

The Institute for Labour Market Policy Evaluation (IFAU) is a research institute under the Swedish Ministry of Industry, Employment and Communications, situated in Uppsala. IFAU's objective is to promote, support and carry out: evaluations of the effects of labour market policies, studies of the functioning of the labour market and evaluations of the labour market effects of measures within the educational system. Besides research, IFAU also works on: spreading knowledge about the activities of the institute through publications, seminars, courses, workshops and conferences; creating a library of Swedish evaluational studies; influencing the collection of data and making data easily available to researchers all over the country.

IFAU also provides funding for research projects within its areas of interest. There are two fixed dates for applications every year: April 1 and November 1. Since the researchers at IFAU are mainly economists, researchers from other disciplines are encouraged to apply for funding.

IFAU is run by a Director-General. The authority has a traditional board, consisting of a chairman, the Director-General and eight other members. The tasks of the board are, among other things, to make decisions about external grants and give its views on the activities at IFAU. Reference groups including representatives for employers and employees as well as the ministries and authorities concerned are also connected to the institute.

Postal address: P.O. Box 513, 751 20 Uppsala

Visiting address: Kyrkogårdsgatan 6, Uppsala

Phone: +46 18 471 70 70

Fax: +46 18 471 70 71

ifau@ifau.uu.se

www.ifau.se

Papers published in the Working Paper Series should, according to the IFAU policy, have been discussed at seminars held at IFAU and at least one other academic forum, and have been read by one external and one internal referee. They need not, however, have undergone the standard scrutiny for publication in a scientific journal. The purpose of the Working Paper Series is to provide a factual basis for public policy and the public policy discussion.

Program evaluation and random program starts*

Peter Fredriksson[†] Per Johansson[‡]

December 17, 2002

Abstract

This paper discusses the evaluation problem using observational data when the timing of treatment is an outcome of a stochastic process. We show that, without additional assumptions, it is not possible to estimate the average treatment effect and treatment on the treated. It is, however, possible to estimate the effect of treatment on the treated up to a certain time point. We propose an estimator to estimate this effect and show that it is possible to test for an average treatment effect.

Keywords: Treatment effects, dynamic treatment assignment, program evaluation, method of matching.

JEL: C14, C41

*Thanks to Kenneth Carling, Paul Frijters, Xavier de Luna, Jeffrey Smith, and Gerard van den Berg for very useful comments. Comments from seminar participants at the conference on "The Evaluation of Labour Market Policies" (Amsterdam, October 2002), Department of Statistics, Umeå university, and IFAU are also gratefully acknowledged.

[†]Department of Economics, Uppsala University, Institute for Labour Market Policy Evaluation (IFAU), and CESifo. Address: Department of Economics, Uppsala University, Box 513, SE-751 20 Uppsala, Sweden. Phone: +46-18-471 11 13. Email: peter.fredriksson@nek.uu.se. Fredriksson acknowledges the financial support from the Swedish Council for Working Life and Social Research (FAS).

[‡]Department of Economics, Uppsala University, and IFAU. Address: IFAU, Box 513, SE-751 20 Uppsala. Phone: +46-18-471 70 86. Email: per.johansson@ifau.uu.se.

1 Introduction

The prototypical evaluation problem is cast in a framework where treatment is offered only once. Thus treatment assignment is a static problem and the information contained in the timing of treatment is typically ignored; see Heckman et al. (1999) for an overview of the literature. This prototype concurs rather poorly with how most real-world programs work. Often it makes more sense to think of the assignment to treatment as a dynamic process, where the start of treatment is the outcome of a stochastic process.

There are (at least) two important implications of taking the timing of events into account. First of all, the timing of events contains additional information which is useful for identification purposes. Indeed, Abbring and van den Berg (2002) have shown that one can identify a causal effect non-parametrically in the Mixed Proportional Hazard model from single-spell duration data without conditional independence assumptions.¹ Second of all, the dynamic assignment process has serious implications for the validity of conditional independence assumptions usually invoked to estimate effects such as treatment on the treated.

The main objective of this paper is to substantiate the second of the above claims. In particular we discuss program evaluations when (i) there are restrictions on treatment eligibility, (ii) no restrictions on the timing of the individual treatment, and (iii) the timing of treatment is linked to the outcome of interest. For instance, this evaluation problem arises when unemployment is a precondition for participation in a labor market program, programs may start at any time during the unemployment spell, and we are interested in employment outcomes. Employment outcomes have increasingly become the focus of the labor market evaluation literature so our analysis should have wide applicability.² We choose to focus on employment outcomes for illustrative purposes but our analysis has implications for all situations when points (ii) and (iii) apply. For

¹At this stage, we are deliberately vague on what causal effect this really is.

²The prime candidate for the shift in emphasis is that the ultimate goal of many labor market programs is to raise the reemployment probability rather than increasing the productivity of the participants. Also, the targets that government agencies responsible for, e.g., training, should fulfill are usually formulated in terms of employment rather than wages. For instance, one of the key targets for evaluating the performance of the Swedish labor market board is that at least 70 percent of participants in labor market training should be regularly employed one year after the end of treatment.

instance, it follows immediately that the points we raise should be taken into consideration in analyses of earnings outcomes.

A second objective of the paper is to bridge some of the gap that exists between the literature on matching and the literature using hazard regressions. In the matching literature one typically considers, e.g., the probability of employment some fixed time period after treatment; Gerfin and Lechner (2002) is a recent example. By assumption, unobserved heterogeneity is not an issue. In the hazard regressions literature, the focus is on the timing of the outflow to a state of interest (e.g. employment). Usually, there is more structure imposed on the form of the hazard but there is also greater concern about unobserved heterogeneity; van den Berg et al. (2004) is an example. Clearly, these outcomes are intimately related and to us the division of the literature seems rather superficial. For instance, with rich data, one might well think of applying a matching approach to estimate the hazard to employment.

Here we assume that we can construct the counterfactual outcome using the method of matching. We take this approach for illustrative purposes – not because we are strong believers in the matching approach. To convey our basic messages as clearly as possible we want to avoid the complications arising from unobserved heterogeneity. Moreover, we want to refrain from making assumptions about the appropriate bivariate distribution for the timing of events. If one is prepared to make assumptions about the functional form of the bivariate distribution, this is an alternative way of attacking the particular evaluation problem that we are considering.

We show that even if we have monozygotic twins and one participates in the program, while the other does not, this is not in general sufficient to obtain unbiased estimates of conventional treatment parameters such as the average treatment effect or treatment on the treated. It is, however, possible to estimate the program effect for those being treated up to a certain time point. Notice that this is the appropriate interpretation of the causal effect estimated in the framework of Abbring and van den Berg (2002). We also show that it is possible to test whether there is an average treatment effect.

The reason why it is difficult to estimate the conventional treatment effects is that in order to get at them one would like to define a comparison group that was never treated. But finding individuals who were never treated involves conditioning on the future since treatment can start at any point in time. By defining the comparison group in this way one is

implicitly conditioning on the outcome variable since those who do not enter in future time periods to a large extent consist of those who have had the luck of finding a job.³ Therefore, the conditional independence assumptions required to estimate the average treatment effect and treatment on the treated do not hold and studies that define the comparison group in this way will generate estimates that are biased towards finding negative treatment effects when, in fact, none exist.

The rest of this paper is structured in the following way. In section 2, we present the evaluation framework. We discuss the potential outcomes of interest, possible estimands, and the specific problem associated with random program starts. Section 3 considers alternative estimators. We propose an estimator of treatment on the treated up to certain point in time. In section 4 we conduct a small Monte Carlo experiment to illustrate the small sample properties of our estimator and to compare it to different estimators available elsewhere in the literature. Section 5, finally, concludes.

2 The framework

We have the following world in mind. Consider a set of individuals who enter unemployment at time 0. At the time of unemployment entry these individuals are identical. Alternatively, we could assume that matching on the observed covariates at unemployment entry is sufficient to take care of any heterogeneity influencing outcomes. We make the assumption that individuals are identical for expositional convenience.

During the unemployment spell they are exposed to two kind of risks: either they get a job offer with instantaneous probability $\tilde{\lambda}_0(t)$ or an offer to participate in a program with probability $\tilde{\gamma}(t)$ per unit time. The instantaneous probability of being offered a job is $\lambda_1(t)$ for treated individuals. Let $I(\cdot)$ denote the indicator function and $v_k(t)$, $k = 0, 1, 2$, the (life-time) utilities associated with open unemployment, program participation and employment, respectively.⁴ The hazard rates to employment are then given by

$$\lambda_0(t) = \tilde{\lambda}_0(t)I(v_2(t) \geq v_0(t))$$

³There is an informal discussion along these lines in Sianesi (2001).

