Dipartimento di Statistica
"Giuseppe Parenti"

# Evaluating the effect of training on wages in the presence of noncompliance and missing outcome data

Paolo Frumento, Fabrizia Mealli, Barbara Pacini, Donald B. Rubin

Università degli Studi di Firenze

WORKING PAPER 2009/05

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

# Evaluating the effect of training on wages in the presence of noncompliance and missing outcome data

Paolo Frumento, Università di Firenze, frumento@ds.unifi.it

Fabrizia Mealli, Università di Firenze, mealli@ds.unifi.it

Barbara Pacini, Università di Pisa, barbara.pacini@sp.unipi.it

Donald B. Rubin, Harvard University, rubin@stat.harvard.edu

*Abstract:* Here, the effects of a training program on employment and wages are evaluated, using data from a randomized study, the National Job Corps Study, and the principal stratification approach to simultaneously address the complications of noncompliance, truncation of wages by nonemployment, and missing outcomes. We conduct a likelihood-based analysis, proposing different ways of improving computational efficiency and identifiability using the theory of finite mixture models and exploiting the EM algorithm. We maintain an IV exclusion restriction assumption on the outcomes, while monotonicity of compliance holds by design. We provide estimates with and without assuming monotonicity of the truncation of wages, under the Missing at Random assumption of outcome missingness. For compliers, results show that the effect on employment is generally positive; however, there is a subgroup of complying participants for whom the program is detrimental on employment in the short term, although this effect becomes negligible in the long term. For the subgroup of always-employed compliers, that is, those who would be employed whether trained or not, the effect of the training on wages is positive, both in the short and long term. According to our estimates, participation in the Job Corps training program is found to be effective: 208 weeks after participation, we found a positive causal effect on employment of about 5%, with an average causal effect on wages for the always employed of about 0.3 $/hour.

**Keywords**: Noncompliance, truncation by death, training, wages, principal stratification, finite mixture models.

# 1 Introduction

Estimating causal effects of interventions is often the focus of empirical studies in medicine and the social sciences. The only generally accepted approach for inferring causality requires that treatment receipt is randomized. Experiments, however, and social experiments in particular, often suffer from a number of complications, most notably noncompliance with assigned treatment, missing outcomes, and "truncation by death" when the outcome is not always well-defined (e.g, quality of life when dead).

We evaluate the effects of a randomized job-training program, Job Corps, which stands out as the largest, most comprehensive US education and job training program for disadvantaged youths between the ages of 16 and 24, using data from the National Job Corps Study, conducted by Mathematica Policy Research, Inc. The study is based on a national random sample of all eligible applicants in late 1994 and 1995. Sampled youths were assigned randomly to a program group or a control group. Consistent with the program's aim, key outcomes of interest are: employment, total earnings, and wages. In the empirical analysis, we focus on the effect of the program on wages and employment.

In the study all three complications mentioned above are present, namely a) compliance with assigned treatment was not perfect, as only 64% of those assigned to the program group effectively enrolled in Job Corps; b) due to attrition or item nonresponse, outcome is missing on some participants in the study; c) wages are truncated by death, meaning wages are not defined for those who are not employed.

Previous studies on these data neglected noncompliance, by focusing on intention-to-treat (ITT) effects of being offered participation in Job Corps (Lee, 2008; Flores-Lagunes et al., 2007; Zhang et al., 2008b). Being in a all-or-none compliance setting, and with access to Job Corps being denied to those assigned to the control group, individuals can be classified as compliers or never-takers (Angrist et al., 1996); in this setting the ITT effect, under a plausible outcome exclusion restriction assumption, can be regarded as being conservative for the effect of treatment receipt, i.e., it is possibly diluted by noncompliance to treatment assignment. That is, if the encouragement to take the job training could be increased, e.g., by advertising its success on those who completed it, the actual positive ITT effect would be larger. This may be a reason why, so far, negligible or small effects of Job Corps were found on employment and wages, especially in the long run (Zhang et al., 2008b; Lee, 2008). Here, we account for all complications, in order to evaluate the effects of Job Corps on those who were assigned and participated in the program, i.e., the compliers.

The framework we adopt uses potential outcomes to define causal effects regardless of the mode of inference, often referred to as the Rubin Causal Model (RCM; Holland, 1986); causal effects are defined by comparisons of potential outcomes on a common set of units (Rubin, 1974, 1978, 2005). We apply Principal Stratification (PS; Frangakis and Rubin, 2002), which was originally introduced to address post-treatment complications within the RCM. PS can be applied in various contexts, leading to parametric, semi and nonparametric inference, depending on the set of assumptions that are maintained, as well as whether point (full) or partial identification is to be achieved. It has been often used to represent and solve single complications. Few papers have dealt with more than one complication simultaneously; in general, the assumptions being considered are more complicated than those in the presence of each of the complications separately, and tend to grow exponentially with the number of distinct complications.

In this paper, we develop a likelihood-based approach to estimate the effect of training on employment and wages for compliers. We conduct a likelihood-based analysis using the EM algorithm, and propose different ways of improving computational efficiency and identifiability using the theory of finite mixture models. We maintain the exclusion restriction on outcomes, and monotonicity of compliance holds by design. We provide estimates with and without assuming monotonicity of the truncation of wages, under the Missing at Random assumption of outcome missingness. Following Frangakis and Rubin (2002), we classify the individuals into principal strata according to the joint values of the potential compliance, employment status, and attrition behaviour when assigned to be trained and when not assigned to be trained. Our primary causal estimands are: the average effect on employment for compliers and the average effect on wages for those compliers who are employed irrespective of treatment assignment.

Results support the claim that the effect on wages is positive for compliers. At week 45, there is a group of complying participants for whom participation is detrimental in terms of employment: in the short run, there are trained units who choose to be not employed and wait for a better job, whereas, if randomized to be not trained, would accept job offer they reject if trained, which results in a negative effect on employment. However, in the long run, this negative effect tends to disappear, and a greater employment rate is observed among the trained units. These results are consistent with the empirical literature, which suggests that almost all programs reduce employment and earnings in the short run. This so-called "lock-in" effect is well documented in many studies and can be also attributed to reduced search intensity of participants or fewer job offers during the program (Lechner and Wunsch, 2007; van Ours, 2004).

We proceed as follows. Section 2 reviews the literature on the three complications: noncompliance with treatment assignment, truncation by death, and missing outcomes. Section 3 discusses the framework needed to address the three issues simultaneously, and Section 4 illustrates the likelihood approach used to estimate the average treatment effects on employment and wages. Section 5 presents the results of the empirical analysis conducted on the Job Corps Study data; concluding remarks are provided in Section 6.

## 2 Analysis of broken randomized studies

A perfect randomized experiment is the accepted tool to infer on causal effects: units are randomly assigned to the treatment or to the control group, which ensures that treated and control units have the same distribution of all pre-randomization individual characteristics. The advantages of randomization are, however, partially lost if some experimental units do not comply with treatment assignment; noncompliance is a common issue, especially in experiments with human subjects. Different approaches have been proposed to deal with the nonrandom treatment receipt due to noncompliance. Comparing subjects by treatment received, rather than by treatment assigned, the "as-treated" analysis, generally leads to biased estimates of treatment effects; this is also true with a "per-protocol" approach, where only units who comply with treatment assignment are included in the analysis (Robins and Greenland, 1994; Sheiner and Rubin, 1995; Barnard et al., 1998). For these reasons, the standard approach to noncompliance is to compare average outcomes by assignment, ignoring the compliance behavior, as if compliance had been perfect; this is often referred to as the intention-to-treat (ITT) analysis (Breslow, 1982; Fisher et al., 1990; Lee et al., 1991; Meier, 1991). Yet this estimates the effect of assignment to treatment, not the effect of receipt of treatment.

Imbens and Angrist (1994) and Angrist et al. (1996) show the assumptions required to estimate the causal effect of treatment receipt, rather than treatment assignment, on the subpopulation of compliers, in the framework of the econometric instrumental variables (IV) method: this is indeed a special case of principal stratification, where principal strata are defined by the post-assignment compliance behavior. Monotonicity and exclusion restriction assumptions play a crucial role in this setting: monotonicity rules out the existence of defiers (units doing the opposite of their assignment), whereas, by the exclusion restriction, assignment is affecting the outcome only through the receipt of the treatment. Some testable constraints are implied by the exclusion restriction (Balke and Pearl, 1997; Imbens and Rubin, 1997b), but in order to relax it, it is useful to make additional assumptions. Various strategies have been proposed in the literature to achieve identification in the absence of exclusion restrictions. Imbens and Rubin (1997a) develop likelihood-based, including fully Bayesian, methods which improve upon conventional econometric IV estimators. Little and Yau (1997) and Hirano et al. (2000) extend the analysis of Imbens and Rubin (1997b) to allow for the presence of pre-treatment covariates; if they are available, more modelling options, other than strictly forcing the exclusion restriction, can be considered to achieve identifiability.

In Hirano et al. (2000), relaxing exclusion restrictions, but maintaining monotonicity within a full Bayesian analysis, allows the estimation of the effect of assignment for various subpopulations defined by compliance status. Covariates can also be exploited to achieve identification and improve efficiency. In the framework of principal stratification, plausible behavioral hypotheses, within or among groups defined by the values of the covariates, can be translated into restrictions on coefficients within or among strata. For some covariates, the same coefficient across strata can be imposed (Frangakis, 2006), or some interaction terms can be excluded (Jo, 2002).

The presence of missing outcomes characterizes both experimental and observational studies. As with noncompliance, outcome missingness is a post-treatment complication which breaks the benefits of randomization. Inference requires some form of imputation of the missing outcome data, either implicit or explicit. The appropriate set of assumptions, however, depends on the scientific setting. An account of the different approaches proposed in the literature to deal with both noncompliance and missing outcomes is in Mealli and Rubin (2002).

