STATISTICS IN TRANSITION-new series, Summer 2013
Vol. 14, No. 2, pp. 249–272

THE EFFECT OF UNEMPLOYMENT BENEFITS ON LABOUR MARKET BEHAVIOUR IN LUXEMBOURG

Nicholas T. Longford[1], **Ioana C. Salagean**[2]

ABSTRACT

We apply the potential outcomes framework to estimate the effect of awarding unemployment benefits on gaining long-term employment after an unemployment spell and on the time it takes to achieve it. We conclude that such awards, regarded as a treatment, are associated with poorer labour force outcomes than no awards.

Key words: administrative register, labour market, propensity matching, unemployment spells.

1. Introduction

The effect of unemployment benefits is extensively studied in the labour-market and other economics literature (Hunt, 1995; Roed and Zhang, 2003; Lalive and Zweimüller, 2004; and Uusitalo and Verho, 2010). The research is stimulated by high expenditure on these benefits, especially at times of higher unemployment. Such expenditure is regarded by some as an investment; the desirable return on it is stable and long-term employment after a short spell of unemployment, possibly with some training and short-term subsidised employment. The key issue is whether and to what extent the benefits have such an effect, and how the rules for awarding benefits should be adjusted to optimally combine the roles of welfare support while encouraging the resumption of employment and being frugal with public funds (Lalive, van Ours and Zweimüller, 2006). Jobs that are created as part of active labour market policies, intended as short-term stop-gaps, or are subsidised by other means, are not regarded as successful outcomes in this context, although they may be effective conduits for such outcomes.

[1] SNTL and Department d'Economia i Empresa, Universitat Pompeu Fabra, c/ Ramon Trias Fargas 25 – 27, 08005 Barcelona, Spain. E-mail: sntlnick@sntl.co.uk.
[2] CEPS/INSTEAD, Avenue de la Fonte 3, L–4364, Esch-Belval, Luxembourg. E-mail: ioana.salagean@ceps.lu.

The methodological challenge related to estimation of the effect of benefits is the impossibility of directly observing the labour market behaviour of two persons identical in all relevant aspects, including their background, who have just become unemployed – one who receives benefits and the other one who does not. This is compounded by the rule-based nature of awarding benefits. In principle, two persons with identical background at the point of becoming unemployed would be treated identically, and either both or neither of them would receive benefits. Nevertheless, we have a clear conception of what is meant by the effect of the benefits, as the difference in the individual's future position in the labour market (employment status and security, income, and the like) under the two conditions: receiving benefits and not receiving them.

The problem is addressed by models that account for the systematic differences between those who receive benefits (the 'treated', a term motivated by the medical statistics literature) and those who do not. An alternative approach is based on matching (Rosenbaum, 2002) – selecting subsets of recipients and non-recipients that for all purposes appear as if they were assigned to these two treatment groups completely at random. Adjustment by regression uses all the data, exploits a powerful modelling framework aided by methods of model selection, but is associated with numerous caveats related to model assumptions. One of them is the assumption of constant (universal) difference in the outcomes for the treated and untreated units, after a suitable adjustment for covariates and allowance for measurement error. In our case, there is no measurement error but the assumption of treatment homogeneity is not realistic.

Influential observations are another concern. They are extreme or outlying observations in the space of the covariates and their removal may alter the model fit much more than the removal of an observation closer to the centre of gravity of the observations would. However, such observations are often least relevant to the comparison of the treated and not treated units, because these two groups usually differ in the tails of their distributions of the covariates (Crump *et al.*, 2009). In estimation with large datasets, our focus should be on reducing the bias of estimators because their variances are small. Model selection procedures balance the effort to reduce the bias (retain more complex models) and variance (reduce model complexity). In contrast, methods based on matching emphasise bias reduction (Rubin, 2006). For discussion of related issues, see Lechner (2002); Hirano, Imbens and Ridder (2003); and Abadie and Imbens (2006).

Another reason for preferring the analysis by matched pairs is that some of the outcome variables we define are not on scales that can be associated with any common distributions. In some cases, the values of the outcome variable permit only partial ordering. That is, the difference of the outcomes cannot be quantified for some pairs of units, or we do not want to commit ourselves to any particular scale on which such differences could be quantified. We prefer to form matched pairs and tabulate the comparisons of their outcomes. There are three possible elementary results for a pair:

- the treated unit has a superior outcome;

- the treated unit has an inferior outcome;
- there is a tie – the units have outcomes that we regard as identical, or we cannot resolve which outcome is superior.

We say that the treated unit is a winner (and the untreated unit in the pair a loser) in the first case, and a loser in the second. The *balance* in a set of matched pairs is defined as the difference of the number of winners and losers in the treated group. We use the balance as an estimator of the average effect of the treatment. We then decide whether the realised balance could differ from zero purely by chance or it reflects the sign of the average treatment effect among the (treated) units that receive unemployment benefit.

The number of ties in the matched pairs also has a role in the analysis. If there are many ties, a more refined definition of the outcome variable may resolve some (or even most) of these ties, resulting in a substantially altered balance of the winners and losers. Desirable are results in which the number of ties is much smaller than the balance, so that the conclusion would not be altered if all the ties were redefined as winners (or losers). In this way, the ties become the element of a sensitivity analysis. Of course, the ties cannot be resolved in some settings when the outcomes are genuinely identical or not comparable.

For a large population, we might associate these three counts with a trinomial distribution. How- ever, our target is the assessment of the treatment effect for a particular set of units – the units that have been treated. We seek to establish what would happen if those who received benefits were denied them. Simplistic views include the suggestion that receiving benefits reduces the urgency of job search, but increases its quality – the recipient is more selective (discriminating) among the options for new employment and more patient in committing him- or herself to a new job. We test this hypothesis by matched-pairs analysis.

The dataset in our analysis is extracted from the databases of the Luxembourg Unemployment Agency (*ADEM – Agence pour le développement de l'emploi*) and the Register of National Insurance Contributions of Luxembourg (IGSS – *Inspection générale de la sécurité sociale*). The extracts are monthly lists of unemployed registered at ADEM, together with some background information, and labour force states at the end of each month inferred from the IGSS database. The lists cover the period from January 2007 to July 2011 (55 months). The records in these lists are linked by the (unique) national insurance number. The lists are reformatted to a dataset in which a record contains the sequence of monthly labour force states, supplemented by background information about the individual. A record is associated with an unemployment spell, and so an individual may have several records in the dataset. Such records are not exact duplicates, because some information, such as level of education and marital status are temporal, especially for younger people. Also, the age of the person at the time of becoming unemployed is an important variable.

The units of the analysis are unemployment (U) spells qualified by the period of the study (January 2008 – December 2010, 36 months) and the age of the

person at the beginning of the U spell (up to 30 years). The information from 2007 is reserved for defining covariates and the information from 2011 (part) for defining outcome variables. The U spell that defines the unit is called *reference*. Apart from the treatment, the reference spell is also associated with a person, the month when it starts and its length (in months) or the month when it ends. The treatment (award of unemployment benefits) is applied to the reference spell. Other U spells of the same person may be treated differently.