⁴The openly unemployed refers to the unemployed who do not participate in a labor market program.

$$\lambda_1(t) = \tilde{\lambda}_1(t)I(v_2(t) \geq v_1(t))$$

for treated and untreated individuals respectively.⁵ The hazard rate to program participation is given by

$$\gamma(t) = \tilde{\gamma}(t)I(v_1(t) \geq v_0(t))$$

Potentially, the utilities associated with each state are random (i.e. $v_k(t) = v_k + \varphi_k(t)$), but in the spirit of the assumption of no heterogeneity, we will assume that the random components ($\varphi_k(t)$) are purely idiosyncratic.

A convenient special case is when the processes determining offer arrival rates have no memory (i.e. they are Poisson). Then unemployment durations are exponentially distributed (with parameter $\exp(\lambda_0)$) and we can represent the potential duration if not treated as

$$\ln T(0) = \lambda_0 + \varepsilon_0, \tag{1}$$

where ε_0 is Type I extreme value distributed.

Further the log of the duration until treatment *start* (T^s) has an analogous representation, i.e.,

$$\ln T^s = \gamma + \epsilon \tag{2}$$

where ϵ is also Type I extreme value distributed. Notice that unemployment duration post treatment entry is simply given by $T_{t^s}^p = \max(T - t^s, 0) = T_{t^s}^p(1)$. Thus, equations (1) and (2) imply a specification for the potential duration over the distribution of t^s if the individual had not been treated at time t^s , $T_{t^s}^p(0)$.

Now that we have introduced some notation let us define the notational convention that we will adopt throughout the paper. Stochastic variables are denoted by upper-case letters (e.g. T and T^s), realizations of the stochastic processes are lower-case (e.g. t and t^s), and potential outcomes are indicated by 0 and 1 (e.g. $T(0)$ and $T(1)$).

Equations (1) and (2) are written in the form of accelerated duration models (ADM); see e.g. Kalbfleisch and Prentice (1980). Of course, the representations in (1) and (2) are unduly restrictive. We have no reason

⁵Throughout we assume that the effect of treatment occurs directly upon enrollment. As long as there is no pre-treatment effect this assumption is not important for the substance of the paper.

to postulate a particular distribution for ε_0 and ε_1 , for instance. Therefore, we will sometimes work with more general forms of the ADM

$$\ln T(0) = \beta_0 + \sigma_0 \varepsilon_0 \quad (3)$$

$$\ln T^s = \beta_1 + \sigma_1 \varepsilon_1 \quad (4)$$

without making distributional assumptions about ε_j . Only if ε_j is extreme value distributed do (3) and (4) imply a proportional hazard representation. In particular, if ε_j is extreme value distributed the durations are Weibull distributed. Other distributional assumptions about ε_j will generate hazards of the non-proportional variety. While it is true that the duration distributions implied by (3) and (4) have considerable generality, we also note that none of our results depend on the additive structure (3) and (4). In fact all of our results hold true so long as the durations are monotonic in ε_j .

It is sometimes convenient to have a particular specification of the data generating process (dgp) to work with. However, most of the time it is sufficient to work with the following dgp

$$D = I(T > t^s) \quad (5)$$

i.e. individuals are observed to take treatment if their unemployment duration (T) is longer than their duration till program start (t^s).

2.1 Objects of evaluation

We would either like to estimate the average treatment effect

$$\Delta^p = E(T^p(1)) - E(T^p(0)) \quad (6)$$

or treatment on the treated

$$\Delta_1^p = E(T^p(1) | D = 1) - E(T^p(0) | D = 1) \quad (7)$$

where $\Delta^p = \Delta_1^p$ in the ideal experimental setting. One of the potential durations in (6) or (7) is of course a missing counterfactual outcome. For example, we observe $T^p(1)$ for a treated individual but we do not observe $T^p(0)$. This is always true, even in experiments.

What makes this problem somewhat special is that in many realistic situations we lack starting dates for those not treated and hence we can

not use the post treatment duration for the untreated to estimate the counterfactual means $E(T^p(0) | D = 1)$ or $E(T^p(0))$. This is different than in the experimental situation, where treatment is offered at some fixed point in time, and the fairly uncommon situation where a program starts after a fixed duration.⁶

For later purposes it is useful to define two *potential* survival functions

$$S_1^p(t) = \exp\left(-\int_{t^s}^t \lambda_1(\tau) d\tau\right)$$

$$S_0^p(t) = \exp\left(-\int_{t^s}^t \lambda_0(\tau) d\tau\right)$$

Then we can define the treatment effect in terms of the difference in the survival functions

$$\Delta^p(t) = S_1^p(t) - S_0^p(t), \quad t \in (t^s, \infty)$$

Defining the treatment effect in this way is useful as the difference in survival functions integrates to the difference in mean duration, i.e.,

$$\int_0^\infty \Delta^p(t) dt = E(T^p(1)) - E(T^p(0)) = \Delta^p$$

Conditioning on $D = 1$ we can calculate treatment on the treated in an analogous fashion.

To estimate (7) the potential outcome of the non-treated should be conditionally (or mean) independent of treatment; using the notation of Dawid (1979), it must be true that

$$T^p(0) \perp\!\!\!\perp D \tag{8}$$

For the evaluation parameter (6) both potential outcomes should be independent of the treatment, i.e.,

$$(T^p(1), T^p(0)) \perp\!\!\!\perp D$$

⁶Of course there are some treatments that start after a fixed point in time. The expiration of UI benefits is a prototypical example. By definition, random program starts is not going to be an issue in an analysis of the effects of a time limit in UI benefit receipt.

2.2 The random start problem

Consider a treated individual. For this individual we observe a realization of the treatment start (t^s). Using the ADM framework we can represent the log of the potential durations if treated and not treated at t^s as

$$\ln T_{t^s}^p(0) = \delta_0 + \sigma_0 \eta_0 \text{ and } \ln T_{t^s}^p(1) = \delta_1 + \sigma_{01} \eta_1,$$

where $\delta_0 = \beta_0 - t^s$ and η_0 is the censored (at $T > t^s$) distribution for ε_0 . The data generating process is thus such that “unlucky” individuals are more likely to enter treatment.⁷ This feature of the problem is what complicates the evaluation.

Now, consider the individual treatment effect. It is given by

$$\delta = (\delta_1 - \delta_0) + (\sigma_{01} \eta_1 - \sigma_0 \eta_0)$$

If $\delta_1 \neq \delta_0$ and/or $\sigma_{01} \eta_1 \neq \sigma_0 \eta_0$ this implies that the outflow rates differ by treatment status. Moreover, if $\eta_0 \neq \eta_1$ the treatment effect varies stochastically over individuals. If there is no treatment effect, i.e. $\lambda_1(t) = \lambda_0(t)$, then $\sigma_0 \eta_0 = \sigma_{01} \eta_1$ and $\delta_1 = \delta_0$.

It is important to realize that the post treatment duration is stochastically dependent on the pre treatment duration even if there is no treatment effect. This follows since η_0 is the censored distribution of ε_0 . Thus, given the data generating process, we need that $T(0) \perp\!\!\!\perp D$ in order for $T_{t^s}^p(0) \perp\!\!\!\perp D$. In turn, this implies that to estimate an average treatment effect one may have to invoke additional identifying assumptions. One option is to postulate a bivariate distribution for the durations T and T^s . Instead of relying on functional form we would like to consider a less structural approach to resolve the problem of inference. One possible way may be to create a duration matched comparison sample to those flowing into treatment, i.e., to condition on all realizations of t^s . We consider this and other approaches in the next section.

3 Potential estimators

In this section we consider alternative strategies to estimate the parameters of interest. Before discussing potential estimators let us introduce

⁷This is of course true even if we postulate that the distribution of ε_j is extreme value such that we have a proportional hazards model with no time dependence.

some notation that we will use throughout. The sample consists of n and N^c treated and non-treated individuals, respectively. We will index a treated individual by i , a non-treated individual by c , and whenever indexing the total sample we will use m ; hence, $i = 1, \dots, n$, $c = 1, \dots, N^c$ and $m = 1, \dots, N$, where $N = n + N^c$.

