

# Endogenous Multi-Valued Treatment Effect Model under Monotonicity <sup>\*</sup>

Denis Nekipelov<sup>†</sup>

October 2007

## Abstract

The local average treatment effect framework (LATE) is considered an effective approach to estimate the effects of binary endogenous regressors. In this paper I generalize the LATE model to a class of non-linear models with an endogenous regressor which can take multiple discrete values in the presence of a discrete-valued instrumental variable. I provide the semiparametric efficiency bound for finite-dimensional parameters in semiparametric moment equations under the generalized LATE assumption. I also suggest estimation methods which produce estimates achieving the semiparametric efficiency bound. I show how my results can be applied to average and quantile treatment effect settings with a multi-valued endogenous treatment variable and propose an efficient estimation procedure for such models. I apply my methodology to evaluate the effect of job attrition on the hourly wage for past applicants for welfare support in Florida using a randomized sample from the Family Transition Program. I find that failure to account for endogeneity of job attrition can substantially understate the effect.

**Keywords:** Treatment effect, multinomial choice, endogeneity, semiparametric efficiency bound, job attrition

---

<sup>\*</sup>Corresponding author: Denis Nekipelov, Department of Economics, Duke University, e-mail: denis.nekipelov@duke.edu

<sup>†</sup>I am deeply indebted to my advisor Han Hong for tremendous help and encouragement. I am grateful to Jane Cooley and Ralph Boleslavsky who provided extremely helpful suggestions on the manuscript. The usual disclaimer applies. I use the data from the Manpower Demonstration Research Corporation, without representing positions by the State of Florida and the MDRC.

# 1 Introduction

Modern analysis of treatment evaluation studies shows the importance of identification assumptions for recovering distributions of treatment outcomes from the data. A common approach to such evaluation studies is to assume that the treatment variable is conditionally independent from the treatment outcomes (Rubin (1974)). A large literature on treatment effects under this assumption studies outcomes of binary treatments. Recent research by Hahn (1998), Hirano, Imbens, and Ridder (2003), Imbens, Newey, and Ridder (2003), and other papers covers the issues of efficient estimation and dimension reduction for propensity score weighting, matching and projections methods. The problem of recovering the distributions of potential outcomes under the unconfoundedness assumption can be extended to the case of multiple-valued treatments. Frolich (2004) provides an overview of estimation methods and models for cases where treatment can take multiple discrete values. This concept is extended to the case of continuous treatments in Hirano and Imbens (2004). Finally, a recent paper Cattaneo (2007) discusses semiparametric efficiency for a class of non-linear models with multiple treatments. In this paper I develop a model where multi-valued treatment variable can be endogenous, provide semiparametric efficiency bound for a finite-dimensional parameter defined by a semiparametric moment equation, and suggest efficient estimation procedures for this parameter.

Endogeneity of treatment status is an essential problem in many treatment evaluation studies. Such endogeneity can occur when selection into treatments is correlated with the unobserved components of treatment outcomes. Nonlinearity in estimated econometric models with endogenous variables can create additional complexity. The structure of identification conditions for particular classes of econometric problems with continuous endogenous regressors has been studied in recent literature. Relevant examples include non-separable systems of moment equations in Imbens and Newey (2002) and Chesher (2003), quantile moment restrictions in Chernozhukov and Hansen (2005), models with panel error structure in Altonji and Matzkin (1997), and censored variable models in Hong and Tamer (2003). In the treatment effect literature, an attractive approach for dealing with endogenous treatments is to use the notion of local average treatment effects (LATE). The structure of the LATE model relies on the presence of an auxiliary binary instrument, which indicates randomized selection between two endogenous treatment options. The model focuses identification on a particular subset of the population called *compliers*. Com-

pliers receive a strictly higher treatment in one program than in the other. In this model the assumption that the treatment status in one treatment program is weakly higher than in the other allows one to recover distributions of latent outcomes for compliers. The structure of the model and possible estimation procedures are considered in Imbens and Rubin (1997), Angrist, Imbens, and Rubin (1996), Abadie, Angrist, and Imbens (2002), Abadie (2003) and Angrist (2004).