The standard ITT analysis based on complete cases leads to an unbiased estimate of the treatment effect only under the very restrictive assumption that the data are Missing Completely at Random (MCAR; Rubin, 1976; Little and Rubin, 1987). This assumption has testable implications and is often rejected by the data. A more convenient assumption is that the outcomes are Missing at Random (MAR; Rubin, 1976); this assumption allows the probability of missingness to depend on observed but, given these, not on unobserved quantities. If MAR holds and the parameters of the missing data process are distinct from those of the outcome distribution, then the missing data process is said to be ignorable, meaning that the missing data values do not affect the probability of response, given the observed quantities; unfortunately, this very attractive assumption is not testable without auxiliary information, because the data cannot provide any evidence against MAR. An assumption that links noncompliance with missing outcomes is Latent Ignorability (LI; Frangakis and Rubin, 1999). Under LI, the missing data process would be ignorable if the compliance behavior were known for all units; because the true compliance type is in general not observed, the missing mechanism is nonignorable. To achieve full identification of the effect for compliers under LI, additional assumptions are required; different forms of response exclusion restriction are discussed in Frangakis and Rubin (1999), Mealli and Rubin (2002), and Mealli et al. (2004).

Some advantages of random assignment are more difficult to realize if the outcome is not defined for all units. This problem has been dubbed "truncation by death" (Zhang and Rubin, 2003; Rubin, 2006), borrowing the term from medical clinical trials where the outcome – e.g., quality of life – is undefined for patients who die. This problem frequently arises also in the evaluation of social policy interventions (McConnell et al., 2008); e.g., in our case wages are neither observed nor well-defined for units who are not employed.

As pointed out in Rosenbaum (1984), a misleading inference could be obtained by simply comparing surviving treated participants and surviving controls; this is because the survivor status is again a post-treatment variable and the characteristics of the surviving units in the two groups are expected to differ (Lalonde, 1995). Rubin (2000) and Zhang and Rubin (2003) argue that, because the outcome is undefined for non survivors, the only meaningful causal effect is the one on individuals who would survive irrespective of whether they receive the intervention, and so estimate the survivor average causal effect, SACE (Rubin, 2000). Zhang and Rubin (2003) derive bounds on the possible range of values within which the SACE must lie, which are similar in spirit but narrower than bounds presented in Horowitz and Manski (2000). They also derive narrower bounds under monotonicity and stochastic dominance assumptions. Similar bounds are used by Lee (2008) in the analysis of the Job Corps data, where wages are truncated by nonemployment. Sharp bounds are derived by Imai (2007a), also for the effect on the always-survived compliers (Imai, 2007b).

The principal stratification approach allows us to explicitly model the truncation process and the outcome (Rubin 2006; Zhang, Rubin, and Mealli, 2008a, 2008b). In this paper, we follow this approach, focussing on identification and estimation of the intervention's effect on the subpopulation of the always-employed compliers, under specific assumptions on the missing mechanism.

# 3 Estimating the effect on wages with noncompliance and missing outcomes under the MAR assumption

Suppose each unit is randomly assigned to a treatment or to a control group: we denote by $Z_i$ the binary treatment assignment (1 = treatment, 0 = control) and by $D_i(Z_i)$ the binary treatment receipt for unit $i$ (1 = treatment, 0 = control). We are thus assuming that noncompliance is all-or-none, although in principle it is possible to generalize to multiple treatments and partial compliance (Jin and Rubin, 2008).

If compliance is perfect, $D_i(Z_i) = Z_i$. The indicators $D_i(z)$ ($z = 0, 1$) describe the compliance behavior and define four subpopulations: compliers ($C$), for whom $D_i(z) = z$; never-takers ($N$), for whom $D_i(z) = 0$; always-takers ($A$), for whom $D_i(z) = 1$; defiers ($D$), for whom $D_i(z) = 1 - z$ (Angrist et al., 1996). Without additional assumptions, the compliance status of unit $i$ is never observed; by virtue of randomization, however, the four types have the same distribution in both treatment groups.

We now define potential outcomes for all the remaining post-treatment variables. Let $S_i(z, d)$, $W_i(z, d)$, and $R_i(z, d)$ represent the potential employment status indicators (1 = employed, 0 = nonemployed), the potential wages for individual $i$, and the potential response indicators, for all possible combinations of treatment assignment and treatment received ($z = 0, 1$; $d = 0, 1$). Note that, given the compliance status, only two of the four potential outcomes are defined, namely, $S_i(z, D_i(z))$, $W_i(z, D_i(z))$, and $R_i(z, D_i(z))$, $z = 0, 1$; the other two potential outcomes are "a priori counterfactuals". In order to avoid the use of such counterfactuals, we only let the post-treatment variables depend on treatment assignment: $S_i(z)$, $W_i(z)$, and $R_i(z)$, $z = 0, 1$. Following Zhang et al. (2008), because wages are well-defined only if $S_i(z) = 1$, we define the wages to be $W_i(z) = *$ when $S_i(z) = 0$, $z = 0, 1$. On the other hand, if $R_i(z) = 0$ then both $S_i(z)$ and $W_i(z)$ are missing. Since potential outcomes have been introduced as a function of treatment assignment for unit $i$ (irrespective of assignment of other units), we have assumed that SUTVA (Stable Unit Treatment Value Assumption; Rubin, 1978, 1980, 1990) holds; this implies that potential outcomes for individual $i$ are unaffected by the treatment assignments of other individuals, i.e., that there is no interference between units, and that there are no hidden versions of the treatment.

If we indicate by $Z_i$ the observed treatment assignment for unit $i$, then the observed data are:

$$Z_i, D_i(Z_i), R_i(Z_i), S_i(Z_i), W_i(Z_i) \quad i = 1, ..., N.$$

In what follows we will maintain some assumptions. The first one is an exclusion restriction type of assumption, which is assumed to hold for both $W$ and $S$. It implies that for never-takers, potential outcomes do not depend on treatment assignment:

*Assumption: Exclusion restriction for $S$ and $W$ for never-takers*

If $D_i(z) = 0$ ($z = 0, 1$), then

$$S_i(0) = S_i(1),$$

$$W_i(0) = W_i(1).$$

We do not impose any exclusion restriction on the potential response indicator: as a consequence, units with the same treatment receipt may have a different missing data mechanism, according to their treatment assignment.

The second assumption, which holds by construction in our empirical setting, because access to treatment is denied for those assigned to control, is a strong form of monotonicity as proposed by Imbens and Angrist (1994):

*Assumption: Strong monotonicity of compliance*

$$D_i(0) = 0 \text{ for all } i.$$

So, with respect to the compliance behavior, the population is only composed of compliers ($C$) and never-takers ($N$).

Within the Rubin Causal Model, individual causal effects are defined as a comparison of potential outcomes, and usually summarized as averages on a common set of units. With respect to the employment status, following Zhang et al. (2008a, 2008b), units can be classified as:

- $EE = \{i : S_i(1) = S_i(0) = 1\}$, those who would be employed regardless of the treatment assignment; for this stratum, $W_i(1)$ and $W_i(0)$ are defined in $\Re^+$;
- $EN = \{i : S_i(1) = 1$ and $S_i(0) = 0\}$, those who would be employed only under treatment; for this stratum, $W_i(1) \in \Re^+$ and $W_i(0) = *$;
- $NE = \{i : S_i(1) = 0$ and $S_i(0) = 1\}$, those who would be employed only if assigned to the control group; for this stratum, $W_i(1) = *$ and $W_i(0) \in \Re^+$;
- $NN = \{i : S_i(1) = S_i(0) = 0\}$, those who would be nonemployed regardless of the treatment assignment; for this stratum, $W_i(1) = W_i(0) = *$.

If we ignore the missingness mechanism, units can be cross-classified by compliance status and employment status: $[\{C, N\} \times \{EE, EN, NE, NN\}]$, into 8 groups.

By virtue of the exclusion restriction on the employment status, we can cross out the $N\&EN$ group and the $N\&NE$ group, which would imply a direct effect of $Z$ on $S$ for never-takers, thereby contradicting exclusion restriction for them: the groups reduce to six:

$$\{C\&EE, C\&EN, C\&NE, C\&NN, N\&EE, N\&NN\}.$$

Given this classification of units in principal (latent) strata, the average treatment effects of interest are the following:

- the average treatment effect of $Z$ on program participation, $D$:

$$\Delta^{(ZD)} = E[D_i(1)] - E[D_i(0)] = P[D_i(1) - D_i(0) = 1]$$

  which equals the proportion of compliers in the population and is assumed to be greater than zero, by the strong monotonicity assumption which holds by design;

- the average treatment effects of $Z$ on employment, $S$, for compliers, which is often interpreted as the effect of $D$ on $S$:

$$\Delta^{(DS)} = E[S_i(1)|D_i(1) - D_i(0) = 1] - E[S_i(0)|D_i(1) - D_i(0) = 1]$$
$$= P[G_i = C\&EN] - P[G_i = C\&NE]$$

- the average treatment effect of $Z$ on wages, $W$, for the always employed compliers, which is often interpreted as the effect of participation on wages for the always-employed:

$$\Delta^{(DW)} = E[W_i(1)|G_i = C\&EE] - E[W_i(0)|G_i = C\&EE]$$

where in the last two formulas we emphasize that the expected values are taken on a subset of the whole population, the compliers and the always-employed compliers, respectively.

Without further assumptions on the missingness mechanism (such as monotonicity of missingness behavior and exclusion restrictions for some subgroup of units, Mealli and Rubin, 2003), each of the above principal strata is a mixture of 4 subgroups, according to the pair of potential reponse indicators $R_i(1)$, $R_i(0)$: units with never missing outcomes (RR) ($R_i(1) = 1$ and $R_i(0) = 1$), units with always missing outcomes (rr), units with outcomes missing only under control (Rr), and units with outcomes missing only under treatment (rR). Units can be now cross classified into 24 latent strata $\{C\&EE, C\&EN, C\&NE, C\&NN, N\&EE, N\&NN\} \times \{RR, rr, Rr, rR\}$. Among the assumptions about the missing data process proposed in the literature to simplify inference with missing values, one which appears to be plausible in our context is Latent Ignorability (LI), originally proposed by Frangakis and Rubin (1999) in a causal inference context with noncompliance. Because in our setting latent strata are defined w.r.t. noncompliance and potential employment status, LI can be reformulated as follows :

$$\Pr(R_i|Z_i, D_i(1), S_i(1), S_i(0), \mathbf{X}_i, W_i) = \Pr(R_i|Z_i, D_i(1), S_i(1), S_i(0), \mathbf{X}_i)$$

where $D_i(1)$ is the true compliance behavior (1 = for compliers, 0 = for never-takers) and $\mathbf{X}_i$ is an optional vector of pre-treatment covariates. Latent Ignorability implies that if we knew the group membership ($G_i$) of each unit (and the parameters of the missing data process were distinct from those of the outcome distribution), the missing mechanism would be ignorable. Under this assumption, given the covariates and the treatment assignment, units with the same compliance behavior and potential employment status have the same distribution of wages, regardless of the missingness behavior, i.e., $\Pr(W_i|Z_i, D_i(1), G_i, \mathbf{X}_i, R_i) = \Pr(W_i|Z_i, D_i(1), G_i, \mathbf{X}_i)$. We define the following response probabilities:

$$\rho_{i:g,z} = P(R_i = 1|Z_i = z, G_i = g, \mathbf{X}_i)$$

that is, $\rho_{i:g,z}$ is the probability of response for unit $i$, conditional on principal stratum membership $g$ and treatment assignment $z$ ($g \in G, z = \{0, 1\}, i = 1, ..., N$). Since the true compliance behavior and the potential employment status are partially unobserved, the missing data process is in fact nonignorable.