3.1 Duration matching

Here we follow the typical approach to evaluating an on-going program. As indicated above, researchers usually impose a “binary framework” even though the timing of events varies. To implement the idea that the assignment to treatment occurs only at a “single point in time” there is typically a classification window of some length (C). Individuals that take up treatment within, say, the first six months of the unemployment spell are defined as the treated ($D(C) = 1$) while those that do not are defined as the non-treated ($D(C) = 0$). Then the typical outcome would be something like the employment status one year after treatment entry (t^s). Thus the starting point for measuring the effect of treatment occurs before the end of the classification window ($t^s < C$).

A practical problem is that those who had the luck of finding a job quickly are more likely to be found in the non-treated group. Thus some trimming of the left-tail of the duration distribution seems to be called for. Here we follow an approach that is akin to the one suggested by Lechner (1999). Before matching on the covariates he proposes a procedure to trim the duration distribution of the non-treated such that he obtains a duration matched comparison sample.

To illustrate the approach as clearly as possible, let us consider the extreme case where $C \rightarrow \infty$. Now, duration matching is an attempt to estimate (7). This requires the CIA (8). The expectation $E(T^p(1)|D = 1)$ can be estimated as

$$\hat{t}^p = \frac{1}{n} \sum_{i=1}^n (t_i - t_i^s)$$

An estimator of the counterfactual outcome, $E(T^p(0)|D = 1)$, is based on random sampling from the inflow distribution, $F(T^s|D = 1)$. For a random draw, t_i^s , an individual from the comparison sample is matched if the unemployment duration for this randomly assigned individual satisfies $t_c > t_i^s$. Applying this procedure we get a duration matched comparison

sample (consisting of n matches) and may calculate

$$\hat{t}_c^p = \frac{1}{n} \sum_{i=1}^n t_{c_i}^p, \quad (9)$$

where $t_{c_i}^p = t_c - t_i^s$ is the observed unemployment duration after t_i^s for a (randomly assigned) matched individual. The treatment effect is then estimated as

$$\widehat{\Delta}_1^p = \hat{t}^p - \hat{t}_c^p \quad (10)$$

Proposition 3.1 *The conditional independence assumption (8) does not hold.*

Proof. To prove this proposition let us consider (3) and (4). Let $T_{\bar{t}}^p(0)$ be the potential post treatment unemployment duration if not treated up to a fixed time period \bar{t} . Consider an individual treated at $t^s = \bar{t}$. For this individual we know that $T > \bar{t}$. For a potential comparison individual we have $\bar{t} < T < T^s$ since this individual was never treated. Thus

$$\ln T_{\bar{t}}^p(0)|(D = 1) = \ln T(0)|(D = 1, T > \bar{t}) - \bar{t} = \beta_0 - \bar{t} + \sigma_0 \varepsilon_0|(T > \bar{t}) \quad (11)$$

$$\ln T(0)|(D = 0, T > \bar{t}) - \bar{t} = \beta_0 - \bar{t} + \sigma_0 \varepsilon_0|(T^s > T > \bar{t}) \quad (12)$$

and hence $T_{\bar{t}}^p(0)D|(T > \bar{t})$. ■

Proposition 3.2 *When there is no treatment effect, the duration matched estimator ($\widehat{\Delta}_1^p$) is positively biased*

Proof. To prove this proposition take the expectations of (11) and (12). Since $E(\varepsilon_0|(T^s > T > \bar{t})) < E(\varepsilon_0|(T > \bar{t}))$ we get $E(\ln T_{\bar{t}}^p(0)|D = 1) > [E(\ln T(0)|D = 0, T > \bar{t}) - \bar{t}] = E(\ln T_{\bar{t}}^p(0)|D = 0)$. ■

Notice that these two results hold for all specifications of the error terms. In particular, the duration matched estimator is biased even though the hazards to employment and treatment are constant.

Proposition 1 follows from the observation that for all classification periods such that $t^s < k$ there is some conditioning on the future involved when defining the potential comparison group for an individual treated at t^s . Given that there is no treatment effect we can also determine the sign of the bias involved in applying this procedure; see Proposition 2. The intuition for the latter result is simply that for the comparison group

we know that (since the individual is not treated) the spell ends with employment, while for the treated group we do not know if the spell ends in employment. Therefore, there is a positive bias in the effect of treatment on post-treatment durations (i.e. there is a bias towards finding negative treatment effects). Let us also make the (perhaps obvious) remark that Propositions 1 and 2 hold if the observations on unemployment durations are censored at, say, \bar{L} , although one would expect the bias to be reduced in magnitude.

To sum up, it is not possible to create a sample of matching individual who do not receive treatment at any point in time. In defining the treated and the comparisons, the sampling is on ε_0 , which in turn determines, for any \bar{t} , the (potential) outcome $T_{\bar{t}}^p(0)$. Thus for those treated we have large ε_0 and hence large $T_{\bar{t}}^p(0)$ while the opposite is true for the untreated. We wish to emphasize that the crux of the problem with this estimator lies in the use of a classification window; it is not due to the trimming procedure. It is the strive to transform a world where treatment assignment is the outcome of two dependent stochastic processes to an idealized world where treatment assignment and outcomes occurs at single points in time that causes the problems.

3.2 The proportional hazard model

A popular approach to estimate the treatment effect is to use the proportional hazard model; see, e.g., Crowley and Hu (1977), Lalive, van Ours and Zweimüller (2002), and Richardsson and van den Berg (2002). Here we examine what happens when we impose a proportional hazard model in our context.

Suppose that the hazard after treatment is given by

$$\lambda_1(t) = h_0(t) \exp(\delta D)$$

where $D = I(T > t^s)$.⁸ If δ estimates the average treatment effect then $\lambda_0(t) = h_0(t)$. So if the model has a proportional hazard specification, the outflow of the treated relative to the non-treated identifies the treatment effect: $\lambda_1(t) = \lambda_0(t) \exp(\delta D)$.

Can we estimate the average treatment effect using this framework? The following proposition provides part of the answer.

⁸Note that this representation has an analogue in the ADM model (1).

Proposition 3.3 *The data generating process $D = I(T > t^s)$ implies that the baseline hazard for the treated is not equal to the baseline hazard in the population, i.e., $h_0(t) \neq \lambda_0(t)$.*

Proof. Proposition 1 implies that $E(T(0)|D = 1) > E(T(0)|D = 0)$. Since this is true for any censoring point $t = c > 0$ the survival function for the treated is larger than the survival function for the non-treated, i.e. $S(t|D = 1) > S(t|D = 0)$. Now,

$$\begin{aligned}
 S(t|D = 1) &> S(t|D = 0) \Leftrightarrow \\
 \ln S(t|D = 1) &> \ln S(t|D = 0) \Leftrightarrow \\
 \int_0^t \frac{d \ln S(s|D = 1)}{ds} ds &> \int_0^t \frac{d \ln S(s|D = 0)}{ds} ds \Leftrightarrow \\
 - \int_0^t \lambda(s|D = 1) ds &> - \int_0^t \lambda(s|D = 0) ds \Leftrightarrow \\
 \int_0^t [\lambda(s|D = 1) - \lambda(s|D = 0)] ds &< 0
 \end{aligned}$$

■

Thus, the mirror image of the fact that those we observe taking treatment have longer expected unemployment duration is that the hazard is lower for treated individuals than non-treated individuals.

We can always write the appropriate baseline hazard as

$$h_0(t) = \lambda_0(t|D = 1) \Pr(D(t) = 1) + \lambda_0(t|D = 0) \Pr(D(t) = 0)$$

Proposition 3 implies that $\lambda_0(t|D = 1) \neq \lambda_0(t|D = 0)$. Further, if $\delta > 0$ it is not possible to identify all components of the baseline hazard using observational data. So estimates of the treatment effect using the proportional hazards specification will, in general, neither estimate the average treatment effect nor treatment on the treated. Can we say anything about the sign of the bias relative to the true parameter, δ ? Proposition 4 outlines the results

Proposition 3.4 *a) If there is no treatment effect ($\delta = 0$), the proportional hazards estimator ($\hat{\delta}_{PH}$) has the property that $\text{plim } \hat{\delta}_{PH} = 0$. b) If $\delta \neq 0$, then $\text{plim } |\hat{\delta}_{PH}| < |\delta|$.*

Proof. See appendix. ■

The intuition for Proposition 4b) is the following. With observational data, the risk set used for estimation includes individuals who are not treated at time t but will be treated at some future time point $s > t$. The inclusion of these individuals (in addition to those who have been treated prior to t and those who are never treated) will lead to attenuation bias.

