline simulations. First, if we under-correct for unobserved heterogeneity (no unobserved heterogeneity) this leads to sizable bias; if we over-correct for unobserved heterogeneity the bias is also large. Second, the ML criterion tends to select models with an overly-dispersed unobserved heterogeneity distribution, which is associated with large bias. Third, the three criteria that penalize parameter abundance (AIC, BIC and HQIC) all perform well, since they lead to models

[23]Note that all information criteria select the number of support points based on the joint assessment of the treatment and outcome equations. This complicates the interpretation of whether a given model fits the unobserved heterogeneity in the best way, since as mentioned we do not know the true unobserved heterogeneity distribution for the outcome equation.

[24]The time-varying unemployment rate is measured at monthly level and varies across counties (län). We refer to it as local unemployment rate. This time-varying covariate was included in the selection model to simulate the placebo treatments. On the other hand, the (time-fixed) local unemployment rate measured at the inflow month was included among the regressors throughout the main analyses.

characterized by low bias.

One important difference compared to the baseline simulations is that the average bias for the BIC and HQIC are now closer to zero. This confirms that exploiting time-varying covariates greatly helps identifying the model parameters. Note that this result holds even though we have generated substantial and complex heterogeneity by omitting a large number of covariates, including a wide range of short- and long-term labor market history variables, as well as firm characteristics and attributes of the last job. This produced substantial bias in the model without unobserved heterogeneity. The importance of variation induced by time-varying covariates echoes the results from Gaure et al. (2007), who reach a similar conclusion, the only difference being that they use calendar-time dummies whereas we exploit time-varying local unemployment rate.

6 Conclusions

In this paper, we have modified a recently proposed simulation technique, the Empirical Monte Carlo approach, to evaluate the Timing-of-Events model. This method allowed us to exploit rich administrative data to generate realistic placebo treatment durations, overcoming the common critique that standard simulation studies are sensitive to the data generating process chosen by the researcher.

For ToE models, one key issue is the specification of the discrete support points distribution for the unobserved heterogeneity. From our simulations, we conclude that information criteria are a reliable way to specify the support points distribution in the form of the number of support points to include in the model. This holds as long as the criteria include a substantial penalty for parameter abundance. Instead, information criteria with little penalty for parameter abundance, such as the ML criterion, should be avoided altogether. Three criteria, which all perform well, are the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the Hannan-Quinn information criterion (HQIC). All three protect both against over-correction for unobserved heterogeneity (due to the inclusion of spurious support points) and against under-correction due to insufficient adjustment for unobserved heterogeneity. On the other hand, we show that no single criterion is superior in all settings. Overall, these results hold under different types of unobserved heterogeneity. The model is also generally able to approximate well the true underlying unobserved heterogeneity distribution of the treatment equation. Another key conclusion is that exogenous variation in the form of calendar-time variation proves to be useful for identification. On the other hand, ToE models that only rely on variation in the observed covariates also tend to

produce good results, as long as an appropriate information criterion is used.

The EMC simulation design was also exploited to inspect which covariates are important confounders that need to be controlled for when estimating selection models. Our results confirm the general lessons learned from Heckman et al. (1996, 1997, 1998) and confirmed by Smith and Todd (2005), among others. We find that it is important to control for socio-economic characteristics, timing of inflow into unemployment, local labor market (regional) indicators and unemployment rate, and short-term labor market history.We also confirm that long-term history variables are generally less important once the above mentioned characteristics are accounted for. Moreover, in line with Lechner and Wunsch (2013), we find that the short-term labor market history characteristics considered should be sufficiently detailed. As in their case, controlling for detailed long-term labor market histories or other (relatively harder to collect) information does not help to further reduce the bias. On the other hand, differently from Lechner and Wunsch (2013), we do not achieve further bias reduction when adding short-term unemployment, out-of-labor force and earnings histories to the short-term labor market history information. Moreover, once we control for the above-mentioned variables, we do not find the information on UI benefits or on the last job to be relevant confounders. In our setting, the joint inclusion of all other controls likely proxies for this information.

We also inspect the relevance of very narrowly defined blocks of employment and unemployment history variables. Here, we confirm the result that in our setting information on unemployment history is generally less relevant that information on employment history. We also show that some specific short-term employment variables, appropriately added to the baseline information defined above, are able to reduce the bias to a very large extent, and in some cases they completely eliminate it. This is for instance the case for the share of time spent in employment in the past two years. Overall, this is good news, especially in view of the fact that other detailed information on parents' income and from linked employer-employee databases that we have access to might be in general hard to retrieve in other settings.

References

Abbring, Jaap H., Jan C. van Ours, and Gerard J. van den Berg (2005). "The effect of unemployment insurance sanctions on the transition rate from unemployment to employment". In: *The Economic Journal* 115.505, pp. 602–630.

Abbring, Jaap H. and Gerard J. van den Berg (2003). "The nonparametric identification of treatment effects in duration models". In: *Econometrica* 71.5, pp. 1491–1517.

Abbring, Jaap H, Gerard J Van Den Berg, and Jan C Van Ours (2001). "Business Cycles and Compositional Variation in U.S. Unemployment". In: *Journal of Business & Economic Statistics* 19.4, pp. 436–448.

