

Program Evaluation and Random Program Starts*

Peter Fredriksson[†] and Per Johansson[‡]

October 14, 2002

(Preliminary. Comments welcome)

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.

1 Introduction

A standard problem in the evaluation literature is the problem with counterfactual outcomes, i.e. the would-be if not treated. This is a missing data problem and the would-be outcome must be created using reasonable assumptions.

In this paper 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

*Thanks to Jeffrey Smith for very useful comments. Also thanks to seminar participants at the Department of Statistics, Umeå university, and IFAU.

[†]Department of Economics, Uppsala University and IFAU. 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).

[‡]Institute for Labour Market Policy Evaluation (IFAU). Phone: +46-18-471 70 86. Email: per.johansson@ifau.uu.se.

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 been increasingly become the focus of the labor market evaluation literature so our analysis should have wide applicability.¹ 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 instance, it follows immediately that the points we raise should be taken into consideration in analysis of earnings outcomes.

The evaluation problem could be addressed by using parametric regression models, where the treatment participation decision is simultaneously modeled with the outcome of interest. However, this approach relies exclusively on the ability to correctly specify the appropriate bivariate distribution. Here, we use a less “heroic” approach and assume that we can construct the counterfactual outcome using the method of matching.

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 the average treatment effect on the unemployed population. It is, however, possible to estimate the program effect for those being treated up to a certain time point. 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 average treatment effect is that in order to get at this effect one would like to define a comparison group who was never treated. But finding individuals who were never treated involves conditioning on the future since programs 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 did not enter the program in the future to a large extent consist of those who had the luck of finding a job.² Therefore, studies that define the comparison group in this way will estimate program effects that are biased downwards.

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 and possible estimands. Section 3 addresses the specific problem

¹The 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.

²There is an informal discussion along these lines in Sianesi (2001).

associated with random program starts. Section 4 considers alternative estimators. We propose an estimator of treatment on the treated up to certain point in time. In section 5 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 6, finally, concludes.

2 The framework

We have the following world in mind. Consider a set of individuals who enter unemployment at time t_0 . At the time of unemployment entry these individuals are identical. Alternatively, we could assume that matching on the observed covariates at the time of 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.³ The hazard rates to employment are then given by

$$\begin{aligned}\lambda_0(t) &= \tilde{\lambda}_0(t)I(v_2(t) \geq v_0(t)) \\ \lambda_1(t) &= \tilde{\lambda}_1(t)I(v_2(t) \geq v_1(t))\end{aligned}$$

for treated and untreated individuals respectively.⁴ 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 (e.g. they might be Poisson). Then unemployment

³The openly unemployed refers to the unemployed who do not participate in a labor market program.

⁴Throughout we assume that the effect of treatment occurs directly upon enrollment. We make this assumption for expositional convenience. It is not important for the substance of the paper.

durations are exponentially distributed (with parameter $\exp(\beta_0)$) and we can represent the potential duration if not treated as

$$\ln T(0) = \beta_0 + \varepsilon_0, \quad (1)$$

where ε_0 is Type I extreme value distributed.

Further the log of the duration until program start (t_0^p) has an analogous representation, i.e.,

$$\ln t_0^p = \beta_1 + \varepsilon_1 \quad (2)$$

where ε_1 is also Type I extreme value distributed. Observe that post program entry duration in unemployment is simply given by $u = t - t_0^p$. So these two specifications also imply a specification for the post treatment duration if not treated $U(0)$.

Equations (1) and (2) are written in the form of accelerated duration models (ADM). Of course, the representations in (1) and (2) are unduly restrictive. We have no reason to postulate a particular distribution for ε_j , $j = 0, 1$, for instance. Therefore, we will sometimes work with more general forms of the ADM

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

$$\ln t_0^p = \beta_1 + \sigma_1 \varepsilon_1 \quad (4)$$

without making distributional assumptions about ε_j . Only if ε_j is extreme value distributed do (3) and (4) imply a proportional hazards 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

$$D = I(t > t_0^p) \quad (5)$$

i.e. individuals are observed to take treatment if their unemployment duration (t) is longer than their duration till program start (t_0^p).

2.1 Objects of evaluation

Let the post treatment potential unemployment duration if treated be $U(1)$ then in the ideal world we would either like to estimate the average treatment

effect

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

or treatment on the treated

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

Thus, we are generally interested in the causal effect of treatment on the post-treatment durations. One of these potential durations is of course a missing counterfactual outcome. For example, we observe $U(1)$ for a treated individual but we do not observe $U(0)$. This is always true, even in experiments.

What makes this problem somewhat special is that we in many realistic situations 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 $E(U(0)|D = 1)$ or $E(U(0))$. 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, $t_0^p = \bar{t}_0^p$ say.⁵

An alternative estimand is then to define the treatment effects in terms of the durations since unemployment entry

$$\Delta = E(T(1)) - E(T(0)) \quad (8)$$

$$\Delta_1 = E(T(1)|D = 1) - E(T(0)|D = 1) \quad (9)$$

In this case we, thus, ask whether treatment affected unemployment duration as a whole. The interpretation of the two “parameters” in (8) and (9) is perhaps not as clean as the ones defined in (6) and (7). Since the treatment occurs after unemployment entry, the effects of treatment will be diluted.

To estimate the evaluation parameters (7) or (9), potential outcomes should be conditionally (or mean) independent of treatment; using the notation of Dawid (1979), it must be true that

$$U(0) \perp\!\!\!\perp D, \text{ or } T(0) \perp\!\!\!\perp D$$

For the evaluation parameters (6) and (8) both potential outcomes should be independent of the treatment, i.e.,

$$(U(1), U(0)) \perp\!\!\!\perp D, \text{ or } (T(1), T(0)) \perp\!\!\!\perp D$$

⁵Of 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.