An alternative missing data assumption, which could be a reasonable one if the set of pre-treatment covariates contains rich information on units, is Missing at Random (MAR; Rubin, 1976), which in our setting can be written as:

$$\Pr(R_i|Z_i, D_i, \mathbf{X}_i, S_i, W_i) = \Pr(R_i|Z_i, D_i, \mathbf{X}_i).$$

Under MAR the response probability is the same for all units with the same treatment assignment, treatment receipt and pre-treatment covariates.

LI and MAR are substantially different assumptions on the missing data mechanism; usually when assuming LI additional restrictions on the response probabilities, $\rho_{i:g,z}$, are imposed to achieve or improve identifiability (Frangakis and Rubin 1999; Mattei and Mealli, 2007).

In our case, we can prove that the missing mechanism is ignorable if, in addition to LI, we impose the following restrictions on the response probabilities:

$$\rho_{i:C\&EE,1} = \rho_{:C\&EN,1} = \rho_{i:C\&NE,1} = \rho_{i:C\&NN,1}, \tag{1}$$

$$\rho_{i:N\&EE,1} = \rho_{i:N\&NN,1},$$

$$\rho_{i:C\&EE,0} = \rho_{i:C\&EN,0} = \rho_{i:C\&NE,0} = \rho_{i:C\&NN,0} = \rho_{i:N\&EE,0} = \rho_{i:N\&NN,0}.$$

These restrictions provide an alternative interpretation of the MAR assumption in this specific context: compliers and never-takers are allowed to have a different response behavior under treatment – since the value of their $D_i$ is different – but not under control, where the two groups have the same observed value of $D_i$. In the next section, we will illustrate how LI and MAR are used in specifying the observed likelihood function.

# 4 Likelihood-based estimation

The estimation issue can be viewed as a missing data problem, because we cannot observe which stratum each unit belongs to; among units with observed outcomes, we can observe the following groups, defined according to different combinations of observed $Z$, $D$ and $S$:

- $O(1, 1, 1) = \{i : Z_i = 1, D_i = 1 \text{ and } S_i = 1\}$, those who are assigned to the treatment group, comply with assignment and are employed; they are a mixture of the two principal strata $C\&EE$ and $C\&EN$;

- $O(1,1,0) = \{i : Z_i = 1, D_i = 1$ and $S_i = 0\}$, those who are assigned to the treatment group, comply with assignment and are nonemployed; they are a mixture of the two principal strata $C\&NN$ and $C\&NE$;

- $O(1,0,1) = \{i : Z_i = 1, D_i = 0$ and $S_i = 1\}$, those who are assigned to the treatment group, do not comply with assignment and are employed; they belong to the principal stratum $N\&EE$;

- $O(1,0,0) = \{i : Z_i = 1, D_i = 0$ and $S_i = 0\}$, those who are assigned to the treatment group, do not comply with assignment and are not employed; they belong to the principal stratum $N\&NN$;

- $O(0,0,1) = \{i : Z_i = 0, D_i = 0$ and $S_i = 1\}$, those who are assigned to the control group and are employed; they are a mixture of the three principal strata $C\&EE, C\&NE, N\&EE$;

- $O(0,0,0) = \{i : Z_i = 0, D_i = 0$ and $S_i = 0\}$, those who are assigned to the control group and are not employed; they are a mixture of the three principal strata $C\&EN, C\&NN, N\&NN$.

For units with missing outcomes, the values of $S$ and $W$ are not observed; the observed groups are defined according to the values of $Z$ and $D$:

- $O'(1,1) = \{i : Z_i = 1$ and $D_i = 1\}$, those who are assigned to the treatment group and comply with assignment; they are a mixture of the four principal strata $C\&EE, C\&EN, C\&NE, C\&NN$;

- $O'(1,0) = \{i : Z_i = 1$ and $D_i = 0\}$, those who are assigned to the treatment group and do not comply with assignment; they are a mixture of the two principal strata $N\&EE, N\&NN$;

- $O'(0,0) = \{i : Z_i = 0$ and $D_i = 0\}$, those who are assigned to the control group; they are a mixture of all the strata in $G$.

Note that, for those who are assigned to the treatment group, the compliance behavior is known; the employment status provides information about the pair $S_i(1)$, $S_i(0)$ and narrows the admissible strata. Among units with missing outcomes, the information on $S$ is not available: as a consequence, within the treatment group, we can only classify the units as compliers (the $O'(1,1)$

group) or as never-takers (the $O'(1,0)$ group), whereas in the control group (the $O'(0,0)$ group) additionally the compliance behavior is not observed, so that we cannot restrict the number of admissible strata.

In order to form the likelihood function, we have to specify a distribution for the potential outcomes conditional on the observed pre-treatment covariates, as well as a model for the principal strata membership $G$. The likelihood function results in a finite mixture model likelihood (see, e.g., McLachlan and Peel, 2000), which can be maximized using the EM (Expectation-Maximization) algorithm (Dempster, Laird and Rubin, 1977), as detailed in the Appendix.

We denote by $G_i$ the principal strata membership for unit $i$; $\mathbf{G} = G_1, ..., G_N$ is the corresponding $N$-dimensional vector. Under the stated assumptions, the principal strata defined w.r.t. noncompliance and employment are $\{C\&EE, C\&EN, C\&NE, C\&NN, N\&EE, N\&NN\}$. An additional assumption, which may improve estimation, is a monotonicity assumption for truncation, which rules out the existence of stratum $C\&NE$; that is, for compliers, treatment is not detrimental in terms of employment. If this assumption holds the number of principal strata reduces to $k = 5$.

To simplify the notation, we assume that $\mathbf{X}$ includes the constant term – that is, a column containing the unit vector. We specify a multinomial logistic model[1] for the $k$-dimensional vector of principal strata memberships:

$$P(G_i = g) = \frac{\exp\{\mathbf{X}_i \boldsymbol{\alpha}_g\}}{\sum_{h=1}^{k} \exp\{\boldsymbol{X}_i \boldsymbol{\alpha}_h\}} = \pi_{i:g}$$

where $g \in G$ and the $k^{th}$ principal stratum ($N\&NN$) is taken as the baseline (that is, $\boldsymbol{\alpha}_k = \mathbf{0}$). We denote by $\pi_{i:g}$ the probability of belonging to stratum $g$ for unit $i$, given the vector of pre-treatment covariates $\mathbf{X}_i$.

We specify a Normal distribution for log-wages:

$$\begin{aligned}
&if\ G_i = C\&EE, && \log[W_i(1)] \sim N(\mathbf{X}_i \boldsymbol{\beta}_{C\&EE,1}, \sigma^2_{C\&EE,1}),\\
& && \log[W_i(0)] \sim N(\mathbf{X}_i \boldsymbol{\beta}_{C\&EE,0}, \sigma^2_{C\&EE,0}),\\
&if\ G_i = C\&EN, && \log[W_i(1)] \sim N(\mathbf{X}_i \boldsymbol{\beta}_{C\&EN,1}, \sigma^2_{C\&EN,1}),\\
&if\ G_i = C\&NE, && \log[W_i(0)] \sim N(\mathbf{X}_i \boldsymbol{\beta}_{C\&NE,0}, \sigma^2_{C\&NE,0}),
\end{aligned}$$

[1] Alternative specifications, such as multinomial probit models, could be implemented.

$$if\ G_i = N\&EE, \qquad \log[W_i(1)] \sim \log[W_i(0)] \sim N(\mathbf{X}_i\boldsymbol{\beta}_{N\&EE}, \sigma^2_{N\&EE}).$$

For the $C\&EE$ group, the parameters of the wage distribution vary across the two treatment groups; for the $C\&EN$ group, wages are only defined if $Z_i = 1$; for the $C\&NE$ group, wages are only defined if $Z_i = 0$. The exclusion restriction implies that for the $N\&EE$ group the parameters of the wage distribution are the same irrespective of treatment assignment. Clearly, for the $C\&NN$ and $N\&NN$ groups, there are no associated wages.