Advani, Arun, Toru Kitagawa, and Tymon Słoczyński (2018). "Mostly Harmless Simulations? On the Internal Validity of Empirical Monte Carlo Studies". In: *IZA Discussion Papers, No. 11862*.

Arni, Patrick, Rafael Lalive, and Jan C. Van Ours (2013). "How effective are unemployment benefit sanctions? Looking beyond unemployment exit". en. In: *Journal of Applied Econometrics* 28.7, pp. 1153–1178.

Baert, Stijn, Bart Cockx, and Dieter Verhaest (2013). "Overeducation at the start of the career: Stepping stone or trap?" en. In: *Labour Economics* 25, pp. 123–140.

Baker, Mikael and Angelo Melino (2000). "Duration dependence and nonparametric heterogeneity: A Monte Carlo study". In: *Journal of Econometrics* 96, pp. 357–393.

Bergemann, Annette, Laura Pohlan, and Arne Uhlendorff (Aug. 2017). "The impact of participation in job creation schemes in turbulent times". en. In: *Labour Economics* 47, pp. 182–201.

Bijwaard, Govert E., Christian Schluter, and Jackline Wahba (July 2014). "The Impact of Labor Market Dynamics on the Return Migration of Immigrants". en. In: *Review of Economics and Statistics* 96.3, pp. 483–494.

Bodory, Hugo et al. (2016). "The finite sample performance of inference methods for propensity score matching and weighting estimators". In: *IZA Discussion Papers, No. 9706*.

Briesch, Richard A., Pradeep K. Chintagunta, and Rosa L. Matzkin (2010). "Nonparametric Discrete Choice Models With Unobserved Heterogeneity". In: *Journal of Business & Economic Statistics* 28.2, pp. 291–307.

Busk, Henna (2016). "Sanctions and the exit from unemployment in two different benefit schemes". In: *Labour Economics* 42, pp. 159–176.

Caliendo, Marco, Steffen Künn, and Arne Uhlendorff (Oct. 2016). "Earnings exemptions for unemployed workers: The relationship between marginal employment, unemployment duration and job quality". en. In: *Labour Economics* 42, pp. 177–193.

Caliendo, Marco, Robert Mahlstedt, and Oscar A. Mitnik (2017). "Unobservable, but unimportant? The relevance of usually unobserved variables for the evaluation of labor market policies". In: *Labour Economics* 46, pp. 14–25.

Crépon, Bruno et al. (2018). "Information shocks and the empirical evaluation of training programs during unemployment spells". en. In: *Journal of Applied Econometrics* 33.4, pp. 594–616.

De Luna, X., Anders Forslund, and Linus Liljeberg (2008). "Effects of vocational labor market training for participants in the period 2002–04 (in Swedish)". In: *IFAU working paper*.

Eberwein, Curtis, John C. Ham, and Robert J. Lalonde (Oct. 1997). "The Impact of Being Offered and Receiving Classroom Training on the Employment Histories of Disadvantaged Women: Evidence from Experimental Data". en. In: *The Review of Economic Studies* 64.4, pp. 655–682.

Fox, Jeremy T. et al. (2012). "The random coefficients logit model is identified". In: *Journal of Econometrics* 166.2, pp. 204–212.

Frölich, Markus, Martin Huber, and Manuel Wiesenfarth (2017). "The finite sample performance of semi- and non-parametric estimators for treatment effects and policy evaluation". en. In: *Computational Statistics & Data Analysis* 115, pp. 91–102.

Gaure, Simen, Knut Røed, and Tao Zhang (2007). "Time and causality: A Monte Carlo assessment of the timing-of-events approach". en. In: *Journal of Econometrics* 141.2, pp. 1159–1195.

Gautier, Eric and Yuichi Kitamura (2013). "Nonparametric Estimation in Random Coefficients Binary Choice Models". In: *Econometrica* 81.2, pp. 581–607.

Goffe, William L., Gary D. Ferrier, and John Rogers (1994). "Global optimization of statistical functions with simulated annealing". In: *Journal of Econometrics* 60.1-2, pp. 65–99.

Gooley, Ted A et al. (1999). "Estimation of failure probabilities in the presence of competing risks: new representations of old estimators". In: *Statistics in Medicine* 18, pp. 695–706.

Harkman, A. and A. Johansson (1999). "Training or Subsidized Jobs–What Works?" In: *Working paper, AMS, Solna*.

Heckman, J. and B. Singer (1984). "A Method for minimizing the impact of distributional assumptions in econometric models for duration data". In: *Econometrica* 52.2, p. 271.

Heckman, J. J. et al. (Nov. 1996). "Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method". en. In: *Proceedings of the National Academy of Sciences* 93.23, pp. 13416–13420.

Heckman, James et al. (1998). "Characterizing Selection Bias Using Experimental Data". In: *Econometrica* 66.5, pp. 1017–1098.

Heckman, James J., Hidehiko Ichimura, and Petra Todd (1997). "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme". en. In: *Review of Economic Studies* 64.4, pp. 605–654.

Holm, Anders et al. (2017). "Employment effects of active labor market programs for sick-listed workers". In: *Journal of Health Economics* 52, pp. 33–44.