Using the ADM framework above we can represent the log of the potential durations if treated and not treated as

$$\ln U(0) = \beta_0 + \sigma_0 \eta_0 \text{ and } \ln U(1) = \delta_1 + \sigma_01 \eta_1,$$

where η_0 is the censored (at $t > t_0^p$) distribution for ε_0 . If $\delta_1 \neq \delta_0$ and/or $\sigma_01 \eta_1 \neq \sigma_0 \eta_0$ this implies that the offer rates differ by treatment status. Using the switching regression framework the observed post treatment duration for the treated can then be written as

$$\ln u = \beta_0 + D([\delta_1 - \beta_0] + [\sigma_01 \eta_1 - \sigma_0 \eta_0]) + \sigma_0 \eta_0$$

where D is a treatment indicator. Observe that if $\eta_0 \neq \eta_1$ this is a random coefficient model. If there is no treatment effect i.e. $\lambda_1(t) = \lambda_0(t)$ above, then $\sigma_0 \eta_0 = \sigma_01 \eta_1$ and $\delta_1 = \beta_0$. We can then we write the observed post treatment duration for the treated as

$$\ln u = \beta_0 + \sigma_0 \eta_0.$$

We can hence see that the post treatment duration is stochastically dependent on the pre treatment duration. In order for $U(0) \perp\!\!\!\perp D$, we generally need that $T(0) \perp\!\!\!\perp D$. This will be dicussed in more detail in the following section.

Let us first digress for a moment and define two *potential* survival functions

$$S_1^p(t) = \exp\left(-\int_{t_0^p}^t \lambda_1(\tau) d\tau\right)$$

$$S_0^p(t) = \exp\left(-\int_{t_0^p}^t \lambda_0(\tau) d\tau\right)$$

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

$$\Delta^p(t) = S_1^p(t) - S_0^p(t), \quad t \in (t_0^p, \infty)$$

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

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

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

3 The random start problem

Here we want to illustrate the complications stemming from the fact that program start is the outcome of a stochastic process. Remember that the population is homogenous with respect to observed and unobserved characteristics at the time of unemployment entry. However, individuals make a draw from the idiosyncratic duration distribution. This feature in combination with the process generating observational data has important implications for the possibility of getting unbiased estimates of the treatment parameters.

To illustrate the problem in a particularly simple way, consider the ADM model in equations (3) and (4) when $\sigma_0 = \sigma_1 = 1$. We observe individuals taking treatment if $t > t_0^p$. Thus

$$D = \begin{cases} 1 & \text{if } \varepsilon_0 - \varepsilon_1 > \beta_1 - \beta_0 \\ 0 & \text{if } \varepsilon_0 - \varepsilon_1 \leq \beta_1 - \beta_0 \end{cases}$$

Therefore $E(\varepsilon_0|D = 1) > E(\varepsilon_0|D = 0)$ since $E(\varepsilon_0|(\varepsilon_0 - \varepsilon_1) > (\beta_1 - \beta_0)) > E(\varepsilon_0|(\varepsilon_0 - \varepsilon_1) \leq (\beta_1 - \beta_0))$. The data generating process is thus such that "unlucky" individuals (those with high ε_0) are more likely to enter treatment.⁶ This implies that to estimate an average treatment effect one has to invoke additional identifying assumptions; one option, which is not particularly attractive, is to postulate a bivariate distribution for the durations T and T_0^p .

Instead of relying on functional form we would like to consider a less structural approach to resolve the problem of inference. One possible way is to create a duration matched comparison sample to those flowing into treatment, i.e., to condition on t_0^p . We consider this and other approaches in the next section.

4 Potential estimators

In this section we consider alternative strategies to estimate the parameters of interest. Before discussing potential estimators let us introduce 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 $N = n + N^c$. When considering the virtues and drawbacks of each potential

⁶This 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.

estimator we will focus mainly on the case where there is no treatment since a minimum requirement on any estimator is that it will be unbiased under the null hypothesis.

4.1 Duration matching

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.

Suppose, for concreteness, that we want to estimate (7). The expectation of the post-treatment duration is given by

$$E(U(1)|D = 1) = E[E(T|T > t_0^p) - t_0^p] = E(T(1)|D = 1) - E(T_0^p(1)|D = 1), \quad (10)$$

where $E(T_0^p(1)|D = 1)$ is the expected duration until program start for those treated. The expectation in (10) can be estimated as

$$\hat{u} = \frac{1}{n} \sum_{i=1}^n (t_i - t_{i0}^p).$$

A potential estimator of the counterfactual outcome, $E(U(0)|D = 1)$, is based on random sampling from the inflow distribution, $F(t_0^p|D = 1)$. For a random draw, t_{i0}^p , an individual from the comparison sample is matched if the unemployment duration for this randomly assigned individual satisfies $t_c > t_{i0}^p$. Applying this procedure we get a duration matched comparison sample (consisting of n matches) and may calculate

$$\hat{u}_c = \frac{1}{n} \sum_{i=1}^n u_{c_i},$$

where $u_{c_i} = t_c - t_{i0}^p$ is the observed unemployment duration after t_{i0}^p for a (randomly assigned) matched individual. The treatment effect would then be estimated as

$$\hat{\Delta}_1^p = \hat{u} - \hat{u}_c. \quad (11)$$

Proposition 1 *The duration matched estimator ($\hat{\Delta}_1^p$) is positively biased*

Proof. To prove this proposition let us again consider the ADM – see equations (3) and (4) – with $\sigma_0 = \sigma_1 = 1$. Let $U_{\bar{t}}(0)$ be the potential duration in unemployment if not treated up to a fixed time period \bar{t} we have

$$E(\ln T(0)|D = 1, t > \bar{t}) = \beta_0 + E(\varepsilon_0 | \ln \bar{t} < \ln t)$$

$$E(\ln T(0)|D = 0, t > \bar{t}) = \beta_0 + E(\varepsilon_0 | \ln \bar{t} < \ln t \leq \beta_1 + \varepsilon_1)$$

Since $E(\varepsilon_0 | \ln \bar{t} < \ln t \leq \beta_1 + \varepsilon_1) < E(\varepsilon_0 | \ln \bar{t} < \ln t)$ we get that $E(\ln T(0)|D = 1, t > \bar{t}) > E(\ln T(0)|D = 0, t > \bar{t})$ and since $T(0) = U_{\bar{t}}(0) + \bar{t}$

$$E(U(0)|D = 1, t > \bar{t}) > E(U(0)|D = 0, t > \bar{t})$$

Thus, we have $U(0) \not\perp D | (t > \bar{t})$. ■

Notice that this result holds 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.