The next section introduces a terminology for sequences and gives more details of the ADEM and IGSS databases. The following section discusses the method of analysis based on the potential outcomes framework (Rubin, 1974; Holland, 1986). Section 4 gives further details specific to the matched-pairs analysis. The application to the resolutions of unemployment spells of young members of the labour force in 2008 – 2010 is discussed in Section 5.

2 Discrete sequences

A record in the original dataset is the sequence of $K = 55$ labour force states of a person. An element of the sequence corresponds to a calendar month, and the sequence to a contiguous set of calendar months, from January 2007 until July 2011. The status has five possible values: employment (E), unemployment (U), economic inactivity (I), transition (T) and absence (A). For example, a person who has been continually employed over a period of K months has the sequence EE ... EE of length K; a person who completed his/her education and found a job three months later, has sequence AA ... AUUUEE ... . Status A in month m results from having a record in neither ADEM nor IGSS at any time prior to and including month m. The complete definitions of the states are given in Appendix.

A sub-sequence is defined as a sequence for a contiguous subset of months. A sequence comprises *spells*. A spell is defined as a sub-sequence that is composed of the same status throughout, and the states at the immediately preceding and succeeding time points are different. A spell is characterised by the status and its length. For example, the sequence of the first person, continually employed, comprises a single spell of E. This spell has length greater than or equal to K, because the person may have been employed in the period immediately preceding the beginning of our records and following the end of the records. The sequence of the second person comprises three spells: A, U and E. The second spell is U and has length 3. Suppose the other two spells of the person have lengths 17 (A) and 35 (E). Then the sequence is completely described as (17A, 3U, 35E). A person may have a more complex description of the sequence, with many spells and several spells in the same state, such as (5A, 4U, 1T, 6E, 2U, 3E, 3I, 1T, 4E, 22E, 4I). The order of the spells is essential, and we do not summarise the sequence by the totals of time points (months) spent in a state, such as 31 in E, 6 in U, and so on. In fact, the longest spell of E, or of U, is of

interest in some analyses. The first and last spells in a sequence are often *censored*. That is, the first spell may have started before the first time point (month) of the sequence. Similarly, the last spell is incomplete if it continues into the future beyond the data horizon (month $K = 55$ in our data). In some instances, the first spell is complete. For example, in the next section we study sequences that start with (the first month of) a U spell.

The *history* of a spell (of a sequence) is defined as a sub-sequence that ends in the month immediately before the beginning of the spell. The history is qualified by its length. For example, suppose a spell of a sequence starts in January 2008. Then the sub-sequence for the 12 months of 2007 is its history of length 12. The future of a spell is defined similarly, although we include the defining spell in the sub-sequence. For example, the future of length 19 of a spell that lasted from January 2010 until February 2011 is the sub-sequence from January 2010 until July 2011.

We distinguish between sets of sequences that have the same delimiting (starting and ending) points (months), such as January 2007 and July 2011, and sets in which the start (or the end) is triggered by an event, such as the beginning of a U spell. In the latter case, some selection may take place, as there may be units (persons) whose sub-sequences do not qualify to the set. In fact, the selection takes place even in our original dataset, because persons with no U spell before their 31st birthday are not included. Further, we draw a distinction between sets of persons (individuals) and sequences. In the former, each person from a given domain is included at most once. In the latter, a person may be included several times. For instance, we consider in the Sections 4 and 5 sets of U spells. A person with several qualifying U spells is included once for each spell. Such spells of a person have different starting months, but differ also in length, as well as in their histories and futures.

2.1. ADEM and IGSS databases

The sequences of labour force states of length 55, covering the period from January 2007 until July 2011, are defined for all the persons who had a U spell in this period (had a case file open in ADEM) and their age was in the range 15–30 years at the beginning of one of their U spells in the period. We focus on the young because many of them who have had a U spell tend to have complex sequences, whereas among older members of the labour force U spells tend to be longer and are often followed by long spells of E or I. Also, the young are substantially over-represented in the U spells; about 43% of all ADEM records are for persons aged 30 or below at the beginning of the U spell.

We combine the information extracted from ADEM records, comprising the beginning and end of each U spell with information from IGSS, from which we infer the states in the other months (E, I, T, A); see Appendix for details. The

sequences are supplemented by background information listed in Table 1. The choice and purpose of the transformed variables and interactions, listed at the bottom part of the table is explained in Section 5. There are 43 825 reference spells (units of analysis) involving 26 835 persons (12 541 women and 14 294 men, with 19 328 and 24 497 spells, respectively). A person has a sequence in the dataset for every U spell in the studied period, subject to the condition of being aged 15 – 30 years at the beginning of the U spell.

The 55-month sequences contain a total of 2.410 million time points; 1.036 million of them are in state E, 351 000 in U, 452 000 in I, 82 000 in T and 489 000 in A. When reduced to unique persons, after discarding duplicate records, the number of time points is $26\,835 \times 55 = 1.476$ million, with 649 000 E's, 169 000 U's, 256 000 I's, 38 000 T's and 365 000 A's.

We discard all the records with reference (U) spells in 2007 and 2011 because the former do not have a history of one year, which we want to use for defining background variables, and the recorded future (up to July 2011) of the latter is too short. Thus, our analysis is reduced to 24 040 reference spells, of 10 346 women and 13 694 men, with 7874 unique women and 9 507 men. Men have higher average number of U spells (1.44) than women (1.31). Of the selected reference spells, 9 094 are associated with benefits (3982 for women and 5112 for men) and 14 946 are not (6364 + 8582).

The top panel of Figure 1 shows the distribution of the reference spells over time (starting months), separately for spells associated with benefits and those without. Every autumn (September and October) there is a peak in the number of failed applications (with no benefits awarded), and a smaller peak in January. The numbers of successful applications are greater in autumn and early winter, with a sharp fall in February. The following three panels display the distribution of ages of the applicants. They show that failed applications dominate among the young, and men in particular, but from about 24 years of age on the success rate it is close to 50%. It increases slowly with age.

There are 4820 persons with multiple reference spells; the highest multiplicity is seven, in three instances, followed by six, in eleven instances. The multiple reference spells are either treated (successful applications) on all occasions, not treated (failures) on any occasions, or have one or several untreated spells followed by one or several treated spells. There are 3424 persons with two reference spells; 1670 of these pairs are both untreated, 1112 are both treated, 642 have an untreated spell followed by a treated spell. Of the 1048 persons with three reference spells, 530 have three untreated spells each, 299 three treated spells each, 93 have a treated spell only in the third U spell and 126 with a sole untreated spell in the first U spell in the designated period. We note that the persons' histories may include U spells that have not been recorded (before 2007).

Table 1. Covariates used in the analysis.