Huber, Martin, Michael Lechner, and Giovanni Mellace (2016). "The finite sample performance of estimators for Mediation Analysis under Sequential Conditional Independence". en. In: *Journal of Business & Economic Statistics* 34.1, pp. 139–160.

Huber, Martin, Michael Lechner, and Conny Wunsch (2013). "The performance of estimators based on the propensity score". en. In: *Journal of Econometrics* 175.1, pp. 1–21.

Huh, Keun and Robin C. Sickles (1994). "Estimation of the duration model by nonparametric Maximum Likelihood, Maximum Penalized Likelihood, and Probability Simulators". In: *The Review of Economics and Statistics* 76.4, p. 683.

Ichimura, Hidehiko and T.Scott Thompson (1998). "Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution". In: *Journal of Econometrics* 86.2, pp. 269–295.

Jahn, Elke and Michael Rosholm (2013). "Is temporary agency employment a stepping stone for immigrants?" en. In: *Economics Letters* 118.1, pp. 225–228.

Kastoryano, Stephen and Bas Van der Klaauw (2011). "Dynamic evaluation of job search assistance". In: *IZA Discussion Papers, No. 5424*.

Kyyrä, Tomi (2010). "Partial unemployment insurance benefits and the transition rate to regular work". en. In: *European Economic Review* 54.7, pp. 911–930.

Kyyrä, Tomi, Pierpaolo Parrotta, and Michael Rosholm (2013). "The effect of receiving supplementary UI benefits on unemployment duration". en. In: *Labour Economics* 21, pp. 122–133.

Lalive, Rafael, Jan C. van Ours, and Josef Zweimüller (2005). "The effect of benefit sanctions on the duration of unemployment". In: *Journal of the European Economic Association* 3.6, pp. 1386–1417.

Lalive, Rafael, Jan C. Van Ours, and Josef Zweimüller (Jan. 2008). "The Impact of Active Labour Market Programmes on the Duration of Unemployment in Switzerland". en. In: *The Economic Journal* 118.525, pp. 235–257.

Lalonde, Robert (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". In: *The American Economic Review* 76.4, pp. 604–620.

Lechner, Michael and Anthony Strittmatter (2017). "Practical procedures to deal with common support problems in matching estimation". en. In: *Econometric Reviews*, pp. 1–15.

Lechner, Michael and Conny Wunsch (2013). "Sensitivity of matching-based program evaluations to the availability of control variables". en. In: *Labour Economics* 21, pp. 111–121.

Lindeboom, Maarten, Ana Llena-Nozal, and Bas van der Klaauw (Dec. 2016). "Health shocks, disability and work". en. In: *Labour Economics* 43, pp. 186–200.

Lindsay, Bruce (1983). "The geometry of Mixture Likelihoods: A general Theory". In: *The Annals of Statistics* 11.1, pp. 86–94.

McVicar, Duncan, Julie Moschion, and Jan C. van Ours (Sept. 2018). "Early illicit drug use and the age of onset of homelessness". en. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Mueser, Peter R., Kenneth R. Troske, and Alexey Gorislavsky (2007). "Using State Administrative Data to Measure Program Performance". In: *Review of Economics and Statistics* 89.4, pp. 761–783.

Muller, Paul, Bas van der Klaauw, and Arjan Heyma (2017). "Comparing econometric methods to empirically evaluate Job-Search Assistance". In: *Unpublished manuscript*.

Narendranathan, Wiji and Mark Stewart B. (1993). "Modeling the Probability of Leaving Unemployment: Competing Risks Models with Flexible Base-line Hazards". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 42.1, pp. 63–83.

Palali, Ali and Jan C. van Ours (Dec. 2017). "Love Conquers all but Nicotine: Spousal Peer Effects on the Decision to Quit Smoking". en. In: *Health Economics* 26.12, pp. 1710–1727.

Richardson, Katarina and Gerard J. van den Berg (2013). "Duration dependence versus unobserved heterogeneity in treatment effects: Swedish labor market training and the transition rate to employment". en. In: *Journal of Applied Econometrics* 28.2, pp. 325–351.

Ridder, Geert (1987). "The sensitivity of duration models to misspecified unobserved heterogeneity and duration dependence". In: *Unpublished manuscript*.

Røed, Knut and Oddbjørn Raaum (2006). "Do labour market programmes speed up the return to work?" In: *Oxford Bulletin of Economics and Statistics* 68.5, pp. 541–568.

Smith, Jeffrey and Petra Todd (2005). "Does matching overcome LaLonde's critique of nonexperimental estimators?" In: *Journal of Econometrics* 125.1-2, pp. 305–353.

Tatsiramos, Konstantinos (2010). "Job displacement and the transitions to re-employment and early retirement for non-employed older workers". en. In: *European Economic Review* 54.4, pp. 517–535.

Van Ours, Jan C. and Jenny Williams (2012). "The effects of cannabis use on physical and mental health". en. In: *Journal of Health Economics* 31.4, pp. 564–577.

— (2009). "Why parents worry: Initiation into cannabis use by youth and their educational attainment". en. In: *Journal of Health Economics* 28.1, pp. 132–142.

Van Ours, Jan C. et al. (May 2013). "Cannabis use and suicidal ideation". en. In: *Journal of Health Economics* 32.3, pp. 524–537.