We denote by $\boldsymbol{\xi} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\rho}\}$ the parameters' vector of this model, where

$$\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_{C\&EE}, \boldsymbol{\alpha}_{C\&EN}, \boldsymbol{\alpha}_{C\&NE}, \boldsymbol{\alpha}_{C\&NN}, \boldsymbol{\alpha}_{N\&EE}\}$$

$$\boldsymbol{\beta} = \{\boldsymbol{\beta}_{C\&EE,1}, \boldsymbol{\beta}_{C\&EE,0}, \boldsymbol{\beta}_{C\&EN,1}, \boldsymbol{\beta}_{C\&NE,0}, \boldsymbol{\beta}_{N\&EE}\}$$

$$\boldsymbol{\sigma} = \{\sigma_{C\&EE,1}, \sigma_{C\&EE,0}, \sigma_{C\&EN,1}, \sigma_{C\&NE,0}, \sigma_{N\&EE}\}$$

and $\boldsymbol{\rho}$ is the vector of the response probabilities. We denote by $N(\mu, \sigma)$ the probability density function of a Normal distribution with mean $\mu$ and variance $\sigma$ evaluated at $\log(W_i)$. Assuming Latent Ignorability, the likelihood function can be written as:

$$L(\boldsymbol{\xi}|\mathbf{Z}, \mathbf{D}, \mathbf{R}, \mathbf{S}, \mathbf{W}, \mathbf{X}) =$$

$$\prod_{i\in O(1,1,1)} \left[\rho_{i:C\&EE,1}\pi_{i:C\&EE}N(\mathbf{X}_i\boldsymbol{\beta}_{C\&EE,1}, \sigma^2_{C\&EE,1}) + \rho_{i:C\&EN,1}\pi_{i:C\&EN}N(\mathbf{X}_i\boldsymbol{\beta}_{C\&EN,1}, \sigma^2_{C\&EN,1})\right]^{\omega_i}$$

$$\times \prod_{i\in O(1,1,0)} \left[\rho_{i:C\&NE,1}\pi_{i:C\&NE} + \rho_{i:C\&NN,1}\pi_{i:C\&NN}\right]^{\omega_i} \times \prod_{i\in O(1,0,1)} \left[\rho_{i:N\&EE,1}\pi_{i:N\&EE}N(\mathbf{X}_i\boldsymbol{\beta}_{N\&EE}, \sigma^2_{N\&EE})\right]^{\omega_i}$$

$$\times \prod_{i\in O(1,0,0)} \left[\rho_{i:N\&NN,1}\pi_{i:N\&NN}\right]^{\omega_i} \times \prod_{i\in O(0,0,1)} \Big[\rho_{i:C\&EE,0}\pi_{i:C\&EE}N(\mathbf{X}_i\boldsymbol{\beta}_{C\&EE,0}, \sigma^2_{C\&EE,0})$$

$$+\rho_{i:C\&NE,0}\pi_{i:C\&NE}N(\mathbf{X}_i\boldsymbol{\beta}_{C\&NE,0}, \sigma^2_{C\&NE,0}) + \rho_{i:N\&EE,0}\pi_{i:N\&EE}N(\mathbf{X}_i\boldsymbol{\beta}_{N:EE}, \sigma^2_{N\&EE})\Big]^{\omega_i}$$

$$\times \prod_{i\in O(0,0,0)} \left[\rho_{i:C\&EN,0}\pi_{i:C\&EN} + \rho_{i:C\&NN,0}\pi_{i:C\&NN} + \rho_{i:N\&NN,0}\pi_{i:N\&NN}\right]^{\omega_i}$$

$$\times \prod_{i\in O'(1,1)} \left[(1-\rho_{i:C\&EE,1})\pi_{i:C\&EE} + (1-\rho_{i:C\&EN,1})\pi_{i:C\&EN} + (1-\rho_{i:C\&NE,1})\pi_{i:C\&NE} + (1-\rho_{i:C\&NN,1})\pi_{i:C\&NN}\right]^{\omega_i}$$

$$\times \prod_{i\in O'(1,0)} \left[(1-\rho_{i:N\&EE,1})\pi_{i:N\&EE} + (1-\rho_{i:N\&NN,1})\pi_{i:N\&NN}\right]^{\omega_i} \times \prod_{i\in O'(0,0)} \left[\sum_{g\in G} (1-\rho_{i:g,0})\pi_{i:g}\right]^{\omega_i}$$

where $\omega_i$ are eventual sampling weights. As said before, additional restrictions are needed to make the missing mechanism ignorable. Assuming that equalities in (1) hold, the likelihood factorizes and the model parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}\}$ can be estimated independently of the response probabilities. The likelihood function simplifies as follows:

$$L(\boldsymbol{\xi}|\mathbf{Z}, \mathbf{D}, \mathbf{R}, \mathbf{S}, \mathbf{W}, \mathbf{X}) \propto$$

$$\prod_{i \in O(1,1,1)} \left[\pi_{i:C\&EE} N(\mathbf{X}_i \boldsymbol{\beta}_{C\&EE,1}, \sigma^2_{C\&EE,1}) + \pi_{i:C\&EN} N(\mathbf{X}_i \boldsymbol{\beta}_{C\&EN,1}, \sigma^2_{C\&EN,1})\right]^{\omega_i}$$

$$\times \prod_{i \in O(1,1,0)} [\pi_{i:C\&NE} + \pi_{i:C\&NN}]^{\omega_i} \times \prod_{i \in O(1,0,1)} \left[\pi_{i:N\&EE} N(\mathbf{X}_i \boldsymbol{\beta}_{N\&EE}, \sigma^2_{N\&EE})\right]^{\omega_i} \times \prod_{i \in O(1,0,0)} [\pi_{i:N\&NN}]^{\omega_i}$$

$$\times \prod_{i \in O(0,0,1)} \left[\pi_{i:C\&EE} N(\mathbf{X}_i \boldsymbol{\beta}_{C\&EE,0}, \sigma^2_{C\&EE,0}) + \pi_{i:C\&NE} N(\mathbf{X}_i \boldsymbol{\beta}_{C\&NE,0}, \sigma^2_{C\&NE,0}) + \pi_{i:N\&EE} N(\mathbf{X}_i \boldsymbol{\beta}_{N\&EE}, \sigma^2_{N\&EE})\right]^{\omega_i}$$

$$\times \prod_{i \in O(0,0,0)} [\pi_{i:C\&EN} + \pi_{i:C\&NN} + \pi_{i:N\&NN}]^{\omega_i}$$

$$\times \prod_{i \in O'(1,1)} [\pi_{i:C\&EE} + \pi_{i:C\&EN} + \pi_{i:C\&NE} + \pi_{i:C\&NN}]^{\omega_i}$$

$$\times \prod_{i \in O'(1,0)} [\pi_{i:N\&EE} + \pi_{i:N\&NN}]^{\omega_i}.$$

The units in the $O'(1,1)$ and $O'(1,0)$ groups provide information on their compliance behavior and affect the estimates of the $\pi_{i:g} (i = 1, ..., N; g \in G)$; the units in the $O'(0,0)$ group are uninformative and disappear from the likelihood function (since $\sum_g \pi_{i:g} = 1$). Note that assuming monotonicity of truncation amounts to setting $\pi_{i:C\&NE} = 0$ and it would force the estimate of the treatment effect on employment to be positive. The complete-data log-likelihood function is shown in the Appendix, together with the derivation of the EM algorithm steps. Note in addition that the likelihood is precisely the one that would be written and used under the MAR assumption, if the parameters of the missing data process are distinct from those of the outcome distributions.

Once parameter estimates are obtained, the causal effects of interest can be evaluated, averaging over the observed distribution of covariates. Following Zhang et al. (2008b), we estimate the proportion of each stratum as

$$\hat{\pi}_g = \frac{\sum_{i=1}^{N} \omega_i \hat{\pi}_{i:g}}{\sum_{i=1}^{N} \omega_i}.$$

The causal effect of $Z$ on $D$ is estimated as the proportion of compliers: $\hat{\Delta}^{(ZD)} = \hat{\pi}_{C\&EE} + \hat{\pi}_{C\&EN} + \hat{\pi}_{C\&NE} + \hat{\pi}_{C\&NN}$. Estimates of the average treatment effects on wages and employment are obtained as:

$$\hat{\Delta}^{(DW)} = \frac{\sum_{i=1}^{N} \omega_i \hat{\pi}_{i:C\&EE} \exp\left\{\mathbf{X}_i \hat{\boldsymbol{\beta}}_{C\&EE,1} + \frac{1}{2}\hat{\sigma}^2_{C\&EE,1}\right\}}{\sum_{i=1}^{N} \omega_i \hat{\pi}_{i:C\&EE}} - \frac{\sum_{i=1}^{N} \omega_i \hat{\pi}_{i:C\&EE} \exp\left\{\mathbf{X}_i \hat{\boldsymbol{\beta}}_{C\&EE,0} + \frac{1}{2}\hat{\sigma}^2_{C\&EE,0}\right\}}{\sum_{i=1}^{N} \omega_i \hat{\pi}_{i:C\&EE}}$$

and

$$\hat{\Delta}^{(DS)} = \hat{\pi}_{C\&EN} - \hat{\pi}_{C\&NE}$$

respectively.

Once the asymptotic covariance matrix of the estimates has been obtained, the asymptotic standard errors of the above quantities may be computed using the Delta method. The analysis could be improved by avoiding asymptotic approximations and employ either a Direct Likelihood approach (as in Zhang et al., 2008b) or a full Bayesian analysis.

In the next Section, this framework will be applied to the Job Corps Study, to evaluate the effects of the program on employment and wages.

# 5 Application to the Job Corps Study

The evaluation of government-sponsored job-training programs is a difficult task, undertaken by a number of authors in last decades (Heckman and Hotz, 1989, Lalonde, 1995, Burghardt et al., 2001, Zhang, Rubin and Mealli, 2008a, 2008b). For our analysis, we use data from the National Job Corps Study (conducted by Mathematica Policy Research, Inc. for the U.S. Department of Labor) and estimate the effect of the program on employment and wages. The data are from a random sample of all selected applicants (N = $15,386$) in 1994 and 1995: among them, units were randomly assigned to the program or to the control group; only those assigned to the program group ($9,409$ units, about 61%) were admitted to enroll in the Job Corps: among them, $6,039$ (64%) complied with the assignment. For all units, pre-treatment covariates (**X**) were collected. In principle, with full compliance and no missing outcomes, including covariates in the analysis is not fundamental for estimating the treatment effects, because – by virtue of the randomization process – the covariates distribution is independent of treatment assignment. However, the covariates are helpful and necessary for three main reasons here: first, they generally improve the model specification and the prediction of the unobserved potential outcomes; second, they allow a more plausible generalization to a population with different characteristics; third, conditioning on covariates is explicitly required by the MAR assumption.

Summary statistics of the pre-treatment covariates ($\mathbf{X}$) are displayed in Table 1 (N = 15,376: we removed 2 observations aged more than 30 – all others units are aged 16-24 – and 8 units with overly large ($> 50,000$) values for earnings in the previous year). Statistics show that there are some missing values among the covariates. We solved this missing data problem by imputing the missing values in $\mathbf{X}$ using the `mice` procedure in R, which generates multiple imputations for incomplete multivariate data by Gibbs Sampling; we used only the baseline covariates as predictors in the chained equations. Linear regression has been used for numerical covariates; binary/multinomial logistic models for dichotomous/polytomous variables. Ten different imputation have been generated, leading to very similar estimates of the model parameters, as found also in Zhang et al. (2008b): for this reason, we only present the results from one single imputed data set.