Variable	Notation	Values	Summary
Age (years)	*age*	15 – 30	Mean 23.7; median 24.0, st. dev. 3.73.
Sex	*sex*	0, 1	0: women (43.3%), 1: men (56.7%).
Marital status	*civstat*	1, 2	1: single (82.8%), 2: other (17.2%).
Nationality	*natio_cat*	1, 3, 7	1: Luxemburgois (44.0%), 3: Portuguese (28.1%), 7: other (27.8%).
Level of education	*edu*	1, 2, 3	1: basic (39%), secondary (46.2%) tertiary (14.8%).
Spell No.	*rk*	1, 2, 3, 4	U spell since Jan. 2007; 1: first (54.1%), 2: second (27.1%), 3: third (11.9%), 4: fourth or later (8.9).
Months since completing education	*dfinetu*	0 – 51.5	Mean 44.3; median 45.5; st. dev. 5.8.
Months since the first record in IGSS	*prem_affil*	22 – 51.6	Mean 43.2; median 44.6; st. dev. 6.8.
Unemployment office (Region)	*ccel*	1, 2, 3, 4	1: Luxembourg City (42.5%); 2: Esch-sur-Alzette (36.5%); 3: Diekirch (14.3%); 4: Wiltz (6.8%)
Employment sector	*cemprec*	1, 2, 4, 5, 8, 9, 10, 11, 12	1: arts and technical; 2: management; 4: sales; 5: agriculture and forestry; 8: crafts and manual; 9: food, chemical, machinery; 10: hotels and restaurants; 11: other services; 12: no profession
History	*pstat*	E, U, I, T, A	12 categorical variables.
Future	*astat*	E, U, I, T, A	25 categorical variables.
Start	*start*	13 – 48	Mean 29.5, median 30.0; st. dev. 9.7.
Duration of the U spell	*dur.UN*	0 – 42	Mean 3.12; median 2.00; st. dev. 4.09.
Transformed variables			
Square of dfinetu	*Sq.dfinetu*		dfinetu^2
Square of prem_affil	*Sq.prem_affil*		prem_affil^2
Square of age	*Sq.age*		age^2
Bi-months in status E in 6 months of history	*Pstat.SumE*	0, 1, 2, 3	Categorical variable (0 – 49.9%, 1 – 10.5%, 2 – 10.7% ,3 – 28.9%)
Interactions			
Age × pstat 1 = E	*Age.Pstat1E*		Age by status E in month −1
Age × sex	*Age.Sex*		Age by level of education 2
Age × education (2)	*Age.Edu2*		
Nationality × pstat 1 = I	*Nat.pstat1I*		Nationality by status I in month −1
Region × sex	*Ccel.Sex*		
Education × sex	*Edu.Sex*		

Note: All the summaries are for the reference spells.

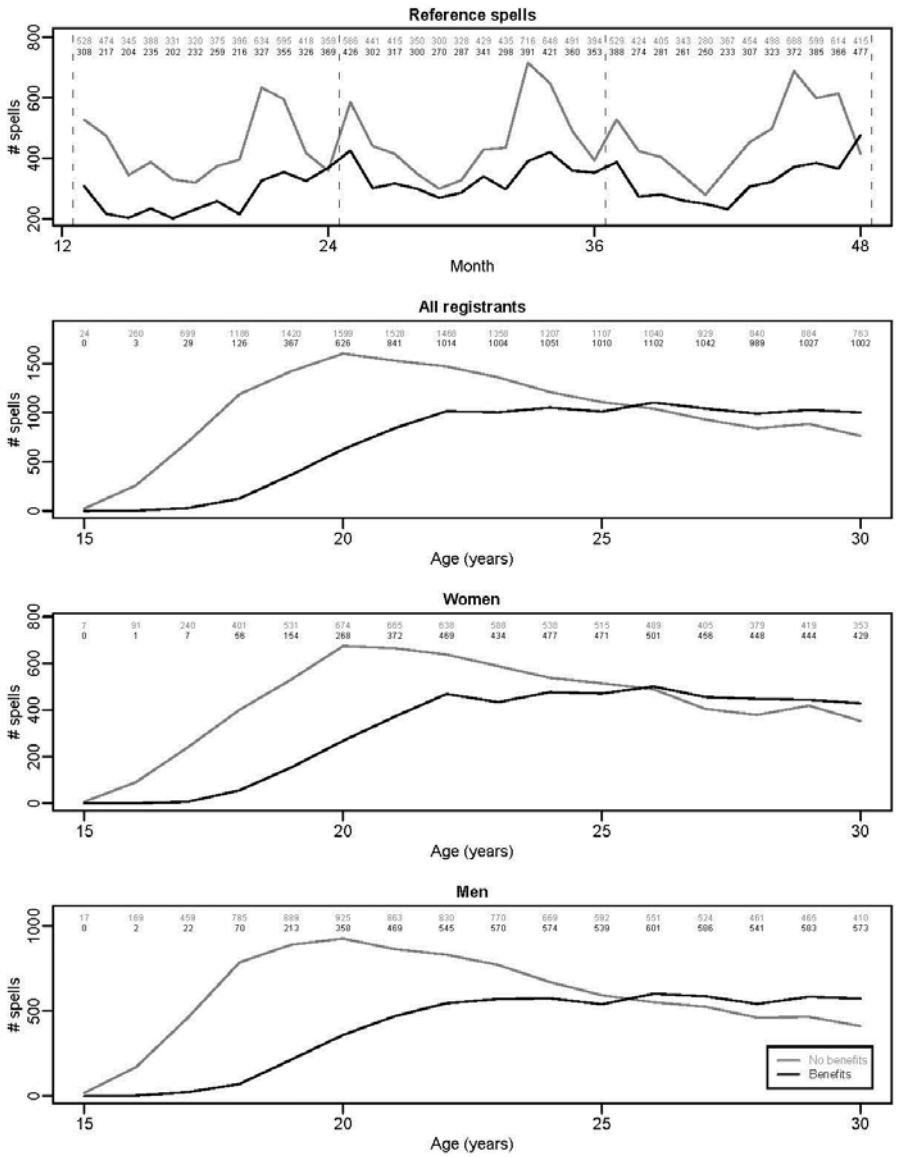

Figure 1. Receipt of unemployment benefits for reference spells, by age and sex.

3. Methods

Our analyses concern the complete enumeration of ADEM registrants in the period from January 2008 to December 2010 (36 months) who were 15–30 years of age at the beginning of a U spell. We have data for an additional year, 2007, but this we reserve for the definition of background variables. The unit of analysis is a U spell, supplemented by its history and future. Thus, the unit is itself a sequence, but a person may be represented in the analysis by several sequences, one for each U spell. Each reference spell is associated with receipt or not of unemployment benefits. We consider the future of the reference spell as the outcome. A person may have U spells (case files in ADEM) additional to the reference, both in its history and future. Two sequences of a person with multiple U spells are linked by histories and futures that are shifted by the distance in time between the two reference spells.

We want to establish how different the futures of the treated reference spells (with receipt of benefits) would be if they were not treated (with no benefits received). Such research questions are regarded as *counterfactual* because the difference considered cannot be observed for any unit. We have no control over the assignment of the benefit status. Hypothetically, if the assignment could have been made at random, applying experimental design, the analysis would be relatively straightforward. Even in that setting we could not establish the difference of the futures of a reference spell under the two conditions – treatment by the benefit, and no treatment – because only one of the conditions could be realised. However, the random assignment, together with a set of assumptions discussed in Section 3.1, would ensure that the average treatment effect is estimated without bias.

In summary, our focal population is the set of all U spells in the designated period that result in a uniquely identified case file (record) and registration in ADEM. The unit of analysis is defined by its reference spell. It has to be distinguished from the other U spells in the history and future of the person's sequence. For the history, we consider the immediately preceding 12 months, and for the future up to 25 months immediately following the start of the reference spell. The futures are censored by the data horizon of July 2011 for all reference spells that started after June 2009.