However, the inclusion of those treated in the future in the risk set is a virtue when $\delta = 0$. The inclusion of these individuals balances the bias that would arise if only the never treated were used as comparisons.

The thrust of Proposition 4 is that the proportional hazards specification is a fertile ground for testing. However, the estimate will be smaller in absolute value than the average treatment effect when a treatment effect exists. Notice also that standard (Wald) tests will not give correct inference since the true model is non-proportional; see DiRienzo and Lagakos (2001).

Abbring and van den Berg (2002) show that the variation in the timing of treatment identifies a causal treatment parameter in the proportional hazard model. This is also true in our case since the model in this subsection is really a stylized version of their more general model. Suppose instead that we define a time-varying treatment indicator $D(s) = I(s > t^s)$. Thus $D(s) = 1$ for individuals who have been treated prior to s and $D(s) = 0$ for individuals who remain untreated at s (but may be treated in the future). Now, consider estimating $\delta(s)$ in

$$\lambda_1(s) = h_0(s) \exp(\delta(s)D(s))$$

It is clearly possible to estimate the causal treatment effect, $\delta(s)$, since $h_0(s)$ is also the baseline hazard for those who have not been treated at s . Thus, taking the timing of treatment seriously allows the identification of a causal parameter. But the interpretation of this parameter is perhaps not standard as we are about to illustrate.

3.3 Matching with a time-varying treatment indicator

The lesson from the above sub-section is that one should take the timing of treatment seriously. However, if we believe in the assumptions that justify matching we have no reason to postulate a proportional hazard. Instead we will introduce a non-parametric matching estimator that takes the timing of events into account but does not rely on proportionality.

For the purpose of introducing this estimator let us move to discrete time. Let us define the time-varying treatment indicator $D(\bar{t})$ such that $D(\bar{t}) = I(T \geq \bar{t} \geq t^s)$.

It is straightforward to show that

Lemma 3.5 *Potential unemployment duration is independent of the treatment indicator $D(\bar{t})$.*

Proof. Consider the ADM model (3). Then

$$\ln T_{\bar{t}}^p(0) = \ln T(0) | (D(\bar{t}) = 1) - \bar{t} = \beta_0 - \bar{t} + \sigma_0 \varepsilon_0 | (T \geq \bar{t})$$

$$\ln T(0) | (D(\bar{t}) = 0, T \geq \bar{t}) - \bar{t} = \beta_0 - \bar{t} + \sigma_0 \varepsilon_0 | (T \geq \bar{t})$$

and hence $T_{\bar{t}}^p(0) \perp\!\!\!\perp D(\bar{t})$. ■

Thus, the gain of introducing the time-varying treatment indicator, $D(\bar{t})$, is immediate: potential unemployment duration is conditionally independent of $D(\bar{t})$.⁹ However, the cost of this procedure is that we estimate a different treatment effect than, e.g., (7). The analogue to treatment on the treated is in this case the effect of entering at \bar{t} or earlier relative to not having done so for individuals who have taken treatment before \bar{t} ; (see Sianesi, 2001, for an analogous definition of the estimand of interest):

$$\Delta_{1\bar{t}}^p = E(T_{\bar{t}}^p(1) | D(\bar{t}) = 1) - E(T_{\bar{t}}^p(0) | D(\bar{t}) = 1) \quad (13)$$

If the effect of entering at \bar{t} is constant over time, estimates of $\Delta_{1\bar{t}}^p$ is lower in absolute value than the original object of evaluation (Δ_1^p).

To obtain a single number one would potentially like to average over the distribution of program starts, i.e., calculate

$$E_{(T^s|D=1)}(\Delta_{1\bar{t}}^p) = E_{(T^s|D=1)} \left[E(T_{\bar{t}}^p(1) | D(\bar{t}) = 1) - E(T_{\bar{t}}^p(0) | D(\bar{t}) = 1) \right] \quad (14)$$

⁹It may be useful to relate this result to the theory of point processes (see e.g. Lancaster, 1990, ch. 5). If we randomly select an individual at \bar{t} from the stock of unemployed individuals, then the stock sampling hazard is equal to

$$\chi(t) = \lambda_0(t) \frac{t}{e(t)} \leq \lambda_0(t), \quad t \geq \bar{t}$$

where $e(t)$ is the expected total duration for an eligible individual given survival up to \bar{t} . This result is denoted length biased sampling in the literature. What we have accomplished by defining the treatment indicator $D(\bar{t})$ is that the hazard, $\chi(t)$, is independent of treatment status. This result does not hold with duration matching.

where $E_{(T^s|D=1)}(\cdot)$ is the expectation with respect to the unemployment duration until program start for those treated. It is important to emphasize that this is *not* an estimate of treatment on the treated – it is just a way of calculating an average of $\Delta_{1\bar{t}}^p$.

If there is no censoring in the data the arguments in (13) or (14) can be estimated with the mean duration for the treated and non-treated at $\bar{t} = 1, \dots, \max(t^s)$. But how should we go about estimating an objective such as (13) if the data are right-censored (at the exogenous date \bar{L})? A natural estimator is to compare the empirical hazard of the $D(\bar{t}) = 1$ group with the $D(\bar{t}) = 0$ group.¹⁰

For an individual who has been treated at t or earlier the empirical hazard at time t is given by

$$\lambda(t, D(t) = 1) = \frac{n^1(t)}{R^1(t)} = \frac{1}{R^1(t)} \sum_{i=1}^{R^1(t)} y_i(t),$$

where $y_i(t) = 1$ if individual i that starts a program in period t or earlier leaves unemployment at t and $R^1(t)$ is the number of individuals with $t^s \leq t$ at risk in t . Hence, $n^1(t) = \sum_{i=1}^{R^1(t)} y_i(t)$ is the number of individuals in the risk set leaving in t . For the comparison group we calculate

$$\lambda(t, D(t) = 0) = \frac{n^0(t)}{R^0(t)}$$

Here $R^0(t)$ is the set of individuals that has not joined the program at t and are at risk of being employed in t ; $n^0(t)$ is the number of individuals in the risk set leaving in t . Under the null hypothesis of no treatment (H_0), $\lambda(t, D(t) = 0)$ is an unbiased estimator of the hazard rate to employment for a randomly chosen individual who did not receive treatment at t .

The survival function conditioning on $D(t) = 1$ is then

$$S(t|D(t) = 1) = \prod_{s=l}^t (1 - \lambda(s, D(s) = 1)), \quad t = l, \dots, \bar{L} \quad (15)$$

and similarly for individuals in the comparison group. The effect of joining the program at t or earlier can then be calculated as the difference between

¹⁰In the following we discuss unbiasedness and consistency neglecting the problem associated with discretizing data when t is truly continuous.

the two survival functions, i.e.

$$\widehat{\Delta}(t) = S(t|D(t) = 1) - S(t|D(t) = 0), t = l, \dots, \bar{L} \quad (16)$$

The change in mean unemployment duration up to \bar{L} can now be calculated as $\widehat{\Delta}_{\bar{L}} = \sum_{t=l}^{\bar{L}} \widehat{\Delta}(t)$.

Let $S_1(t|D(t) = 1)$ be the survival function for the treated population and let $S_0(t|D(t) = 1)$ be the counterfactual survival function for this population. Observe that $S(t|D(t) = 1)$ is the maximum likelihood estimator (MLE) of $S_1(t|D(t) = 1)$; see Kalbfleisch and Prentice (1980) ch. 4. Therefore, $\text{plim} S(t|D(t) = 1) = S_1(t|D(t) = 1)$. We can now make a statement about the virtue of (16)

Proposition 3.6 $\text{plim } \widehat{\Delta}(t) = S_1(t|D(t) = 1) - S_0(t|D(t) = 1)$.

Proof. Since $T(0) \perp\!\!\!\perp D(\bar{t}) | (t \geq \bar{t})$, $S(t|D(t) = 0)$ is the MLE of $S_0(t|D(t) = 1)$. Hence, $\text{plim } S(t|D(t) = 0) = S_0(t|D(t) = 1)$ and the proposition follows. ■

It should be clear that both estimators $S(t|D(t) = 1)$ and $S(t|D(t) = 0)$ are biased estimators of the population survival functions $S_1(t)$ and $S_0(t)$ as well as the survival functions for the selected population $S_1(t|D = 1)$ and $S_0(t|D = 1)$. From the above analysis we know that the hazard rate of those entering treatment is lower than the hazard rate for randomly assigned individuals; thus, $S_0(t|D = 1) > S_0(t)$ and $S_1(t|D = 1) > S_1(t)$. It is difficult to make a statement about the relationship between $S_0(t|D(t) = 1)$ and, e.g., $S_0(t|D = 1)$ or $S_1(t|D(t) = 1)$ and, e.g., $S_1(t|D = 1)$ or $S_1(t)$. Accordingly we cannot generally determine how (16) relates to the average treatment effect and treatment on the treated. If the treatment effects do not change sign over time, the sign of $\widehat{\Delta}(t)$ is equal to the sign of the average treatment effect and treatment on the treated at t .