Van den Berg, Gerard J. and Sumedha Gupta (Mar. 2015). "The role of marriage in the causal pathway from economic conditions early in life to mortality". en. In: *Journal of Health Economics* 40, pp. 141–158.

Van den Berg, Gerard J. and Jan C. van Ours (Jan. 1996). "Unemployment Dynamics and Duration Dependence". en. In: *Journal of Labor Economics* 14.1, pp. 100–125.

Van den Berg, Gerard J. and Johan Vikström (2014). "Monitoring job offer decisions, punishments, exit to work, and job quality". In: *The Scandinavian journal of economics* 116.2, pp. 284–334.

Van den Berg, Gerard J. Van and Jan C. van Ours (Mar. 1994). "Unemployment Dynamics and Duration Dependence in France, the Netherlands and the United Kingdom". en. In: *The Economic Journal* 104.423, p. 432.

Vikstrom, Johan and Gerard J. van den Berg (2017). "Långsiktiga effekter av arbetsmarknadsutbildning (in Swedish)". In: *IFAU rapport 2017:17*.

Tables and Figures

Table 1: Sample statistics and estimates from the selection model using the full sample of actual treated and non-treated

	Treated	Control	Selection model	
			Est.	Std. Er.
Number of observations	76,302	2,564,561	2,640,863	
Panel A: Baseline socio-economic characteristics				
Country of origin: Not Europe	0.20	0.16	0.0910***	(0.0120)
Age 25-29	0.23	0.26	0.1366***	(0.0126)
Age 30-34	0.20	0.20	0.1188***	(0.0117)
Age 40-44	0.16	0.15	-0.0363***	(0.0123)
Age 45-49	0.12	0.11	-0.1441***	(0.0137)
Age 50-54	0.09	0.09	-0.3510***	(0.0160)
Male	0.67	0.51	0.4719***	(0.0091)
Married	0.35	0.34	0.0017	(0.0089)
Children: At least one	0.43	0.43	0.1265***	(0.0100)
Children: No. of children in age 0-3	0.20	0.20	0.0565***	(0.0116)
Education: Pre-high school	0.18	0.17	-0.1432***	(0.0253)
Education: High school	0.57	0.50	0.0624**	(0.0248)
Education: University College or higher	0.22	0.31	-0.0490**	(0.0250)
Panel B: Inflow time and regional information				
Beginning of unemployment: June-August	0.26	0.30	-0.0135	(0.0084)
Inflow year: 2003-2005	0.30	0.35	-0.3952***	(0.0217)
Inflow year: 2006-2007	0.16	0.18	-0.2562***	(0.0230)
Inflow year: 2008-2009	0.23	0.18	-0.3304***	(0.0233)
Inflow year: 2010-2011	0.18	0.17	-0.2455***	(0.0240)
Region: Stockholm	0.13	0.21	-0.3412***	(0.0158)
Region: Gothenborg	0.13	0.16	-0.3634***	(0.0127)
Region: Skane	0.12	0.14	-0.2910***	(0.0129)
Region: Northern parts	0.21	0.15	0.1647***	(0.0112)
Region: Southern parts	0.14	0.12	0.0111	(0.0126)
Monthly regional unemployment rate	10.54	9.77	0.0234***	(0.0021)
Panel C: Short–term employment history (2 years) and employment duration				
Time employed in last spell	859.82	831.20	0.0000	(0.0000)
Missing time employed in last spell	0.20	0.17	0.0493***	(0.0150)
Months employed in last 6 months	3.37	3.54	-0.0003	(0.0039)
Months employed in last 24 months	12.79	13.50	0.0040***	(0.0013)
No employment in last 24 months	0.22	0.19	-0.1354***	(0.0250)
Time since last employment if in last 24 months	2.31	2.42	-0.0069***	(0.0015)
Number of employers in last 24 months	1.66	1.79	0.0115***	(0.0035)
Employed 1 year before	0.59	0.59	0.0353***	(0.0122)
Employed 2 years before	0.59	0.59	0.0207*	(0.0122)
Panel D: Short–term unemployment history (2 years) and unemployment duration				
Time unemployed in last spell	107.11	89.43	0.0000	(0.0000)
Missing time unemployed in last spell	0.53	0.51	0.0213*	(0.0130)
Days unemployed in last 6 months	18.94	14.79	0.0008***	(0.0002)
Days unemployed in last 24 months	143.53	120.87	0.0003***	(0.0000)
No unemployment in last 24 months	0.44	0.44	-0.0511***	(0.0150)
Days since last unemployment if in last 24 months	15.12	14.76	0.0001	(0.0001)