The intuition for Proposition 1 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)

Remark 1 *The bias of the duration matched estimator with censored observations on unemployment duration is positive but smaller in magnitude than the estimator based on completed spells.*

Proof. Suppose that unemployment duration is censored at c . Then for each random draw \bar{t} from the distribution $F(t_0^p | D = 1)$ $D = 1$ if $t_0^p < c$ and $t_0^p < t$ while $D = 0$ if $\ln \bar{t} < \ln t$ and $\min(\ln t, \ln c) < \beta_1 + \varepsilon_1$. Therefore

$$E(\ln T(0)|D = 1, t > \bar{t}) = \beta_0 + E(\varepsilon_0 | \ln \bar{t} < \ln t, \ln \bar{t} < \ln c)$$

$$E(\ln T(0)|D = 0, t > \bar{t}) = \beta_0 + E(\varepsilon_0 | \ln \bar{t} < \ln t, \min(\ln t, \ln c) < \beta_1 + \varepsilon_1)$$

So $E(U(0)|D = 1, t > \bar{t}) > E(U(0)|D = 0, t > t_0^p)$ by the same argument as above. ■

Thus, censoring implies that the bias will be reduced in magnitude. But this procedure of defining the comparison group will still yield an upward biased estimate of the treatment effect on unemployment duration.

To sum up, it is not possible to create a sample of matching individual who do not receive treatment at any point time. In defining the treated and the comparisons, the sampling is on ε_0 , which in turn determines the (potential) outcome $U(0)$. Thus for those treated we have large ε_0 and hence large $U(0)$ while the opposite is true for the untreated.

4.2 The proportional hazards 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 van den Berg and Richardsson (2002). The last two papers use the approach suggested in Abbring and van den Berg (2002). The main thrust of the paper by Abbring and van den Berg is that variation in the timing of treatment identifies the coefficient of interest in a proportional hazards model. Here we examine what happens when we impose a proportional hazard model in our context.⁷

The hazard after treatment is assumed to be

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

where $D = I(t > t_0^p)$. If δ estimates the average treatment effect then $\tilde{\lambda}_0(t) = h_0(t)$. So if the model has a proportional hazards specification, the outflow of the treated relative to the non-treated identifies the treatment effect: $\lambda_1(t) = \tilde{\lambda}_0(t) \exp(\delta D)$.⁸

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

Proposition 2 *The data generating process $D = I(t > t_0^p)$ implies that the baseline hazard for the treated is not equal to the baseline hazard in the population, i.e., $h_0(t) \neq \tilde{\lambda}_0(t)$.*

Proof. Proposition 1 implies that $E(T(0)|D = 1) > E(T(0)|D = 0)$. Since this is true for any censoring point $t = c > 0$ (see Remark 1) we have $S(t|D = 1) > S(t|D = 0)$. Now,

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

⁷The framework in Abbring and van den Berg (2002) is much more general than the model presented here.

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

■

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

$$h_0(t) = \tilde{\lambda}_0(t|D=1) \Pr(D(t)=1) + \tilde{\lambda}_0(t|D=0) \Pr(D(t)=0)$$

Proposition 2 implies that $\tilde{\lambda}_0(t|D=1) \neq \tilde{\lambda}_0(t|D=0)$. 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 these two treatment parameters? Proposition 3 outlines the results

Proposition 3 *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 3b) 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 $t > s$. 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.

To sum up, Proposition 3 shows that the proportional hazards specification is a fertile ground for testing (See DiRienzo and Lagakos, 2001, however, for tests of treatment effects using misspecified proportional hazard models). However, the estimate will be smaller in absolute value than the average treatment effect when a treatment effect exists.

4.3 Estimation based on right-censoring

As we have emphasized repeatedly it is the fact that program starts are random that complicates the evaluation. Suppose instead that we discretize time and right-censor the data at a fixed time point \bar{t} . We then define $D(\bar{t}) = I(t_0^p \leq \bar{t} \leq t)$.

It is straightforward to show that

Lemma 1 *Potential unemployment duration is independent of the treatment indicator $D(\bar{t})$.*

Proof. Consider the ADM model (3) with $\sigma_0 = \sigma_1 = 1$. Then

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

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

and hence $T(0) \perp\!\!\!\perp D(\bar{t})|(t \geq \bar{t})$ and $U(0) \perp\!\!\!\perp D(\bar{t})|(t \geq \bar{t})$. ■

The gain of right-censoring the data is immediate: potential unemployment duration is conditionally independent of the treatment indicator.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):

$$\Delta_{1\bar{t}}^p = E(U(1)|D(\bar{t}) = 1) - E(U(0)|D(\bar{t}) = 1) \quad (12)$$

If the effect of entering at \bar{t} is constant over time one would expect that 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

$$E_{(T_0^p|D=1)}(\Delta_{1\bar{t}}^p) = E_{(T_0^p|D=1)} [E(U(1)|D(\bar{t}) = 1) - E(U(0)|D(\bar{t}) = 1)] \quad (13)$$

where $E_{(T_0^p|D=1)}(\cdot)$ is the expectation with respect to the unemployment duration until program start for those treated.

If the data are not censored the arguments in (12) or (13) can be estimated with the mean duration for the treated and non-treated at $\bar{t} = 1, \dots, \max(t_0^p)$.

How should we go about estimating an objective such as (12) if the data are right-censored (at the exogenous date \bar{L})? A natural estimator is to

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

$$\chi(t) = \tilde{\lambda}_0(t) \frac{t}{e(t)} \leq \tilde{\lambda}_0(t), \quad t \geq \bar{t}$$

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.

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 $t_0^p \leq t$ 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 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

$$S(t|D(t) = 1) = \prod_{s=l}^t (1 - \lambda(s, D(s) = 1)), \quad t = l, \dots, \bar{L} \quad (14)$$

and similarly for individuals in the comparison group.¹¹ The effect of joining the program at t or earlier can then be calculated as the difference between the two survival functions, i.e.

$$\widehat{\Delta}(t) = S(t|D(t) = 1) - S(t|D(t) = 0), \quad t = l, \dots, \bar{L} \quad (15)$$

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 (up to t) 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, $p \lim S(t|D(t) = 1) = S_1(t|D(t) = 1)$ (in discrete time). We can now make a statement about the virtue of (15)

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

¹¹Notice that if $\delta = 0$, the survival function based on estimating the proportional hazards model is equal in expected value to the non-parametric one in (14).

Proposition 4 $p \lim \widehat{\Delta}(t) = S_1(t|D(t) = 1) - S_0(t|D(t) = 1)$.

Proof. Since $T(0) \perp\!\!\!\perp D(\bar{t}) | (t \geq \bar{t})$, $S(t|D(t) = 0)$ is the MLE of $S_0(t|D(t) = 1)$ and $p \lim S(t|D(t) = 0) = S_0(t|D(t) = 1)$ (in discrete time) 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 $S_1(s)$ and $S_0(s)$ 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 (15) 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 .

4.3.1 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, 2001, 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 .¹²

Since this evaluation problem is analogous as 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.

Let $Y = 1$ if the individual is employed C periods after program start and $Y = 0$ otherwise. Also, let $Y(1)$ and $Y(0)$ be the associated potential outcomes. From the above discussion it follows that the estimand of interest is:

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

Let us consider the estimation of the components of $\mu(\bar{t})$. The estimator

¹²We 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.

of the job finding probability if $t_0^p \leq \bar{t}$ 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 $t \geq \bar{t}$. In this expression, $N_C(\bar{t}) = \sum_{j=1}^{N(\bar{t})} y_j$ 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 then calculate the average of these effects as

$$\begin{aligned} \widehat{\Delta}_C &= \sum_{\bar{t}=l}^{\bar{L}} [\bar{y}_C(D(\bar{t}) = 1) - \bar{y}_C(D(\bar{t}) = 0)] \Pr(t_0^p = \bar{t}) \\ &= \frac{1}{n} \sum_{\bar{t}=l}^{\bar{L}} \bar{y}_C(D(\bar{t}) = 1) \frac{n(\bar{t})}{n} - \sum_{\bar{t}=l}^{\bar{L}} \bar{y}_C(D(\bar{t}) = 0) \frac{n(\bar{t})}{n} \\ &= \pi_1 - \sum_{\bar{t}=l}^{\bar{L}} \frac{N_T(\bar{t})}{N(\bar{t})} \frac{n(\bar{t})}{n}, \end{aligned} \tag{16}$$

where π_1 is the proportion of treated individuals employed C periods after treatment.

5 Monte Carlo simulation

Here we illustrate the method suggested above and contrast this with alternative methods in the literature. 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_0^p as

$$\ln t_i = b_0 + b_1 x_i + \delta I(t_i > t_{0i}^p) + \sigma_0 \varepsilon_{0i}$$

and

$$\ln t_{0i}^p = a_0 + a_1 x_i + \sigma_1 \varepsilon_{1i},$$

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_0^p are Weibull distributed. The hazards to employment and programs are then equal to

$$\lambda_0(t) = \alpha_0 t^{\alpha_0 - 1} e^{-\alpha_0(b_0 + b_1 x_i)} \text{ and } \gamma(t_0^p = t) = \alpha_1 t^{(\alpha_1 - 1)} e^{-\alpha_1(a_0 + \alpha_1 x_i)},$$

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 $\sigma_1 = 3$, $a_0 = b_0 = 3$, $a_1 = 1$, $b_1 = (0, 1)$ and $\delta = (0, 0.2, 0.4)$. The sample size is at three levels $N = 500, 1000$ and 1500 .¹³ In the homogenous setting (i.e. $b_1 = 0$), around 20 percent of the sample is treated while for the $b_1 = 1$ setting, 28 percent of the sample is treated. Since $\sigma_0 = 1.2$ we have a decreasing hazard to employment. Under homogeneity, the expected length in open unemployment is 22 months ($E(T) = \exp(b_0)\Gamma(1 + \sigma_0)$, see e.g., Lancaster (1990, ch. 2)) and the expected length under observed heterogeneity is approximately 27 months.

From the above specification it seems possible to write

$$\lambda_1(t) = \lambda_0(t)e^{-\delta}$$

However this is not true. Instead $\lambda_1(t|D = 1) = \lambda_0(t|D = 1)e^{-\delta}$ where $\lambda_0(t|D = 1)$ is the hazard rate for those selected into treatment.

We begin by considering the performance of the proportional hazards specification; see section 5.1. The estimates of the treatment effects is then based on the partial MLE (see, e.g., equation 25). In the following section, we study the survival function estimator. Lastly, we consider the properties of estimators based on a fixed evaluation period. The analysis in sections 5.2 and 5.3 is based on discretizing data to monthly intervals (j) as follows: $j = j \leq t < j + 1$, $j = 1, \dots, \bar{L}$.

5.1 The proportional hazards specification

A proportional hazards model is estimated using the partial MLE. The results are displayed in Table 1 for the case where $b_1 = 0$. As a reference case we also provide the result from the partial MLE when individuals are randomly assigned to treatment. The proportional hazards estimator applied to the setting with observational data is denoted $\hat{\delta}_{PH}$; the estimator applied to the data with random assignment is denoted $\hat{\delta}_E$.

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.

Table 1: Bias, mean square error (MSE) and power (size) of a test of $H_0 : \delta = 0$

	$\delta = 0$			$\delta = 20$			$\delta = 40$		
	bias	MSE	size	bias	MSE	power	bias	MSE	power
$N = 500$									
$\hat{\delta}_{PH}$	0.15	0.70	0.8	-12.13	2.13	0.8	-24.24	6.35	4.8
$\hat{\delta}_E$	-0.29	0.80	4.8	0.26	0.00	51.0	-0.40	0.81	98.7
$N = 1000$									
$\hat{\delta}_{PH}$	0.00	0.36	0.7	-12.65	1.94	2.00	-24.17	6.50	24.1
$\hat{\delta}_E$	-0.38	0.40	6.1	-0.09	0.40	86.4	0.12	0.41	100
$N = 1500$									
$\hat{\delta}_{PH}$	-0.05	0.23	0.4	-12.79	1.85	4.3	-24.17	6.05	53.8
$\hat{\delta}_E$	-0.03	0.28	6.0	0.17	0.28	96.4	-0.20	0.28	100

Note: All figures are multiplied with 100.

Results From Table 1 we see that the bias and mean square error (MSE) in the observational and experimental settings are comparable when there is no effect, i.e., when $\delta = 0$. We report Wald tests of the hypothesis $\delta = 0$, which are based on the negative inverse of the Hessian. The sizes of these tests are too small when using observational data. This is so, since the true model is non-proportional, even under H_0 . This small sample result confirms analytical results (see DiRienzo and Lagakos, 2001). When $\delta \neq 0$ we see a fairly substantial downward bias in the estimates based on observational data. Moreover, the power of the Wald test is low.

5.2 Non-parametric specification

5.2.1 The survival function

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

$$\hat{\Delta}_k(t) = S(t|D(t) = 1) - S(t|D(t) = 0), \quad t = l, \dots, \bar{L} - 1, \quad k = 0, 1. \quad (17)$$

in both the homogenous and heterogenous settings. Here the subindex $k = 0$ refers to the naive estimator and $k = 1$ refers to the matching estimator.

Results The results from these experiments are displayed in Figure 1-3.¹⁴ In Figure 1 and 2 we also display the average treatment effect (ATE) and treatment on the treated (SATE). 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 $S_0(t) = \exp(-(t \exp(b_0 + b_1 \bar{x}_1))^{\alpha_0})$ and the survival function if treated by $S_1(t) = \exp(-(t \exp(b_0 + b_1 \bar{x}_1 - \delta))^{\alpha_0})$. SATE is calculated as the average difference in the conditional survival functions over 1000 replications.

Figure 1 shows the bias of the estimators under H_0 , i.e. $\delta = 0$, in the case with heterogeneity, $b_1 = 1$ and an evaluation period of $\bar{L} = 240$. It is evident that the matching estimator $\hat{\Delta}_1(t)$ is an unbiased estimator of ATE and reduces the bias that occurs from the naive estimator. The degree of bias is independent of the censoring date, \bar{L} . In the situation without heterogeneity (i.e. $b_1 = 0$) the bias is less for both estimators and generally smaller for the naive estimator $\hat{\Delta}_0(t)$.

Figure 2 displays the result when $\delta = 0.2$ and $b_1 = 1$ for $\bar{L} = 240$.¹⁵ Since $\delta > 0$, the treatment effect is negative, i.e., program participation prolongs durations. The $\hat{\Delta}_1(t)$ estimator is almost always larger than the ATE. The $\hat{\Delta}_1(t)$ estimator is larger than SATE during the initial quarter of the evaluation and lower thereafter. The change in mean unemployment duration up to \bar{L} ($\hat{\Delta}_{\bar{L}} = \sum_{t=l}^{\bar{L}} \hat{\Delta}_1(t)$) is 10.7 “months”. The SATE and ATE up to \bar{L} are respectively equal to 14.1 and 7.6 “months”. Thus for this specific application the $\hat{\Delta}_{\bar{L}}$ 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 $\hat{\Delta}_1(t)$ in the $b_1 = 1$ setup. The Wald test is calculated as

$$\hat{\Delta}_1(t) / \sqrt{\text{Var}(\hat{\Delta}_1(t))},$$

where $\text{Var}(\hat{\Delta}_1(t))$ is calculated as $\text{Var}(S(t|D(t) = 1) + \text{Var}(S(t|D(t) = 0))$ and the variance for the estimated survival function is equal to (see, e.g., Lancaster, 1990)

$$\text{Var}(S(t|D(t) = j) = S(t|D(t) = j)^2 \sum_{s=l}^t \frac{n^j(s)}{(R^j(s) - n^j(s))R^j(s)}. \quad (18)$$

¹⁴To restrict the length of the paper we do not present all available graphs. Some results are discussed without graphs. These unreported graphs can be obtained from the authors upon request.

¹⁵When $b_1 = 0$ both graphs are very similar to the situation with the matching estimator and $b_1 = 1$.

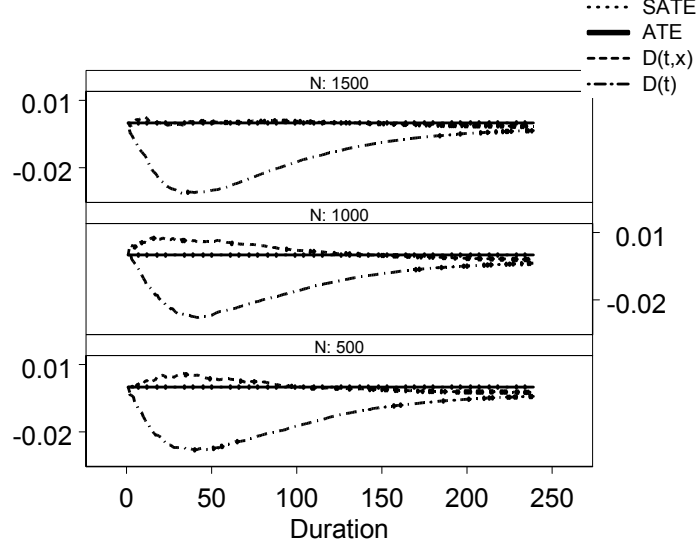


Figure 1: The bias of the survival function estimators $\hat{\Delta}_0(t) = D(t)$ and $\hat{\Delta}_1(t) = D(t, x)$ and the ATE and SATE when there is no treatment ($\delta = 0$) and heterogeneity ($b_1 = 1$) and evaluation period $\bar{L} = 240$ months.

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

5.2.2 The outcome at a fixed evaluation period

The outcome variable is the average probability of employment one "year" after the start of treatment. To make the results conformable to the earlier ones we set \bar{L} to 48 and 240. The estimator of the treatment on the treated effect up to \bar{L} is given in (16) and repeated here for convenience

$$\hat{\Delta}_C = \sum_{\bar{t}=l}^{\bar{L}} [\bar{y}_C(D(\bar{t}) = 1) - \bar{y}_C(D(\bar{t}) = 0)] \Pr(t_0^p = \bar{t}). \quad (19)$$

The corresponding matching estimator is given by

$$\hat{\Delta}_C(x) = \sum_{\bar{t}=l}^{\bar{L}} \left[\frac{1}{n(\bar{t})} \sum_{i=1}^{n(\bar{t})} [y_i - y_{c_{i\bar{t}}}] \right] \Pr(t_0^p = \bar{t}), \quad (20)$$

where $c_{i\bar{t}}$ is obtained from (27) and $y_m = I(t_m - \bar{t} \leq C)$, $m = i, c_{i\bar{t}}$.

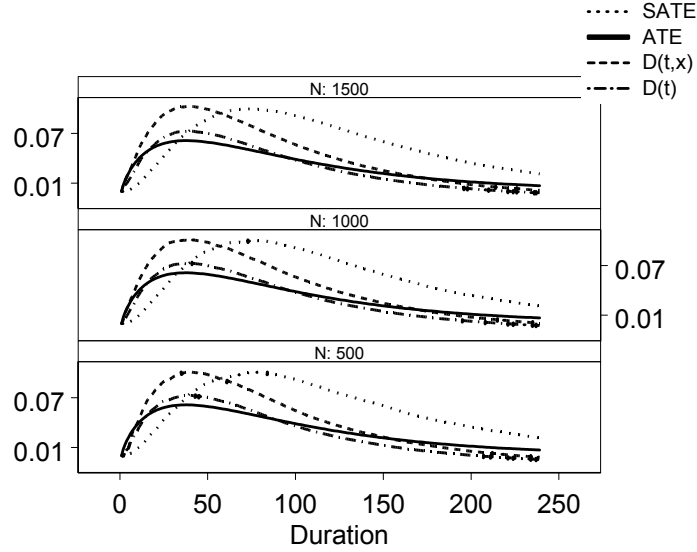


Figure 2: The estimators $\widehat{\Delta}_0(t) = D(t)$ and $\widehat{\Delta}_1(t) = D(t, x)$ and the ATE and SATE when there is a treatment effect ($\delta = 0.2$) heterogeneity ($b_1 = 1$) and evaluation period $\bar{L} = 240$ months.

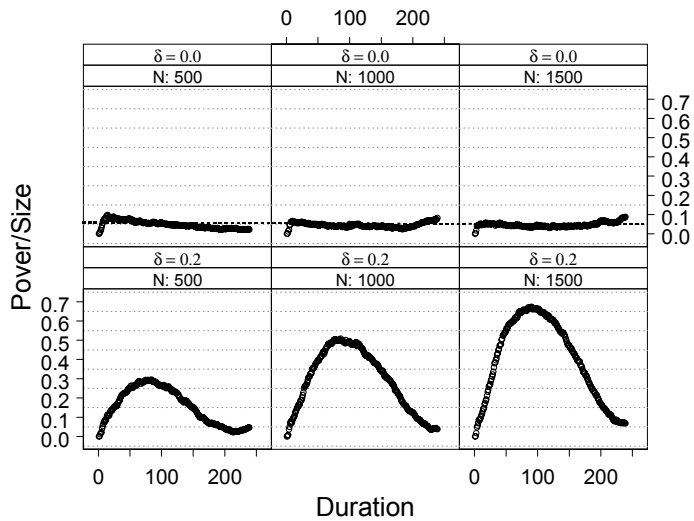


Figure 3: The power of the Wald test based on $\widehat{\Delta}_1(x)$ estimator when there is heterogeneity ($b_1 = 1$) and evaluation period $\bar{L} = 240$ months.

The variances are estimated as

$$\begin{aligned}\text{Var}(\widehat{\Delta}_C) &= \frac{\pi_1(1 - \pi_1)}{n} + \text{Var}_{NT} \\ \text{Var}(\widehat{\Delta}_C(x)) &= \frac{\pi_1(1 - \pi_1) + \pi_0(1 - \pi_0)}{n}\end{aligned}$$

where

$$\text{Var}_{NT} = \sum_{\bar{t}=l}^{\bar{L}} \frac{[(\bar{y}_C(D(\bar{t}) = 0)(1 - \bar{y}_C(D(\bar{t}) = 0)))]}{N(\bar{t})} \Pr(t_0^p = \bar{t})^2$$

and $\pi_0 = \frac{1}{n} \sum_{i=1}^n y_{c_i}$.

This estimator is contrasted with the estimator used in a series of papers with Lechner as one of the authors (see. Lechner, 1999, 2000, and Gerfin and Lechner, 2001) and Larsson (2000). The estimator in, e.g., Gerfin and Lechner (2001) is based on the approach sketched in section 3.1.¹⁶ 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, \bar{t}_r , from the distribution $F(t_0^p | D = 1)$, a randomly drawn individual in the comparison sample is retained if $t > \bar{t}_r$, otherwise (s)he is removed from the sample

Given a unique match¹⁷ 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 $(\bar{y}(1 - \bar{y}) + \bar{y}_c(1 - \bar{y}_c))/n$.

We also perform estimation without matching on the covariates. In this case we randomly find a matched individual satisfying $t \geq t_{0i}^p$ for a treated individual entering at t_{0i}^p . The treatment effect is then estimated as

$$\widehat{\nabla}_C = \frac{1}{n_1} \sum_{i=1}^{n_1} [y_i - y_c(t \geq t_{0i}^p)] \quad (22)$$

where n_1 is the number of individuals in treatment which are matched with non-treated individuals. The variance is calculated as $(\bar{y}_1(1 - \bar{y}_1) + \bar{y}_c^1(1 - \bar{y}_c^1))/n_1$, where $\bar{y}_1 = n_1^{-1} \sum_{i=1}^{n_1} y_i$ and $\bar{y}_c^1 = n_1^{-1} \sum_{i=1}^{n_1} y_c(t \geq t_{0i}^p)$

¹⁶In Lechner (1999) he specifies three estimators, partial, random and inflated. He states that the random estimator (described below) performs best.

¹⁷Lechner bases his 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.

Results The results from the Monte Carlo simulation for $\bar{L} = 48$ are shown in Tables 2 and 3, respectively.¹⁸ In Table 2, the results from the experiment with no treatment effect is given while Table 3 gives the result for the $\delta = 0.2$ treatment.

We start by commenting on Table 2 where we present the bias, variance (Var.) and the size (nominal level 5%) of the Wald test of a treatment effect. Columns 1 to 3 display the results when there is no heterogeneity. The $\hat{\Delta}_C$ estimator in (19) is the best estimator (considering all criteria) although (as expected) the $\hat{\Delta}_C(x)$ estimator in (20) becomes almost as good when the sample size increases. The sizes of the tests conform with the nominal level.

The estimators \hat{V}_C and $\hat{V}_C(x)$ (defined in equations 22 and 21) perform rather poorly. Both estimators are always significantly negative, suggesting a 4 percent decrease in the employment probability 12 months after entering the program. The sizes of the Wald tests are increasing with N and always too large. It is noteworthy that the bias is larger for the $\hat{V}_C(x)$ than for \hat{V}_C estimator. This has to do with the adjustment for the inflow into treatment. With the $\hat{V}_C(x)$ estimator we only adjust the left tail of the inflow distribution. Hence, if we have many long unemployment spells for the treated sample, then these individuals remain in the matched sample. For the \hat{V}_C estimator, however, a treated individual i is removed if there is no comparable individual in the comparison sample, i.e. if there is no individual in the comparison sample such that $t_c > t_{i0}^p$.

Columns 4 to 6 present the results with heterogeneity. Here the $\hat{\Delta}_C(x)$ is the only estimator with satisfying properties. The $\hat{V}_C(x)$ estimator suggests that employment is reduced by three percent as a result of treatment.

We now turn to the experiment with a negative treatment effect displayed in Table 3. Here we present the estimate (Est.) variance (Var.) and the power of the Wald test. In addition we present estimates (based on 1000 replications) of the average treatment effect (ATE) and treatment on the treated (SATE). It appears that the \hat{V}_C and $\hat{V}_C(x)$ estimates does comparatively well in terms of estimating treatment on the treated. However, we would argue that this is mostly a fluke. If we would consider the case with no heterogeneity and an evaluation period of $\bar{L} = 240$, then SATE equals -15.85 . The estimates $\hat{\Delta}_C$ then is -13.5 while \hat{V}_C equals -22 . Moreover, if we would consider the case of a positive average treatment effect the power of \hat{V}_C and $\hat{V}_C(x)$ would be substantially lower.¹⁹

¹⁸We focus on $\bar{L} = 48$ since this is closest to the evaluation period used in most empirical applications.

¹⁹Remaining results for the case $\bar{L} = 240$ can briefly be summarized as follows. When there is no treatment effect, the bias of the \hat{V}_C and $\hat{V}_C(x)$ estimators increases. As we

Table 2: Bias, variance (Var.) and size (nominal level, 5 percent) from the Monte Carlo simulation without a treatment effect ($\delta = 0$). Maximum observation length $\bar{L} = 48$. Percent.

Heterogeneity	no ($b_1 = 0$)			yes ($b_1 = 1$)		
	bias	Var.	size	bias	Var.	size
$N = 500$						
$\hat{\Delta}_C(x)$	-0.06	0.56	4.6	-0.21	0.36	3.7
$\hat{\Delta}_C$	-0.99	0.29	4.8	1.68	0.19	4.8
$\hat{\nabla}_C(x)$	-4.25	0.60	8.8	-3.39	0.41	9.3
$\hat{\nabla}_C$	-4.06	0.50	7.2	-1.22	0.39	6.1
$N = 1000$						
$\hat{\Delta}_C(x)$	-0.17	0.28	4.7	0.28	0.20	5.5
$\hat{\Delta}_C$	-0.17	0.16	5.6	1.95	0.10	7.8
$\hat{\nabla}_C(x)$	-4.05	0.31	12.2	-2.97	0.20	10.0
$\hat{\nabla}_C$	-3.78	0.31	10.9	-0.65	0.19	5.0
$N = 1500$						
$\hat{\Delta}_C(x)$	0.31	0.19	5.1	0.11	0.12	4.6
$\hat{\Delta}_C$	0.03	0.10	5.1	1.78	0.06	8.9
$\hat{\nabla}_C(x)$	-4.04	0.22	16.0	-3.02	0.13	10.9
$\hat{\nabla}_C$	-3.58	0.19	13.7	-0.83	0.13	5.0

Table 3: Estimate (Est.), variance (Var.) and power (nominal level, 5 percent) from the Monte Carlo simulation with a negative treatment effect ($\delta = 0.2$). Maximum observation period $\bar{L} = 48$. Percent.

Heterogeneity	no ($b_1 = 0$)			yes ($b_1 = 1$)		
	Est.	Var.	power	Est.	Var.	power
$N = 500$						
ATE and SATE	-3.39 and -15.97			-4.48 and -13.35		
$\hat{\Delta}_C(x)$	-9.96	0.55	29.6	-8.00	0.35	27.7
$\hat{\Delta}_C$	-10.03	0.25	51.8	-6.15	0.17	38.4
$\hat{\nabla}_C(x)$	-16.80	0.49	62.2	-12.81	0.39	56.9
$\hat{\nabla}_C$	-16.13	0.52	60.6	-10.62	0.35	44.3
$N = 1000$						
ATE and SATE	-3.34 and -15.85			-4.48 and -13.31		
$\hat{\Delta}_C(x)$	-9.71	0.24	45.5	-7.69	0.17	45.6
$\hat{\Delta}_C$	-10.04	0.11	80.7	-6.11	0.08	59.0
$\hat{\nabla}_C(x)$	-16.54	0.26	88.9	-12.66	0.19	84.0
$\hat{\nabla}_C$	-15.90	0.24	87.2	-10.72	0.17	71.9
$N = 1500$						
ATE and SATE	-3.36 and -15.91			-4.48 and -13.29		
$\hat{\Delta}_C(x)$	-9.81	0.17	66.5	-7.91	0.11	64.0
$\hat{\Delta}_C$	-10.09	0.09	92.5	-6.18	0.05	74.4
$\hat{\nabla}_C(x)$	-16.65	0.18	96.9	-12.82	0.12	96.1
$\hat{\nabla}_C$	-16.20	0.18	96.5	-10.76	0.11	88.6

5.3 Summary

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

- The proportional hazards model provides unbiased estimates of the constant treatment effect only when there is no treatment effect. Standard Wald tests do not give correct inference when testing for a treatment effect. Instead other testing procedures should be used (see DiRenzo and Lagakos, 2001).
- 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 severe negative bias in the matching approach applied by, e.g., Gerfin and Lechner (2001) 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.

6 Discussion

In this paper we have considered the evaluation problem using observational data when program start is the outcome of a stochastic process. The evaluation problem is a difficult one but it is commonly encountered in practice.

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. In the simulation study, however, the sizes of the Wald tests were too low. Thus, other testing procedures should be used instead; see DiRenzo and Lagakos (2001). Another approach is to test for a treatment effect using the non-parametric survival matching estimator proposed in this paper.

increase the evaluation period the bias of $\hat{\Delta}_C$ and $\hat{\Delta}_C(x)$ are reduced.

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. 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.
- DiRenzo, 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 (2001), A Microeconomic Evaluation of the Active Labour Market Policy in Switzerland, manuscript, SIAW, University of St. Gallen. (Forthcoming *Economic Journal*.)
- 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, manuscript, SIAW, University of St. Gallen. (Forthcoming *Review of Economics and Statistics*.)
- 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

Appendix: Proof of proposition 3

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

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

and

$$\begin{aligned} E((1 - D_{(j)}) R_{(j)}(1)) &= E(R_{(j)}(1) | D_{(j)} = 0) \Pr(D_{(j)} = 0) \\ &= E(R_{(j)}(1)) \Pr(D = 0) \end{aligned} \quad (24)$$

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

$$\begin{aligned} L(\delta) &= \prod_{j=1}^k \left(\frac{\exp(\delta D_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\delta D_l)} \right) \\ &= \prod_{j=1}^k \left(\frac{\exp(\delta D_{(j)})}{R_{(j)}(1) \exp(\delta) + R_{(j)}(0) + R_{(j)}(0|1)} \right) \end{aligned} \quad (25)$$

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 at the probability to enter treatment for a randomly chosen individual at $t_0 = 0$. 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 and hence $p \lim \hat{\delta}_{PH} = 0$. 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 $p \lim \hat{\delta}_E = \delta$. This estimator is only based on the rank orders of the treated relative to the rank orders for those not treated.²⁰ 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 $p \lim |\hat{\delta}_{PH}| < |\delta|$.

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

$$U(0) \perp\!\!\!\perp D(\bar{t}) | \mathbf{x}, (t \geq \bar{t}) \quad (26)$$

²⁰Note 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 DiRenzo and Lagakos, 2001). Wald tests of a treatment effect are biased, however.

This assumption guarantees that

$$\begin{aligned} E_{(T_0^p|D=1)} [U(0)|D(\bar{t}) = 1] &= E_{(T_0^p|D=1)} E_{\mathbf{X}}[E(U(0)|D(\bar{t}) = 0, \mathbf{x})] \\ &= E_{(T_0^p|D=1)} E_{\mathbf{X}}[E(U(0)|D(\bar{t}) = 1, \mathbf{x})], \end{aligned}$$

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 $E_{(T_0^p|D=1)} [U(0)|D(\bar{t}) = 1]$.

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} .²¹ By (26) it then holds that (see Rosenbaum and Rubin, 1983)

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

So, under the CIA (26), the counterfactual can be estimated as

$$\begin{aligned} E_{(T_0^p|D=1)} [U(0)|D(\bar{t}) = 1] &= E_{(T_0^p|D=1)} E_e[E(U(0)|D(\bar{t}) = 0, e(\mathbf{x}))] \\ &= E_{(T_0^p|D=1)} E_e[E(U(0)|D(\bar{t}) = 1, e(\mathbf{x}))], \end{aligned}$$

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, e.g., from 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:

$$c_{i\bar{t}} = \arg \min_{c \in N(\bar{t})} |\hat{\omega}(i) - \hat{\omega}(c)|, \quad (27)$$

where $\hat{\omega}(c)$ 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.

²¹This means that for each \mathbf{x} satisfying the CIA there must be individuals in both states.

With a complete set of pairs of treated and non-treated individuals the estimators (13) and (16) are given by

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

$$\widehat{\Delta}_C(x) = \sum_{\bar{t}=l}^{\bar{L}} \left[\frac{1}{n_{cs}(\bar{t})} \sum_{i=1}^{n_{cs}(\bar{t})} [y_i - y_{c_{i\bar{t}}}] \right] \Pr(T_0^p = \bar{t}), \quad \bar{t} = l, \dots, \bar{L}$$

while the estimator (15) is given by

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

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 .