3.3.1 A fixed evaluation period

In the evaluation literature, it is common to use the probability of employment after a fixed time period C (e.g. one year) after the start of the program (cf. Gerfin and Lechner, 2002, and Larsson, 2000). The advantage of this approach is that treatment is allowed to affect the separation margin as well. The drawback is that there is some arbitrariness

in determining C .¹¹

Since this evaluation problem is analogous to the one we have considered above, it should be obvious that it is impossible to estimate the average treatment effect (and treatment on the treated) without additional assumptions on the process governing the inflow into treatment. The insights from the above analysis apply directly.

To illustrate the analysis a problem featuring a fixed evaluation period let us introduce the following notation. Let $Y = 1$ if the individual is employed C periods after program start and $Y = 0$ otherwise. Define $Y(1)$ and $Y(0)$ to be the associated potential outcomes. The estimand of interest is:

$$\mu(\bar{t}) = E(Y(1) - Y(0)|D(\bar{t}) = 1)$$

Consider the estimation of the components of $\mu(\bar{t})$. The estimator of the job finding probability if $t^s \leq \bar{t}$ is

$$\bar{y}_C(D(\bar{t}) = 1) = \frac{n_C(\bar{t})}{n(\bar{t})} = \frac{1}{n(\bar{t})} \sum_{i=1}^{n(\bar{t})} y_i, \bar{t} = l, \dots, \bar{L} - C$$

where $y_i = I(t_i - \bar{t} \leq C)$. The number of treated individuals at \bar{t} leaving before C is $n_C(\bar{t}) = \sum_{i=1}^{n(\bar{t})} y_i$. For the comparison group we calculate

$$\bar{y}_C(D(\bar{t}) = 0) = \frac{N_C(\bar{t})}{N(\bar{t})},$$

for individuals such that $t \geq \bar{t}$. Here, $N_C(\bar{t}) = \sum_{j=1}^{N(\bar{t})} y_j$ is the number of individuals not in treatment at \bar{t} leaving to employment before C . Note that $\bar{y}_C(D(\bar{t}) = 0)$ is an unbiased estimator of $E(Y(0)|D(\bar{t}) = 1)$. We can

¹¹We would argue is inherently more informative to estimate the survival functions, since we can always complement the analysis by looking at, e.g., the probability of reentry into the unemployment pool.

then calculate the average of these effects as

$$\begin{aligned}
\widehat{\Delta}_C &= \sum_{\bar{t}=l}^{\bar{L}} [\bar{y}_C(D(\bar{t}) = 1) - \bar{y}_C(D(\bar{t}) = 0)] \Pr(t^s = \bar{t}) \\
&= \frac{1}{n} \sum_{\bar{t}=l}^{\bar{L}} \bar{y}_C(D(\bar{t}) = 1) \frac{n(\bar{t})}{n} - \sum_{\bar{t}=l}^{\bar{L}} \bar{y}_C(D(\bar{t}) = 0) \frac{n(\bar{t})}{n} \\
&= \pi_1 - \sum_{\bar{t}=l}^{\bar{L}} \frac{N_{\bar{T}}(\bar{t})}{N(\bar{t})} \frac{n(\bar{t})}{n}, \tag{17}
\end{aligned}$$

where $\Pr(t^s = \bar{t}) = n(\bar{t})/n$ is the empirical distribution of the inflow into treatment and π_1 is the proportion of treated individuals employed C periods after treatment.

4 Monte Carlo simulation

Here we illustrate the method suggested above and contrast this with the traditional duration matching approach. To add some realism to this exercise we also consider heterogeneity at this stage. In the appendix we give a brief account of the required CIA assumption and the matching protocol.

For the purpose of the Monte Carlo simulation we generate both T and T^s as

$$\ln t_i = b_0 + x_i + \delta I(t_i > t_i^s) + \sigma_0 \varepsilon_{0i}$$

and

$$\ln t^s = a_0 + x_i + \sigma_1 \varepsilon_{1i},$$

where the density function of $\eta_h = \exp(\varepsilon_h)$, $h = 0, 1$, is the standard exponential distribution, $f(\eta_h) = \exp(-\eta_h)$. Hence both t and t^s are Weibull distributed. The hazards to employment and programs are then equal to

$$\lambda_0(t) = \alpha_0 t^{\alpha_0-1} e^{-\alpha_0(b_0+x_i)} \text{ and } \gamma(t^s = t) = \alpha_1 t^{(\alpha_1-1)} e^{-\alpha_1(a_0+x_i)},$$

where $\sigma_0^{-1} = \alpha_0$ and $\sigma_1^{-1} = \alpha_1$. x is taken to be uniformly distributed and fixed in repeated samples. $\sigma_0 = 1.2$ and $\sigma_1 = 3$, $a_0 = b_0 = 3$, and $\delta = (0, 0.2, 0.4)$.¹² The sample size is set at three levels $N = 500, 1000$ and

¹²The Monte Carlo simulation when $\delta > 0$ is performed in the following manner: If $\ln t_i = b_0 + x_i + \sigma_0 \varepsilon_{0i} > \ln t_s$ then $\ln t_i$ is increased with δ units.

1500.¹³ Throughout, the number of replications is set to 1000. In this setting, 28 percent of the sample is treated. Since $\sigma_0 = 1.2$ we have a decreasing hazard to employment. The expected length of unemployment is approximately 27 months.

We begin by studying the properties of the survival function estimator. Then we move on to consider estimators based on a fixed evaluation period.¹⁴ Throughout we discretize data to monthly intervals (j) as follows: $j = j \leq t < j + 1, j = 1, \dots, \bar{L}$.

4.1 The survival function estimator

Here we calculate the difference between the Kaplan Meier survival functions, i.e.,

$$\widehat{\Delta}(t) = S(t|D(t) = 1) - S(t|D(t) = 0), \quad t = l, \dots, \bar{L} - 1 \quad (18)$$

The results from these experiments are displayed in Figure 1-3. In Figures 1 and 2 we also display the average treatment effect (ATE) and treatment on the treated (TT). ATE is calculated as

$$\Delta(t) = S_1(t) - S_0(t), \quad t = l, \dots, \bar{L} - 1,$$

where the survival function if not treated is given by $S_0(t) = \exp(-(t \exp(b_0 + \bar{x}_1))^{\alpha_0})$ and the survival function if treated by $S_1(t) = \exp(-(t \exp(b_0 + \bar{x}_1 - \delta))^{\alpha_0})$. TT is calculated as the average difference in the conditional survival functions over the 1000 replications.

Figure 1 shows the bias of the estimators under H_0 , i.e., $\delta = 0$, in the case with an evaluation period of $\bar{L} = 240$. The figure shows that the matching estimator $\widehat{\Delta}(t)$ is an unbiased estimator of ATE. We have also examined the bias with a shorter evaluation period. The degree of bias is independent of the censoring date, \bar{L} .

¹³The parameters have been chosen with an eye towards the situation in Sweden during the early 90's (see Fredriksson and Johansson, 2002, for an application). In these data, about three quarters of the treated enroll during the first year of an unemployment spell and approximately 26 percent take part in training during the maximum of five years that we observe the individuals.

¹⁴In previous versions of the paper we have also considered a proportional hazard specification. These results basically confirm what we have already established in section 3.2. The proportional hazards estimate of δ is biased downwards in absolute value if $\delta > 0$. Moreover, the Wald test is severely undersized. These results are available on request.

Figure 2 displays the result when $\delta = 0.2$ and $\bar{L} = 240$. Since $\delta > 0$, program participation prolongs durations. $\widehat{\Delta}(t)$ is almost always larger than ATE. Moreover, $\widehat{\Delta}(t)$ is larger than TT during the initial quarter of the evaluation and lower thereafter. The change in mean unemployment duration up to \bar{L} ($\widehat{\Delta}_{\bar{L}} = \sum_{t=l}^{\bar{L}} \widehat{\Delta}_1(t)$) is 10.7 “months”. The TT and ATE up to \bar{L} are respectively equal to 14.1 and 7.6 “months”. Thus for this specific application the $\widehat{\Delta}_{\bar{L}}$ estimate is in between these two measures.