In the linear predictor, the education (number of years of schooling) has been included as a dummy variable (= 1 if greater than the sample median); we also collapsed the information on the marital status in the dummy "with a partner".

Table 2 presents summary statistics for the outcome variables (employment, total earnings and weekly hours at $45^{th}$, $135^{th}$ and $208^{th}$ week after treatment). In our analysis, inadmissible outcome values (units with more than 84 weekly hours, employed people with zero weekly earnings or hours) are considered missing.

To simplify the model, we assumed that treatment assignment for compliers enters in the linear predictor without interactions with the covariates; that is, $\boldsymbol{\beta}_{C\&EE,1}$ and $\boldsymbol{\beta}_{C\&EE,0}$ only differ in the intercept, so that

$$\mathbf{X}_i\boldsymbol{\beta}_{C\&EE,1} = \mathbf{X}_i\boldsymbol{\beta}_{C\&EE,0} + \gamma$$

for each $i$. We also assumed that the treatment receipt in the $C\&EE$ group has no effect on the variance: this implies $\sigma^2_{C\&EE,0} = \sigma^2_{C\&EE,1}$. The exclusion restriction is always maintained – that is, we constrained the causal effects for never-takers to be zero. Violations of the exclusion restriction have no testable consequences and – in this case – we believe that this assumption is plausible; however, units who refused the enroll in Job Corps could regret the vanished opportunity: we do not know if this eventuality has had some consequences in terms of potential employment status and potential wages.

The model has now 221 parameters; the monotonicity of truncation assumption rules out the $C\&NE$ group and reduces the number of parameters to 175. For each of the 3 weeks under study separately, we estimated a model with and without monotonicity assumption for truncation.

Tables 3 and 4 present results without and with monotonicity of truncation, respectively. Without imposing monotonicity (Table 3), for week 45 we found evidence that all latent strata exist; for compliers, we estimated a negative treatment effect on employment ($-8.22\%$), whereas for the always employed compliers, the effect on wages is found to be positive (about 0.276 \$/hour). For weeks 135 and 208, a positive treatment effect was found on both employment (+4.87% and +4.85%, respectively) and wages (0.210 and 0.337 \$/hour, respectively). The estimated probability of the $C\&NE$ group was found to be very high (15.47%) at week 45, but negligible at the subsequent weeks (1.37% at week 135, 1.35% at week 208).

Because of this lack of evidence of the $C\&NE$ group, we also estimated the model imposing monotonicity of truncation; results are displayed in Table 4. With respect to Table 3, completely different estimates are found for week 45: the effect on employment is constrained to be positive and very different probabilities are obtained for the $C\&EE$, $C\&EN$ and $C\&NN$ strata. According to the AIC (Akaike's Information Criterion) the monotonicity of truncation should be rejected for week 45. In week 135 and 208, the estimates under monotonicity are very similar to those of Table 3; in both cases, the BIC suggests that the model with monotonicity should be preferred, whereas the opposite conclusion is drawn according to the AIC; however, for weeks 135 and 208 this assumption appears to be quite reasonable.

In the short run (after 45 weeks), there are trained units who choose to be nonemployed and wait for a better job: this results in a negative effect on employment. In the long run, the $C\&NE$ group tends to disappear and a greater employment rate is observed among trained units. These results are consistent with the empirical literature on the effect of active labor market policies, which suggests that almost all programs reduce employment and earnings in the short run. This so-called “lock-in” effect is well documented in many studies and can be also attributed to reduced search intensity of participants or fewer job offers during the program (Lechner and Wunsch, 2007; van Ours, 2004).

For the always employed compliers, the effect on wages is found to be higher in week 45 than in week 135; this corroborates the above arguments: in the short run, treated units can either find a good job or choose to remain not employed; in the long run, job selection probably becomes less strict: this may be a reason why a still positive but lower effect on wages (together with a positive effect on employment) is found at week 135. However, the gap between treated and control units seems to increase over time, as suggested by our estimates of the treatment effect on wages for the always employed compliers at week 208. Note, however, that the different effect on wages found at the three different weeks could also be due to the fact that the unobserved group of the always employed compliers may include different units at different weeks.

These results also demonstrate how crucial is the monotonicity of truncation in this case: for weeks 135 and 208, it seems reasonable that this assumption holds, because there is a very weak evidence of existence of the $C\&NE$ group, whereas at week 45 monotonicity is not a plausible assumption.

The obtained results may be sensitive to our working assumptions; in particular, the exclusion restriction could be questionable, because units who refused the treatment could regret the vanished opportunity; however, removing this assumptions may be detrimental in terms of model identification. A possible strategy to improve identification is to use a multivariate model, e.g., a bivariate normal distribution for the couple ($\log(W)$, $\log(H)$) for each week, where $W$ denotes the hourly wage and $H$ the weekly working hours, or a joint dynamic model over the three weeks under study for employment status and wages. With a multivariate approach, increased efficiency could be achieved; with the aim of decomposing a finite mixture with a great number of components, using a multiple classification criterion is also expected to reduce the occurrence of spurious optimizers and local maxima, which represents a serious problem in the estimation of the model we presented here.

## 6 Concluding remarks

We evaluated the effects of the Job Corps training program on employment and wages, using data from a randomized study, the National Job Corps Study, and the principal stratification approach to simultaneously address the issues of noncompliance, truncation of wages and missing outcomes. This approach consists in estimating the causal effects of interest for a common set of units: the average treatment effect on employment is estimated on the subpopulation of compliers; among them, only for the always employed the effect on wages is defined and estimated in a meaningful way.

Pre-treatment covariates were used in the prediction of the potential outcomes, the compliance behavior, and missing outcomes; we focused our analysis on the observed outcomes at $45^{th}$, $135^{th}$ and $208^{th}$ weeks after participation in the program. The exclusion restriction on outcomes for never-takers was maintained, whereas the presence of always-takers and defiers is excluded by design. Some restrictions on covariates reduced the number of parameters and improved model identification.

The treatment effects were found to be increasing over time; the effect on employment was negative at week 45, whereas the estimated effect on wages is always positive. A critical role is played by the monotonicity of truncation assumption, which rules out those who would be nonemployed if treated and employed if not treated: this assumption does not seem to hold at week 45, but becomes more plausible at weeks 135 and 208. We may argue that, in a short run, there are trained units who choose to be not employed, waiting for a "good" job; in the long run, trained units are more likely to find a job, and a positive treatment effect on employment is found. These results are consistent with the empirical literature on the effect of active labor market policies, which suggests that almost all programs reduce employment and earnings in the short run (the so-called "lock-in" effect). Finally, the effect on wages reflects the increase in the human capital due to the program participation.

The obtained results may be sensitive to our working assumptions; however, relaxing some hypotheses may weaken identification and lead to poor estimates in terms of efficiency. In order to overcome the difficulties inherent in the lack of full identification when the exclusion restriction is relaxed, a possible strategy could be the simultaneous modeling of more than one outcome; indeed, the use of multivariate models generally provides a greater discriminant power in disentangling mixtures.

# References


[1] Angrist, J.D., Imbens, G.W., Rubin, D.B. (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91, 444-455.

[2] Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. Review of Economics and Statistics, 60, 47-57.

[3] Ashenfelter, O., Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect on training programs. Review of Economics and Statistics, 67, 648-660.

[4] Balke, E., Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. Journal of the American Statistical Association, 92, 1171-1176.

[5] Barnard, J., Du, J., Hill, J.L., Rubin, D.B. (1998). A broader template for analyzing broken randomized experiments. Sociological Methods and Research, 27, 285-317.

[6] Breslow, N.E. (1982). Clinical trials. In Encyclopedia of Statistical Sciences, 2, 13-21. Wiley, New York.

[7] Burghardt, J., McConnell, S., Schochet, P., Johnson, T., Gritz, M., Glazerman, S., Homrighausen, J. (2001). Does Job Corps work? Summary of the National Job Corps Study. Document No. PR01-50, Princeton, NJ: Mathematica Policy Research, Inc.

[8] Burghardt, J., McConnell, S., Schochet, P. (2003). National Job Corps Study: findings using administrative earnings records data. Final report. Document No. PR03-92, Princeton, NJ: Mathematica Policy Research, Inc.

[9] Card, D., Sullivan, D. (1988). Measuring the effect of subsidized training programs on movements in and out of employment. Econometrica, Vol. 56, 3, 497-530

[10] Dempster, A.P., Laird, N., Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data using the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B 39, 1-38.

[11] Fisher, R.A. (1925). The Design of Experiments. Oliver and Boyd, London, 1st edition.

[12] Fisher, L., Dixon, D., Herson, J., Frankowski, R., Hearron, M., Peace, K. (1990). Intention to treat in clinical trials. In Statistical Issues in Drug Research and Development (K. Peace, ed.), 331-350. Dekker, NewYork.

[13] Flores-Lagunes, A., Gonzalez, A., Neumann, T. (2007). Estimating the effects of length of exposure to a training program: the case of Job Corps. IZA Discussion Paper, No. 2846.

[14] Fraker, T., Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. Journal of Human Resources, Vol. 22, 2, 194-227.

[15] Frangakis, C.E., Rubin, D.B. (2002). Principal stratification in causal inference. Biometrics, 58, 21-29.

[16] Frangakis, C.E. (2006). Comment to "A. Forcina, Causal effects in the presence of noncompliance: a latent variable interpretation". Metron, Vol 64, 3, 1-27.

[17] Frumento, P. (2009). Finite Mixture Models. Some computational and theoretical developments with applications. PhD Thesis, University of Florence.

[18] Heckman, J.J. (1979). Sample selection bias as a specification error. Econometrica, 47,153-162.

[19] Heckman, J., Robb, R. (1985). Alternative methods for evaluating the impact of interventions. In Heckman and Singer (eds.), Longitudinal analysis of labor market data, Cambridge, Cambridge University Press.

[20] Heckman, J.J., Vytlacil, E.J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proceedings of the National Academy of Sciences, USA, 96, 4730-4734.

[21] Hirano, K., Imbens, G., Rubin, D.B., Zhou, X. (2000). Assessing the effect of an influenza vaccine in an encouragement design. Biostatistics, 1, 69-88.

[22] Holland, P. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81, 945-960.

[23] Horowitz, J. Manski, C. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. Journal of the American Statistical Association.