Continue to next page

Table 1 – continued from previous page

	Treated	Control	Selection model	
			Est.	Std. Err.
Number of unemployment spells in last 24 months	0.82	0.88	0.0033	(0.0060)
Unemployed 6 months before	0.20	0.16	0.0171	(0.0151)
Unemployed 24 months before	0.24	0.22	-0.0327***	(0.0121)
Any program in last 24 months	0.03	0.02	0.0579**	(0.0291)
Panel E: Short–term welfare history (2 years)				
Welfare benefits -1 year	4928.00	3742.27	0.0318***	(0.0078)
Welfare benefits -2 years	4258.73	3542.66	0.0075	(0.0095)
On welfare benefits -1 year	0.19	0.14	0.0028	(0.0166)
On welfare benefits -2 years	0.17	0.14	-0.0720***	(0.0163)
Panel F: Earnings history (2 years)				
Earnings 1 year before	111684.78	110247.91	0.0095*	(0.0055)
Earnings 2 years before	111858.48	110612.95	-0.0157*	(0.0094)
Panel G: Long-term employment history (10 years)				
Months employed in last 10 years	58.19	62.91	-0.0022***	(0.0002)
Number of employers in last 10 years	4.72	5.12	0.0119***	(0.0012)
Cumulated earnings 5 years before	533484.45	530466.42	0.0629***	(0.0114)
Panel H: Long-term unemployment history (10 years)				
Days unemployed in last 10 years	788.31	693.41	-0.0001***	(0.0000)
No unemployment in last 10 years	0.18	0.17	-0.0890***	(0.0158)
Days since last unemployment if in last 10 years	256.77	290.49	-0.0000***	(0.0000)
Number of unemployment spells in last 10 years	3.63	3.83	0.0074***	(0.0018)
Average unemployment duration	95.31	90.15	-0.0001***	(0.0000)
Duration of last unemployment spell	180.26	154.83	-0.0001***	(0.0000)
Any program in last 10 years	0.15	0.12	0.0348	(0.0227)
Any program in last 4 years	0.06	0.05	0.0509**	(0.0243)
Number of programs in last 10 years	0.19	0.15	0.0342**	(0.0157)
Panel I: Long-term welfare history, out-of-labor-force (10 years)				
Yearly average welfare benefits last 4 years	4239.77	3533.38	-0.0213	(0.0142)
Yearly average welfare benefits last 10 years	3918.49	3448.42	-0.0828***	(0.0086)
No welfare benefits last 4 years	0.69	0.75	-0.0824***	(0.0150)
No welfare benefits last 10 years	0.51	0.59	-0.0946***	(0.0109)
Panel J: Characteristics of the last job				
Wage	18733.31	18860.58	-0.0597***	(0.0052)
Wage missing	0.54	0.52	-0.0215	(0.0337)
Occupation:				
Manager	0.04	0.07	-0.3102***	(0.0388)
Requires higher education	0.04	0.06	-0.1240***	(0.0375)
Clerk	0.04	0.05	-0.0037	(0.0374)
Service, care	0.09	0.13	-0.0047	(0.0357)
Mechanical, transport	0.13	0.07	0.2107***	(0.0352)
Building, manufacturing	0.06	0.05	0.0597	(0.0371)
Elementary occupation	0.05	0.05	-0.0044	(0.0375)
Panel K: Characteristics of the last firm				
Firm size	2523.01	3873.70	0.0000**	(0.0000)
Age of firm	12.95	14.13	0.0006	(0.0009)
Average wage	21588.62	21517.77	0.0007	(0.0048)
Wage missing	0.62	0.58	-0.0459	(0.0541)

Continue to next page

Table 1 – continued from previous page

	Treated	Control	Selection model	
			Est.	Std. Err.
Mean tenure of employees	3.43	3.68	-0.0029	(0.0024)
Age of employees	27.74	29.44	-0.0033***	(0.0009)
Share of immigrants	0.12	0.13	-0.1709***	(0.0255)
Share of females	0.26	0.34	-0.4736***	(0.0236)
No previous firm	0.28	0.24	-0.4104***	(0.0428)
Most common occupation:				
Manager	0.04	0.06	-0.1260**	(0.0571)
Higher education	0.04	0.04	-0.0294	(0.0572)
Clerk	0.03	0.03	0.0633	(0.0579)
Service, care	0.10	0.17	0.0396	(0.0554)
Building, manufacturing	0.04	0.03	-0.0574	(0.0574)
Mechanical, transport	0.11	0.06	0.0581	(0.0554)
Elementary occupation	0.02	0.02	-0.0817	(0.0602)
Industry:				
Agriculture, fishing, mining	0.01	0.01	-0.0906**	(0.0406)
Manufacturing	0.17	0.10	0.2257***	(0.0253)
Construction	0.05	0.06	-0.2065***	(0.0292)
Trade, repair	0.06	0.07	-0.1552***	(0.0270)
Accommodation	0.02	0.03	-0.2239***	(0.0336)
Transport, storage	0.06	0.04	0.1663***	(0.0278)
Financial, real estate	0.08	0.08	-0.0127	(0.0265)
Human health, social work	0.06	0.12	-0.1581***	(0.0298)
Other - public sector	0.04	0.08	-0.2254***	(0.0308)
Other	0.06	0.07	-0.1207***	(0.0277)
Panel L: Unemployment insurance				
UI: Daily benefit level in SEK	384.11	277.33	0.2316***	(0.0118)
UI: Eligible	0.84	0.83	-0.0134	(0.0136)
UI: No benefit claim	0.37	0.54	0.2181***	(0.0238)
UI 1 year before	12712.71	13211.32	-0.0086	(0.0054)
UI 2 years before	12779.13	13181.89	0.0056	(0.0059)
Cumulated UI 5 years before	62624.69	63758.25	-0.0929***	(0.0075)
Panel M: Parents' previous income				
Mother's past income (age 35-55)	659.10	772.63	-0.0061	(0.0052)
Father's past income (age 35-55)	856.04	1039.85	-0.0505***	(0.0055)
Missing mother's past income	0.39	0.34	0.0185	(0.0138)
Missing father's past income	0.47	0.42	-0.0517***	(0.0137)
Panel N: Duration dependence				
Baseline hazard, part 2			0.2653***	(0.0186)
Baseline hazard, part 3			0.5528***	(0.0161)
Baseline hazard, part 4			0.6408***	(0.0169)
Baseline hazard, part 5			0.6466***	(0.0178)
Baseline hazard, part 6			0.6843***	(0.0166)
Baseline hazard, part 7			0.5186***	(0.0171)
Baseline hazard, part 8			-0.0601***	(0.0162)