Figure 3 presents the power and size (nominal level 5%) of the Wald test for the matching estimator $\widehat{\Delta}(t)$. The Wald test is calculated as

$$\widehat{\Delta}(t)/\sqrt{\text{Var}(\widehat{\Delta}(t))},$$

where $\text{Var}(\widehat{\Delta}(t))$ is calculated as $\text{Var}(\widehat{\Delta}(t)) = \text{Var}(S(t|D(t) = 1) + \text{Var}(S(t|D(t) = 0))$ and the variance for the estimated survival function is equal to (see, e.g., Lancaster, 1990)

$$\text{Var}(S(t|D(t) = j) = S(t|D(t) = j)^2 \sum_{s=l}^t \frac{n^j(s)}{(R^j(s) - n^j(s))R^j(s)}. \quad (19)$$

Figure 3 shows that the size of the test is satisfactory. The shape of the power functions do not cause concern.

4.2 The outcome at a fixed evaluation period

The outcome variable is the average probability of employment one “year” after the start of treatment. The matching estimator is given by

$$\widehat{\Delta}_C(x) = \sum_{\bar{t}=l}^{\bar{L}} \left[\frac{1}{n(\bar{t})} \sum_{i=1}^{n(\bar{t})} [y_i - y_{c_{i\bar{t}}}] \right] \frac{n(\bar{t})}{n}, \quad (20)$$

where $c_{i\bar{t}}$ is obtained from (26) and $y_m = I(t_m - \bar{t} \leq C)$, $m = i, c_{i\bar{t}}$.

The variance is estimated as

$$\text{Var}(\widehat{\Delta}_C(x)) = \frac{\pi_1(1 - \pi_1) + \pi_0(1 - \pi_0)}{n}$$

where $\pi_0 = \frac{1}{n} \sum_{i=1}^n y_{c_{i\bar{t}}}$.

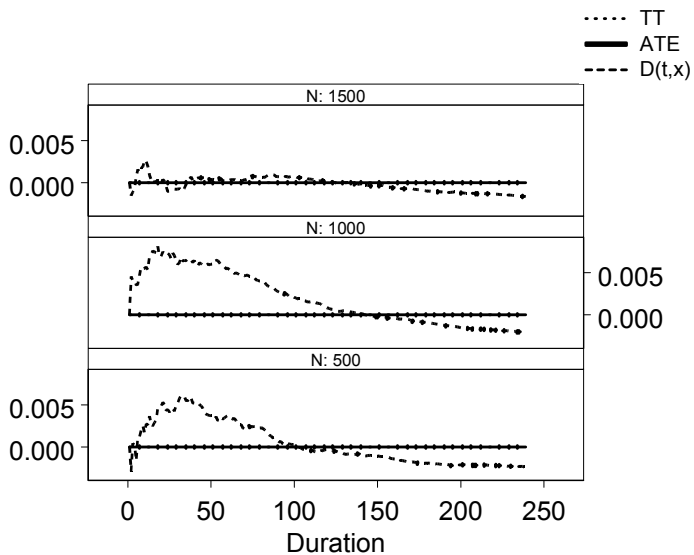


Figure 1: The bias of the survival function estimators $\hat{\Delta}(t) = D(t, x)$, ATE and TT with no treatment ($\delta = 0$) and an evaluation period of $\bar{L} = 240$ months.

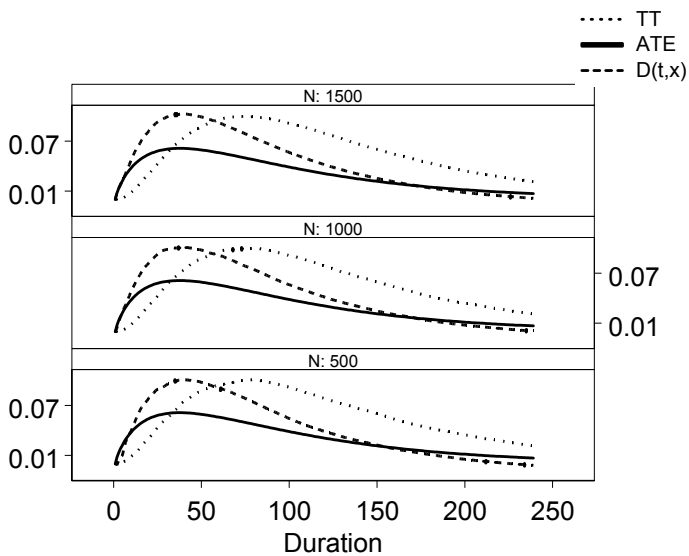


Figure 2: $\widehat{\Delta}_1(t) = D(t, x)$, ATE and TT with a treatment ($\delta = 0.2$) and evaluation period of $\bar{L} = 240$ months.

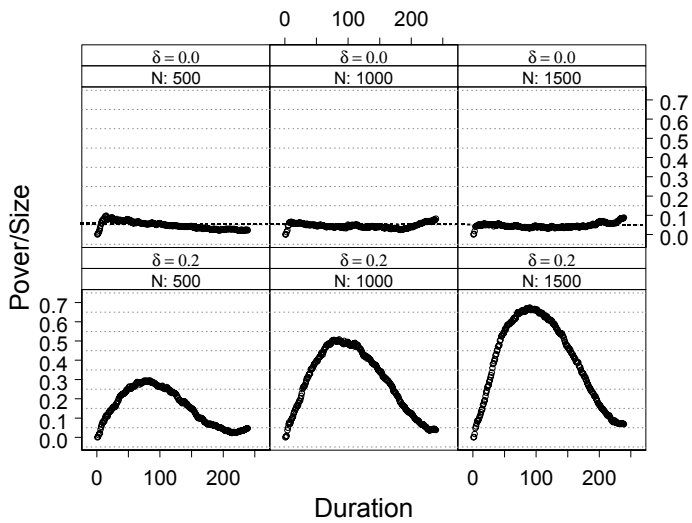


Figure 3: The power of the Wald test based on the $\widehat{\Delta}(t)$ estimator with and evaluation period of $\bar{L} = 240$ months.

This estimator is contrasted with the estimator in Lechner (1999, 2000), Gerfin and Lechner (2002) and Larsson (2000).¹⁵ The estimator in, e.g., Gerfin and Lechner (2002) is based on the approach sketched in section 3.1. First an adjusted sample of N_i^c individuals, mimicing the duration distribution of the treated, is created by randomly drawing individuals in the comparison sample. For a random draw, t_r^s , from the distribution $F(T^s | D = 1)$, a randomly drawn individual in the comparison sample is retained if $t > t_r^s$, otherwise (s)he is removed from the sample

Given a unique match¹⁶ for a treated individual, the estimator is

$$\widehat{\nabla}_C(x) = \bar{y} - \bar{y}_c \quad (21)$$

where $\bar{y} = n^{-1} \sum_{i=1}^n y_i$, $\bar{y}_c = n^{-1} \sum_{i=1}^n y_{c_i}$ and $y_m = I(t_m - \bar{t} \leq C)$, $m = i, c$. The variance is estimated as $(\bar{y}(1 - \bar{y}) + \bar{y}_c(1 - \bar{y}_c))/n$.

The results from the Monte Carlo simulation with a classification window of $C = 12$ and a maximum observation length of $\bar{L} = 48$ are shown in Table 1.¹⁷ In columns 2-4, the results from the experiment with no treatment effect is given while columns 5-7 gives the result for the $\delta = 0.2$ treatment.

We start by commenting on columns 2-4 where we present the bias, variance, and the size (nominal level 5 percent) of the Wald test of a treatment effect. The $\widehat{\Delta}_C(x)$ estimator performs satisfactory while the $\widehat{\nabla}_C(x)$ estimator suggests that employment is reduced (the estimate is significant in about 10 percent of the cases) by three percent as a result of treatment.

We now turn to the experiment with a negative treatment effect displayed in columns 5-7. Here we present the estimate, variance, and the power of the Wald test. In addition we present estimates (based on the 1000 replications) of the average treatment effect (ATE) and treatment on the treated (TT). It seems like the $\widehat{\nabla}_C(x)$ estimator does comparatively

¹⁵Lechner (1999) specifies three estimators, partial, random and inflated. He states that the random estimator (described below) performs best.