[24] Imai, K. (2007a). Sharp bounds on the causal effects in randomized experiments with "truncation-by-death". Statistics & Probability Letters, 78, 144-149.

[25] Imai, K. (2007b). Identification analysis for randomized experiments with noncompliance and truncation by death. Technical Report, Department of Politics, Princeton University.

[26] Imbens, G.W., Angrist, J. (1994). Identification and estimation of local average treatment effects. Econometrica, 62, 467-476.

[27] Imbens, G.W., Rubin, D.B. (1997a). Bayesian inference for causal effects in randomized experiments with noncompliance. The Annals of Statistics, Vol. 25, 1, 305-327.

[28] Imbens, G.W., Rubin, D.B. (1997b). Estimating outcome distributions for compliers in instrumental variables models. Review of Economic Studies, 64, 555-574.

[29] Jin, H., Rubin, D.B. (2008). Principal stratification for causal inference with extended partial compliance. Journal of the American Statistical Association, 103, 101-111.

[30] Jo, B. (2002). Estimation of intervention effects with noncompliance: alternative model specifications. Journal of Educational and Behavioral Statistics,Vol. 27,No. 4 (Winter, 2002), 385-409.

[31] Jo, B. (2008). Bias mechanisms in intention-to-treat analysis with data subject to treatment noncompliance and missing outcomes. Published on behalf of American Educational Research Association, http://jeb.sagepub.com/cgi/content/abstract/33/2/158.

[32] Lalonde, R.J. (1986). Evaluating the econometric evaluations of training programs with experimental data. American Economic Review, 76, 604-620.

[33] Lalonde, R.J. (1995). The promise of public sector-sponsored training programs. The Journal of Economic perspectives, 9, 149-168.

[34] Lechner, M., Wunsch, C. (2007). Are training programs more effective when unemployment is high? IAB Discussion Paper 200707.

[35] Lee, D.S. (2008). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. Review of Economic Studies, forthcoming.

[36] Lee, Y., Ellenberg, J., Hirtz, D., Nelson, K. (1991). Analysis of clinical trials by treatment actually received: is it really an option? Statistics in Medicine, 10, 1595-1605.

[37] Little, R., Yau, L. (1998). Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. Psychological Methods, 3, 147-159.

[38] Manski, C. (1990). Nonparametric bounds on treatment effects. American Economic Review Papers and Proceedings, 80, 319-323.

[39] Mattei, A., Mealli, F. (2007). Application of the prinicipal stratification approach to the Faenza randomized experiment on breast self-examination. Biometrics, 63, 2, 437-446.

[40] Mealli, F., Rubin, D.B. (2002). Assumptions when analyzing randomized experiments with noncompliance and missing outcomes. Health Services and Outcomes Research Methodology, Springer Netherlands, Vol. 3, Numbers 3-4, 225-232.

[41] McConnell, S., Stuart, E.A., Devaney, B.m. (2008). The Truncation-by-Death Problem. What To Do in an Experimental Evaluation When the Outcome Is Not Always Defined. Evaluation Review, 32, 2 April, 157-186.

[42] McLachlan, G., Peel, D. (2000). Finite Mixture Models. Wiley Series in Probability and Statistics, New York, USA.

[43] Meier, P. (1991). Comment on "Compliance as an explanatory variable in clinical trials" by B. Efron and D. Feldman. Journal of the American Statistical Association, 86, 19-22.

[44] Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles. Translated in Statistical Science, 5, 465-480, 1990.

[45] Robins, J.M., Greenland, S. (1994). Adjusting for differential rates of prophylaxis therapy for PCP in high- versus low-dose AZT treatment arms in an AIDS randomized trial. Journal of the American Statistical Association, 89, pp 737-749.

[46] Rosenbaum, P. (1984). The Consequences of adjustement for a concomitant variable that has been affected by the treatment. Journal of the Royal Statistical Society, Series A, 147, 656-666.

[47] Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and non randomized studies. Journal of Educational Psychology, 66, 688-701.

[48] Rubin, D.B. (1976). Inference and missing data. Biometrika, 63, 581-592.

[49] Rubin, D.B. (1978). Bayesian inference for causal effects. Annals of Statistics, 6, 34-58.

[50] Rubin, D.B. (1980). Discussion of "Randomization analysis of experimental data: the Fisher randomization test" by D. Basu. Journal of the American Statistical Association, 75, 591-593.

[51] Rubin, D.B. (1990). Formal modes of statistical inference for causal effects. Journal of Statistical Planning and Inference, 25, 279-292.

[52] Rubin, D.B. (2000). The utility of counterfactuals for causal Inference – Discussion of "Causal inference without counterfactuals" by A.P. Dawid. Journal of the American Statistical Association, 95, 435-438.

[53] Rubin, D.B. (2005). Causal Inference using potential outcomes: design, modeling, decisions. Journal of the American Statistical Association, 100, 322-331.

[54] Rubin, D.B. (2006). Causal inference through potential outcomes and principal stratification: application to studies with censoring due to death. Statistical Science, 21, 299-321.

[55] Schochet, P.Z., Bellotti, J., Cao, R.J., Glazerman, S., Grady, A., Gritz, M., McConnell, S., Johnson, T., Burghardt, T. (2003). National Job Corps Study: data documentation and public use files: Volume I. Mathematica Policy Research, Inc.

[56] Sheiner, L.B., Rubin, D.B. (1995). Intention-to-treat analysis and the goals of clinical trials. Clinical Pharmacology and Therapy, 57, 6-10.

[57] Titterington, D.M., Smith, A.F.M., Makov, U.E. (1985). Statistical analysis of Finite Mixture Distributions. New York: Wiley.

[58] Venables,W. N., Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

[59] Van Buuren, S., Oudshoorn, C.G.M. (2007). mice: Multivariate Imputation by Chained Equations. R package version 1.16. http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm.

[60] van Ours, J. (2004). The locking-in effect of subsidized jobs. Journal of Comparative Economics, 32, 37-52.

[61] Zhang, J.L., Rubin, D.B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". Journal of Educational and Behavioral Statistics, 28, 353-368.

[62] Zhang, J.L., Rubin, D.B., Mealli, F. (2008a). Evaluating the effects of job training programs on wages through principal stratification. In Modelling and Evaluating Treatment Effects in Econometrics, D.L. Millimet, J.A. Smith, E.J. Vytlacil, eds., Elsevier, 117-145.

[63] Zhang, J.L., Rubin, D.B., Mealli, F. (2008b). Likelihood-based analysis of the causal effects of job-training programs using principal stratification. Journal of the American Statistical Association, forthcoming.

# Appendix

We provide the steps of the EM algorithm used in the estimation of the model described in Section 4. The complete-data log-likelihood function may be written as follows:

$$
\begin{aligned}
& l(\boldsymbol{\xi}|\mathbf{Z}, \mathbf{D}, \mathbf{R}, \mathbf{S}, \mathbf{W}, \mathbf{X}, \mathbf{G}) = \\
& \sum_{i \in O(1,1,1)} \omega_i I(G_i = C\&EE) \log\left[\pi_{i:C\&EE} N(\mathbf{X}_i \boldsymbol{\beta}_{C:EE,1}, \sigma^2_{C\&EE,1})\right] \\
& + \sum_{i \in O(1,1,1)} \omega_i I(G_i = C\&EN) \log\left[\pi_{i:C\&EN} N(\mathbf{X}_i \boldsymbol{\beta}_{C\&EN,1}, \sigma^2_{C\&EN,1})\right] \\
& + \sum_{i \in O(1,1,0)} \omega_i I(G_i = C\&NE) \log\left[\pi_{i:C\&NE}\right] + \sum_{i \in O(1,1,0)} \omega_i I(G_i = C\&NN) \log\left[\pi_{i:C\&NN}\right] \\
& + \sum_{i \in O(1,0,1)} \omega_i I(G_i = N\&EE) \log\left[\pi_{i:N\&EE} N(\mathbf{X}_i \boldsymbol{\beta}_{N\&EE}, \sigma^2_{N\&EE})\right] \\
& + \sum_{i \in O(1,0,0)} \omega_i I(G_i = N\&NN) \log\left[\pi_{i:N\&NN}\right] \\
& + \sum_{i \in O(0,0,1)} \omega_i I(G_i = C\&EE) \log\left[\pi_{i:C\&EE} N(\mathbf{X}_i \boldsymbol{\beta}_{C\&EE,0}, \sigma^2_{C\&EE,0})\right] \\
& + \sum_{i \in O(0,0,1)} \omega_i I(G_i = C\&NE) \log\left[\pi_{i:C\&NE} N(\mathbf{X}_i \boldsymbol{\beta}_{C\&NE,0}, \sigma^2_{C\&NE,0})\right] \\
& + \sum_{i \in O(0,0,1)} \omega_i I(G_i = N\&EE) \log\left[\pi_{i:N\&EE} N(\mathbf{X}_i \boldsymbol{\beta}_{N\&EE}, \sigma^2_{N\&EE})\right] \\
& + \sum_{i \in O(0,0,0)} \omega_i I(G_i = C\&EN) \log\left[\pi_{i:C\&EN}\right] \\
& + \sum_{i \in O(0,0,0)} \omega_i I(G_i = C\&NN) \log\left[\pi_{i:C\&NN}\right] \\
& + \sum_{i \in O(0,0,0)} \omega_i I(G_i = N\&NN) \log\left[\pi_{i:N\&NN}\right] \\
& + \sum_{i \in O'(1,1)} \omega_i I(G_i = C\&EE) \log\left[\pi_{i:C\&EE}\right] + \sum_{i \in O'(1,1)} \omega_i I(G_i = C\&EN) \log\left[\pi_{i:C\&EN}\right] \\
& + \sum_{i \in O'(1,1)} \omega_i I(G_i = C\&NE) \log\left[\pi_{i:C\&NE}\right] + \sum_{i \in O'(1,1)} \omega_i I(G_i = C\&NN) \log\left[\pi_{i:C\&NN}\right] \\
& + \sum_{i \in O'(0,0)} \omega_i I(G_i = N\&EE) \log\left[\pi_{i:N\&EE}\right] + \sum_{i \in O'(0,0)} \omega_i I(G_i = N\&NN) \log\left[\pi_{i:N\&NN}\right] + l(\boldsymbol{\theta})
\end{aligned}
$$

where $I(\cdot)$ is the general indicator function and $l(\boldsymbol{\theta})$ contains the parameters of the missing data process; since $\boldsymbol{\theta}$ is not of interest under the MAR assumption, we focus on the remaining parameters. Once an initial value $\boldsymbol{\xi}^{(0)}$ for the parameters vector has been chosen, the E-step of the EM algorithm computes the conditional probabilities of each stratum, given the current estimates $\boldsymbol{\xi}^{(t)}$:

- for $i \in O(1,1,1)$

$$P^{(t)}(G_i = C\&EE) = \frac{\pi^{(t)}_{i:C\&EE} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{C\&EE,1}, \sigma^{2(t)}_{C\&EE,1})}{\pi^{(t)}_{i:C\&EE} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{C\&EE,1}, \sigma^{2(t)}_{C\&EE,1}) + \pi^{(t)}_{i:C\&EN} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{C\&EN,1}, \sigma^{2(t)}_{C\&EN,1})}$$

$$P^{(t)}(G_i = C\&EN) = \frac{\pi^{(t)}_{i:C\&EN} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{C\&EN,1}, \sigma^{2(t)}_{C\&EN,1})}{\pi^{(t)}_{i:C\&EE} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{C\&EE,1}, \sigma^{2(t)}_{C\&EE,1}) + \pi^{(t)}_{i:C\&EN} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{C\&EN,1}, \sigma^{2(t)}_{C\&EN,1})}$$

$$P^{(t)}(G_i = C\&NE) = P^{(t)}(G_i = C\&NN) = P^{(t)}(G_i = N\&EE) = P^{(t)}(G_i = N\&NN) = 0$$

- for $i \in O(1,1,0)$

$$P^{(t)}(G_i = C\&NE) = \frac{\pi^{(t)}_{i:C\&NE}}{\pi^{(t)}_{i:C\&NE} + \pi^{(t)}_{i:C\&NN}}$$

$$P^{(t)}(G_i = C\&NN) = \frac{\pi^{(t)}_{i:C\&NN}}{\pi^{(t)}_{i:C\&NE} + \pi^{(t)}_{i:C\&NN}}$$

$$P^{(t)}(G_i = C\&EE) = P^{(t)}(G_i = C\&EN) = P^{(t)}(G_i = N\&EE) = P^{(t)}(G_i = N\&NN) = 0$$

- for $i \in O(1,0,1)$

$$P^{(t)}(G_i = N\&EE) = 1$$

$$P^{(t)}(G_i = C\&EE) = P^{(t)}(G_i = C\&EN) = P^{(t)}(G_i = C\&NE)$$

$$= P^{(t)}(G_i = C\&NN) = P^{(t)}(G_i = N\&NN) = 0$$

- for $i \in O(1,0,0)$

$$P^{(t)}(G_i = N\&NN) = 1$$

$$P^{(t)}(G_i = C\&EE) = P^{(t)}(G_i = C\&EN) = P^{(t)}(G_i = C\&NE)$$

$$= P^{(t)}(G_i = C\&NN) = P^{(t)}(G_i = N\&EE) = 0$$

- for $i \in O(0,0,1)$

$$P^{(t)}(G_i = C\&EE) =$$

$$= \frac{\pi^{(t)}_{i:C\&EE} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{C\&EE,0}, \sigma^{2(t)}_{C\&EE,0})}{\pi^{(t)}_{i:C\&EE} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{C\&EE,0}, \sigma^{2(t)}_{C\&EE,0}) + \pi^{(t)}_{i:C\&NE} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{C\&NE,0}, \sigma^{2(t)}_{C\&NE,0}) + \pi^{(t)}_{i:N\&EE} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{N\&EE}, \sigma^{2(t)}_{N\&EE})}$$

$$P^{(t)}(G_i = C\&NE) =$$

$$= \frac{\pi^{(t)}_{i:C\&NE} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{C\&NE,0}, \sigma^{2(t)}_{C\&NE,0})}{\pi^{(t)}_{i:C\&EE} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{C\&EE,0}, \sigma^{2(t)}_{C\&EE,0}) + \pi^{(t)}_{i:C\&NE} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{C\&NE,0}, \sigma^{2(t)}_{C\&NE,0}) + \pi^{(t)}_{i:N\&EE} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{N\&EE}, \sigma^{2(t)}_{N\&EE})}$$

$$P^{(t)}(G_i = N\&EE) =$$

$$= \frac{\pi^{(t)}_{i:N\&EE} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{N\&EE}, \sigma^{2(t)}_{N\&EE})}{\pi^{(t)}_{i:C\&EE} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{C\&EE,0}, \sigma^{2(t)}_{C\&EE,0}) + \pi^{(t)}_{i:C\&NE} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{C\&NE,0}, \sigma^{2(t)}_{C\&NE,0}) + \pi^{(t)}_{i:N\&EE} N(\mathbf{X}_i \boldsymbol{\beta}^{(t)}_{N\&EE}, \sigma^{2(t)}_{N\&EE})}$$

$$P^{(t)}(G_i = C\&EN) = P^{(t)}(G_i = C\&NN) = P^{(t)}(G_i = N\&NN) = 0$$

- for $i \in O(0,0,0)$

$$P^{(t)}(G_i = C\&EN) = \frac{\pi^{(t)}_{i:C\&EN}}{\pi^{(t)}_{i:C\&EN} + \pi^{(t)}_{i:C\&NN} + \pi^{(t)}_{i:N\&NN}}$$

$$P^{(t)}(G_i = C\&NN) = \frac{\pi^{(t)}_{i:C\&NN}}{\pi^{(t)}_{i:C\&EN} + \pi^{(t)}_{i:C\&NN} + \pi^{(t)}_{i:N\&NN}}$$

$$P^{(t)}(G_i = N\&NN) = \frac{\pi^{(t)}_{i:N\&NN}}{\pi^{(t)}_{i:C\&EN} + \pi^{(t)}_{i:C\&NN} + \pi^{(t)}_{i:N\&NN}}$$

$$P^{(t)}(G_i = C\&EE) = P^{(t)}(G_i = C\&NE) = P^{(t)}(G_i = N\&EE) = 0$$

- for $i \in O'(1,1)$

$$P^{(t)}(G_i = C\&EE) = \frac{\pi^{(t)}_{i:C\&EE}}{\pi^{(t)}_{i:C\&EE} + \pi^{(t)}_{i:C\&EN} + \pi^{(t)}_{i:C\&NE} + \pi^{(t)}_{i:C\&NN}}$$

$$P^{(t)}(G_i = C\&EN) = \frac{\pi^{(t)}_{i:C\&EN}}{\pi^{(t)}_{i:C\&EE} + \pi^{(t)}_{i:C\&EN} + \pi^{(t)}_{i:C\&NE} + \pi^{(t)}_{i:C\&NN}}$$

$$P^{(t)}(G_i = C\&NE) = \frac{\pi^{(t)}_{i:C\&NE}}{\pi^{(t)}_{i:C\&EE} + \pi^{(t)}_{i:C\&EN} + \pi^{(t)}_{i:C\&NE} + \pi^{(t)}_{i:C\&NN}}$$

$$P^{(t)}(G_i = C\&NN) = \frac{\pi^{(t)}_{i:C\&NN}}{\pi^{(t)}_{i:C\&EE} + \pi^{(t)}_{i:C\&EN} + \pi^{(t)}_{i:C\&NE} + \pi^{(t)}_{i:C\&NN}}$$

$$P^{(t)}(G_i = N\&EE) = P^{(t)}(G_i = N\&NN) = 0$$

- for $i \in O'(1,0)$

$$P^{(t)}(G_i = N\&EE) = \frac{\pi^{(t)}_{i:N\&EE}}{\pi^{(t)}_{i:N\&EE} + \pi^{(t)}_{i:N\&NN}}$$

$$P^{(t)}(G_i = N\&NN) = \frac{\pi^{(t)}_{i:N\&NN}}{\pi^{(t)}_{i:N\&EE} + \pi^{(t)}_{i:N\&NN}}$$

$$P^{(t)}(G_i = C\&EE) = P^{(t)}(G_i = C\&EN) = P^{(t)}(G_i = C\&NE) = P^{(t)}(G_i = C\&NN) = 0$$

The above conditional probabilities are the estimates of the unknown indicator functions $I(\cdot)$ in the complete-data log-likelihood function; replacing the $I(G_i = g)$ with the $P^{(t)}(G_i = g)$ we obtain the expected log-likelihood $l_E(\boldsymbol{\xi}|\mathbf{Z}, \mathbf{D}, \mathbf{R}, \mathbf{S}, \mathbf{W}, \mathbf{X})$. The M-step consists in optimizing $l_E(\cdot)$ with respect to the parameters vector $\boldsymbol{\xi}$, leading to a new estimate $\boldsymbol{\xi}^{(t+1)}$: to update $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$, a standard routine for linear regression models can be used; a procedure for multinomial logistic models is needed in estimating $\boldsymbol{\alpha}$, given the current posterior probabilities. As showed in Dempster et al. (1977), iterating this process monotonically increases the likelihood function, or at least leaves it unchanged; the algorithm runs until a stopping criterion has been satisfied.

As in any finite mixture of Normal distributions, the log-likelihood function is unbounded and the EM algorithm may fall in a spurious maximum: in this case, the procedure must be restarted with new starting values. In addition, there often exist other solutions which may be regarded as spurious, lying very close to the edge of the parameter space: this happens when a component with very small variance is fitted; usually, this component density constitutes a cluster containing a few data points, very close together or almost lying in the same subspace. Such estimate tends to "interpolate" a local pattern and provides a bad fit for the remaining observations; as a consequence, the fitted model is not of practical use in inference.

Another complication is that the log-likelihood function presents an unknown number of local solutions: the best one – that is, the one with the higher log-likelihood value – is usually chosen as the MLE. For this reason, a great number of different starting values for the EM algorithm should be used.