Notes: Columns 1-2 report sample averages for the full sample with actual treated and non-treated. Columns 3-4 estimates and standard errors from the corresponding selection model. *, ** and *** denote significance at the 10, 5 and 1 percent levels. All earnings and benefits are in SEK and inflation-adjusted.

Table 2: Bias of the training effect when including different sets of covariates

Number of observations	Est.	Std. Err.
	2,564,561	
Panel A: Baseline		
Baseline socio-economic characteristics	0.0693***	(0.00241)
Calendar time (inflow dummies)	0.1107***	(0.00239)
Region dummies	0.0912***	(0.00240)
Local unemployment rate	0.1174***	(0.00239)
All the above	0.0616***	(0.00243)
Panel B: Baseline and:		
Employment history (last 2 years) and duration	-0.0144***	(0.00244)
Unemployment history (last 2 years) and duration	0.0503***	(0.00243)
Earnings history (last 2 years)	0.0401***	(0.00243)
Welfare benefit history (last 2 years)	0.0469***	(0.00243)
All of the above	-0.0228***	(0.00244)
Panel C: Baseline, short-term history and:		
Employment history (last 10 years)	-0.0239***	(0.00244)
Unemployment history (last 10 years)	-0.0289***	(0.00244)
Welfare benefit history (10 years)	-0.0190***	(0.00244)
All of the above	-0.0241***	(0.00244)
Panel D: Baseline, short-term history, long-term history and:		
Last wage	-0.0266***	(0.00244)
Last occupation dummies	-0.0246***	(0.00244)
Firm characteristics (last job)	-0.0228***	(0.00245)
Unemployment benefits	0.0153***	(0.00244)
Parents income	-0.0231***	(0.00244)
All of the above	0.0090***	(0.00246)

Notes: Estimated biases using the full sample with placebo treated and placebo non-treated adjusting for different sets of covariates. Hazard rate estimates for time in unemployment using a parametric proportional hazard model with piecewise constant baseline hazard (8 splits). *, ** and *** denote significance at the 10, 5 and 1 percent levels.

Table 3: Bias of the effect of training adjusting for short-term labor market histories

Number of observations	Est.	Std. Err.
	2,564,561	
Baseline	0.0616***	(0.00243)
Panel A: Employment duration and baseline		
Time employed in last spell	0.0394***	(0.00243)
Panel B: Short-term employment rates (2 years) and baseline		
Months employed in last 6 months	0.0168***	(0.00243)
Months employed in last 24 months	0.0091***	(0.00243)
No employment in last 24 months	0.0121***	(0.00243)
All variables	-0.0004	(0.00244)
Panel C: Other short-term employment history (2 years) and baseline		
Employed 1 year before	0.0160***	(0.00243)
Employed 2 years before	0.0265***	(0.00243)
Time since last employment if in last 24 months	0.0598***	(0.00243)
Number of employers in last 24 months	0.0427***	(0.00243)
All variables	0.0022	(0.00243)
Panel D: Unemployment duration and baseline		
Time unemployed in last spell	0.0547***	(0.00243)
Panel E: Short-term unemployment rates (2 years) and baseline		
Days unemployed in last 6 months	0.0632***	(0.00243)
Days unemployed in last 24 months	0.0616***	(0.00243)
No unemployment in last 24 months	0.0611***	(0.00243)
All variables	0.0564***	(0.00243)
Panel F: Other short-term unemployment history (2 years) and baseline		
Days since last unemployment if in last 24 months	0.0616***	(0.00243)
Number of unemployment spells in last 24 months	0.0560***	(0.00243)
Unemployed 6 months before	0.0632***	(0.00243)
Unemployed 24 months before	0.0590***	(0.00243)
Any program in last 24 months	0.0618***	(0.00243)
All variables	0.0539***	(0.00243)

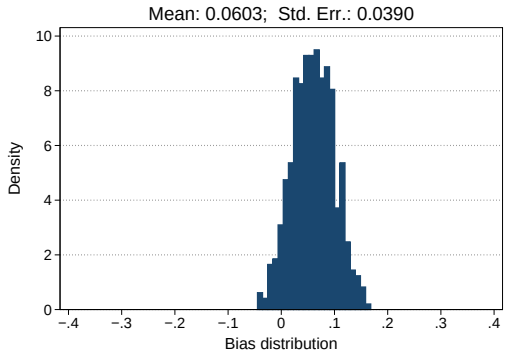
Notes: Estimated biases using the full sample with placebo treated and placebo nontreated adjusting for different sets of covariates. Hazard rate estimates for time in unemployment using a parametric proportional hazard model with piecewise constant baseline hazard (8 splits). The baseline model includes baseline socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

Table 4: Bias and variance of the estimated treatment effect for a pre-specified number of support points and support points according to model selection criteria.