We distinguish three categories of variables:

- treatment indicators
- outcomes
- covariates

A treatment indicator T is a binary variable with values zero or unity; $T = 1$ for one treatment and $T = 0$ for the other. Every unit is assumed to have been administered one of the two treatments. The only pair of treatments we consider is the receipt of benefits and not receiving them. We elect $T = 1$ for the receipt and $T = 0$ for the complement. A person may have several U spells in the

designated period, between January 2008 and the earlier of July 2011 and the 31st birthday, and each of them is the reference spell of a sequence (the unit of analysis).

We specify several outcome variables that summarise a particular aspect of the future of the reference spell. We regard the resolution (closure) of an ADEM case file that coincides with the reference spell as a success if the future contains an E spell of length at least L. We set $L = 12$, regarding as a landmark uninterrupted employment over a period of one year after a U spell. Insisting on a longer spell, such as $L = 18$, would lead to excessive censoring by the data horizon in July 2011 and too few successes would be recorded. Another variable is the length of the longest E spell in the sequence. Comparisons of the values of this variable for sequences with different starting points are problematic because of the fixed data horizon. However, comparisons for two sequences with the same starting point have face validity. If two sequences have the same maximum length of E spell, then we regard as superior the sequence in which the first E spell of maximum length was achieved earlier. Some other problems (ambiguities) with this definition are discussed in Section 5.

3.1. The potential outcomes framework

An outcome variable Y is associated with two potential outcome variables; $Y^{(t)}$ is the variable defined as the outcome assuming that treatment $t = 0$ or $t = 1$ was applied. Our goal is to compare the values of $Y^{(1)}$ and $Y^{(0)}$ on the set of units (the futures of the reference spells) that received treatment $t = 1$. The variable Y can be expressed as a composition

$$Y^{(T)} = (1 - T)Y^{(0)} + TY^{(1)},$$

and its distribution is a *mixture* of the distributions of $Y^{(0)}$ and $Y^{(1)}$. Inferences about $Y^{(1)} - Y^{(0)}$ are relatively easy to make when T is independent of $Y^{(0)}$ and $Y^{(1)}$, as it would be if the treatment were randomised. Potential versions, associated with the two treatments, can be defined for any variable. Of course, the two versions may be identical. Background variables, defined prior to administering the treatment, which could not have been informed by the value of T, are obvious examples.

We use as covariates an extensive set of variables listed in Table 1. The qualifying attribute for a covariate is that its two potential versions are identical − that its values would not be altered if the treatment assignment (the values of T) were changed. Variables defined prior to the beginning of the reference spell are covariates. We discount the possibility that the history of the reference spell is influenced in any way by the anticipated value of T in any future U spell. The covariates include variables defined on the 12 months of labour market history (labour force states). This is based on a pragmatic decision not to lose in the analysis too many spells for which this history is not completely recorded.

The potential outcomes framework has the assumption of stable unit-treatment variable, referred to by the acronym SUTVA. It can be summarised as follows: in hypothetical replications of the study, with the same set of units, the same sets of values of all variables except the treatment, but a replicate realisation of the treatment assignment process (mechanism), the outcome for unit i is determined entirely by the treatment assigned; $y_i^{(0)}$ if $t_i = 0$ and $y_i^{(1)}$ if $t_i = 1$.

Close scrutiny of this condition reveals that it is far from trivial, and in several aspects contentious in our setting. First, it implies that the units do not interfere with one another. That is, the outcome of one unit is unaffected by the treatment assigned to another unit. In our case, it entails the assumption that the future of one U spell is unrelated to the future of a later U spell of the same person. This is patently false, especially for two short U spells that are separated by a short time. However, the dependence is difficult to describe and relates to a small fraction of the units. Next, the rules for awarding benefits are well known to all parties involved, and so the conduct of the persons threatened by unemployment is affected by the anticipated (possibility of) loss of job. Unemployment often arises in quanta (several persons losing their jobs) and is anticipated. A lot of unemployment arises after the conclusion of fixed-term contracts that cover the same period of time and are awarded to a set of workers at the same time. Every person's behaviour is affected by the experiences of acquaintances, and family members in particular, so the two potential outcomes do not describe the entire range of possible values of the outcome of a person. Further, some unemployed and those about to become unemployed are advised by agents, such as union representatives. As a consequence, the conduct of some unemployed may be coordinated. A principled solution to this problem is to consider all configurations of plausible treatment assignments and define a potential outcome variable for every one of them. However, the number of such variables would proliferate and become unmanageable even in some simple scenarios, so pursuing this approach is not feasible.

3.2. The treatment effect

With two potential outcome variables, the unit-level treatment effect is defined as the difference of the outcomes under the alternative treatments:

$$\Delta Y = Y^{(1)} - Y^{(0)},$$

with value Δy_i for unit i. We impose no conditions on the distribution of this variable. In particular, we do not assume that ΔY is constant. The average treatment effect for a set of units u is defined as the mean of the (unit-level) treatment effects:

$$\Delta \bar{Y}_u = \frac{1}{n_u} \sum_{i \epsilon u} \Delta y_i,$$

where n_u is the number of units in u. The average treatment effect is qualified by the set u. In our analysis, u is the set of U spells for which $T = 1$ was applied.

Another factor relevant to our analysis is the treatment assignment process (mechanism). It is defined by the joint distribution of the treatments assigned to the units in u. In a typical observational study this distribution is not known, and inferences about it are difficult to make because we have only one realisation of this distribution – the realised assignment of the treatments to u. It is more practical to think about treatment assignments in alternative replications of the study. We cannot rule out the possibility that certain units would never receive a particular treatment. For them the unit-level treatment effect ΔY is not defined. It might be constructive to exclude such units from u because the underlying question (What is the difference...?) about them is meaningless.

3.3. The missing-data perspective

If the values of $Y^{(0)}$ and $Y^{(1)}$ were observed for all units in u, the analysis would be straightforward, evaluating the mean of the differences in (1). In the established terminology for missing data, the set of $n_u \times 2$ values of the potential outcomes $Y^{(0)}$ and $Y^{(1)}$ is referred to as the *complete dataset*. The observed dataset, comprising a subset of n_u values (determined by the realised treatment assignment), is called the *incomplete dataset*. The set-difference of the two datasets is the set of *missing values*.

An analyst's first instinct might be to impute a value for each missing item. Such an imputation (data completion) results in a *completed* dataset. The theory of multiple imputation (Rubin, 2002), documents the problematic nature of this approach. If we analyse a completed dataset as if it were the complete dataset, we obtain inferences that project too much confidence because in the analysis we pretend to have more information than was collected. The problem does not arise when the statistic of interest, such as an estimator, is a linear function of the missing values. However, we also want to assess the quality of the estimator by estimating its sampling variance (or standard error), and that is a distinctly nonlinear function of the missing values. Note also that we have to distinguish between the sampling variance of the estimator that relates to complete data and the sampling variance for the combination of the processes of data generation and nonresponse (missingness). The former sampling variance is zero because we study a fixed set of units and assume their potential outcomes to be fixed.