Efficiency results for a linear model of the average treatment effect under the LATE assumptions are provided in Frolich (2006). Hong and Nekipelov (2007) develops a general theory of efficient estimation for the non-linear LATE models generated by arbitrary (conditional and unconditional) moment conditions. The authors consider both models generated by moment functions of observable outcomes and a class of separable models with moments generated by latent potential outcomes. Hong and Nekipelov (2007) provide two alternative methodologies for efficient estimation in the LATE context. One suggested approach is based on an efficiently re-weighted moment function using inverse probability weights. An alternative approach forms the objective function as a linear combination of conditional expectations of moment function given observable variables. An attractive feature of analysis in Hong and Nekipelov (2007) is that the authors provide an explicit expression for the semiparametrically defined weighting matrix for an over-identified moment function. This expression allows one to transform both an over-identified unconditional moment equation and a conditional moment equation to an exactly identified optimal moment vector. In this case computation of the asymptotic variance of the estimator reduces to computation of the variance of the moment function. The asymptotic variance of such an estimator coincides with the semiparametric efficiency bound.

In this paper I propose a generalization of the LATE model to the case where both the treatment and the instrument can take multiple discrete values. I provide a minimum set of conditions for identification of the model from the observed distributions. I also provide semiparametric efficiency results for moment-based models in the context of the generalized LATE model and describe the properties of the efficient estimation procedure. In addition, develop a set estimation procedure for a finite-dimensional parameter in the moment equation when the monotone local instrument model is only set identified. My set identification and estimation approach has the attractive feature that it is easily obtainable from a point identified

submodel. My efficiency calculations use existing approaches for binary models described in Hahn (1998), Chen, Hong, and Tamer (2005), Chen, Hong, and Tarozzi (2007) for models under unconfoundedness, and Hong and Nekipelov (2007) specifically considering the binary LATE framework. My methodology extends the current literature on estimation of non-linear models with endogenous regressors and can be applied to the analysis of endogenous quantile treatment effects, as well as non-linear ATE and ATT models on compliers. Microeconomic models provide a natural application of the generalized LATE approach, which is designed for estimation of non-linear models (such as quantile regression) in the presence of discrete endogenous regressors.

I demonstrate the application of my efficient estimation methodology by analyzing the effect of job attrition on the wage using the results from a randomized welfare experiment from Florida. The dataset comes from the Manpower Demonstration Research Corporation. It contains information on participants in the randomized experiment conducted from 1994 to 1999 in Florida, where welfare applicants were randomly assigned to either the conventional welfare program, the Aid to Families with Dependent Children program, or the experimental Family Transition Program (FTP). The FTP was aimed at developing specific skills for more effective job performance. In terms of intensity of training the FTP significantly exceeds the quality of the standard support offered under the AFDC.

I study the relationship between the hourly wage on the primary job and job attrition, i.e. observed changes in employment. Indicators of attrition can be endogenous because, on the one hand, job search effort (and thus the propensity to change jobs) depend on the wage rate. On the other hand, endogeneity of attrition can be induced by the unobserved heterogeneity across individuals, e.g. indicating their ability. Bottom-censoring of hourly wages of individuals in the data further transforms the problem into a complex non-linear problem where the distribution of random shocks is not parametrically specified, the outcome variable is bottom-censored and the regressor of interest is endogenous. Such a framework serves as a natural application of the generalized LATE model that I offer in this paper. Random assignment to the two welfare programs serves a natural binary instrument. Existence of this instrument allows me to correct for endogeneity of the attrition indicators and estimate the returns to attrition for the subpopulation of the generalized compliers, defined as individuals for whom participation in the

FTP strictly increases job attrition.

The proposed generalized approach can be extended to cases where instruments are not available for all discrete values of the endogenous regressors. If it is reasonable to assume that some values of the regressor of interest are independent (its covariance matrix is block-diagonal), the independence assumption can be combined with the LATE assumption to allow for "correlation nests" in the support of the regressor. If such an independence assumption is not reasonable, one can use set inference methods for the generalized LATE model, which I briefly discuss in this paper.

The structure of the paper is the following. In Section 2, I describe the primitive assumptions of the generalized LATE model and prove identification of distributions of latent treatment outcomes. I also set up the semiparametric inference problem for estimation of a finite-dimensional parameter in a conditional moment equation<sup>1</sup>. In Section 3, I derive the semiparametric efficiency bound for the conditional-moment-based model and describe the structure of the optimal non-linear instrument. In Section 4, I provide alternative estimation procedures which produce semiparametrically efficient parameter estimates and prove their consistency and asymptotic normality given particular regularity conditions. In Section 5, I demonstrate that a moment-based model under the generalized LATE assumptions can be used in the traditional treatment effect setting such as estimation of average and quantile treatment effects. In Section 6, I apply my methodology to recover the effect of job attrition on the hourly wage for the data from the Family Transition Program in Florida. Section 7 concludes.

## 2 Model and Identification

### 2.1 Structure of the Model

In this paper I define a generalized local treatment effect model with multi-valued treatments. The structure of the multi-valued treatment effect models follows the setup in Frolich (2004) while the monotonicity in the structure of treatments is in line with the literature on binary

---

<sup>1</sup>In Appendix A.2 I demonstrate that if the identifying assumptions for the generalized LATE model are not attractive for a particular problem, one can use set inference techniques. Parameter estimates in the set-identified case can be constructed from the point estimates under the generalized LATE assumptions.

endogenous treatments in Angrist, Imbens, and Rubin (1996), Imbens and Rubin (1997), Abadie, Angrist, and Imbens (2002), Abadie (2003), Angrist (2004), and Hong and Nekipelov (2007). I consider a model with a vector of random potential outcomes  $\mathbf{Y} = (Y_1, \dots, Y_K)$  and a vector of discrete random treatment variables  $\mathbf{S} = (S_1, \dots, S_M)'$  such that  $S_k$  takes positive integer values and  $1 \leq S_k \leq K$ . I will use the term "treatment selection rule" interchangeably with the term "treatment variable" throughout the paper. In addition, I consider an instrument  $Z$  which takes positive integer values such that  $1 \leq Z \leq M$ . Throughout the paper I assume that the number of values of  $Z$  is smaller or equal to the number of the treatment selection rules. The joint distribution of  $(\mathbf{Y}, \mathbf{S}, \mathbf{Z})$  depends on a set of covariates  $X \in \mathcal{X} \subset \mathbb{R}^k$ . I introduce the following notion of generalized compliers for multi-valued treatments.

**Definition 1** *Generalized compliers are the sub-population for which the alignment of potential treatment outcomes is strictly monotonic:  $S_1 < S_2 < \dots < S_M$ .*

Definition 1 implies that the generalized compliers receive strictly different treatments for each two arbitrary selected treatment rules. The object of interest is the collection of conditional distributions of potential treatments for groups of generalized compliers:

$$Y_k | p = S_1 < \dots < S_M.$$

Note that if the support of the treatment rules  $S_k$  is sufficiently rich, there will be multiple sets of generalized compliers for  $1 \leq S_1 \leq K - M$ . In this framework there can also be multiple groups of generalized compliers corresponding to the same value of  $S_1$ . For instance, if  $K > M$ , then for  $S_1 = 1$  both sequences of observations  $S_1 = 1, S_2 = 2, \dots, S_M = M$  and  $S_1 = 1, S_2 = 3, \dots, S_M = M + 1$  satisfy the definition of generalized compliers. Below I list assumptions necessary to identify the model under consideration. In further sections I show how these assumptions can be relaxed to build a set-identified model.

- Assumption 1**
1. (Validity of instrument)  $(\mathbf{Y}, \mathbf{S}) \perp Z \mid X = x$ .
  2. (Non-degenerate treatment selection)  $\Pr\{Z = k \mid X\} \in (0, 1)$  for  $k = 1, \dots, M$ .
  3. (Non-degenerate treatment choice)  $\Pr\{S_k = p \mid X\} \in (0, 1)$  for  $k = 1, \dots, M$  and  $p = 1, \dots, K$ .

4. **(Convexity)** Selection rules  $S_k$  for  $k = 1, \dots, M$  and  $S_k < K$  are contained in weakly increasing convex sequences of treatment selection choices, moreover this ranking is preserved almost everywhere in  $\mathcal{X}$ :

$$\Pr \{0 \leq S_k - S_{k-1} \leq S_{k+1} - S_k \leq 1 | X = x\} = 1,$$

for all  $k = 2, \dots, M - 1$ .

5. **(Separability)** For  $k = 2, \dots, M - 1$  and  $1 < S_k < K$

$$f(y_p, S_{k+1} - S_k \neq S_k - S_{k-1}, x) = 0.$$

6. **(Boundary conditions)** For  $1 < p < K$   $f(y_p, S_{M-1} < S_M = p, x) = 0$ . If for some  $k$   $S_k = 1$  then  $S_{k-j} = 1$  for all  $j = 1, \dots, k - 1$ . Similarly, if  $S_k = K$  then  $S_{k+j} = K$  for all  $j = 1, \dots, M - k$ .

Assumptions 1.1 - 1.6 impose a rigid structure on the primitives of the model. Assumption 1.1 is equivalent to the standard assumption on instruments in the regular IV model, requiring conditional independence of the joint distribution of potential outcomes and treatment selection rules from the instrument. Assumptions 1.2 and 1.3 assert that all treatment selection rules and all treatment choices for each treatment selection rule are chosen with non-zero probability. Therefore, the data contain enough information to recover marginal distributions of treatment choices.

Assumption 1.4 is similar in spirit to the monotonicity assumption in the standard LATE model (where both the instrument and the treatment have binary support:  $M = K = 2$ ). It is called "convexity" because the sets of treatment selection choices  $S_1, \dots, S_M$ , which satisfy this condition, will be represented by convex line graphs on the plane  $(Z, S)$ . To see this, note that, according to this assumption, the difference between treatment selection choices for treatment rules  $k + 1$  and  $k$  cannot be smaller than the difference between choices for treatment rules  $k$  and  $k - 1$ . On the other hand, the difference between two consecutive treatment selection choices is bounded between 0 and 1. As a result, if  $S_k = S_{k-1} + 1$  for some treatment rule  $k$  then  $S_l = S_{l-1} + 1$  for all  $l \geq k + 1$ . Therefore, if a sequence of treatment selection choices  $S_k, S_{k+1}, S_{k+2}, \dots$  is increasing, it cannot become "flat" (i.e., for instance  $S_{k+2} = S_{k+3}$  is not allowed in this sequence). In practice, Assumption 1.4 implies that the treatment effect for latent

treatment choices can exhibit a threshold property: treatment status can remain constant for treatment selection rules from 1 to  $k$  and then monotonically increases for selection rules from  $k + 1$  to  $M$ . This situation might arise if the treatment program can set individual-specific "minimum" treatments.

Assumption 1.5 does not have an analog in the binary LATE model. This assumption states that the treatment outcome is never observed for non-monotone sequences of treatment selection choices. In fact the sequences of treatment choices where  $S_{k+1} - S_k = S_k - S_{k-1}$  satisfying Assumption 1.4 are either sequences where  $S_1 = \dots = S_M$  or where  $S_M = S_{M-1} + 1 = \dots = S_1 + M - 1$ . This suggests that although threshold treatment rules with  $S_1 = \dots = S_k < S_{k+1} < \dots < S_M$  are possible, they are never applied or their outcome is never observed. The treatments which produce observable outcomes are either the same across treatment rules, or different for all treatment rules (provided that convexity assumption 1.4 is satisfied).

While Assumption 1.5 imposes restrictions on observable distributions when both the treatment choice and the instrument are away from boundary points on their support, Assumption 1.6 provides identifying conditions on the boundary. The first part of Assumption 1.6 states that treatment outcomes generated by the extreme treatment choices are not observed for compliers when the instrument takes the maximum value. This assumption is important when the number of possible treatment choices is bigger than the number of possible values of the instrument. In such situation there will be multiple groups of compliers such that  $p = S_1 < S_2 < \dots < S_M$  for  $p = 1, \dots, K - M + 1$ . The first part of Assumption 1.6 deals with groups of compliers where  $2 \leq p \leq K - M$ . In this case, each pair of values of the treatment variable and the instrument generates two conditional outcomes. The first possible outcome is generated given that the observation is treated uniformly across treatment rules ( $S_1 = \dots = S_M$ ). The second outcome is generated given that observation is never treated equally across treatment rules ( $S_1 < \dots < S_M$ ). In order to disentangle the two components from the outcome, it is necessary to allow them to be observed separately. The first part of Assumption 1.6 generates such separability of observed outcomes. This first part is not needed for elements of the population with the highest and the lowest values of treatments given Assumption 1.4. In particular, if  $S_M = 1$  and  $Z = M$ , then from Assumption 1.4 it follows that  $S_1 = \dots = S_M = 1$ . Therefore, observation with  $S_M = 1$  and  $Z = M$  can only correspond to the case with uniform treatment (i.e.

non-compliers). Similarly, if  $S_1 = K$  and  $Z = 1$ , this observation also can only be attributed to non-compliers.

Once a sequence of treatment choices achieves the minimum or the maximum treatment, it cannot continue monotonically. Assumption 1.6 fixes this problem, by allowing the treatment rules to generate uniformly the smallest treatment values once a sequence of treatment choices achieves the minimum treatment value. Similar situation occurs on the upper boundary where treatment rules give maximum treatment once a particular treatment rule produces the maximum treatment choice. On the upper boundary I allow convexity assumption to be violated (Assumption 1.4, however, is stated in a non-contradictory way and does not extend to the boundary of support of treatments). This part of Assumption 1.4 is inherent to the structure of the model, and it arises because the range of offered treatments is finite.

**Example 1**

Consider an example where  $K = M > 2$ . In this case there is only one group of compliers with  $1 = S_1 < \dots < S_M = M$ . All monotone sequences of treatment selection choices with  $S_1 > 1$  achieve the maximum treatment for treatment selection rule  $k < M$ . Therefore, all such sequences are non-convex. To disentangle distributions of unobservable outcomes for compliers treated with maximum and minimum treatments one needs to apply only Assumptions 1.1 -1.4 and the second part of Assumption 1.6. The separability assumption is not required for these points because there is only one strictly monotone sequence of treatment outcomes, which corresponds to the set of compliers. To extract the treatment outcome distributions for compliers not treated with maximum and minimum treatments requires the full set of Assumptions 1.1-1.6.

**2.2 Observable variables and data structure**

The potential treatment selection rules, treatment choices and potential treatment outcomes are not always observed. Observable characteristics include the actual realization of treatment choices corresponding to the treatment selection rule indicated by the instrument. Similarly, potential outcomes are not observed either. I denote the observed treatment status  $W_2$  and

observed treatment outcome  $W_1$ . They can be expressed through unobservable variables as

$$W_1 = \sum_{k=1}^K \mathbf{1}(W_2 = k) Y_k,$$

$$W_2 = \sum_{m=1}^M \mathbf{1}(Z = m) S_m.$$

Observable variables, therefore, include  $W_1$ ,  $W_2$ , instrument  $Z$ , and a vector of covariates  $X$ . In the further discussion I use the notation  $W$  to denote the pair of variables  $(W_1, W_2)$ . The values of instrument  $Z$  and covariates  $X$  are assumed to be always observed. The data consist of a cross-section of i.i.d. realizations  $(w_{1i}, w_{2i}, z_i, x_i)$  for  $i = 1, \dots, N$  corresponding to realizations of a vector of random variables  $(\mathbf{Y}, \mathbf{S}, Z, X)$ . The problem of identification is to recover the conditional distribution of latent treatment outcomes from the observable distribution of realized outcomes  $(W, Z, X)$ . To simplify further manipulations, I introduce additional notation corresponding to observable distributions:

$$f_j(w_1, p) = f_x(w_1 | w_2 = p, z = j),$$

$$\mathcal{P}_j(p) = P_x(w_2 = p | z = j),$$

$$\mathcal{Q}_j = P_x(z = j),$$

$$\mathbf{P}_m = \mathbf{Pr}\{w_2 = m | x\}.$$

I also denote  $P_x(j_1, \dots, j_M) = \mathbf{Pr}(S_1 = j_1, \dots, S_M = j_M | X)$ . Finally, I introduce the notation for the population proportion of a particular group of generalized compliers as:

$$P_{>}(p) = \mathbf{Pr}(p = S_1 < \dots < S_M | x),$$

and the density of the treatment outcome for compliers as

$$f_{>}(w_1, k, p) = f(y_k | p = S_1 < \dots < S_M, x).$$

I omit indexation by covariates  $X$ , while it is understood throughout this paper that all relevant distributions depend on covariates.

### 2.3 Identifying treatment choice probabilities for generalized compliers

Probabilities of observing groups of generalized compliers are identified by Assumptions 1.1 - 1.4 and 1.6. Given non-degenerate distribution of treatment choices over the support for

all treatment selection rules, Assumptions 1.4 and 1.6 determine plausible configurations of treatment choices in the model. The identification conditions can be visualized on the  $(Z, S)$  plane. Sequences of treatment selection choices satisfying these assumptions take the form

$$S_1 = \dots = S_k = p, S_{k+1} = p + 1, \dots, S_M = M - k + p.$$

This structure of sequences of treatment selection choices allows me to recover probabilities of interest  $P_{>}(p)$ . To recover these probabilities note that

$$\mathcal{P}_1(p) = P_{>}(p) + \sum_{k=2}^M P_x(p = s_1 = \dots = s_k < s_{k+1} < \dots < s_M).$$

On the other hand, considering the probability in the "adjacent node" we obtain that

$$\mathcal{P}_2(p) = P_{>}(p-1) + \sum_{k=2}^M P_x(p = s_1 = \dots = s_k < s_{k+1} < \dots < s_M).$$

As a result, we obtain a simple relation between proportions of two consecutive groups of compliers in the population:

$$P_{>}(p) = P_{>}(p-1) + \mathcal{P}_1(p) - \mathcal{P}_2(p).$$

For the first group of compliers such expression gives an explicit representation, due to the presence of boundary conditions

$$P_{>}(1) = \mathcal{P}_1(1) - \mathcal{P}_2(1).$$

Then all subsequent complier proportions are determined recursively as

$$P_{>}(p) = \sum_{k=1}^p (\mathcal{P}_1(k) - \mathcal{P}_2(k)). \tag{1}$$

This expression uniquely determines the probability of treatment choices for compliers. The general result is summarized in the following theorem.

**Theorem 1** *Suppose that Assumptions 1.1-1.4 and 1.6 are satisfied. Then complier probabilities  $P_{>}(p)$  are exactly identified from the data, i.e. the proposed assumptions provide a minimum set of conditions for identification.*

Identification conditions impose rigid restrictions on the probabilities in the analyzed model. It is necessary to develop a constructive proof of identification first in order to understand how the identifying assumptions limit the considered types of distributions. Second, a constructive proof will make it possible to consider the robustness of the model with respect to relaxing these assumptions by allowing set identification. A complete proof of identification is provided in Appendix A.1.