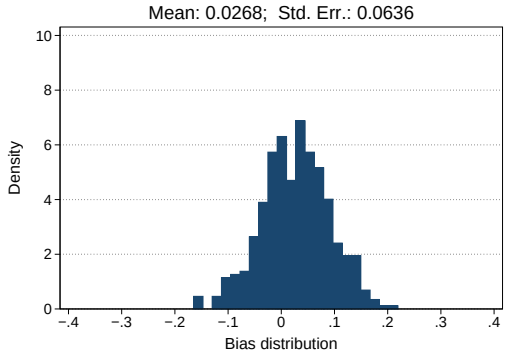
	Sample size								
	10,000			40,000			160,000		
	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Number of pre-specified support points									
1	0.060	(0.039)	0.0052	0.057	(0.020)	0.0037	0.058	(0.009)	0.0034
2	0.027	(0.064)	0.0048	0.022	(0.031)	0.0014	0.023	(0.014)	0.0007
3	0.046	(0.089)	0.0101	0.030	(0.042)	0.0026	0.028	(0.019)	0.0011
4	0.057	(0.098)	0.0128	0.035	(0.043)	0.0031	0.032	(0.021)	0.0015
5	0.062	(0.097)	0.0133	0.037	(0.044)	0.0033	0.033	(0.021)	0.0015
6	0.064	(0.099)	0.0138	0.037	(0.044)	0.0033	0.033	(0.021)	0.0015
Panel B: Model selection criteria									
ML	0.064	(0.099)	0.0139	0.037	(0.044)	0.0033	0.033	(0.021)	0.0015
AIC	0.032	(0.076)	0.0068	0.024	(0.036)	0.0018	0.026	(0.018)	0.0010
BIC	0.027	(0.064)	0.0048	0.022	(0.031)	0.0014	0.023	(0.014)	0.0007
HQIC	0.027	(0.064)	0.0048	0.022	(0.031)	0.0014	0.023	(0.014)	0.0007
Panel C: Average # support points, by selection criteria									
ML		4.11			3.99			4.10	
AIC		2.14			2.21			2.53	
BIC		1.99			2.00			2.00	
HQIC		2.01			2.00			2.04	

Notes: Estimated bias, variance and mean squared error of the treatment effect from a ToE model with different specifications of the discrete support point distribution. Simulations using 500 replications with random drawings from the full sample with placebo treated and placebo non-treated. Hazard rate estimates for time in unemployment. Each model uses a piecewise constant baseline hazard (8 splits) and the observed covariates include socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate.

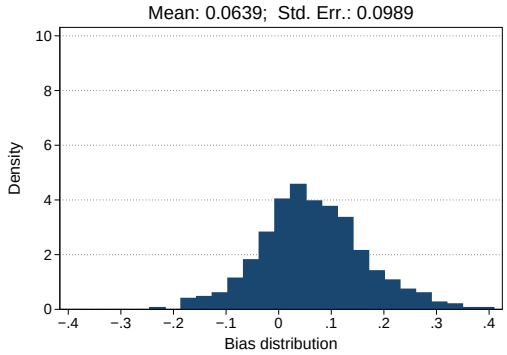
Figure 1: Distribution of the bias of the estimated treatment effect for a pre-specified number of support points, by number of support points.



(a) 1 support point



(b) 2 support points



(c) 6 support points

Note: Distribution of the estimated bias of the treatment effect from a ToE model with different specifications of the discrete support point distribution. Simulations using 500 replications with 10,000 random drawings from the full sample of placebo treated and placebo non-treated. Hazard rate estimates for time in unemployment. Each model uses a piecewise constant baseline hazard (8 splits) and the observed covariates include socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate.

Table 5: Bias and variance of the estimated treatment effect when *excluding different sets of covariates*, by model selection criteria and sample size

Specification	Exclude more covariates			Exclude fewer covariates		
	Bias (1)	SE (2)	MSE (3)	Bias (4)	SE (5)	MSE (6)
Panel A: 10,000 observations						
ML	0.091	(0.162)	0.0344	0.073	(0.122)	0.0201
AIC	0.029	(0.010)	0.0108	0.035	(0.114)	0.0142
BIC	0.024	(0.067)	0.0051	0.005	(0.063)	0.0039
HQIC	0.024	(0.068)	0.0052	0.013	(0.091)	0.0085
Average # support points, by selection criteria						
ML		4.78			5.20	
AIC		2.34			3.12	
BIC		2.00			2.20	
HQIC		2.01			2.62	
Panel B: 40,000 observations						
ML	0.025	(0.068)	0.0053	0.049	(0.060)	0.0060
AIC	0.009	(0.049)	0.0025	0.029	(0.062)	0.0047
BIC	0.019	(0.034)	0.0015	0.005	(0.039)	0.0016
HQIC	0.018	(0.036)	0.0016	0.010	(0.050)	0.0026
Average # support points, by selection criteria						
ML		4.88			5.59	
AIC		2.65			4.22	
BIC		2.00			3.16	
HQIC		2.04			3.62	

Notes: Estimated bias, variance and mean squared error of the treatment effect from a ToE model with different specifications of the discrete support point distribution. Simulations using 500 replications with random drawings from the full sample with placebo treated and placebo non-treated. Hazard rate estimates for time in unemployment. Each model uses a piecewise constant baseline hazard (8 splits). The baseline model include baseline socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate. The "exclude more covariates" model excludes baseline socio-economic characteristics, and the "exclude fewer covariates" adds control for short-term earnings history.