Multiple imputation is a flexible alternative for addressing this problem. We generate several sets of plausible values for the missing items according to a model specified for them. In the imputations, we reflect the uncertainty about the model parameters, as well as the uncertainty that would be present in a model even if all its parameters were known. In our setting, the treatment assignment is the sole source of uncertainty.

4. Matched pairs analysis

Our analysis proceeds by the following steps. First we estimate the propensity of receiving a treatment as a function of the background variables. Next we form pairs of units, one treated and one not treated, that have similar propensities. Then we estimate the average treatment effect by averaging the within- pair contrasts (comparisons). In the final step, we estimate the sampling variance of this estimator by replicating the processes of matching and estimation. The following sections give details of the steps.

4.1. Propensity scoring

In this step, we fit a logistic regression model in which the outcome variable is the treatment applied and the covariates are all the variables listed in Table 1. The propensity is defined as the conditional probability of treatment 1 given the values of the covariates:

$$p(\mathbf{x}) = \mathrm{P}(T = 1 \mid \mathbf{X} = \mathbf{x}).$$

The propensity for unit i, with the vector of covariates $\mathbf{x}_i$, is defined as $p(\mathbf{x}_i)$. The propensities are used for forming pairs of units (reference spells), one with $T = 0$ and the other with $T = 1$, by matching on their values of $p(\mathbf{x})$. That is equivalent to matching on a (strictly) monotone transformation of the propensities.

In a perfectly implemented experiment, the propensities are known. For example, if units are assigned to the two treatment groups completely at random with probabilities equal to 0.5, then $p(\mathbf{x}) = 0.5$ for all $\mathbf{x}$. Problems arise when the protocol of the experiment is not adhered to the letter, as when some units depart from the assigned treatment regimen or drop out from the study.

Our approach can be described as selecting from the observed units a subset that resembles as closely as can be arranged, in all features that can be checked, a dataset generated by a (perfectly executed) experimental design. The main criterion for this is that the two groups in the selected subset are balanced – have near-identical profiles of the covariates. When there are several covariates, such a balance is very difficult to arrange. Rosenbaum and Rubin (1983) showed that arranging this balance can be replaced by the task of matching on the values of the propensity or its monotone transformation. Using the logit (log-odds) scale for the scores is an obvious choice. The logit is defined as $\mathrm{logit}(p) = \log(p) - \log(1 - p)$.

The propensity is a variable defined on the interval [0, 1]. It usually attains many distinct values, most of them with small frequency or even uniquely, and therefore matching on its values exactly would yield too few pairs. In practice, the range [0, 1] is split into a number of intervals and matches are sought within them. Further, the propensity is not established with precision, but merely

estimated. Rubin and Thomas (1996) showed that the uncertainty about the propensity can be ignored, and matching based on the estimated propensity (or its monotone transformation) is sufficient.

Apart from the propensity score, we match also on the date when the U spell starts. This is referred to as *blocking* – matching separately within subsets of the units; in our case, the subsets are defined by the starting date (month). This is essential for some of the outcome variables which are affected by censoring and for which a comparison of two units with different starting dates is problematic.

4.2. Matching and complete-data analysis

We divide the fitted propensities into $H = 10$ intervals so that each interval contains approximately the same number of units. These intervals are further subdivided into groups according to the starting date of the reference spell. There are up to $H \times H'$ matching groups, where $H' = 36$ is the number of distinct starting dates (months). A combination of intervals and dates may contain no units or only units with one value of T. Within every one of the groups in which both treatments are represented, we form matched pairs by the following process. Suppose group h has n_{ht} units with treatment $t = 0, 1$. If $n_{h0} \leq n_{h1}$, then we match each untreated unit ($t = 0$) in the group with a treated unit ($t = 1$), selected at random and without replacement. This is implemented in practice by sorting the n_{h0} untreated units, selecting n_{h0} units sequentially from the set of n_{h1} treated units in the group, and forming pairs in the obvious manner.

Let y_{htj} be the value of the outcome variable for the unit which represents treatment t in pair j of group h. For each pair we evaluate the sign of the contrast $\Delta y_{hj} = y_{h1j} - y_{h0j}$, or the sign (positive, negative or tie) is established without evaluating Δy_{hj} when the difference is not well defined. The results are then tallied across the pairs, to obtain the counts of winners, losers and ties. These tables (triplets), denoted by Δy_h, are further tallied over the matching groups, to obtain a single triplet Δy of counts of pairs in which the representative of the treatment is the winner, loser, or there is a tie.

As an alternative the triplets can be tallied with weights that account for the treated units that have not been matched. If group h has fewer untreated than treated units, $r_h = n_{h1}/n_{h0} > 1$, then $n_{h1} - n_{h0}$ treated units are not matched. We compensate for their loss by increasing the contribution of group h from Δy_h to $r_h \Delta y_{0h}$, where $r_h = n_{h1}/n_{h0}$. For groups in which $n_{h0} \geq n_{h1}$, the contribution to Δy is not changed; r_h is set to unity. There may be one or several groups in which there are no untreated units, $n_{h0} = 0$, so none of the treated units are matched. In every matching exercise, we monitor the number of such units. If it is excessive, then we revise the definition of the matching groups. With fewer groups defined, fewer such failures to match are likely.

The number of treated units involved in such failures can be included in a sensitivity analysis. Consider the following two extremes for the set of all treated units without a match:

1. if they were matched they would all be winners;
2. if they were matched they would all be losers.

If in both of these scenarios we obtain the same inference, then the failures are immaterial. Of course, if there are many such failed matches, then we are threatened with an impasse, when different conclusions are arrived with the two extreme assumptions. The ties between the units in a pair can be dealt with similarly.

4.3. Sampling variation

Regarding the set of focal units (spells) u for which we want to estimate the average effect of the treatment as fixed, the entire process of generating the propensity scores, matching and evaluation entails a single source of uncertainty, namely, the forming of matched pairs. The logistic (or another) regression used for generating the propensity scores would be associated with variation *if* we considered the focal set of units and their assignment to treatment as a realisation of a random process. However, our analysis is conditioned on the treatment assignment and the units included in the analysis, and so each propensity score is without any sampling variation.

We estimate the sampling variation associated with matching by simply replicating it several times. When we do not want to refer to a scale for the outcomes we evaluate the variances of the counts of positive and negative within-pair contrasts, and make inferences about the expectation of the average contrasts, $E(\Delta y)$.

5. Application

We implemented the matched pairs analysis and associated diagnostics in a customised set of R functions. In the first step, we fit a logistic regression of the treatment indicator on all the covariates. The regression parameter estimates or the quality of the fit are of no interest because the sole purpose of the fit is to form matched pairs based on the estimated propensities. We have a lot of observations, 24 040, so we are concerned principally about bias resulting from poor matching. That is why we prefer to err on the side of specifying a richer propensity model; little efficiency is lost by including in the model a few redundant covariates. The key criterion for the appropriateness of the model is that the pairs matched on propensities are also balanced with respect to their distributions of the background variables.

The propensity model with the original variables, involving 71 parameters, 55 of them related to the history of the reference spell, turned out to be unsatisfactory, because the pairs formed were poorly matched on the dispersions of several continuous variables. We supplemented the model with the covariates listed at the bottom of Table 1, involving 16 further parameters. The transformations and interactions were identified principally by trial and error. Note that the addition of a covariate to the model does not always result in an improved balance of the other variables, and may even be detrimental to the balance of some of them.