For our work, we used the statistical package R. A genetic algorithm was used in the search of the absolute maximum (among local maxima) of the log-likelihood function (Frumento, 2009); the EM algorithm stopped when the maximum absolute change in the parameters vector between two consecutive iterations was smaller than 0.0001. For each model, the asymptotic covariance matrix was obtained by analytical evaluation of the Hessian of the log-likelihood function. The causal effects on employment and wages have been computed and approximate standard errors have been obtained by means of the Delta method; also in this case, analytical derivatives were used.

| | Treatment | | | Control | | | Difference | |
|---|---|---|---|---|---|---|---|---|
| Variable | Prop. non-miss. | Mean | Std. Dev. | Prop. non-miss. | Mean | Std. Dev. | Diff. | Std. Err. |
| Female | 0.96 | 0.41 | 0.49 | 0.95 | 0.40 | 0.49 | 0.00 | 0.01 |
| Age at baseline | 0.96 | 18.85 | 2.18 | 0.95 | 18.82 | 2.15 | 0.03 | 0.04 |
| White, non-Hispanic | 1.00 | 0.30 | 0.46 | 1.00 | 0.30 | 0.46 | 0.00 | 0.01 |
| Black, non-Hispanic | 1.00 | 0.46 | 0.50 | 1.00 | 0.45 | 0.50 | 0.01 | 0.01 |
| Hispanic | 1.00 | 0.17 | 0.37 | 1.00 | 0.17 | 0.38 | 0.00 | 0.01 |
| Other race | 1.00 | 0.07 | 0.26 | 1.00 | 0.07 | 0.26 | 0.00 | 0.00 |
| Never married | 0.94 | 0.91 | 0.28 | 0.92 | 0.91 | 0.28 | 0.00 | 0.00 |
| Married | 0.94 | 0.02 | 0.14 | 0.92 | 0.02 | 0.14 | 0.00 | 0.00 |
| Living together | 0.94 | 0.04 | 0.20 | 0.92 | 0.04 | 0.20 | 0.00 | 0.00 |
| Separated | 0.94 | 0.02 | 0.16 | 0.92 | 0.02 | 0.14 | 0.00 | 0.00 |
| With a partner | 0.94 | 0.06 | 0.24 | 0.92 | 0.06 | 0.24 | 0.00 | 0.00 |
| Has children | 0.99 | 0.17 | 0.38 | 0.99 | 0.17 | 0.38 | 0.00 | 0.01 |
| Number of children | 0.99 | 0.24 | 0.61 | 0.98 | 0.23 | 0.59 | 0.01 | 0.01 |
| Education | 0.93 | 10.06 | 1.52 | 0.92 | 10.07 | 1.52 | −0.01 | 0.03 |
| Mother's education | 0.76 | 11.52 | 2.56 | 0.74 | 11.53 | 2.62 | −0.01 | 0.05 |
| Father's education | 0.58 | 11.47 | 2.87 | 0.56 | 11.57 | 2.84 | −0.10 | 0.06 |
| Ever arrested | 0.94 | 0.26 | 0.44 | 0.92 | 0.26 | 0.44 | 0.00 | 0.01 |
| Household Inc. < 3000 | 0.59 | 0.26 | 0.44 | 0.59 | 0.25 | 0.43 | 0.01 | 0.01 |
| 3000 − 6000 | 0.59 | 0.20 | 0.40 | 0.59 | 0.21 | 0.41 | −0.01 | 0.01 |
| 6000 − 9000 | 0.59 | 0.11 | 0.32 | 0.59 | 0.11 | 0.31 | 0.00 | 0.01 |
| 9000 − 18000 | 0.59 | 0.25 | 0.43 | 0.59 | 0.25 | 0.43 | 0.00 | 0.01 |
| > 18000 | 0.59 | 0.19 | 0.39 | 0.59 | 0.19 | 0.39 | 0.00 | 0.01 |
| Personal Inc. < 3000 | 0.87 | 0.79 | 0.41 | 0.86 | 0.79 | 0.40 | 0.00 | 0.01 |
| 3000 − 6000 | 0.87 | 0.13 | 0.33 | 0.86 | 0.13 | 0.33 | 0.00 | 0.01 |
| 6000 − 9000 | 0.87 | 0.05 | 0.22 | 0.86 | 0.04 | 0.20 | 0.01 | $0.00^{(*)}$ |
| > 9000 | 0.87 | 0.03 | 0.18 | 0.86 | 0.03 | 0.18 | 0.00 | 0.00 |
| At baseline: | | | | | | | | |
| Have Job | 0.92 | 0.21 | 0.41 | 0.91 | 0.21 | 0.41 | 0.00 | 0.01 |
| Had Job, prev. yr. | 0.94 | 0.65 | 0.48 | 0.92 | 0.64 | 0.48 | 0.01 | 0.01 |
| Months empl., prev. yr. | 0.89 | 3.77 | 4.26 | 0.88 | 3.75 | 4.30 | 0.01 | 0.07 |
| Earnings, prev. yr. | 0.87 | 2859.89 | 4210.62 | 0.86 | 2868.57 | 4350.31 | 8.69 | 74.16 |
| N | 9409 | | | 5977 | | | | |

Table 1: Summary statistics of the pre-treatment covariates; in the last column, $(*)$ denotes that the difference between the treatment and the control group is statistically significant at 0.05 level (all statistics have been computed before the imputation). All computations use design weights.

| Variable | Treatment | | | Control | | | Difference | |
|---|---|---|---|---|---|---|---|---|
| | Prop. non-miss. | Mean | Std. Dev. | Prop. non-miss. | Mean | Std. Dev. | Diff. | Std. Err. |
| Week 45 | | | | | | | | |
| Employed | 0.88 | 0.35 | 0.48 | 0.85 | 0.43 | 0.49 | −0.08 | 0.01$^{(*)}$ |
| Weekly earnings | 0.88 | 89.19 | 154.32 | 0.85 | 103.39 | 150.82 | −14.19 | 2.64$^{(*)}$ |
| Weekly hours | 0.88 | 14.49 | 21.83 | 0.85 | 17.49 | 22.52 | −3.01 | 0.38$^{(*)}$ |
| Week 135 | | | | | | | | |
| Employed | 0.76 | 0.54 | 0.49 | 0.76 | 0.52 | 0.50 | 0.03 | 0.01$^{(*)}$ |
| Weekly earnings | 0.76 | 182.16 | 217.8 | 0.76 | 164.24 | 201.59 | 17.92 | 3.88$^{(*)}$ |
| Weekly hours | 0.76 | 23.92 | 24.31 | 0.76 | 22.51 | 24.07 | 1.41 | 0.45$^{(*)}$ |
| Week 208 | | | | | | | | |
| Employed | 0.67 | 0.60 | 0.49 | 0.68 | 0.56 | 0.50 | 0.04 | 0.01$^{(*)}$ |
| Weekly earnings | 0.67 | 220.15 | 240.66 | 0.68 | 194.88 | 219.51 | 25.27 | 4.52$^{(*)}$ |
| Weekly hours | 0.67 | 26.54 | 24.12 | 0.68 | 24.41 | 23.86 | 2.13 | 0.47$^{(*)}$ |

Table 2: Summary statistics of the outcome variables; in the last column, $(*)$ denotes that the difference between the treatment and the control group is statistically significant at 0.05 level. All computations use design weights.

| week | $\pi_{C\&EE}$ | $\pi_{C\&EN}$ | $\pi_{C\&NE}$ | $\pi_{C\&NN}$ | $\pi_{N\&EE}$ | $\pi_{N\&NN}$ | $\Delta^{(DS)}$ | $\Delta^{(DW)}$ | BIC | AIC |
|---|---|---|---|---|---|---|---|---|---|---|
| 45 | 0.1643 | 0.0725 | 0.1547 | 0.3223 | 0.1205 | 0.1656 | −0.0822 | 0.2757 | 28553.4 | 26864.8 |
| | (0.0059) | (0.0054) | (0.0072) | (0.0074) | (0.0041) | (0.0048) | (0.0072) | (0.0523) | | |
| 135 | 0.3328 | 0.0624 | 0.0137 | 0.2894 | 0.1636 | 0.1381 | 0.0487 | 0.2099 | 28292.9 | 26604.4 |
| | (0.0057) | (0.0049) | (0.0016) | (0.0051) | (0.0045) | (0.0045) | (0.0050) | (0.0576) | | |
| 208 | 0.3789 | 0.0620 | 0.0135 | 0.2549 | 0.1713 | 0.1194 | 0.0485 | 0.3374 | 25821.2 | 24132.6 |
| | (0.0060) | (0.0049) | (0.0015) | (0.0052) | (0.0051) | (0.0050) | (0.0051) | (0.0668) | | |

Table 3: Maximum likelihood estimates – adjusted for covariates and without monotonicity of truncation – of the average treatment effects on employment ($\Delta^{DS}$) and wages ($\Delta^{DW}$) for week 45, 135 and 208 (asymptotic standard errors between brackets). For each week, we provide the estimated proportion of each principal stratum; the BIC and AIC are returned for a comparison with results in Table 4.

| week | $\pi_{C\&EE}$ | $\pi_{C\&EN}$ | $\pi_{C\&NN}$ | $\pi_{N\&EE}$ | $\pi_{N\&NN}$ | $\Delta^{(DS)}$ | $\Delta^{(DW)}$ | BIC | AIC |
|---|---|---|---|---|---|---|---|---|---|
| 45 | 0.2468 | 0.0294 | 0.4382 | 0.1333 | 0.1523 | 0.0294 | 0.1693 | 28505.7 | 27191.5 |
| | (0.0045) | (0.0025) | (0.0048) | (0.0041) | (0.0043) | (0.0025) | (0.0426) | | |
| 135 | 0.3389 | 0.0592 | 0.3001 | 0.1652 | 0.1364 | 0.0592 | 0.2070 | 28127.5 | 26813.3 |
| | (0.0057) | (0.0047) | (0.0049) | (0.0046) | (0.0045) | (0.0047) | (0.0571) | | |
| 208 | 0.3865 | 0.0559 | 0.2672 | 0.1724 | 0.1181 | 0.0559 | 0.3217 | 25696.1 | 24381.9 |
| | (0.0059) | (0.0047) | (0.0050) | (0.0052) | (0.0049) | (0.0047) | (0.0669) | | |

Table 4: Maximum likelihood estimates – adjusted for covariates and assuming monotonicity of truncation – of the average treatment effects on employment ($\Delta^{DS}$) and wages ($\Delta^{DW}$) for week 45, 135 and 208 (asymptotic standard errors between brackets). For each week, we provide the estimated proportion of each principal stratum; the BIC and AIC are returned for a comparison with results in Table 3.