Table 6: Bias and variance of the estimated treatment effect when augmenting the baseline model with *covariates more or less correlated* with those left in the error term.

Degree of correlation	Positive			Small positive			Negative		
	Bias (1)	SE (2)	MSE (3)	Bias (4)	SE (5)	MSE (6)	Bias (7)	SE (8)	MSE (9)
Correlation	0.278			0.049			-0.257		
Panel A: 10,000 observations									
ML	0.063	(0.093)	0.0127	0.063	(0.100)	0.0140	0.044	(0.099)	0.0119
AIC	0.035	(0.076)	0.0070	0.033	(0.087)	0.0087	0.021	(0.081)	0.0070
BIC	0.027	(0.060)	0.0043	0.028	(0.070)	0.0057	0.019	(0.065)	0.0046
HQIC	0.027	(0.060)	0.0043	0.029	(0.071)	0.0059	0.017	(0.066)	0.0046
Average # support points, by selection criteria									
ML	4.19			4.48			4.27		
AIC	2.17			2.28			2.20		
BIC	2.00			1.99			1.95		
HQIC	2.01			2.01			2.01		
Panel B: 40,000 observations									
ML	0.042	(0.041)	0.0034	0.036	(0.047)	0.0035	0.019	(0.046)	0.0025
AIC	0.025	(0.036)	0.0019	0.025	(0.045)	0.0026	0.011	(0.039)	0.0016
BIC	0.022	(0.029)	0.0013	0.024	(0.034)	0.0018	0.013	(0.032)	0.0012
HQIC	0.022	(0.030)	0.0014	0.024	(0.035)	0.0018	0.013	(0.032)	0.0012
Average # support points, by selection criteria									
ML	3.99			4.62			4.34		
AIC	2.24			2.62			2.28		
BIC	2.00			2.00			2.00		
HQIC	2.01			2.04			2.01		

Notes: Estimated bias, variance and mean squared error of the treatment effect from a ToE model with different specifications of the discrete support point distribution. Simulations set as for Table 4. The three model specifications correspond to the baseline model of Table 4 augmented with: *Welfare benefit history (last 2 years)*, *Previous firm most common occupation dummies*, and *Last occupation dummies*, respectively. Correlation coefficients computed from the outcome model using all actual treated and control units, by correlating the linear predictor of the covariates pulled-out from the error term with that of all covariates left in the error term.

Table 7: Unobserved heterogeneity approximation, treatment equation

	Mean $\exp(V_p)$	Std. Err. $\exp(V_p)$
Panel A: Data generating process		
	0.00056	0.00023
Panel B: Number of pre-specified support points		
2	0.00047	0.00003
3	0.00047	0.00020
4	0.00046	0.00023
5	0.00047	0.00027
6	0.00047	0.00031
Panel C: Model section criteria		
ML	0.00047	0.00030
AIC	0.00047	0.00003
BIC	0.00047	0.00010
HQIC	0.00047	0.00003

Notes: Mean and standard error of the treatment unobserved heterogeneity relative to the baseline specification (baseline characteristics, timing of inflow and regional information), in the population of actual treated and controls and as approximated by the ToE mass points and respective probabilities (averaged across 500 replications, each with a sample of 10,000 units). Both the actual and approximated unobserved heterogeneity distributions include the constant.

Table 8: Bias and variance of the estimated treatment effect with *exogenous variation*, by model selection criteria and sample size

Specification	Time-varying local unemployment rate		
	Bias (1)	SE (2)	MSE (3)
Number of pre-specified support points			
1	0.056	(0.039)	0.0046
2	0.016	(0.066)	0.0046
3	0.056	(0.100)	0.0132
4	0.074	(0.109)	0.0174
5	0.082	(0.108)	0.0185
6	0.084	(0.109)	0.0189
Model selection criteria			
ML	0.084	(0.109)	0.0189
AIC	0.033	(0.090)	0.0093
BIC	0.016	(0.066)	0.0046
HQIC	0.017	(0.069)	0.0051
Average # support points, by selection criteria			
ML		4.46	
AIC		2.25	
BIC		1.99	
HQIC		2.01	

Notes: Estimated bias, variance and mean squared error of the treatment effect from a ToE model with different specifications of the discrete support point distribution. Simulations using 500 replications with random drawings from the full sample with placebo treated and placebo non-treated. Hazard rate estimates for time in unemployment. Each model uses a piecewise constant baseline hazard (8 splits). The baseline model include baseline socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate.