The Analysis of Randomized Experiments with Missing Data

Guido W. Imbens and William A. Pizer

May 2000 • Discussion Paper 00–19

Resources for the Future
1616 P Street, NW
Washington, D.C. 20036
Telephone: 202–328–5000
Fax: 202–939–3460
Internet: http://www.rff.org

Discussion papers are research materials circulated by their authors for purposes of information and discussion. They have not undergone formal peer review or the editorial treatment accorded RFF books and other publications.

The Analysis of Randomized Experiments with Missing Data

Guido W. Imbens
UCLA

William A. Pizer
Resources for the Future

August 1999

Abstract

The otherwise straightforward analysis of randomized experiments is often complicated by the presence of missing data. In such situations it is necessary to make assumptions about the dependence of the selection mechanism on treatment, response, and covariates. The widely used approach of assuming that the data is missing at random conditional on treatment and other fully observed covariates is shown to be inadequate to describe data from a randomized experiment when partially observed covariates are also present.

This paper presents an alternative to the missing at random model (MAR) which is both consistent with the data and preserves the appeal of MAR. In particular, the proposed family of models minimize the discrepancy with MAR while explaining observed deviations. We apply this approach to data from the Restart job training program in the United Kingdom as well as an artificial data set. Evaluation of the Restart program is not affected by the assumption of MAR; both approaches suggest that the program increased the chances of exiting unemployment by around 9% within six months. However, analysis of the artificial data demonstrates that assuming MAR can easily lead to erroneous conclusions.

Guido Imbens is a professor at the University of California at Los Angeles, and William A. Pizer is a fellow in the Quality of the Environment Division at Resources for the Future. The authors would like to thank Don Rubin for comments. Imbens would also like to thank the National Science Foundation for financial support under grant SBR 9511718 and the Alfred L. Sloan Foundation for a research fellowship.

1. Introduction

Randomized experiments offer many benefits to the researcher. The randomization of treatment assignment ensures that treatment and control groups are comparable, and therefore causal inferences regarding the average effect of the treatment of interest can be drawn without additional assumptions. Specifically, randomization avoids the need for modeling the outcome distributions because it ensures that average causal effects can be estimated by the difference between average treatment outcomes and average control outcomes.

These benefits, however, require that we have complete data on treatment and response for all units. When follow-up surveys are used to collect data from an otherwise randomized experiment, non-response frequently leads to missing data. For example, in the randomized job-training experiment Restart discussed in this paper and analyzed previously by Dolton and O'Neill (1996a, 1996b), the response variable and most covariates are missing for almost half the sample due to nonresponse in a subsequent survey. Similarly, Imbens, Rubin, and Sacerdote (1999) survey lottery winners to study the effect of an infusion of money on labor supply, but although the lottery itself is random, response to the study's survey is not. In such cases where covariates are missing for some units, one cannot avoid making assumptions regarding the dependence of the missing data mechanism on both treatment assignment and the values of missing variables.

What distinguishes this problem from most missing data problems (e.g., Gourieroux and Monfort 1981), is that there is indirect information relevant for the missing data process, namely random assignment. In this paper we develop a framework for estimating average treatment effects in randomized experiments where information on outcome and covariates or pretreatment variables is missing for some units, but data on treatment assignment (and possibly other covariates) is always available. In particular, we extend the standard approach to modeling missing data in order to explain the observed covariate distribution, given an initially random assignment of treatment.

The standard approach (Rubin 1976; Little and Rubin 1987) assumes that, conditional on treatment and any fully-observed covariates, the data are *missing at random* (MAR) or, alternatively, the missing data process is *ignorable.* With missing covariates, however, such assumptions are not necessarily adequate to describe the data. The reason is that the two assumptions, (i) random assignment (RA) of treatment, and (ii) missing at random, have implications that can be in conflict. Specifically, if we observe that among complete-data Observations, those assigned treatment have different covariate distributions than those assigned control, we can deduce that the missing data are not missing at random.

Motivated by this conflict, we develop alternative models for the analysis of data from randomized experiments with missing pretreatment variables and outcomes. The families of models we develop have two key properties. First, they are identified, meaning that for each family in large samples there will always be a unique member of the family consistent with both the distribution of the observed data as well as with the restrictions implied by random assignment. Second, the estimated model can be interpreted as the model consistent with the restriction implied by RA that is as close to MAR as possible. In other words, the imputed distribution for the missing pretreatment variables will be as close to the distribution of the observed pretreatment variables as is consistent with random assignment.

In the next section we set up the basic problem. In Section 3 we abstract from the presence of the outcome variable and focus solely on imputing a single binary pretreatment variable. Properly estimating the marginal distribution of the covariate will be central to our approach to the general problem. In Section 4 we discuss an alternative derivation of this solution which follows previous analyses of estimating probabilities in a two-way classification with known marginals (e.g., Little and Wu 1991). A key difference is that rather than knowing the marginal distribution of some variables as in Little and Wu, we know that some variables are independent. Section 5 extends the basic model to include the response variable and more general pretreatment variables. Section 6 contains an illustration of the techniques using data from an experimental evaluation of a job training program in the United Kingdom.[1]

[1] We are grateful to P. Dolton for making this data available to us.

Section 7 concludes.

2. Randomized Experiments with Missing Data

Consider a randomized experiment with N units, indexed by $i = 1, \ldots, N$, and with two treatment conditions denoted by $T_i = 1$ (treatment) and $T_i = 0$ (control). For each unit i there are two outcomes, $Y_i(0)$ for the response with control, and $Y_i(1)$ for the response with treatment. The actual, or observed, response is denoted by $Y_i = Y_i(T_i) = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0)$. For each unit i there is a vector of pretreatment variables X_i. The missing data indicator is D_i. We observe for each unit in the population the quintuple $(D_i, T_i, Y_i \cdot D_i, X_i \cdot D_i)$. In other words, for all units we observe T_i and D_i, but only for units with $D_i = 1$ do we observe Y_i and X_i.

Completely random assignment (RA) implies that the population covariate X_i is independent of the treatment indicator T_i or $X_i \perp T_i$. Suppose X_i is missing at random (MAR) so that the selection indicator D_i is independent of X_i after conditioning on treatment T_i, that is, $D_i \perp X_i | T_i$. This implies that the conditional distribution of X_i given T_i is the same as the conditional distribution of X_i given T_i and $D_i = 1$, or $X_i | (T_i, D_i = 1) \sim X_i | T_i$. Combining the implications of RA and MAR therefore implies that $X_i \perp T_i | D_i = 1$. This is a testable independence restriction because for all units with $D_i = 1$ we observe X_i and T_i. That is, the joint hypothesis of RA and MAR has testable implications which, if rejected, force one to go beyond models characterized by MAR if RA is maintained.

In Table 1 the data from the actual Restart program as well as an artificial data set are presented. We focus on the subgroup of males aged 20 to 50, isolated from the original Restart data involving unemployed individuals in the United Kingdom. In principle unemployed individuals are obliged, after a certain period of unemployment, to have a discussion with officials from the local unemployment office about job search strategies and training opportunities. This is the treatment we wish to evaluate. In the Restart study a random sample from this population was exempt from this obligation; this group serves as the control. The

Table 1: Restart Data and Artificial Data

				number of individuals	
T_i	D_i	X_i	Y_i	Restart	Artificial
0	0	–	–	133	1330
0	1	0	0	57	1310
0	1	0	1	7	160
0	1	1	0	56	20
0	1	1	1	30	10
1	0	–	–	1814	18140
1	1	0	0	755	2180
1	1	0	1	324	940
1	1	1	0	630	11130
1	1	1	1	371	470
	Total			4177	41770

$T_i = 1$ for those obliged to receive the interview; $D_i = 1$ for those who responded to the follow-up survey; $X_i = 1$ if the individual had a driver's license; $Y_i = 1$ if the individual successfully exited unemployment after six months.

treatment indicator is equal to one for those individuals obliged to receive the interview and zero for those exempted. The outcome of interest is whether or not individuals successfully exited unemployment within six months of being randomized to either receive or not receive the discussion with a local employment official. The single pretreatment variable is whether or not the individual has a driving license. The outcome and pretreatment variables are only observed if the individual responded to a survey conducted six months after the interview. For details of the data set and the training program see Dolton and O'Neill (1996a, 1996b).

In Table 2 we calculate a number of sample proportions. First, the marginal distribution of assignment and the missing data indicator, $q_{td} = Pr(T_i = t, D_i = d)$. Second, we calculate the mean of X_i given T_i and D_i, $p_{td} = Pr(X_i = 1|T_i = t, D_i = d)$. For $d = 0$ this mean cannot be calculated from the sample, and this is indicated in the table by giving the range of values consistent with both the data and random assignment, ignoring sampling variation. This is similar to the bounds calculated by Horowitz and Manski (1995) and Manski (1995).[2] Third, we calculate the joint distribution of T_i and X_i given $D_i = 1$, $\pi_{tx|1} = Pr(T_i = t, X_i = x|D_i = 1)$. Fourth, we calculate the joint distribution of T_i and X_i without conditioning on $D_i = 1$, $\pi_{tx} = Pr(T_i = t, X_i = x)$. Again we cannot calculate these probabilities exactly from the data but present ranges consistent with both the raw data and RA, again ignoring sampling variation.

When $p_{11} \neq p_{01}$, the assumptions of random assignment and missing at random conflict. RA implies $T_i \perp X_i$. MAR implies $D_i \perp X_i|T_i$. Together they imply that $T_i \perp X_i|D_i = 1$ and thus $Pr(X_i = 1|T_i = 1, D_i = 1) = Pr(X_i = 1|T_i = 0, D_i = 1)$, or $p_{11} = p_{01}$. In the Restart data this is contradicted by the unequal values of $p_{11} = 0.481$ and $p_{01} = 0.573$. In the artificial data, the contradiction is more pronounced with $p_{11} = 0.850$ and $p_{01} = 0.020$. We now examine more flexible alternatives to the MAR model.

3. A Family of Nonignorable Missing Data Models

In this section we look at the problem where we always observe the random assignment T_i, the missing data indicator D_i, but only observe a single binary covariate X_i if $D_i = 1$. This ignores both the outcome variable and other covariates, allowing us to focus attention on the missing data mechanism in the simplest possible case of interest. We also assume the sample is large so that we can ignore sampling variation. In terms of the notation established in the preceeding section, this implies we know the population values $Pr(T_i = t, D_i = d)$, denoted by q^*_{td}, for $t, d = 0, 1$ and the population values of $Pr(X_i = 1|T_i = t, D_i = 1)$, denoted by p^*_{t1} for $t = 0, 1$. We do not know the values of p_{t0} for $t = 0, 1$ because we never observe X_i if $D_i = 0$. MAR implies that $p_{t0} = p^*_{t1}$ for $t = 0, 1$.[3]

[2] Note that these ranges are more restrictive than simply $p_{t0} \in (0, 1)$. This restriction arises because RA implies $Pr(X_i = 1|T_i = 0)$ must equal $Pr(X_i = 1|T_i = 1)$ so that choosing either p_{00} or p_{10} subsequently identifies the other. Extreme values of one parameter can result in $Pr(X_i = 1|T_i = 0) \gtrless Pr(X_i = 1|T_i = 1)$ for all values (between zero and one) of the other and is therefore inconsistent with RA.

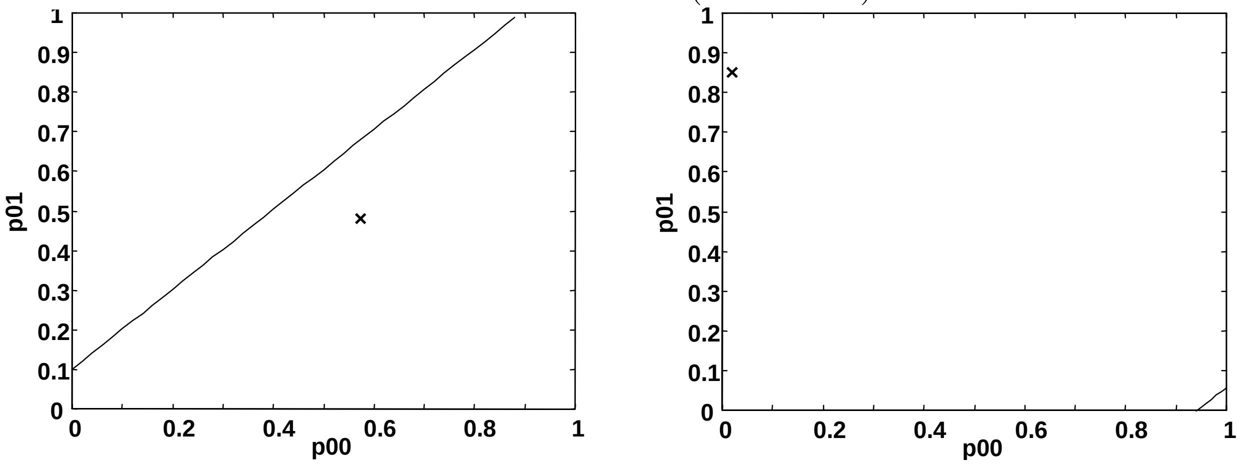
[3] Here and throughout this section we use stars to denote population values of directly estimable parameters; i.e., the parameters whose values we could deduce from an infinitely large sample (q^*_{td} for $t, d = 0, 1$ and p^*_{td} for $t = 0, 1$ and $d = 1$).

Table 2: Sample Proportions and Ranges for Restart and Artificial Data Sets

	t	d	x	Restart Data	Artificial Data
$q_{td} = Pr(T_i = t, D_i = d)$	0	0		0.032	0.032
	0	1		0.036	0.036
	1	0		0.434	0.434
	1	1		0.498	0.498
$p_{td} = Pr(X_i = 1 \| T_i = t, D_i = d)$	0	0		(0.000,0.892)	(0.944,1.000)
	0	1		0.573	0.020
	1	0		(0.100,1.000)	(0.000,0.057)
	1	1		0.481	0.850
$\pi_{tx\|1} = Pr(T_i = t, X_i = x \| D_i = 1)$	0		0	0.029	0.066
	0		1	0.039	0.001
	1		0	0.484	0.140
	1		1	0.449	0.793
$\pi_{tx} = Pr(T_i = t, X_i = x)$	0		0	(0.019,0.047)	(0.035,0.037)
	0		1	(0.021,0.049)	(0.031,0.033)
	1		0	(0.258,0.649)	(0.484,0.509)
	1		1	(0.283,0.674)	(0.448,0.423)

For an explanation of variables see Table 1.

FIGURE 1: GRAPHICAL IMPLICATIONS OF MISSING AT RANDOM (×) AND RANDOM ASSIGNMENT (SOLID LINE)



(a) Restart data

(b) Artificial data

The randomization of the treatment assignment T_i implies a restriction on the joint distribution of (T_i, D_i, X_i), namely that T_i is independent of X_i, or $T_i \perp X_i$, which can be written as a single restriction on the probabilities:

$$Pr(X_i = 1|T_i = 1) = Pr(X_i = 1|T_i = 0).$$

Given the population values of the directly estimable parameters, this restriction can be written as a single linear restriction on the remaining parameters p_{00} and p_{10}:

$$\frac{p^*_{11} \cdot q^*_{11} + p_{10} \cdot q^*_{10}}{q^*_{11} + q^*_{10}} = \frac{p^*_{01} \cdot q^*_{01} + p_{00} \cdot q^*_{00}}{q^*_{01} + q^*_{00}}. \tag{1}$$

This restriction is not necessarily satisfied when we substitute the MAR values for the inestimable parameters, p_{00} and p_{10}, namely $p_{00} = p^*_{01}$ and $p_{10} = p^*_{11}$. Figures 1a and 1b illustrate this for the Restart and artificial data sets by plotting in (p_{00}, p_{10}) space the set of values satisfying the restriction implied by RA and shown in Equation (1) (indicated by the solid line) alongside the values implied by MAR (indicated by the "×").

The next step is to develop a family of selection models that allows for nonignorably missing data. We wish to develop families of models satisfying two conditions:

i. coupled with RA, the selection model should be exactly identified by the observed data; and

ii. the selection model should encompass MAR as a special case.

Condition (i) indicates our interest in selection models leading to unique solutions that are always consistent with both random assignment and all the observed data. Such a solution includes a complete set of parameter values for the selection model and the data distribution (e.g., including those parameters which cannot be observed directly, p_{00} and p_{10}). Condition (ii) implies that when MAR is in fact consistent with RA, the unique solution identified according to Condition (i) will be a MAR missing data model.

Note that the aim is not to find the true values (p_{00}, p_{10}) that generated the data. Such a search would be futile because the data do not contain enough information to uniquely determine (p_{00}, p_{10}). Rather, we wish to develop a rule for picking a point in the intersection of the set of parameter values consistent with the distribution of the observed variables and the set of values consistent with independence of T_i and X_i (RA). In the simple context studied in this section, this intersection is the solid line in Figures 1a and 1b, and the model choice amounts to a rule for choosing a point along that line segment.[4]

We start by considering general missing data models. With both T_i and X_i binary, the general form for these models can be captured by a four parameter specification with

$$Pr(D_i = 1|T_i = t, X_i = x) = g(\alpha_0 + \alpha_1 \cdot t + \alpha_2 \cdot x + \alpha_3 \cdot t \cdot x) \qquad (2)$$

for a known, continuous, and increasing function $g(\cdot)$ satisfying $\lim_{a\to-\infty} g(a) = 0$ and $\lim_{a\to\infty} g(a) = 1$ (assuming all four probabilities are between zero and one). Within this family of models the members with MAR are characterized by $\alpha_2 = \alpha_3 = 0$. The parameters

[4] In some cases it may be of interest to find the entire set of parameter values consistent with the data and RA; that is, the solid line in Figures 1a and 1b. Horowitz and Manski (1995) and Manski (1995) advocate such a strategy. In the current context with all variables binary, this set is straightforward to identify. However, in more complex situations with multiple, continuous-valued covariates, it could be difficult to find the entire set of parameter values consistent with both the data and RA. In such cases it may be necessary to settle for identifying one element of that set.

of the missing data model, α_0, α_1, α_2 and α_3, and the remaining parameters p_{00} and p_{10} of the conditional distribution of X_i given $D_i = 0$ are related by the following definition of the selection probability:

$$\begin{aligned} Pr(D_i = 1|T_i = t, X_i = x) &= g(\alpha_0 + \alpha_1 \cdot t + \alpha_2 \cdot x + \alpha_3 \cdot t \cdot x) \\ &= \frac{q^*_{t1} \cdot {p^*_{t1}}^x \cdot (1 - p^*_{t1})^{1-x}}{q^*_{t1} \cdot {p^*_{t1}}^x \cdot (1 - p^*_{t1})^{1-x} + q^*_{t0} \cdot p^x_{t0} \cdot (1 - p_{t0})^{1-x}}, \end{aligned} \tag{3}$$

for $t, x = 0, 1$ and the restriction implied by random assignment, Equation (1).

This general model, for a given choice of $g(\cdot)$, is not identifiable. For every pair of values $(p_{00}, p_{10}) \in (0, 1) \times (0, 1)$ consistent with restriction (1) there is a unique quadruple of values of $(\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ such that the other four restrictions implied by (3) are satisfied. We therefore need at least one additional restriction on the four parameters α_0, α_1, α_2 and α_3 to be able to identify the remaining parameters.

We propose the restriction $\alpha_2 = -\alpha_3 \cdot E[T_i]$. The missing data model incorporating this restriction can then be written as

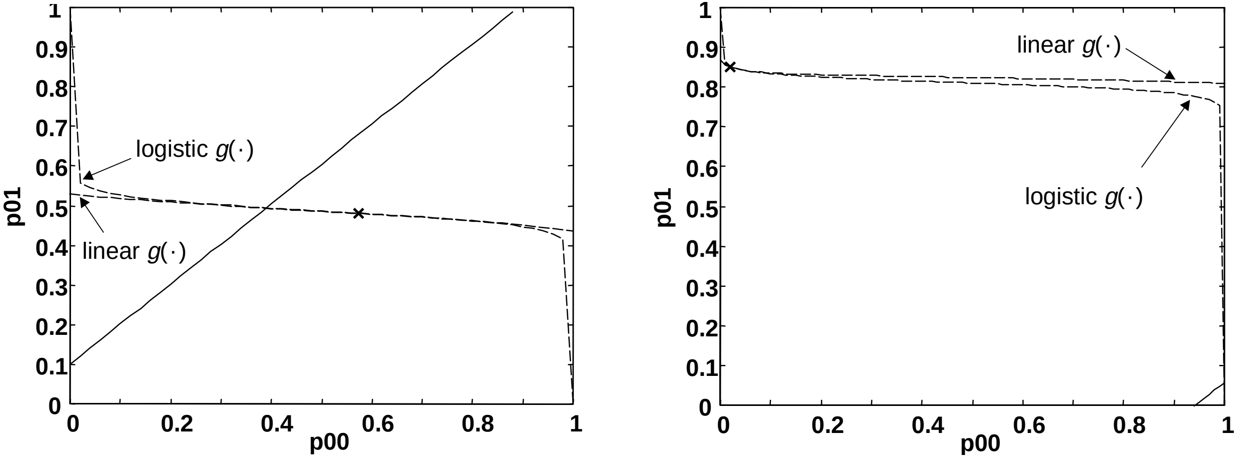
$$Pr(D_i = 1|X_i = x, T_i = t) = g(\alpha_0 + \alpha_1 \cdot t + \alpha_2 \cdot x \cdot (t - E[T_i])). \tag{4}$$

For the two data sets, Figures 2a and 2b illustrate that there is a unique pair (p_{00}, p_{10}) consistent with both the distribution of the observed variables as well as with the restriction implied by random assignment. The dashed lines for the logistic and linear $g(\cdot)$ are the points consistent with the observed values of $\{p^*_{td}\}_{t,d=1}$ and $\{q^*_{td}\}_{t,d}$ and the particular missing data model.[5]

It is clear that this selection model, for any choice of $g(\cdot)$, encompasses all possible MAR models. Specifically, setting $\alpha_2 = 0$ generates the most general MAR model with selection only depending on treatment T_i. For the first condition above to be satisfied, however, it has

[5] Formally the linear probability model does not satisfy the asymptotic conditions on $g(\cdot)$, but there exist $g(\cdot)$ functions satisfying the restrictions which are arbitrarily close to the linear probability model. In particular, such functions would trace the linear probability model up to the boundary of $(p_{00}, p_{10}) \in [0, 1]^2$, then trace the boundary of the space to the corners $(0, 1)$ and $(1, 0)$. For the artificial data set, this is where the solution in fact lies. As with the Restart data, the logistic and linear models yield nearly identical results.

Figure 2: Random assignment (solid) versus selection model with $\alpha_2 = -\mathrm{E}[T_i]\alpha_3$ (dashed)



(a) Restart data (b) artificial data

to be determined, first, whether there always is at least one member of this family consistent with both the data and the restriction implied by RA, and second, whether this member is unique. The following lemma states that this is indeed the case.

Lemma 1 *For all strictly increasing and continuous* $g(\cdot)$ *satisfying* $\lim_{a\to-\infty} g(a) = 0$ *and* $\lim_{a\to\infty} g(a) = 1$, *and for all* $p^*_{tx} \in (0,1)$ *and* $q^*_{tx} \in (0,1)$, $\sum q^*_{tx} = 1$, *satisfying*

$$\frac{p^*_{11}\cdot q^*_{11} + p^*_{10}\cdot q^*_{10}}{q^*_{11}+q^*_{10}} = \frac{p^*_{01}\cdot q^*_{01} + p^*_{00}\cdot q^*_{00}}{q^*_{01}+q^*_{00}}, \tag{5}$$

there is a unique solution $(\alpha_0, \alpha_1, \alpha_2, p_{00}, p_{10})$ *such that*

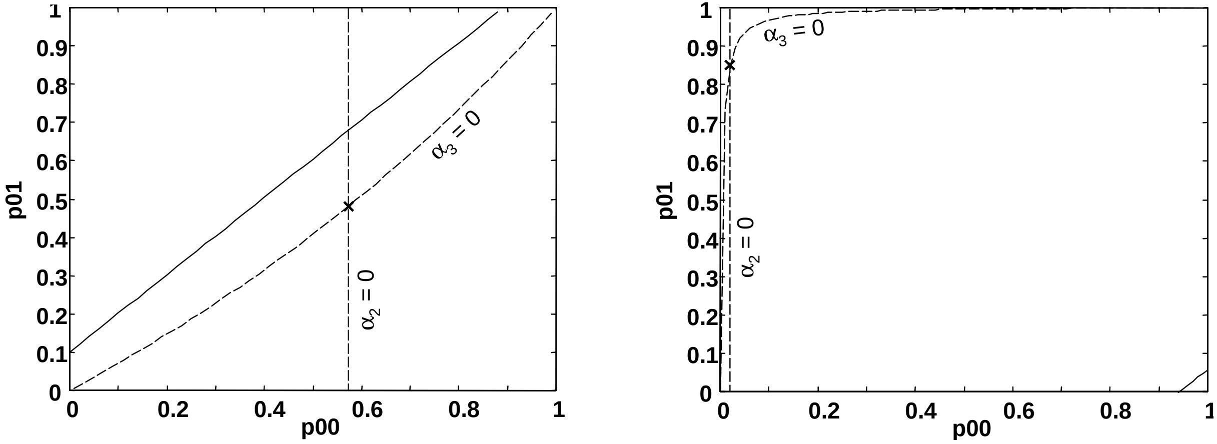
$$\frac{p^*_{11}\cdot q^*_{11} + p_{10}\cdot q^*_{10}}{q^*_{11}+q^*_{10}} = \frac{p^*_{01}\cdot q^*_{01} + p_{00}\cdot q^*_{00}}{q^*_{01}+q^*_{00}}, \tag{6}$$

and for $t, x = 0, 1$

$$Pr(D_i = 1|T_i = t, X_i = x) = g(\alpha_0 + \alpha_1\cdot t + \alpha_2\cdot x\cdot(t - q^*_{10} - q^*_{11})) \tag{7}$$

$$= \frac{q^*_{t1}\cdot(1-p^*_{t1})^{(1-x)}\cdot(p^*_{t1})^x}{q^*_{t1}\cdot(1-p^*_{t1})^{(1-x)}\cdot(p^*_{t1})^x + q^*_{t0}\cdot(1-p_{t0})^{1-x}\cdot p^x_{t0}}.$$

Figure 3: Random assignment (solid) versus selection model with $\alpha_2 = 0$ and $\alpha_3 = 0$ (dashed)



(a) Restart data (b) artificial data

Proof: See appendix.

An obvious alternative to our proposed restriction is to allow for a main effect of the missing variable and restrict the interaction of T_i and X_i to be zero, setting $\alpha_3 = 0$. A third possibility is to restrict the main effect to be zero and allow only for an interaction effect, setting $\alpha_2 = 0$. Both restrictions however imply that for some values of the directly estimable parameters, $\{p^*_{td}\}_{t,d=1}$ and $\{q^*_{td}\}_{t,d=0,1}$, there are no values of α and p_{00} and p_{10} consistent with both the selection model and the restriction implied by random assignment. To illustrate this point, Figure 3a and 3b plot the sets of values of (p_{00}, p_{10}) consistent with the distribution of observed variables (for both Restart and artificial data sets) and the missing data model with $\alpha_3 = 0$ or $\alpha_2 = 0$ (dashed lines). The choice for $g(a)$ is the logistic function, $g(a) = \exp(a)/(1+\exp(a))$. While both curves go through the MAR point (p^*_{01}, p^*_{11}), neither is consistent with the restrictions implied by independence of X_i and T_i (the solid line) for the artificial data, and the $\alpha_3 = 0$ curve is not consistent with the independence restriction for the Restart data.[6]

[6] For example, the restriction $\alpha_2 = 0$ implies that $Pr(D_i = 1|T_i = 0, X_i = 0) = Pr(D_i = 1|T_i = 0, X_i = 1)$, and therefore $q^*_{01}(1-p^*_{01})/(q^*_{01}(1-p^*_{01}) + q^*_{00}(1-p_{00}))$ must equal $q^*_{01}p^*_{01}/(q^*_{01}p^*_{01} + q^*_{00}p_{00})$. This in turn implies $p_{00} = p^*_{01}$.

In addition to the existence and uniqueness properties given in Lemma 1, our choice of the restriction $\alpha_2 = \alpha_3 \cdot E[T_i]$ also has an appealing interpretation as being as close as possible to MAR while remaining consistent with RA. This interpretation is discussed next.

4. A Connection with Estimation of Contingency Tables with Known Marginals

In this section we link the model developed in the previous section with models used to estimate cell probabilities in a contingency table with known marginals. This connection is useful as it highlights the fact that the solution proposed in Lemma 1 can be viewed as being the set of missing probability values closest to MAR while remaining consistent with auxiliary information. The choice of $g(\cdot)$ operationalizes the measure of closeness. In Section 5, however, we will see that one advantage of the earlier approach is that, unlike the contingency table approach, it extends easily to the continuous covariate case.

4.1 Estimation of Contingency Tables with Known Marginals

A number of estimators have been suggested for the problem of estimating cell probabilities in a two-way classification with known marginal distributions (Deming and Stephan 1942; Ireland and Kullback 1968; Little and Wu 1991). Here we are particularly concerned with the interpretation of these estimators when the marginal distributions fail to correspond with the sampled row and column frequencies. Little and Wu (1991) show that the various estimators can in that case be interpreted as corresponding to different models for the relation between the target population (to which the marginal distribution refers) and the sampled population (to which the cell frequencies refer). In our terminology, it is the selection model which identifies this relation. We show that by modifying the known marginals problem – now imposing a marginal distribution for one variable along with independence between the two variables – the previously developed estimators for the known marginals problem lead to the models developed in Section 2. As in the previous section, we continue to focus on large sample issues and ignore estimation problems.

At this point it is convenient to reparameterize in terms of the joint distribution of T_i and X_i. As in Table 2, let $\pi_{tx} = Pr(T_i = t, X_i = x)$ be the parameters of the joint distribution of (T_i, X_i). In terms of the earlier notation,

$$\pi_{tx} = p_{t0}^x(1 - p_{t0})^{1-x}q_{t0} + p_{t1}^x(1 - p_{t1})^{1-x}q_{t1}.$$

In addition, let $\pi_{tx|1} = Pr(T_i = t, X_i = x|D_i = 1)$ be the parameters of the conditional distribution of (T_i, X_i) given $D_i = 1$. As shown in Table 2, in large samples we can estimate the $\pi_{tx|1}$ precisely, but we can only determine ranges for the π_{tx}.

A simple version of the problem considered by Little and Wu (1991) is that of determining π_{tx} given a sample of (T_i, X_i) randomly drawn conditional on $D_i = 1$ and given knowledge of the two marginal distributions. The marginals are summarized by the two parameters $\pi_{1\cdot} = Pr(T_i = 1) = \pi_{10} + \pi_{11}$ and $\pi_{\cdot 1} = Pr(X_i = 1) = \pi_{01} + \pi_{11}$. The solutions proposed in the literature all amount to choosing π_{tx} as close as possible to $\pi_{tx|1}$ while validating the known marginals. They differ in their choice of the measure of closeness. One solution, based on a likelihood metric, estimates π by solving:

$$\max_{\pi_{tx}} \sum_{t,x} \pi_{tx|1} \ln \pi_{tx}, \tag{8}$$

$$\text{subject to } \sum_{t,x} \pi_{tx} = 1, \qquad \pi_{1\cdot} = \pi_{10} + \pi_{11}, \qquad \pi_{\cdot 1} = \pi_{01} + \pi_{11},$$

where $\pi_{tx|1} = \sum_{i|D_i=1} 1\{T_i = t, X_i = x\} \Big/ \sum_{i|D_i=1} 1$ and $1\{\}$ is the indicator function, assuming a value of one when the specified condition is true and zero otherwise. The data (T_i, X_i) are given along with $\pi_{1\cdot}$ and $\pi_{\cdot 1}$. This leads to a solution for π_{tx} which can be written as

$$\pi_{tx} = \pi_{tx|1} \Big/ \Big[1 + \lambda_1 \cdot (t - \pi_{1\cdot}) + \lambda_2 \cdot (x - \pi_{\cdot 1})\Big],$$

with λ_1 and λ_2 the Lagrange multipliers for the restrictions on $\pi_{1\cdot}$ and $\pi_{\cdot 1}$, respectively.

4.2 Imposing Independence

In our problem with missing covariates in randomized trials, the restrictions are not on the two marginal distributions. Rather, we can view the sample as containing a random sample of (T_i, X_i) given $D_i = 1$ combined with knowledge of: (i) the known distribution of T_i, $\pi_{10} + \pi_{11} = q^*_{10} + q^*_{11}$; and (ii) the independence of T_i and X_i, $Pr(X_i = 1) \cdot Pr(T_i = 1) = Pr(T_i = 1, X_i = 1)$, or $(\pi_{01} + \pi_{11}) \cdot (\pi_{10} + \pi_{11}) = \pi_{11}$. We can impose these two restrictions using the likelihood metric and solve the same maximization program as (8) but with different restrictions:

$$\max_{\pi_{tx}} \sum_{t,x} \pi_{tx|1} \ln \pi_{tx}, \tag{9}$$

$$\text{subject to } \sum_{t,x} \pi_{tx} = 1, \qquad \pi_{10} + \pi_{11} = q^*_{01} + q^*_{11}, \qquad (\pi_{01} + \pi_{11}) \cdot (\pi_{10} + \pi_{11}) = \pi_{11}.$$

This leads to solutions for π_{tx} which can be written as the solution to seven equations; the first four from the first order conditions for π_{tx}:

$$\pi_{tx} = \pi_{tx|1} \Big/ \Big[\mu + \lambda_1 \cdot t + \lambda_2 \cdot (x \cdot (\pi_{10} + \pi_{11}) + t \cdot (\pi_{01} + \pi_{11}) - x \cdot t)\Big] \tag{10}$$

for $t, x = 0, 1$, and three equations from the restrictions:

$$\pi_{00} + \pi_{01} + \pi_{10} + \pi_{11} = 1,$$

$$\pi_{10} + \pi_{11} = q^*_{10} + q^*_{11},$$

$$(\pi_{01} + \pi_{11}) \cdot (\pi_{10} + \pi_{11}) = \pi_{11},$$

where μ, λ_1, and λ_2 are the Lagrange multipliers for the three restrictions.

To connect this with the model analyzed in the previous section, consider the solution $(\alpha_0, \alpha_1, \alpha_2, p_{00}, p_{10})$ in Lemma 1 corresponding to the linear probability specification $g(a) = a$. Since $g(\alpha_0 + \alpha_1 \cdot t + \alpha_2 \cdot x \cdot (t - E[T_i])) = Pr(D_i = 1|T_i = t, X_i = x)$ according to (4), the implied solution for the parameters π_{tx} in terms of $(\alpha_0, \alpha_1, \alpha_2)$, denoted by $\tilde{\pi}_{tx}$ is

$$\begin{aligned}\tilde{\pi}_{tx} &= \pi_{tx|1} \cdot Pr(D_i = 1)/Pr(D_i = 1|T_i = t, X_i = x) \\ &= \pi_{tx|1} \cdot (q^*_{01} + q^*_{11})/g(\alpha_0 + \alpha_1 \cdot t + \alpha_2 \cdot x \cdot (t - q^*_{10} - q^*_{11}))\end{aligned}$$

With $g(\cdot)$ linear, this expresion simplifies to

$$\tilde{\pi}_{tx} = \pi_{tx|1} \cdot (q^*_{01} + q^*_{11})/(\alpha_0 + \alpha_1 \cdot t + \alpha_2 \cdot x \cdot (t - q^*_{10} - q^*_{11})).$$

By setting $\mu = \alpha_0/(q^*_{01} + q^*_{11})$, $\lambda_1 = (\alpha_1 + \alpha_2(\pi_{01} + \pi_{11}))/(q^*_{01} + q^*_{11})$, and $\lambda_2 = -\alpha_2/(q^*_{01} + q^*_{11})$ in (10) and checking that $\tilde{\pi}_{tx}$ satisfies the restrictions $\sum \pi_{tx} = 1$, $\pi_{10} + \pi_{11} = q^*_{10} + q^*_{11}$, and $(\pi_{01} + \pi_{11}) \cdot (\pi_{10} + \pi_{11}) = \pi_{11}$, it follows that $\tilde{\pi}_{tx}$—based on Lemma 1 with $g(a) = a$— is also the solution to optimization program (9).

Using a different $g(\cdot)$ function in the model of Section 3 corresponds to using a different metric in (9). This is formalized in the following lemma:

Lemma 2 *For given q^*_{td}, p^*_{t1}, and a continuous and increasing function $g(\cdot)$ satisfying $\lim_{a\to-\infty} g(a) = 0$, $\lim_{a\to\infty} g(a) = 1$, let $(p_{00}, p_{10}, \alpha_0, \alpha_1, \alpha_2)$ be the unique solution to (6)-(7) in Lemma 1 with the implied solution for π_{tx} equal to*

$$\pi_{tx} = p^x_{t0}(1 - p_{t0})^{(1-x)} q^*_{t0} + (p^*_{t1})^x (1 - p^*_{t1})^{1-x} q^*_{t1}. \tag{11}$$

Furthermore, let $h(\cdot)$ be a continuously differentiable and convex function with the inverse of its derivative denoted by $k(\cdot)$; let $\pi_{tx|1} = (p^*_{t1})^x(1 - p^*_{t1})^{1-x} q^*_{t1}$; and let π_{tx} be the solution to:

$$\max_{\pi_{tx}} \sum_{t,x} \pi_{tx|1} h(\pi_{tx}/\pi_{tx|1}), \tag{12}$$

$$\text{subject to } \sum_{t,x} \pi_{tx} = 1, \qquad \pi_{10} + \pi_{11} = q^*_{10} + q^*_{11}, \qquad (\pi_{01} + \pi_{11}) \cdot (\pi_{10} + \pi_{11}) = \pi_{11}.$$

Then, if $k(\cdot) = (q^*_{01} + q^*_{11})/g(\cdot)$, the solution for π_{tx} is the same as that in Lemma 1:

$$\pi_{tx} = p^x_{t0}(1 - p_{t0})^{(1-x)} q^*_{t0} + (p^*_{t1})^x (1 - p^*_{t1})^{1-x} q^*_{t1}.$$

Proof: See appendix.

A popular choice for the convex function $h(z)$ is the likelihood metric $h(z) = \log(z)$ corresponding to the linear probability function $g(y) = y \cdot (q^*_{01} + q^*_{11})$. More generally one can use the function corresponding to the Cressie–Read (1984) family of divergence measures (see also Baggerly 1995 and Corcoran 1995):

$$h(z) = -(z^{-\lambda} - 1)/(\lambda \cdot (\lambda + 1)),$$

corresponding to

$$g(y) = (q^*_{01} + q^*_{11})(y \cdot (\lambda + 1))^{1/(1+\lambda)},$$

for values of λ on the real line. Another interesting metric is obtained by using the limit of the Cressie–Read divergence measure

$$\max_{\pi_{tx}} \lim_{\lambda \to -1} \sum_{t,x} \pi_{tx|1} \frac{1}{\lambda \cdot (\lambda + 1)} \cdot \left[\left(\frac{\pi_{tx}}{\pi_{tx|1}} \right)^{-\lambda} - 1 \right],$$

which reduces to the Kullback–Leibler criterion:

$$\max_{\pi_{tx}} \sum_{t,x} \pi_{tx} \log(\pi_{tx|1}/\pi_{tx}).$$

This criterion leads to the selection model $g(y) = \exp(-y) \cdot (q^*_{01} + q^*_{11})$. Finally, choosing $h(z) = -(z - q^*_{01} - q^*_{11}) \log(z - q^*_{01} - q^*_{11}) + z - q^*_{01} - q^*_{11}$ corresponds to the logistic selection model $g(y) = \exp(y)/(1 + \exp(y))$. It is interesting to note (see also Little and Wu 1991) that conventional choices for the discrepancy measure (e.g., the likelihood metric) do not correspond to conventional choices for the selection probability (e.g., logistic).

5. Response Variables and General Pretreatment variables

In the previous sections the analysis was limited to the case without a response variable and with a single binary pretreatment variable. Extending the basic approach to more general cases is straightforward and will be discussed briefly in this section.

Suppose that the partially observed pretreatment variable X_i takes on $K + 1$ different values. For notational convenience we assume these values are $0, 1, \ldots, K$. We extend the notation from the previous sections by defining $p_{xtd} = Pr(X_i = x|T_i = t, D_i = d)$. Independence of T_i and X_i then implies K restrictions of the type $Pr(X_i = x|T_i = 1) =$

$Pr(X_i = x|T_i = 0)$ for $x = 0, 1 \ldots, K$. Imposing these restrictions using the approach presented in the last section leads to specifications of the missing data model of the form

$$Pr(D_i = 1|T_i = t, X_i = x) = g(\psi + \alpha_x \cdot (t - E[T_i])).$$

The following lemma shows that this model has the same properties as the model for binary pretreatment variables discussed in the previous section.

Lemma 3 *Let $X_i \in \{0, 1, \ldots, K\}$ take on $K+1$ different values. Let $g(\cdot)$ be a continuous, strictly increasing function satisfying $\lim_{a\to\infty} g(a) = 1$ and $\lim_{a\to-\infty} g(a) = 0$. For any $q^*_{td} \in (0,1)$ with $\sum_{t,d} q^*_{td} = 1$, and any $p^*_{xtd} \in (0,1)$ satisfying*

$$\frac{p^*_{x01}q^*_{01} + p^*_{x00}q^*_{00}}{q^*_{01} + q^*_{00}} = \frac{p^*_{x10}q^*_{10} + p^*_{x11}q^*_{11}}{q^*_{10} + q^*_{11}},$$

there is a unique solution $(\psi, \alpha_0, \ldots, \alpha_K, p_{000}, \ldots, p_{K00}, p_{010} \ldots, p_{K10})$ such that for all t and x

$$g(\psi + \alpha_x \cdot (t - q^*_{10} - q^*_{11})) = q^*_{t1}p^*_{xt1}/(q^*_{t1}p^*_{xt1} + q^*_{t0}p_{xt0}),$$

and

$$\frac{p^*_{x01}q^*_{01} + p_{x00}q^*_{00}}{q^*_{01} + q^*_{00}} = \frac{p_{x10}q^*_{10} + p^*_{x11}q^*_{11}}{q^*_{10} + q^*_{11}}.$$

Proof: See appendix.

The second step is to allow for pretreatment variables that are always observed. This implies that the parameters ψ and α_x can depend on the value of the additional pretreatment variable X_{1i}. The final extension is to allow for the presence of the response variable Y_i. There is no direct evidence that the missing data mechanism depends on the outcome variable and we assume, as in the MAR approach, that the probability of the data missing does not depend on the value of Y_i. This leads to the following selection model for the general case:

$$Pr(D_i = 1|Y_i = y, T_i = t, X_{1i} = x_1, X_{2i} = x_2) = g(\psi_{x_1} + \alpha_{x_1,x_2} \cdot (t - E[T])),$$

where the parameter ψ_{x_1} can depend in an unrestricted way on X_{1i}, and the parameter α_{x_1,x_2} on both X_{1i} and X_{2i} (X_{2i}, previously referred to as X_i, is the partially observed pretreatment variable). The response variable is, in turn, specified as a probabilistic function of all the covariates and the treatment.

6. An Application

In this section we use the aforementioned methods to estimate the effect of a training program in the United Kingdom. Randomly chosen unemployed individuals were either required to have a conversation with an official from the local employment office about job search strategies ($T_i = 1$), or not ($T_i = 0$). We are interested in the effect this conversation has on future employment. The outcome is whether the individual has successfully exited from unemployment within six months from the date of randomization ($Y_i = 1$) or not ($Y_i = 0$). The covariate is whether the individual has a driving licence ($X_i = 1$) or not ($X_i = 0$). Both covariate and outcome are only observed for individuals who filled in a survey six months after the randomization date ($D_i = 1$). We have data on 8,189 individuals and focus our attention on the subset of 4,177 males ages 20 to 50. Among the 4,177 observations, 3,894 received the search strategy conversation; 283 did not. Only 2,230 individuals returned the survey revealing both the outcome (exit from unemployment) Y_i and the covariate (possession of a driver's license) X_i.[7] We also analyze an artificial data set that is similar to the original data set but exacerbates the MAR and RA conflict (Table 1 summarizes both sets of data).

We specify the model as follows:

$$Pr(T_i = 1) = \mu_t,$$

$$Pr(X_i = 1) = \mu_x,$$

and for $t, x = 0, 1$, the conditional distribution of the response variable,

$$Pr(Y_i = 1|X_i = x, T_i = t) = \frac{\exp(\beta_0 + \beta_1 \cdot x + \beta_2 \cdot t + \beta_3 \cdot x \cdot t)}{1 + \exp(\beta_0 + \beta_1 \cdot x + \beta_2 \cdot t + \beta_3 \cdot x \cdot t)}, \tag{13}$$

[7] Gender and age information are available for all individuals.

and, for $t, x, y = 0, 1$, the probability of responding to the survey,

$$Pr(D_i = 1|Y_i = y, T_i = t, X_i = x) = \frac{\exp(\alpha_0 + \alpha_1 \cdot t + \alpha_2 \cdot x \cdot (t - \mu_t))}{1 + \exp(\alpha_0 + \alpha_1 \cdot t + \alpha_2 \cdot x \cdot (t - \mu_t))}, \tag{14}$$

Note that the selection model (14) follows the form given in Lemma 1 with logistic $g(\cdot)$. We are interested in the population average effect of the employment conversation on future employment probability. In terms of the parameters defined above, this is

$$\tau = \mu_x \cdot \left(\frac{\exp(\beta_0 + \beta_1 + \beta_2 + \beta_3)}{1 + \exp(\beta_0 + \beta_1 + \beta_2 + \beta_3)} - \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} \right) \tag{15}$$

$$+ (1 - \mu_x) \cdot \left(\frac{\exp(\beta_0 + \beta_2)}{1 + \exp(\beta_0 + \beta_2)} - \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \right).$$

In other words, we wish to compute a population effect averaged over licensed ($X_i = 1$) and unlicensed ($X_i = 0$) individuals. In addition we may be interested in the parameters of the missing data mechanism, particularly in deviations from the missing at random assumption. This is captured by non-zero values of the parameter α_2 in Equation (13).

We estimate the model in two parts. First we multiply impute the missing values Y_i and X_i when $D_i = 0$ (Rubin, 1987, 1996), creating a number of simulated data sets with complete data on X_i and Y_i. Then, for each complete data set we estimate the average treatment effect using standard methods. The variance of the estimate of the average treatment effect is then estimated as the average of the complete data variances plus the variance between the estimates over the complete data sets.

To impute the missing X_i and Y_i we use the (DA) Data Augmentation algorithm proposed by Tanner and Wong (1990). Given initial estimates of the full parameter vector, denoted by $\theta = (\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \beta_3, \mu_t, \mu_x)$, the conditional distribution of X_i given T_i and $D_i = 0$ is Bernoulli with probability

$$Pr(X_i = 1|T_i = t, D_i = 0, \theta) =$$

$$\frac{\mu_x/(1 + \exp(\alpha_0 + \alpha_1 \cdot t + \alpha_2 \cdot (t - \mu_t)))}{\mu_x/(1 + \exp(\alpha_0 + \alpha_1 \cdot t + \alpha_2 \cdot (t - \mu_t))) + (1 - \mu_x)/(1 + \exp(\alpha_0 + \alpha_1 \cdot t))}.$$

With X_i imputed we then impute the missing values for Y_i by using the fact Y_i and D_i are independent conditional on X_i and T_i. Conditional on θ, X_i, T_i and $D_i = 0$ the distribution of Y_i is Bernoulli with probability

$$Pr(Y_i = 1|X_i = x, T_i = t, D_i = 0, \theta) = Pr(Y_i = 1|X_i = x, T_i = t, \theta)$$
$$= \frac{\exp(\beta_0 + \beta_1 \cdot x + \beta_2 \cdot t + \beta_3 \cdot x \cdot t)}{1 + \exp(\beta_0 + \beta_1 \cdot x + \beta_2 \cdot t + \beta_3 \cdot x \cdot t)}.$$

Given imputed values for X_i and Y_i, we then sequentially draw parameters in the vector θ based on their conditional posterior distributions. For the population proportions μ_x and μ_t, the distribution is Beta. For the logistic regression parameters $(\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \beta_3)$, the distribution is non-standard and we use the Metropolis-Hasting algorithm (Metropolis and Ulam 1949; Metropolis et al. 1953; Hastings 1970). The entire process is repeated, generating a Markov chain of parameter estimates which converges to the posterior distribution.

Table 3 summarizes the estimation results for four cases. We estimate both the nonignorable model described above and the MAR model for the Restart and artificial data sets. Estimating the MAR model involves restricting the selection parameter α_2 to be zero and simultaneously ignoring random assignment by permitting $Pr(X_i = 1|T_i = 1) \neq Pr(X_i = 1|T_i = 0)$ (these are parameters p_{1x} and p_{0x} in Table 3). The prior distributions for α and β in all four cases are of the conjugate form discussed by Clogg, Rubin, Schenker, Schultz, and Weidman (1991) and Rubin and Schenker (1987). This is equivalent to "adding" observations of the 2^4 different binary (Y_i, X_i, T_i, D_i) combinations. We choose to add 2.5 observations of each combination which, for the Restart data at least, should have little influence on the posterior. Priors for the remaining parameters (μ_x and μ_t for the non-ignorable model and p_{0x}, p_{1x} and μ_t for the ignorable model) were similarly modeled as conjugate Beta(2,2).

The parameters we wish to focus attention on are τ, describing the average treatment effect of the interview, and α_2, describing the degree to which MAR is violated. For the Restart data, the treatment effect is significantly positive and essentially the same for both the ignorable and non-ignorable models. Receiving the interview increases the probability

TABLE 3: ESTIMATION RESULTS
(posterior means and standard deviations)

	Restart Data		Artificial Data	
parameter	non-ignorable	ignorable	non-ignorable	ignorable
α_0	–0.217	0.116	36.346	0.120
	(0.192)	(0.117)	(0.459)	(0.038)
α_1	0.378	–0.021	–37.825	0.016
	(0.204)	(0.121)	(0.468)	(0.039)
α_2	–0.753	–	43.176	–
	(0.366)		(0.542)	
β_0	–1.873	–1.881	–2.092	–2.091
	(0.356)	(0.359)	(0.083)	(0.082)
β_1	1.031	1.039	1.252	1.250
	(0.362)	(0.366)	(0.091)	(0.091)
β_2	1.287	1.285	1.477	1.449
	(0.417)	(0.420)	(0.394)	(0.356)
β_3	–0.971	–0.974	–1.167	–1.139
	(0.428)	(0.430)	(0.396)	(0.359)
p_{0x}	–	0.571	–	0.021
		(0.040)		(0.004)
p_{1x}	–	0.481	–	0.850
		(0.011)		(0.002)
μ_x	0.488	0.487	0.553	0.794
	(0.011)	(0.011)	(0.005)	(0.003)
μ_t	0.932	0.932	0.932	0.932
	(0.004)	(0.004)	(0.001)	(0.001)
τ	0.089	0.091	0.094	0.056
	(0.034)	(0.034)	(0.047)	(0.061)

of exiting from unemployment within six months by 8.9%, with a 95% probability interval of 1.8% to 15.3%.

This similarity is perhaps surprising given the inconsistency of the MAR and RA assumptions observed in the data. Estimating the MAR model, we see that $p_{0x} \neq p_{1x}$. That is, the probability of having a license is different among the treatment and control groups—a violation of random assignment. When we estimate the non-ignorable selection model, we similarly find a statistically significant estimate of -0.753 for α_2—indicating a violation of missing at random.

Despite these differences, the estimate of μ_x remains the same in both models. This is important. The distinction between the models is how they impute the demographic features of the population, not how they compute the conditional treatment effect (the βs are identical). In this case, the only demographic variable is the fraction of the population that possesses a driver's license.

Looking at Figure 2a, we could surmise this result. As we are drawn away from the point representing MAR using the logistic model, the value of p_{10} (probability of possessing a driver's license among the missing treatment observations) remains unchanged until the boundary of the diagram. Since 92% of the population receive the treatment, this is the bulk of the missing observations. If the imputed value of their likelihood of possessing a license is unchanged, the value for the entire population will be unchanged.

This is a more general consequence when we consider the discrepancy-based approach. Since the measure being minimized is weighted by the probabilities observed in the sample, $\pi_{tx|1}$ in Equation (12), the model places greater weight on maintaining the characteristics of larger subgroups. This inevitably preserves the marginal probabilities such as μ_x as much as possible.

Therefore, to see a difference between the models we have to exacerbate the selection bias to the point where the value of p_{10} is no longer roughly 50%—namely, the border of Figure 2a. In terms of the underlying data, described by q_{td} and p_{t1}, we have to exacerbate the fact that $p_{11} \neq p_{01}$. This is exactly the approach we took with the artificial data where $p_{11} = 0.85$, $p_{01} = 0.02$, but the other data frequencies remain unchanged (see Table 2).

The only other difference between the Restart and artificial data is the total number of observations. In order to avoid having priors influence the results, we increase the number of observations by a factor of ten (see Table 1).[8] The results are shown in the right half of Table 3.

Now, we see a much larger contradiction between MAR and RA. The unequal values of p_{0x} and p_{1x} in the MAR model refute RA, while the significant value of α_2 in the non-ignorable model refutes MAR—in both cases by much larger margins than observed in the original Restart data. More importantly, this leads to a significantly different estimate of μ_x: 0.55 in the non-ignorable model versus 0.79 under MAR. When we use these different values of μ_x to weight the treatment effect among individuals with and without driver's licenses, we come to qualitatively different conclusions. The treatment effect under the ignorable model appears insignificant while the effect under the non-ignorable model is significant. Again, we note that the estimates of β and μ_t are unchanged (compared to the Restart data) since we preserved the remaining features of the data.

What are the essential elements that create a distinction between the two models? The first is that there must be a significant covariate-treatment interaction. In the Restart data, the discussion concerning job search strategies raises the likelihood of successfully exiting unemployment by 18% *among individuals without a driver's license.* There is virtually no effect for individuals possessing a license. Since we care about the *average* treatment effect, however, we need to weight these two effects by the fraction of the population with and without licenses. This is where the missing data assumptions enter: The second requirement for a distinction between the models is that they must result in significantly different estimates

[8] The priors place equal probability—equivalent to 2.5 observations—on each of the $2^4 = 16$ possible (Y_i, X_i, T_i, D_i) combinations. Since some categories would have only one or two observations based on the artificial sample frequencies and the original sample size of 4,177, the priors would tend to make the extreme data frequencies appear less extreme.

of the population value of the covariate. Not only must the MAR and RA assumptions be in conflict, they must conflict to the point that the imputed covariates lead to different marginal distributions among the two models.

7. Conclusion

Missing data complicates the simple and straightforward estimation of average treatment effects afforded by experiments with randomized assignment (RA). The standard approach to modeling the selection process (MAR) is inadequate in cases where available covariates are also missing. Such an approach fails to explain why the observed covariate distribution may be different for the treatment and control groups, in spite of randomized assignment. The proposed non-ignorable selection model developed in this paper attempts to preserve the spirit of MAR while explaining observed deviations. It does so by minimizing the discrepancy with MAR while incorporating our knowledge of randomized assignment as explained in Section 4.

Applied to data on a U.K. job-training program, we find a statistically significant conflict between MAR and RA. However, this has no practical consequence for the estimated treatment effect using the more general model developed herein compared to the standard MAR approach. Both lead us to conclude that the program raised the probability of exiting unemployment within six months by about 9%. However, we show that for similar data with a more significant MAR/RA conflict, the two selection models can generate qualitatively different results, as evidenced by the analysis of an artificial data set. The most effective means of ascertaining such a discrepancy is to estimate both models. We therefore recommend our more general selection model for analyzing missing data in the context of a randomized experiment with missing covariates.

Appendix

Proof of Lemma 1: We prove this in three steps. First, we show that there is an implicit function $p_{10} = h(p_{00})$ defined by this model, with $\lim_{p \to 0} h(p) = 1$, $\lim_{p \to 1} h(p) = 0$, and $h(p)$ strictly decreasing in p. Then, we show that the restriction implied by random assignment defines an implicit function $p_{10} = f(p_{00})$ which is increasing, and which is defined for $p_{00} \in [a, b]$ where either $a = 0$ or $f(a) = 0$, and either $b = 1$ or $f(b) = 1$. With both functions continuous there is always a unique solution to $f(p) = h(p)$, with $p_{00} = p \in [0, 1]$ and $p_{10} = h(p) \in [0, 1]$, which gives the uniqueness for (p_{00}, p_{10}). Finally we show uniqueness of the solutions for α_0, α_1 and α_2.

By assumption $g(\cdot)$ is invertible, and therefore we can simplify the four restrictions in (7) by concentrating out α_0, α_1, and α_2 to get

$$f_1(p_{10}) = f_0(p_{00}), \tag{16}$$

where

$$\begin{aligned} f_0(p_{00}) = {} & (q_{10} + q_{11})^{-1} \cdot g^{-1}\left(\frac{q_{01} \cdot (1 - p_{01})}{q_{01} \cdot (1 - p_{01}) + q_{00} \cdot (1 - p_{00})} \right) \\ & - (q_{10} + q_{11})^{-1} \cdot g^{-1}\left(\frac{q_{01} \cdot p_{01}}{q_{01} \cdot p_{01} + q_{00} \cdot p_{00}} \right), \\ f_1(p_{10}) = {} & (q_{00} + q_{01})^{-1} \cdot g^{-1}\left(\frac{q_{11} \cdot p_{11}}{q_{11} \cdot p_{11} + q_{10} \cdot p_{10}} \right) \\ & - (q_{00} + q_{01})^{-1} \cdot g^{-1}\left(\frac{q_{11} \cdot (1 - p_{11})}{q_{11} \cdot (1 - p_{11}) + q_{10} \cdot (1 - p_{10})} \right). \end{aligned}$$

Because $g(\cdot)$ is increasing, so is its inverse $g^{-1}(\cdot)$. It therefore follows that $f_1(\cdot)$ is strictly decreasing and $f_0(\cdot)$ is strictly increasing in their respective arguments, implying that the implicit function $p_{10} = h(p_{00})$ defined by $f_1(p_{10}) = f_0(p_{00})$ is strictly decreasing in its argument.

The second part of this first step is to show that for all $\varepsilon > 0$ there is a solution (p_{00}, p_{10}) to the equation $f_0(p_{00}) = f_1(p_{10})$ with $0 < p_{00} \leq \varepsilon$ and $1 - \varepsilon \leq p_{10} < 1$, as well as a solution with $1 - \varepsilon \leq p_{00} < 1$ and $0 < p_{10} \leq \varepsilon$. This would prove that $\lim_{p \to 0} h(p) = 1$ and $\lim_{p \to 1} h(p) = 0$. Suppose that $f_1(1-\varepsilon) < f_0(\varepsilon)$. Then, since $\lim_{p_{00} \to 0} f_0(p_{00}) = (q_{10} + q_{11})^{-1} g^{-1}(q_{01}(1-p_{01})/(q_{01}(1-p_{01}) + q_{00})) - (q_{10} + q_{11})^{-1} \lim_{a \to 1} g^{-1}(a) = -\infty$, there must be a solution (p_{00}, p_{10}) with $p_{10} = 1 - \varepsilon$ and $p_{00} < \varepsilon$. Similarly if $f_1(1-\varepsilon) > f_0(\varepsilon)$, there must be a solution with $p_{10} > 1 - \varepsilon$ and $p_{00} = \varepsilon$. Hence there always is a solution with $p_{10} \geq 1 - \varepsilon$ and $p_{00} \leq \varepsilon$. A similar argument can be used to show that there is always a solution with $p_{00} \geq 1 - \varepsilon$ and $p_{10} \leq \varepsilon$. □

From the proof of the lemma it can readily be seen that the missing data model defines an implicit function $p_{10} = p_{10}(p_{00})$ that is downward sloping in p_{00} and passes through the MAR point (p^*_{01}, p^*_{11}). This is true regardless of the choice of $g(\cdot)$ or the values of the observed variables q^*_{td} and p^*_{t1}. This implies that not all points in the set of (p_{00}, p_{10}) consistent with independence of T_i and X_i are consistent with a missing data model based on alternative functions $g(\cdot)$. Specifically, one cannot have both p_{00} and p_{10} larger than the values implied by MAR, and one cannot have both p_{00} and p_{10} smaller than the values implied by MAR.

There is a second restriction implied by the selection model (4). Consider ranking the four selection probabilities, $Pr(D_i = 1 | T_i = t, X_i = x)$ for $X_i = \{0, 1\}$ and $T_i = \{0, 1\}$. Any ranking implies a set of three inequality relations, for example,

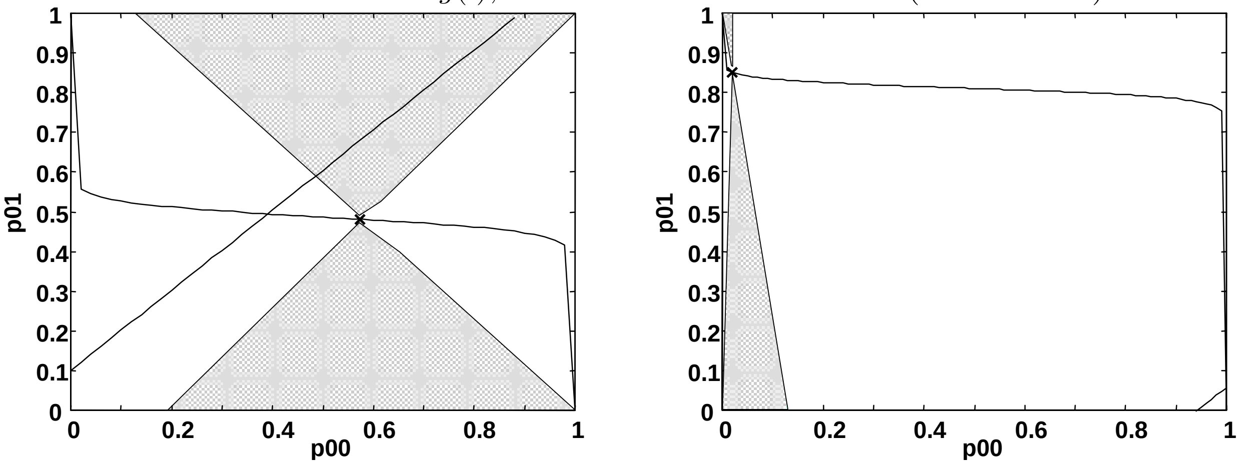
$$\begin{array}{c} Pr(D_i = 1 | T_i = 1, X_i = 0) > \\ Pr(D_i = 1 | T_i = 0, X_i = 0) > \\ Pr(D_i = 1 | T_i = 0, X_i = 1) > \\ Pr(D_i = 1 | T_i = 1, X_i = 1) \end{array} \tag{17}$$

The monotonicity of $g(\cdot)$ maps such relations into the two-dimensional parameter space (α_1, α_2). The parameter α_0 has no effect on the ranking. The above ranking leads to

$$\begin{array}{c} \alpha_1 > 0 \\ 0 > \alpha_2 \cdot (-q^*_{10} - q^*_{11}) \\ \alpha_2 \cdot (-q^*_{10} - q^*_{11}) > \alpha_1 + \alpha_2 \cdot (1 - q^*_{10} - q^*_{11}) \end{array} \tag{18}$$

Since three inequality relations in a two-dimensional space can easily have an empty intersection, certain orderings of the selection probabilities are simply inconsistent with our selection

FIGURE 4: RESTRICTIONS IMPLIED BY SECOND SIDE CONDITION (SHADED AREAS), LOGISTIC CHOICE OF $g(\cdot)$, AND RANDOM ASSIGNMENT (SOLID LINES)



(a) Restart data (b) artificial data

model – for any $g(\cdot)$ function. The relations in (18) demonstrate this: The first relation indicates that α_1 is positive. The second relation indicates that α_2 is positive. However, the third relation indicates that $\alpha_1 + \alpha_2$ is negative. It is therefore impossible to generate the ordering in (17) using the proposed selection model (4).

Such a restriction against certain orderings of selection probabilities translates into a set of inequality restrictions on p_{00} and p_{10} through the relations (3). For the Restart and artificial data given in Table 1, the regions where this restriction eliminates potential solutions are shown in Figures 4a and 4b.

The restriction that p_{00} and p_{10} cannot be either simultaneously larger or smaller than the values implied by MAR, as well as the more complicated restrictions shown in Figure 4, represents side consequences of the identifying assumption (4), much like the absence of an interaction between X_i and T_i is a consequence of setting $\alpha_3 = 0$ in (2). However, these side restrictions also turn out to be sufficient to characterize the set of possible solutions corresponding to choices of $g(\cdot)$. This is formalized in the following lemma.

Lemma 1a *For all $p^*_{tx} \in (0,1)$ and $q^*_{tx} \in (0,1)$ satisfying (5), there exists a monotonic* *function $g(\cdot)$ with $\lim_{a\to\infty} g(a) = 1$ and $\lim_{a\to-\infty} g(a) = 0$ such that the unique solution $(\alpha_0, \alpha_1, \alpha_2, p_{00}, p_{10})$ satisfying* (6) *and* (3) *has $p_{00} = p^*_{00}$ and $p_{10} = p^*_{10}$ if the following two conditions are satisfied:*

i. $\mathrm{sgn}(p^*_{11} - p^*_{10}) = -\,\mathrm{sgn}(p^*_{01} - p^*_{00})$.

ii. Assume $p_t > 0.5$. p^*_{10} cannot be either simultaneously greater or simultaneously smaller than the following four expressions:

(a) $1 - \frac{q^*_{01}\cdot q^*_{10}}{q^*_{00}\cdot q^*_{11}} \cdot \frac{1-p^*_{11}}{p^*_{01}} \cdot p^*_{00}$

(b) $\frac{q^*_{11}\cdot q^*_{00}}{q^*_{01}\cdot q^*_{10}} \cdot \frac{p^*_{11}}{1-p^*_{01}} \cdot (1 - p^*_{00})$

(c) $\frac{q^*_{11}\cdot q^*_{00}}{q^*_{01}\cdot q^*_{10}} \cdot \frac{p^*_{01}}{p^*_{11}} \cdot p^*_{00}$

(d) $1 - \frac{q^*_{11}\cdot q^*_{00}}{q^*_{01}\cdot q^*_{10}} \cdot \frac{1-p^*_{11}}{1-p^*_{01}} \cdot (1 - p^*_{00})$

Proof of Lemma 1a: We prove this lemma by constructing a function $g(\cdot)$ which, when coupled with the independence restriction and the observable quantities $\{q^*_{00}, q^*_{01}, q^*_{10}, q^*_{11}, p^*_{01}, p^*_{11}\}$, exactly identifies any (feasible) chosen set of true values $\{p^*_{00}, p^*_{10}\}$.

From the proof of Lemma 1, we know that a function $g(\cdot)$, which is monotonic and satisfies $\lim_{a\to-\infty} g(a) = 0$ and $\lim_{a\to+\infty} g(a) = 1$, identifies a downward sloping implicit function $p_{10} = f(p_{00})$ given by (16). Defining

$$r_{tx}(p_{t0}) = Pr(D_i = 1|X_i = x, T_i = t) = \frac{(p^*_{t1})^x(1-p^*_{t1})^{1-x}q^*_{t1}}{(p^*_{t1})^x(1-p^*_{t1})^{1-x}q^*_{t1} + (p_{t0})^x(1-p_{t0})^{1-x}q^*_{t0}},$$

this can be rewritten compactly as

$$p^*_t \cdot g^{-1}(r_{11}(p_{10})) + (1-p^*_t)\cdot g^{-1}(r_{01}(p_{00})) = p^*_t \cdot g^{-1}(r_{10}(p_{10})) + (1-p^*_t) g^{-1}(r_{00}(p_{00})) \tag{19}$$

where $p^*_t = q^*_{10} + q^*_{11} = Pr(T_i = 1)$. If our constructed $g(\cdot)$ satisfies (19) at $\{p^*_{00}, p^*_{10}\}$ and those values also satisfy the independence restriction, Lemma 1 proves that this solution is unique and therefore identified by our choice of $g(\cdot)$.

The proof will be expedited by rewriting the side restrictions in terms of r_{tx}. Defining $r^*_{1x} = r_{1x}(p^*_{10})$ and $r^*_{0x} = r_{0x}(p^*_{00})$,

i. $\text{sgn}(r^*_{11} - r^*_{10}) = -\text{sgn}(r^*_{01} - r^*_{00})$

ii. assuming $p_t > 0.5$, the following four inequalities (or the four reciprocal inequalities) cannot simultaneously hold true:

   (a) $r^*_{10} > r^*_{01}$

   (b) $r^*_{00} > r^*_{11}$

   (c) $r^*_{01} > r^*_{11}$

   (d) $r^*_{10} > r^*_{00}$

We construct $g(\cdot)$ by first defining values of $g^{-1}(r^*_{tx})$ which satisfy the monotonicity assumption on $g(\cdot)$ and by extension $g^{-1}(\cdot)$. The remainder of $g^{-1}(\cdot)$ is then defined as a piecewise linear function between the values r^*_{tx} combined with end pieces such that $g^{-1}(b) \rightarrow \pm\infty$ as $b \rightarrow 1$ or 0. The monotonic function $g(\cdot)$ is then formed by inverting $g^{-1}(\cdot)$.

WLOG, assume $p_t > 0.5$ and $|r^*_{11} - r^*_{01}| > |r^*_{10} - r^*_{00}|$ (if not, redefine the treatment/control groups or covariate groups as necessary). Set $g^{-1}(r^*_{00}) = 0$ and define $s = \text{sgn}(r_{10} - r_{00})$. Set $g^{-1}(r^*_{10}) = s$.

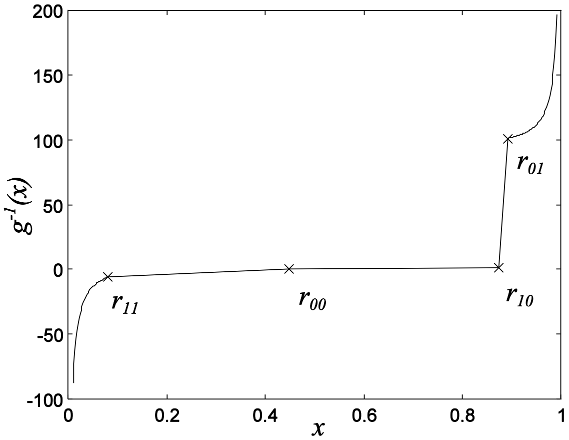
We now consider all possible cases given by the side restrictions,

i. Suppose $s \cdot r^*_{01} > s \cdot r^*_{11}$. With this assumption, coupled with $|r^*_{11} - r^*_{01}| > |r^*_{10} - r^*_{00}|$ and $\text{sgn}(r^*_{11} - r^*_{10}) = -\text{sgn}(r^*_{01} - r^*_{00})$, there are three possible orderings (×2 values of s):

$$\begin{array}{l} s \cdot r^*_{10} > s \cdot r^*_{01} > s \cdot r^*_{00} > s \cdot r^*_{11} \\ s \cdot r^*_{01} > s \cdot r^*_{10} > s \cdot r^*_{00} > s \cdot r^*_{11} \\ s \cdot r^*_{01} > s \cdot r^*_{10} > s \cdot r^*_{11} > s \cdot r^*_{00} \end{array} \tag{20}$$

However, since two of the four inequalities, (iic) and (iid), run in the same direction, at least one of the two remaining inequalities, (iia) or (iib), must run in the opposite

FIGURE 5: CONSTRUCTION OF $g(\cdot)$ FOR PARTICULAR CHOICE OF $\{p^*_{00}, p^*_{10}\}$



direction. This rules out the first ordering. To construct the rest of the $g^{-1}(\cdot)$ function, set $g^{-1}(r^*_{01}) = s \cdot (1+\phi)$ and $g^{-1}(r^*_{11}) = s \cdot (1 - \frac{1-p^*_t}{p^*_t}(1+\phi))$. Choose $\phi > 0$ such that $1 - \frac{1-p^*_t}{p^*_t}(1+\phi) \gtrless 0$ as $r^*_{11} \gtrless r^*_{00}$; this is possible since $p^*_t > 1 - p^*_t$ by assumption. Hence, $g^{-1}(\cdot)$ is monotone for either of the two possible orderings in (20). These values also satisfy (19) and, in fact, we have $\alpha_0 = 0$, $\alpha_1 = s$, $\alpha_2 = -s \cdot \frac{1+\phi}{p^*_t}$ as the unique solution for any monotone function $g(\cdot)$ passing through these four points.

The remainder of the function is constructed by making $g^{-1}(\cdot)$ linear between these four points and defining $g^{-1}(a) = g^{-1}(r_{\min}) + \frac{1}{r_{\min}} - \frac{1}{a}$ for $a < r_{\min}$ and $g^{-1}(a) = g^{-1}(r_{\max}) - \frac{1}{1-r_{\max}} + \frac{1}{1-a}$ for $a > r_{\max}$ where $r_{\min} = \min_{t,x\in\{0,1\}}(r^*_{tx})$ and $r_{\max} = \max_{t,x\in\{0,1\}}(r^*_{tx})$. This construction is illustrated in Figure 5 for the Restart data in Table 1 and $p^*_{00} = 0.100$, $p^*_{10} = 0.804$ with $\phi = 100$.

ii. Suppose instead $s \cdot r^*_{01} < s \cdot r^*_{11}$. Coupled with $|r^*_{11} - r^*_{01}| > |r^*_{10} - r^*_{00}|$ and $\operatorname{sgn}(r^*_{11} - r^*_{10}) = -\operatorname{sgn}(r^*_{01} - r^*_{00})$, there is only one possibility:

$$s \cdot r^*_{11} > s \cdot r^*_{10} > s \cdot r^*_{00} > s \cdot r^*_{01}.$$

To construct the rest of $g^{-1}(\cdot)$, set $g^{-1}(r^*_{01}) = -s$ and $g^{-1}(r_{11}) = s \cdot (1 + \frac{1-p^*_t}{p^*_t})$. These values of $g(\cdot)$ satisfy both monotonicity and (19). They lead to unique values of $\alpha_0 = 0$, $\alpha_1 = s$, $\alpha_2 = \frac{s}{p_t^*}$ for any monotonic $g(\cdot)$ passing through these four points. The remainder of $g^{-1}(\cdot)$ can be constructed as described in the previous paragraph.□

Note that the first condition prevents p_{00}^* and p_{10}^* from being simultaneously greater or smaller than the MAR values, $p_{00}^* = p_{01}^*$ and $p_{10}^* = p_{11}^*$. The second condition reflects the more complex restrictions on ordering. For example, the ordering given in (17) leads to p_{10}^* being greater than each the four expressions given in condition (*ii*) as represented by the upper shaded area in Figure 4a (e.g., a value of $(p_{00}^*, p_{10}^*) = (0.4, 0.9)$). Such a combination cannot be identified for any choice of monotonic $g(\cdot)$ along with (4).

Proof of Lemma 2: The first order conditions for (12) are

$$h'\left(\frac{\pi_{00}}{\pi_{00|1}}\right) - \mu = 0$$

$$h'\left(\frac{\pi_{01}}{\pi_{01|1}}\right) - \mu - \lambda_2(\pi_{10} + \pi_{11})$$

$$h'\left(\frac{\pi_{10}}{\pi_{10|1}}\right) - \mu - \lambda_1 - \lambda_2(\pi_{01} + \pi_{11})$$

$$h'\left(\frac{\pi_{11}}{\pi_{11|1}}\right) - \mu - \lambda_1 - \lambda_2(2\pi_{11} + \pi_{01} + \pi_{10} - 1)$$

combined with the restrictions

$$\sum_{t,x} \pi_{tx} = 1, \qquad \pi_{10} + \pi_{11} = \pi_{1\cdot} = q_{10}^* + q_{11}^*, \qquad (\pi_{01} + \pi_{11})(\pi_{10} + \pi_{11}) = \pi_{11}.$$

We now argue that the implied solution (11) for π_{tx} from Lemma 1 satisfies these seven conditions. By construction, the last three conditions are trivially satisfied based on (11) and (6) from Lemma 1. Using the assumption that $h'(z) = g^{-1}((q_{01}^* + q_{11}^*)/z)$, the first-order conditions become equivalent to (7), setting $\mu = \alpha_0$, $\lambda_2 = -\alpha_2$ and $\lambda_1 = \alpha_1 + \alpha_2(\pi_{01} + \pi_{11})$, and are also satisfied by the solution to Lemma 1.□

Proof of Lemma 3: First we show that there is a unique solution to the raking form of the problem, and then we show that the raking solution translates into a solution for the problem at hand.

Define

$$\pi_{xt|1} = p_{xt1}q_{t1}/(q_{01} + q_{11}).$$

Then there is a unique solution to the set of equations

$$\pi_{tx} = \frac{\pi_{tx|1} \cdot (q_{01} + q_{11})}{g(\mu + \lambda_x(t - q_{10} - q_{11}))},$$

and

$$\frac{\pi_{1x}}{q_{10} + q_{11}} = \frac{\pi_{0x}}{q_{00} + q_{01}},$$

for all t and x.

Substituting for π_{tx} we have, for $x = 0, 1, \dots, K$,

$$\frac{\pi_{0x|1} \cdot (q_{01} + q_{11})}{g(\mu - \lambda_x \cdot (q_{10} + q_{11})) \cdot (q_{00} + q_{01})} = \frac{\pi_{1x|1} \cdot (q_{01} + q_{11})}{g(\mu + \lambda_x \cdot (1 - q_{10} - q_{11})) \cdot (q_{10} + q_{11})}. \tag{21}$$

Given μ, there is a unique solution $\lambda_x(\mu)$ because the left-hand side is strictly increasing in λ_x, going to infinity as λ_x goes to infinity, and the right-hand side is strictly decreasing in λ_x, with limit infinity as λ_x goes to minus infinity.

To establish uniqueness of the solution for μ we need to bound the derivative of λ_x from below. Specifically, we need the result that $(t - q_{1\cdot})\frac{\partial \lambda_x}{\partial \mu} > -1$ for $t = 0, 1$ (where $q_{1\cdot} = q_{10} + q_{11}$). To establish this, take derivatives of both sides of (21) with respect to μ and solve for $\partial \lambda_x / \partial \mu$ to get

$$\frac{\partial \lambda_x}{\partial \mu} = \frac{\pi_{1x|1} \cdot (1 - q_{1\cdot})g'(\mu - \lambda_x q_{1\cdot}) - \pi_{0x|1} q_{1\cdot} g'(\mu + \lambda_x(1 - q_{1\cdot}))}{\pi_{1x|1} q_{1\cdot}(1 - q_{1\cdot})g'(\mu - \lambda_x q_{1\cdot}) + \pi_{0x|1} q_{1\cdot}(1 - q_{1\cdot})g'(\mu + \lambda_x(1 - q_{1\cdot}))}$$

Because $g'(\cdot)$, the derivative of $g(\cdot)$, is positive, it follows that $(t - q_{1\cdot})\partial \lambda_x / \partial \mu > -1$ for $t = 0, 1$.

The equation characterizing μ is $\sum \pi_{tx} = 1$. Substituting for π_{tx} we get

$$\sum_{t,x} \frac{\pi_{tx|1} \cdot (q_{01} + q_{11})}{g(\mu + \lambda_x(\mu)(t - q_{10} - q_{11}))} = 1.$$

The derivative of the left–hand size with respect to μ is

$$-\sum_{tx} \frac{\pi_{tx|1} \cdot (q_{01} + q_{11}) \cdot g'(\mu + (t - q_{10} - q_{11})\lambda_x(\mu)) \cdot (1 + (t - q_{10} - q_{11}) \cdot \partial\lambda_x \partial\mu)}{(g(\mu + \lambda_x(\mu)(t - q_{10} - q_{11})))^2} < 0.$$

Because the limit as μ goes to minus infinity is infinity, and the limit as μ goes to infinity is $q_{01} + q_{11} < 1$, there is a unique solution to the equation, and therefore a unique solution for λ_x and π_{tx}.

Given the solution π_{tx} let

$$p_{xt0} = (\pi_{tx} - p_{xt1} q_{t1})/q_{t0}.$$

The two remaining parts of the proof show that: (1) the p_{xt0} thus defined satisfy the independence conditions, and (2) $g(\cdot)$ is the conditional probability of $D_i = 1$.

For the first part we need to show that

$$\frac{p_{x01} q_{01} + p_{x00} q_{00}}{q_{01} + q_{00}} = \frac{p_{x11} q_{11} + p_{x10} q_{10}}{q_{11} + q_{10}}.$$

Substitute $(\pi_{tx} - p_{xt1} - q_{t1})/q_{t0}$ for p_{xt0} to get

$$\frac{p_{01} + q_{01} + q_{00}(\pi_{0x} - p_{x01} q_{01})/q_{00}}{q_{01} + q_{00}} = \frac{p_{11} + q_{11} + q_{10}(\pi_{1x} - p_{x11} q_{11})/q_{10}}{q_{11} + q_{10}},$$

which simplifies to

$$\frac{\pi_{0x}}{q_{00} + q_{01}} = \frac{\pi_{1x}}{q_{10} + q_{11}},$$

which is one of the restrictions imposed in the definition of π_{tx}, and therefore satisfied by assumption.

For the second part, consider the conditional probability of $D_i = 1$ given $T_i = t$ and $X_i = x$:

$$Pr(D_i = 1 | T_i = t, X_i = x) = \frac{q_{t1} p_{xt1}}{q_{t1} p_{xt1} + q_{t0} p_{xt0}}.$$

Substituting for p_{xt0} leads to

$$\frac{q_{t1}p_{xt1}}{q_{t1}p_{xt1} + q_{t0}(\pi_{tx} - q_{t1}p_{xt1})/q_{t0}}.$$

Substituting for π_{tx} then gives

$$\frac{q_{t1}p_{xt1}g(\mu + \lambda_x(t - q_{10} - q_{11}))}{\pi_{tx|1}(q_{01} + q_{11})}.$$

Finally, substituting $P_{xt1}q_{t1}/(q_{01} + q_{11})$ for $\pi_{tx|1}$ gives the result that

$$Pr(D_i = 1|T_i = t, X_i = x) = g(\mu + \lambda_x(t - q_{10} - q_{11})).$$

□

References

BAGGERLY, K., (1995), "Empirical Likelihood as a Goodness of Fit Measure", Los Alamos National Laboratory, University ofCalifornia.

CLOGG, C.C., D.B. RUBIN, N. SCHENKER, B. SCHULTZ, AND L. WEIDMAN, (1991), "Multiple Imputation of Industry and Occupational Codes in Census Public-use Samples Using Bayesian Logistic Regression," *Journal of the American Statistical Association*, Vol. 86, no. 413, 68–78.

CORCORAN, S.A., (1995), "Empirical Discrepancy Theory" Manuscript, Department of Statistics, Oxford University.

CRESSIE, N. AND T.R. READ, (1984), "Multinomial Goodness-of-Fit Tests," *Journal of the Royal Statistics Society, Ser. B*, Vol 46, 440–464

DEMING, W. E., AND F. F. STEPHAN, (1942), "On a Least Squares Adjustment of a Sampled Frequency When the Expected Marginal Tables are Known," *Annals of Mathematical Statistics*, Vol 11, 427–444.

DOLTON, P. AND D. O'NEILL, (1996a), "The Restart Effect and the Return to Full-time Stable Employment," *Journal of the Royal Statistical Society* A, 159, 275–288.

DOLTON, P. AND D. O'NEILL, (1996b), "Unemployment Duration and the Restart Effect: Some Experimental Evidence," *The Economic Journal*, 106, 387–400.

GOURIEROUX, C. AND A. MONFORT, (1981), "On the Problem of Missing Data in Linear Models," *The Review of Economic Studies*, 48, no 4, 579–86.

HASTINGS, W., (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109

HOROWITZ, J., AND C. MANSKI, (1995), "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63, no 2, 281–302.

IMBENS, G., D. RUBIN AND SACERDOTE, (1999), "Estimating the Effect of Unearned Income on Labor Supply, Earnings, Savings and Consumption: Evidence from a Survey of Lottery Players", NBER working paper 7001.

Ireland, C. T., and S. Kullback, (1968), “Contingency Tables with Known Marginals,” *Biometrika*, 55, 179–188.

Little, R. J. A., and D. B. Rubin, (1987), *Statistical Analysis with Missing Data*, Wiley: New York.

Little, R., and M. Wu, (1991), “Models for Contingency Tables with Known Margins When Target and Sampled Populations Differ, *Journal of the American Statistical Association*, 86, no 413, 87–95.

Manski, C., (1995), *Identification Problems in the Social Sciences*, Harvard University Press: Cambridge.

Metropolis, N. and S. Ulam, (1949), “The Monte Carlo Method,” *Journal of the American Statistical Association*, 44, 335–341.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, (1953), “Equation of State Calculations by Fast Computing Machines,” *Journal of Chemical Physics*, 21, 1087–1092.

Rubin, D. B., (1976), “Inference and Missing Data”, *Biometrika*, Vol 63, 581–592.

Rubin, D. B., (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley: New York.

Rubin, D. B., (1996), “Multiple Imputation after 18+ Years”, *Journal of the American Statistical Association*.

Rubin, D. B. and N. Schenker, (1987), “Logit-Based Interval Estimation for Binomial Data Using the Jeffrey’s Prior”, *Sociological Methodology 1987*, ed. C.C. Clogg, Washington, D.C.: American Sociological Association.

Tanner, M. and W. Wong, (1990), “The calculation of posterior distributions by data augmentation” (with discussion). *Journal of the American Statistical Association*, 82, 528–550.