There are 14 946 units without benefits (untreated) and 9 094 treated units. The distribution of the fitted propensities $\hat{p}$ within the treatment groups is displayed in Figure 2. The diagram shows that a lot of untreated units have very small propensities. Only a small fraction of them will be used in matches with the far fewer treated units that have similar values of the fitted propensity. Most treated units with $\hat{p}$ in the range 0.1–0.5 will be matched because there are sufficiently many untreated units within every narrow range of propensities (the width of a bar in the histogram). For treated units

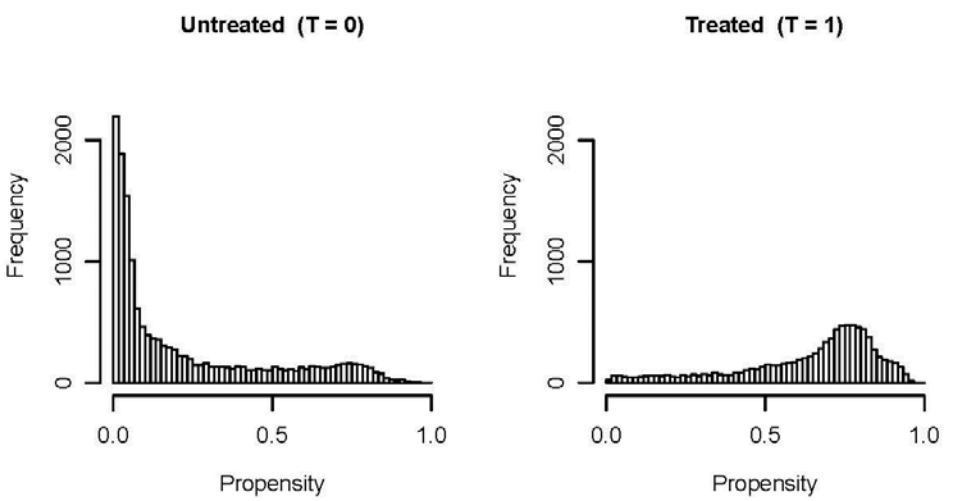

Figure 2. The fitted propensities for the two treatment groups. The vertical dashes mark the deciles of the within-group distributions.

with $\hat{p} > 0.5$, a substantial percentage of treated units will not be matched because they are in a majority.

We match on the fitted propensity and the start (month) of the reference spell. The propensities are split by their deciles into ten groups with approximately equal numbers of units in each. Each of these groups is then split into subgroups

according to the start of the U spell. There are 36 distinct months (13–48). Of the 360 combinations of propensity group and month, 40 groups contain no match (one or both treatment groups are not represented in it), 21 groups have one matched pair each, 19 groups have two pairs, and the largest numbers of pairs matched in a group are 47 and 48, in one instance each. In total, 4631 matched pairs have been formed; they contain 50.9% of the treated units.

The set of matched pairs is satisfactory if they are close to balance, as they would be in an experiment with random treatment assignment. The left-hand panel of Figure 3 displays the contrasts of the means of the continuous variables for the two treatment groups. A horizontal segment is drawn for each variable between the contrast for the unmatched (original) groups and its negative. The value of the contrast is marked by a vertical tick. The contrast for the matched pairs with the original model is marked by a gray disc and the contrast for the extended model by a smaller black disc. The

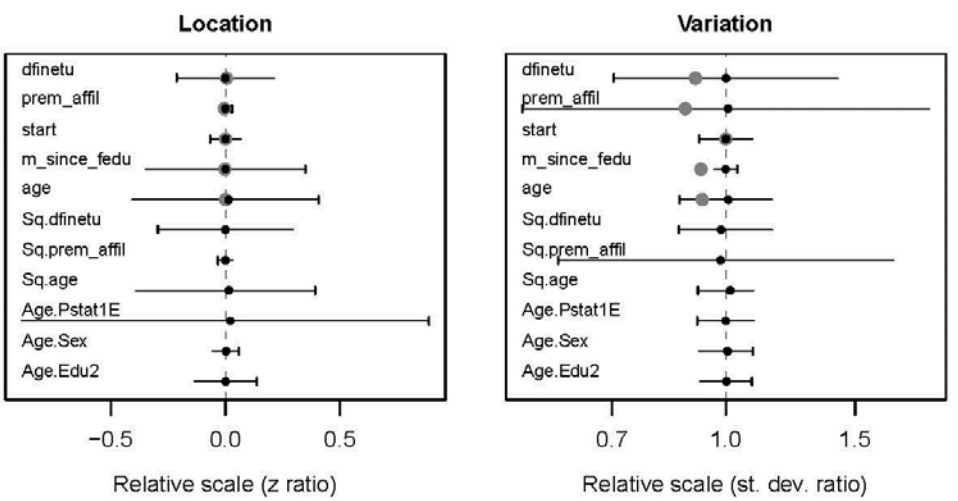

Figure 3. Check on the balance of the continuous covariates for the matched pairs.

gray discs are drawn only for the five variables in the original model. Owing to blocking, the balance for *start* is perfect by design. Note that we broke some rules of invariance, such as not including both variables defined by the interaction of age and the three categories of education. If we followed the rules, a slightly worse balance would be obtained.

The matched pairs should be balanced not only in the within-group means, but in the within-group distributions in general. The right-hand plot displays the ratios of the standard deviations of the continuous covariates for the unmatched and matched groups. The horizontal segments are drawn between s and $1/s$, where s is the ratio for the unmatched groups. The ratios for the matched groups are marked by gray and black discs for the original and extended propensity models,

respectively. The horizontal axis is on the log-scale. The ratios for the matched groups deviate from the ideal of 1.0 only slightly, and the extended model yields a better balance for all four original variables (except for *start*, for which the balance is perfect by design). The balance is nearly perfect for the six added variables.

Figure 4 presents the corresponding plot for the contrasts of the categorical variables. A variable with C categories is represented by $C - 1$ horizontal segments connecting the unmatched balances with their negatives. The black and gray discs mark the matched balances based on the original and extended propensity models. The matched balances are uniformly closer to zero for both the original

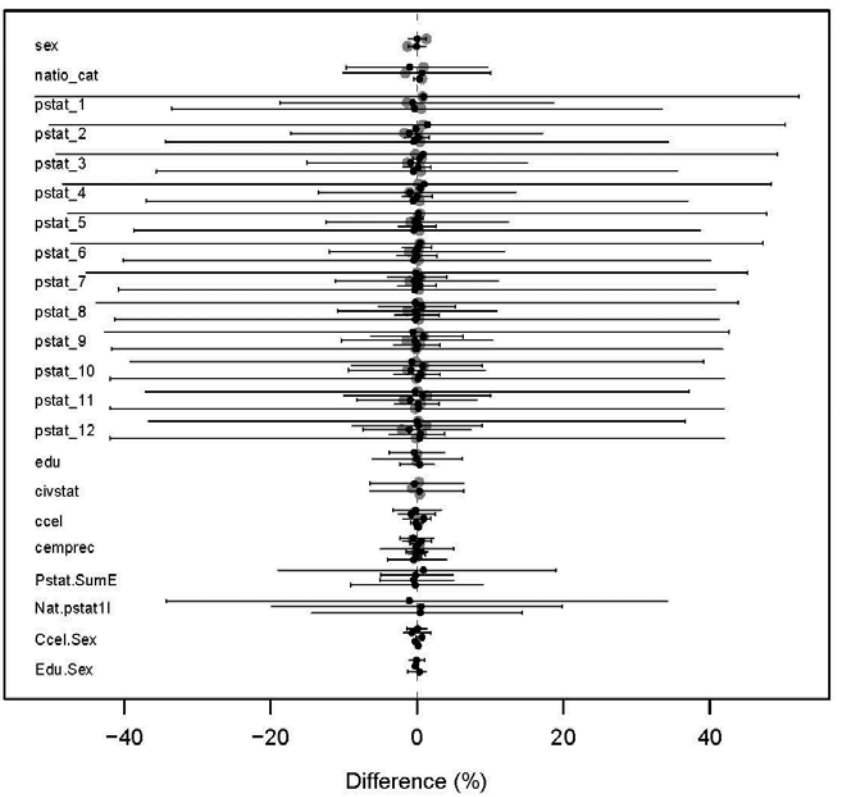

Figure 4. Check on the balance of the categorical covariates for the matched pairs.

model and for the model with four additional variables (15 additional parameters), displayed at the bottom of the diagram.

Having accepted the match as satisfactory, the remainder of the analysis entails comparing the numbers of winners and losers in the within-pair contests. We score the future of each unit as a success when it contains a 12-month (or longer) spell of E, and as a failure otherwise. In a matched pair, the treated

unit is a winner if it is a success and the untreated unit is a failure. The treated unit is a loser if it is a failure and the untreated unit is a success. The result is a tie if the outcome is the same for both units in the pair (successes or failures).

With these rules for scoring, the treated unit is a winner in 831 pairs and a loser in 1169 pairs, and 2631 pairs are tied. Thus, the estimate of the treatment effect is negative, with a balance of 338 units in favour of the untreated group. Some of the ties can be resolved if we compare the lengths of the longest spell within the pairs that are successes. Then the treated group has 1509 losers, 976 winners and there are 2146 ties. The balance, 533, is increased substantially. If we compare the lengths of the longest spells without the qualification of L = 12 months, the treated group has 2434 losers, 1698 winners (positive balance of 736 units) and there are only 499 ties; the treated group would have more losers (237) even if all the ties were conceded to the untreated group.

We can compensate for the failure to match some treated units by weighting the within-pair contrasts. If in matching group *h* there are more treated units than untreated, $n_{1h} > n_{0h}$, we apply weights $r_h = n_{1h}/n_{0h}$ to each within-pair contrast formed in this group. Otherwise, when all treated units are matched, the weight r_h is set to unity. When an E spell of at least 12 months in the future is regarded as a success, the treated group has score 1582.1 and the untreated group 2563.7, and the ties account for 4948.2 points. The total of these three scores is 9094. It coincides with the number of treated units in the (unmatched) sample, because there happen to be no matching groups in which all the units are treated. For such a group, with n_{0h} = 0, the weight r_h would not be defined. The balance, 2563.7 − 1582.1 = 981.6, is much greater than its counterpart without weights, 338. However, the sampling variation of the balance is increased because some within-pair contrasts have large weights. The largest weights are 13.5 for four pairs and 12.4 for nine pairs each, and further 190 pairs have weights greater than 5.0. In total, 2538 pairs (54.8%) have weights greater than unity; 1046 of them are smaller than 2.0. Weights greater than unity are for pairs from matching groups in which treated units have a majority. For the other two outcome variables that score the future of the reference spell, the treated group has far fewer winners than losers and the balance is much greater than in the analysis without weights. The results are summarized in Table 2. The rows are for the analyses in which only E spells of one year or longer count as successes (analysis A), only such spells count but two such spells are compared by their lengths (analysis B), and when the longest E spell of the future counts (even if shorter than one year), and these maxima are compared by their lengths and recency (analysis C).

By comparing the two treatment groups without matching, we obtain the following results. In analysis A, 17.7% of the treated units have an E spell of length 12 or longer in their futures, compared to 18.9% of the untreated. In analysis B, the mean scores of the treated and untreated groups are 3.40 and 3.59, respectively. In analysis C, the treated group has mean 4.54 (months) and the untreated

Table 2. Matched-pairs analysis of the employment futures

	No weights			Weights		
	Winners	Ties	Losers	Winners	Ties	Losers
A. E spell length ≥ 12 months	831	2631	1169	1582.1	4948.2	2563.7
B. E spell length ≥ 12 months +	976	2146	1509	1925.8	3848.4	3319.8
C. Longer E spell (earlier)	1698	499	2434	3275.0	839.9	4979.1

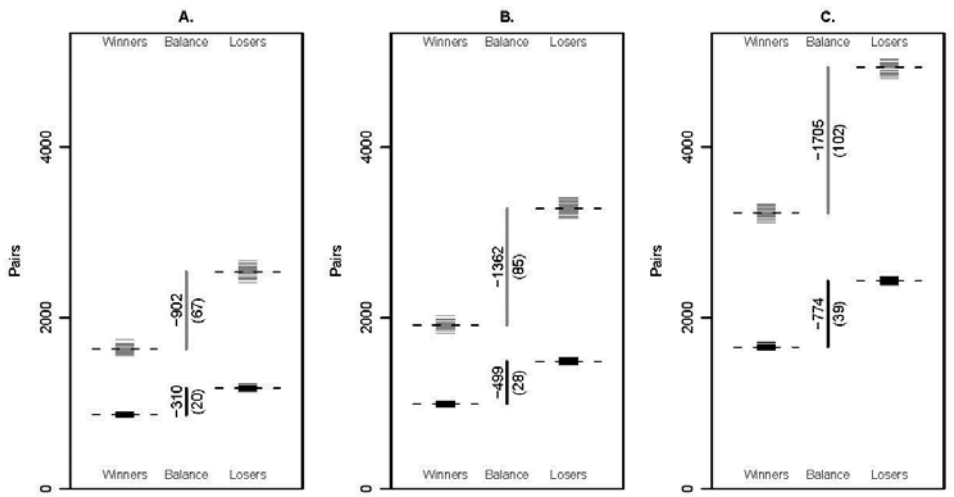

Figure 5. Summary of a set of 50 replicates of the analyses in Table 2.

4.94. Thus, the matched-pairs analysis is in agreement with the 'raw' comparison for all three outcome variables.

The sampling variance of the balances is estimated by replicating the matching process. A set of 50 replications is summarised in Figure 5. The panels A – C correspond to the rows of Table 2. The replicate counts of winners and losers are marked by thin horizontal segments, black for unweighted and gray for weighted analysis. Their means are marked by horizontal dashes and the balances are indicated by vertical segments with their values printed to the left and standard errors to the right in parentheses.

If the estimates of the balance and their standard errors are taken at face value, there is very strong evidence of negative balance for all three outcome variables – benefits are associated with less desirable outcomes of reference spells

(E spells, if any, that are shorter and achieved later). The checks on the balance of the covariates in Figures 3 and 4 are not a complete diagnostic because they cannot indicate that some important covariates have been omitted (not recorded). We can only argue that the list in Table 1 is quite exhaustive.

As another limitation, we highlight a problem in the definition of the outcomes in analysis C, and partly also in B. Suppose both units in a pair have outcome L, but one unit is the winner because his or her E spell of length L was realised earlier. Suppose further that the qualifying E spell of the losing unit ended at the end of the recorded future. It is therefore likely that the losing unit would have had a superior outcome (longer E spell of maximum length), had we had longer recorded future. If the losing unit had shorter maximum E spell than the winner, but this spell were at the end of the recorded future, there is still a likelihood, albeit smaller, that the losing unit would have had a superior outcome if the records of the future were longer.

We refer to such cases as ambiguities. Formally, a matched pair is said to have an ambiguity of order $M = 0, 1, \ldots$, if the difference of their outcomes is M, and the loser's qualifying E spell is at the end of the recorded future. In the 50 replications, the numbers of such ambiguities are in the range 119 – 157, with mean 137. The numbers of ambiguities of order 0 range from 44 to 74, of order 1 from 17 to 42, and they decline rapidly with the order. For example, they range from 3 to 14 for order 5. Also the likelihood that the loser would have a superior outcome if we had extended records (futures) diminishes with the order. The numbers of ambiguities are distributed evenly between treated and untreated units.

We can incorporate the ambiguities in a sensitivity analysis by reclassifying every treated loser in a pair with ambiguity as the winner, and the untreated winner as the loser. Such an analysis can be refined by allowing a realistic percentage of losers to remain losers, but this is not necessary in our case. Even if every ambiguity in the treated group is reclassified, the balance is reduced by approximately the number of ambiguities in the unweighted analysis C. This would not alter the conclusion that the treated units have fewer winners. For the weighted analysis, the ambiguities should be counted with weights; the same conclusion is arrived at.

Other analyses are in accord with the results related to the longest E spell. For example, the mean lengths of the reference spells are 5.17 months in the treated group and 1.88 in the untreated. In the matched-pairs analysis, this difference is reduced only slightly; the respective means are 4.89 and 1.71 months. The counts of winners and losers are in accord with this result; the treated group has 889 winners and 3112 losers, with 630 ties. Some caution is called for in interpreting this analysis, because administrative procedures may be delayed and, in case of short reference spells in particular, the award of an unemployment benefit may be linked with an incorrect U spell. However, we cannot condition the analysis on the length of the reference spell, because it is an outcome of the treatment.

6. Conclusion

We applied the potential outcomes framework to estimate the effect of awarding unemployment benefits on the resolution of the unemployment spell. Our conclusions are unequivocally negative about the benefit. Even after detailed matching on an exhaustive list of background variables, the outcomes are far superior on average for those not awarded benefits, both in terms of the length of the longest E spell in the recorded future and the speed of achieving it. Unemployment spells associated with benefits tend to be longer, even after matching on background.

The principal modelling effort is in the propensity analysis, which does not involve the outcome variable. Therefore, model selection does not introduce any bias. Concerns that one might have about the distribution of the outcome do not arise in our approach; the outcome is defined on a scale that best reflects our purpose, with no regard for the distribution of the values. We avoided defining a scale altogether and based the analysis on the within-pair comparisons. One might argue that some efficiency is lost in the process. However, the sample size in the analysis is so large that variance reduction is of secondary importance to combatting all possible sources of bias, among which selection bias (non-ignorable assignment to the treatment groups) presents the greatest threat. This can be interpreted as insisting on a higher standard for comparing like with like.

The agreement of the results of raw comparisons with matched-pairs analysis is no licence to apply the simple method. The 'raw' comparisons have no credibility for any causal analysis, and the agreement of their results in our analyses is no indication that such an agreement may arise in a similar context.

We studied the effect of unemployment benefits. A more detailed and arguably more relevant issue is the effect of a change in the rules for awarding benefits. Such a study is feasible only if the planned or contemplated changes are implemented, either by an experiment or administratively; see Card and Levine (2000); Carling, Holmlund and Vejsiu (2001); and Lalive, van Ours and Zweimu¨ller (2006) for examples.

REFERENCES

ABADIE, A., IMBENS, G., (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* 74, 235–267.

CARD, D., LEVINE, P. B., (2000). Extended benefits and the duration of UI spells: evidence from the New Jersey extended benefit program. *Journal of Public Economics* 78, 107–138.

CARLING, K., HOLMLUND, B., VEJSIU, A., (2001). Do benefit cuts boost job finding? Swedish evidence from the 1990s. *Economic Journal*, 111, 766–790.

CRUMP, R.K., HOTZ, V. J., IMBENS, G. W., MITNIK, O. A., (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96, 187–199.

HIRANO, K., IMBENS, G., RIDDER, G., (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 1161–1189.

HOLLAND, P. W., (1986). Statistics and causal inference. *Journal of the American Statistical Association* 81, 945–970.

HUNT, J., (1995). The effect of unemployment compensation on unemployment duration in Germany. *Journal of Labor Economics* 13, 88–120.

LALIVE, R., ZWEIMÜLLER, J., (2004). Benefit entitlement and unemployment duration: The role of policy endogeneity. *Journal of Public Economics* 88, 2587–2616.

LALIVE, R., VAN OURS, J., ZWEIMÜLLER, J., (2006). How changes in financial incentives affect the duration of unemployment. *Review of Economic Studies* 73, 1009–1038.

LECHNER, M., (2002). Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *Journal of the Royal Statistical Society* Series A 165, 59–82.

LONGFORD, N. T., NICODEMO, C., NÚÑEZ, M., NÚÑEZ, E., (2011). Well-being and obesity of rheumatoid arthritis patients. *Health Services and Outcomes Research Methodology* 11, 27–43.

ROED, K., ZHANG, T., (2003). Does unemployment compensation affect unemployment duration? *Economic Journal* 113, 190–206.

ROSENBAUM, P. R., (2002). *Observational Studies*, 2nd ed. Springer-Verlag, New York.

ROSENBAUM, P. R., RUBIN, D. B., (1983). On the central role of the propensity score in matching. *Biometrika* 70, 41–55.

RUBIN, D. B., (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 688–701.

RUBIN, D. B., (2002). *Multiple Imputation for Nonresponse in Surveys*. Wiley and Sons, New York.

RUBIN, D. B., (2006). Matched Sampling for Causal Effects. Wiley and Sons, New York.

RUBIN, D. B., THOMAS, N., (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics* 52, 249–264.

UUSITALO, R., VERHO, J., (2010). The effect of unemployment benefits on re-employment rates: Evidence from the Finnish unemployment insurance reform. *Labour Economics* 17, 643–654.

APPENDIX

Definitions of the labour force states

The labour force status of a resident of Luxembourg is established at the end of each month by the following rules:

- employed (E) – an entry in IGSS, but no entry in ADEM in the month;
- unemployed (U) – an entry in ADEM, but no entry in IGSS in the month;
- in transition (T) – entries in both ADEM and IGSS in the month;
- economically inactive (I) – entry in neither ADEM nor IGSS in the month, but an entry in either database in an earlier month;
- absent (A) – entry in neither ADEM nor IGSS in the month, and no entry in either of them at any time in the past.

The status A is relevant only to persons who appear in a dataset in a later month.