¹⁶Gerfin and Lechner (2002) base their inference on matching with replacement. When CIA holds matching with replacement reduces the bias but increases the variance in comparison to an estimator not based on replacement. We do not match with replacement but this has no bearing on the results.

¹⁷We focus on a shorter evaluation period in this instance since this is closer to the typical empirical application.

Table 1: Bias, estimate, variance, size (nominal level, 5 percent) and power in percent. Maximum observation period $\bar{L} = 48$.

	$\delta = 0$			$\delta = 0.2$		
	Bias	Variance	Size	Estimate	Variance	Power
$N = 500$						
ATE and TT				-4.48 and -13.35		
$\hat{\Delta}_C(x)$	-0.21	0.36	3.7	-8.00	0.35	27.7
$\hat{\nabla}_C(x)$	-3.39	0.41	9.3	-12.81	0.39	56.9
$N = 1000$						
ATE and TT				-4.48 and -13.31		
$\hat{\Delta}_C(x)$	0.28	0.20	5.5	-7.69	0.17	45.6
$\hat{\nabla}_C(x)$	-2.97	0.20	10.0	-12.66	0.19	84.0
$N = 1500$						
ATE and TT				-4.48 and -13.29		
$\hat{\Delta}_C(x)$	0.11	0.12	4.6	-7.91	0.11	64.0
$\hat{\nabla}_C(x)$	-3.02	0.13	10.9	-12.82	0.12	96.1

well in terms of estimating TT. However, we would argue that this is a fluke. If we would consider the case with an evaluation period of $\bar{L} = 240$, then TT equals -13.26 . In this case, $\hat{\Delta}_C(x)$ equals -11.61 , while $\hat{\nabla}_C(x)$ equals -21.74 . Moreover, if we would consider the case of a positive average treatment effect ($\delta < 0$) the power of $\hat{\nabla}_C(x)$ would be substantially lower.

4.3 Summary

So let us sum up what we have learned from the Monte Carlo simulation.

- The estimator we propose to estimate the effect of treatment on the treated up to t seems to be reliable in terms of testing for a treatment effect. But it does not seem to give much guideline about the size of the treatment effect. This is by construction, however, as we estimate a different parameter.
- Under the null hypothesis of no treatment, there is a substantial negative bias in the matching approach applied by, e.g., Gerfin and

Lechner (2002) to estimate the average treatment effect. The bias is, as expected, increasing in \bar{L} . Also, the sizes of the Wald tests are too large. Therefore, we reject the null hypothesis too often and may even find statistically significant negative treatment effects. The estimator that we propose suffers from no bias (under H_0) and the small sample performance of the Wald test gives the correct size.

5 Discussion

In this paper we have considered the evaluation problem using observational data when the program start is the outcome of a stochastic process. We have shown that without strong assumptions about the functional form of the two processes generating the inflow into program and employment it is only possible to estimate the effect of treatment on the treated up to a certain time point. It is, however, possible to test for the existence of an average treatment effect. The test can, e.g., be implemented by assuming a proportional hazards model. Another approach is to test for a treatment effect using the non-parametric survival matching estimator proposed in this paper.

We have assumed that selection is purely based on observables (the Conditional Independence Assumption, CIA). Whether CIA is reasonable assumption depends crucially on the richness of the information in the data. Even if we assume that unobserved heterogeneity is not an issue, the evaluation problem is demanding on the data. In order to construct the comparison population we need longitudinal data where we can observe the duration path up to a fixed censoring time. Knowing the entire path is crucial as we need to screen it during the evaluation time in order to define the non-treated population up to a certain time period, \bar{t} .

We think that the issues we have raised applies fairly generally to evaluations of on-going labor market programs. The problems associated with estimating the average treatment effect and treatment on the treated affect all outcomes that are functions of the outflow to employment. Hence, it applies directly when the outcome of interest is employment (or annual earnings) some time after program start. Moreover, if skill loss increases with unemployment duration, as suggested by the recent analysis in Edin and Gustavsson (2001), one should be careful when estimating the effect of treatment on wages. Although it may be tempting to screen the future in order to find individuals who did not take part in the program during

some window there is a definite risk associated with doing this. It is more probable that individuals who, by the luck of the dice, found employment are included in the comparison group. But if there is skill loss, this lucky draw will in turn spill over onto wages yielding a negative bias in the estimates of the treatment effects. Thus the issues we have raised here may be important also for studies examining the treatment effects on wages.

References

- Abbring, J.H. and G.J. van den Berg (2002), The Non-parametric Identification of Treatment Effects in Duration Models, manuscript, Free University of Amsterdam.
- Crowley, J. and M. Hu (1977), Covariance Analysis of Heart Transplant Survival Data, *Journal of the American Statistical Association*, **72**, 27-36.
- Dawid, A.P. (1979). Conditional Independence in Statistical Theory, *Journal of the Royal Statistical Society Series B*, **41**, 1-31.
- DiRienzo, A.G. and S.W. Lagakos (2001), Effects of Model Misspecification on Tests of no Randomization Treatment Effect Arising from Cox's Proportional Hazard Model. *Journal of the Royal Statistical Society Series B*, **63**, 745-757.
- Edin, P-A. and M. Gustavsson (2001), Time out of Work and Skill Depreciation, mimeo, Department of Economics, Uppsala University.
- Gerfin M. and M. Lechner (2002), A Microeconomic Evaluation of the Active Labour Market Policy in Switzerland, *Economic Journal*, **112**, 854-893.
- Heckman, J.J., R.J. Lalonde, J.A. Smith (1999), The Economics and Econometrics of Active Labor Market Programs, in O. Ashenfelter and D. Card (eds) *Handbook of Labor Economics* vol. 3, North-Holland, Amsterdam.
- Kalbfleisch, J.D. and R.L. Prentice (1980). *The Statistical Analysis of Failure Time Data*, New York: Wiley.
- Lalive, R, J. van Ours and J. Zweimüller (2002), The Impact of Active Labor Market Programs on the Duration of Unemployment, IEW Working Paper No. **51**, University of Zurich.
- Lancaster, T. (1990). *The Econometric Analysis of Transition Data*, Cambridge: Cambridge University Press.
- Larsson, L. (2000), Evaluation of Swedish Youth Labour Market Programmes, Working Paper 2000:6, Department of Economics, Uppsala University. (Forthcoming *Journal of Human Resources*.)

- Lechner, M. (1999), Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification, *Journal of Business and Economic Statistics*, **17**, 74-90.
- Lechner, M. (2000), Programme Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labour Market Policies, *Review of Economics and Statistics*, **84**, 205-220.
- Richardsson, K. and G.J. van den Berg (2002), The Effect of Vocational Employment Training on the Individual Transition Rate from Unemployment to Work, Working Paper 2002:8, Institute for Labour Market Policy Evaluation, Uppsala.
- Rosenbaum, P.R. (1995). *Observational Studies (Springer Series in Statistics)*, Springer Verlag. New york.
- Rosenbaum, P.R and D.B. Rubin (1983), The Central Role of the Propensity Score in Observational Studies for Causal Effect, *Biometrika*, **70**, 41 – 55.
- Sianesi, B. (2001), An Evaluation of the Active Labour Market Programmes in Sweden, Working Paper 2001:5, Institute for Labour Market Policy Evaluation, Uppsala
- van den Berg, G.J., B. van der Klaauw, and J.C. van Ours (2004), Punitive Sanctions and the Transition from Welfare to Work, forthcoming *Journal of Labor Economics*.

Appendix: Proof of proposition 4

It is helpful to first consider the experimental estimate $\hat{\delta}_E$. Suppose we were to conduct an experiment where at $t = 0$ individual are randomly assigned to a treatment ($D = 1$) and a comparison (control) group ($D = 0$). To simplify the exposition, assume that we observe k unique durations after randomization. Order the k survival times such that $t_{(1)} < t_{(2)} < \dots < t_{(k)}$. Associate a treatment indicator with each unique duration such that $D_{(j)} = 1$ if the individual has been treated in period $t \leq t_{(j)}$ and $D_{(j)} = 0$ otherwise. Now, consider the partial likelihood

$$L(\delta) = \prod_{j=1}^k \left(\frac{\exp(\delta D_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\delta D_l)} \right) = \prod_{j=1}^k \left(\frac{\exp(\delta D_{(j)})}{R_{(j)}(1) \exp(\delta) + R_{(j)}(0)} \right)$$

where $R_{(j)}(1)$ and $R_{(j)}(0)$ denote the number of treated and non-treated in the risk-set respectively. The maximum likelihood estimator of δ under random sampling is given as

$$\hat{\delta}_E = \ln \left(\sum_{j=1}^k D_{(j)} R_{(j)}(0) \right) - \ln \left(\sum_{j=1}^k R_{(j)}(1) (1 - D_{(j)}) \right).$$

If there is no treatment effect then

$$\begin{aligned} E(D_{(j)} R_{(j)}(0)) &= E(R_{(j)}(0) | D_{(j)} = 1) \Pr(D_{(j)} = 1) \\ &= E(R_{(j)}(0)) \Pr(D = 1) \end{aligned} \quad (22)$$

and

$$\begin{aligned} E((1 - D_{(j)}) R_{(j)}(1)) &= E(R_{(j)}(1) | D_{(j)} = 0) \Pr(D_{(j)} = 0) \\ &= E(R_{(j)}(1)) \Pr(D = 0) \end{aligned} \quad (23)$$

and hence $\hat{\delta}_E \xrightarrow{P} 0$. If $\delta > 0$ then, $R_{(j)}(1)$ and $D_{(j)}$ are no longer independent and $\Pr(D_{(j)}) \neq \Pr(D)$.

Now consider the partial likelihood in the observational setting

$$\begin{aligned} L(\delta) &= \prod_{j=1}^k \left(\frac{\exp(\delta D_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\delta D_l)} \right) \\ &= \prod_{j=1}^k \left(\frac{\exp(\delta D_{(j)})}{R_{(j)}(1) \exp(\delta) + R_{(j)}(0) + R_{(j)}(0|1)} \right) \end{aligned} \quad (24)$$

The difference compared with the partial likelihood in the experimental setting is the inclusion of $R_{(j)}(0|1)$, which is the number of individuals that have not been treated at $t \leq t_{(j)}$ but will be treated in the future. The estimator for the observational data is equal to

$$\hat{\delta}_{PH} = \ln \left(\sum_{j=1}^k D_{(j)}(R_{(j)}(0) + R_{(j)}(0|1)) \right) - \ln \left(\sum_{j=1}^k R_{(j)}(1)(1 - D_{(j)}) \right),$$

If there is no treatment effect (i.e. $\delta = 0$) then, as above, $\Pr(D_{(j)}) = \Pr(D)$; that is, the probability to enter treatment at duration $t_{(j)}$ is the same as the probability to enter treatment for a randomly chosen individual at $t = 0$. This means that the probability to belong to the comparison group is not dependent on the order (j) of the durations and as a result we get the same expressions as above; hence, $\text{plim} \hat{\delta}_{PH} = 0$. The inclusion of those treated in the future in the risk-set, i.e. $R_{(j)}(0|1)$, balances the bias that would result if only the never treated are used as comparisons.

If $\delta \neq 0$ then $\text{plim} \hat{\delta}_E = \delta$. This estimator is only based on the rank orders of the treated relative to the rank orders for those not treated.¹⁸ In the observational setting the only change (from the case without a treatment effect) in rank order is for the individuals who are never treated and the estimator $\hat{\delta}_{PH}$ will be biased downwards in absolute terms; hence $\text{plim} |\hat{\delta}_{PH}| < |\delta|$.

¹⁸Note that the rank statistic is sufficient to yield consistent estimates of the parameters in the proportional hazards model without knowledge of $\lambda_0(\cdot)$. This is also true if the true model is of the non-proportional variety (see DiRienzo and Lagakos, 2001). Wald tests of a treatment effect are biased, however.

Appendix: Matching with heterogeneity

We consider only the conditions for unbiased estimation in a time invariant setting (i.e., $\mathbf{x}_{mt} = \mathbf{x}_m \forall t \leq \bar{t}, m = i, c$).

The required conditional independence assumption (CIA) is

$$T_{\bar{t}}^p(0) \perp\!\!\!\perp D(\bar{t}) | \mathbf{x} \quad (25)$$

This assumption guarantees that

$$\begin{aligned} E_{(T^s|D=1)} \left[T_{\bar{t}}^p(0) | D(\bar{t}) = 1 \right] &= E_{(T^s|D=1)} E_{\mathbf{X}} [E(T_{\bar{t}}^p(0) | D(\bar{t}) = 0, \mathbf{x})] \\ &= E_{(T^s|D=1)} E_{\mathbf{X}} [E(T_{\bar{t}}^p(0) | D(\bar{t}) = 1, \mathbf{x})], \end{aligned}$$

where $E_{\mathbf{X}}$ is the expectation with respect to \mathbf{X} . Thus conditional on \bar{t} and \mathbf{x} we can use unemployment duration for individuals not treated at \bar{t} to estimate $E_{(T^s|D=1)} \left[T_{\bar{t}}^p(0) | D(\bar{t}) = 1 \right]$.

Let the conditional probability of being treated at \bar{t} given \mathbf{x} be given by $e(\mathbf{x}) = \Pr(D(\bar{t}) = 1 | \mathbf{x})$ and let $0 < e(\mathbf{x}) < 1$ for all \mathbf{x} .¹⁹ By (25) it then holds that (see Rosenbaum and Rubin, 1983)

$$\mathbf{x} \perp\!\!\!\perp D(\bar{t}) | e(\mathbf{x}).$$

So, under the CIA (25), the counterfactual can be estimated as

$$\begin{aligned} E_{(T^s|D=1)} \left[T_{\bar{t}}^p(0) | D(\bar{t}) = 1 \right] &= E_{(T^s|D=1)} E_e [E(T_{\bar{t}}^p(0) | D(\bar{t}) = 0, e(\mathbf{x}))] \\ &= E_{(T^s|D=1)} E_e [E(T_{\bar{t}}^p(0) | D(\bar{t}) = 1, e(\mathbf{x}))], \end{aligned}$$

where E_e is the expectation with respect to $e(\mathbf{x})$.

A matching algorithm We use a one-to-one matching procedure based on estimated propensity scores $\hat{\omega}_m = e(\mathbf{x}_m, \hat{\beta})$, where $\hat{\beta}$ is an estimated parameter vector from, e.g., a logit maximum likelihood estimator. Let treated individuals at \bar{t} be indexed by i and individuals in the comparison group at \bar{t} by c . The unique match (for each \bar{t}) is found by minimizing the distance between the estimated propensity scores:

$$c_{i\bar{t}} = \arg \min_{c \in N(\bar{t})} |\hat{\omega}(i) - \hat{\omega}(c)|, \quad (26)$$

¹⁹This means that for each \mathbf{x} satisfying the CIA there must be individuals in both states.

where $\widehat{\omega}(c)$ is the $(N(\bar{t}) \times 1)$ vector of estimated propensity scores at time \bar{t} . After finding a match for individual i , the process starts over again until $n_{cs}(\bar{t})$ comparable individuals is found in the comparison sample. Here $n_{cs}(\bar{t})$ is the number of individuals on the common support.

The process is started by randomly drawing an individual in the treatment sample, then one should make another random draw from the remaining $n_{cs}(\bar{t}) - 1$ treated individuals and so on until $n_{cs}(\bar{t})$ matching individuals are found.

With a complete set of pairs of treated and non-treated individuals the estimators (14) and (17) are given by

$$\widehat{\Delta}_{i\bar{t}}^p(x) = \sum_{\bar{t}=1}^{\bar{L}} \left(\frac{1}{n_{cs}(\bar{t})} \sum_{i=1}^{n_{cs}(\bar{t})} [t_i - t_{c_{i\bar{t}}}] \right), \bar{t} = l, \dots, \bar{L}$$

$$\widehat{\Delta}_C(x) = \sum_{\bar{t}=l}^{\bar{L}} \left[\frac{1}{n_{cs}(\bar{t})} \sum_{i=1}^{n_{cs}(\bar{t})} [y_i - y_{c_{i\bar{t}}}] \right] \Pr(T^s = \bar{t}), \bar{t} = l, \dots, \bar{L}$$

while the estimator (16) is given by

$$\widehat{\Delta}(t, x) = S_x(t|D(t) = 1) - S_x(t|D(t) = 0), t = l, \dots, \bar{L}$$

where $S_x(t|D(t) = 0) = \prod_{s=l}^t (1 - \lambda_x(s, D(s) = 0))$ and

$$\lambda_x(s, D(s) = 0) = \frac{1}{R_{\bar{t}}^1(s)} \sum_{i=1}^{R_{\bar{t}}^1(t)} y_{c_{i\bar{t}}}(s),$$

where $R_{\bar{t}}^1(s)$ is the risk set for the matched individuals at \bar{t} still at risk in time period s .