In the proof I consider model behavior under a significantly milder assumption: weak monotonicity allowing sequences of treatment choices with  $S_1 \leq \dots \leq S_M$ . I describe the structure of the matrix  $A$  which transforms the unobserved joint probabilities of treatment choices into observed outcome probabilities. I show that, except for the binary case, this matrix will be rectangular and joint choice probabilities are not identified. Then I demonstrate that this matrix has rank  $(K - 1)M$  by showing how to construct a non-singular transformation of  $A$  to a sparse matrix containing a  $(K - 1)M \times (K - 1)M$  identity submatrix. The matrix of identifying equations under Assumptions 1 has a 2-diagonal structure and can be transformed to a matrix with  $(K - 1)M \times (K - 1)M$  identity submatrix as well (the rest of the elements are equal to zero). As a result, the set of linearly independent columns of  $A$  can be transformed to the matrix of coefficients for identifying equations under Assumptions 1.1 - 1.6 by a non-singular transformation.

This theorem demonstrates that the set of identifying conditions imposed by Assumptions 1.1 - 1.6 allows one to recover the maximal possible set of joint probabilities of treatment selection choices from the data. Next I will consider the problem of identifying conditional distributions of potential treatment outcomes from the data.

## 2.4 Identifying distributions of treatment outcomes

Identification of the distributions of the treatment outcomes is achieved due to Assumptions 1.1 - 1.6. The intuitive identification argument is provided by the separability Assumption 1.5. In my model each possible treatment outcome  $Y_k$  with the instrument equal to  $m < M$  belongs to only one group of generalized compliers (those for whom  $k - m = S_1 < \dots < S_M$ ) and only one group of non-compliers ( $k = S_1 = \dots = S_M$ ). From the boundary condition, it follows that the boundary values of the instrument ( $Z = M$ ) generate the treatment outcome only for

non-compliers. Therefore, we can identify the corresponding density of the treatment outcome for compliers by subtracting the component corresponding to non-compliers for  $Z = M$  from each distribution component for  $Y_k$  given that  $Z = m < M$ . As a result, for each distribution of  $Y_k | p = S_1 < \dots < S_M$  there will be exactly one element of the observable distribution corresponding to  $w_2 = k$  and  $z = k - p + 1$  which identifies the latent distribution of interest.

Formally, the identifying equation will take the form

$$f_{>}(w_1, k, p) P_{>}(p) + f_x(y_k = w_1 | k = S_1 = \dots = S_M) P_x(k, \dots, k) = f_{k-p+1}(w_1, k) \mathcal{P}_{k-p+1}(k),$$

$$f_x(y_k = w_1 | k = S_1 = \dots = S_M) P_x(k, \dots, k) = f_M(w_1, k) \mathcal{P}_M(k),$$

for  $p = 1, \dots, K - M$  and  $k = p, \dots, p + M - 1$ . Note that the treatment outcome  $Y_p$  cannot be identified for a group of compliers for which  $p \notin [S_1, S_M]$ . The reason for this result is that this treatment outcome is never observed for this group of compliers, while the structure of the assumptions does not provide the possibility to recover this distribution from observations for different treatment outcomes. The final expression for the density of the outcome distribution for a group of compliers (given  $k < K$ ) can be written as:

$$f_{>}(w_1, k, p) = \frac{f_{k-p+1}(w_1, k) \mathcal{P}_{k-p+1}(k) - f_M(w_1, k) \mathcal{P}_M(k)}{\sum_{j=1}^p (\mathcal{P}_1(j) - \mathcal{P}_2(j))}.$$

I set boundary conditions for the entire support of binary regressor except  $S_m = K$ . On the upper boundary of the support of the treatment selection choices, there is only one point belonging to the subset of compliers, where  $Z = M$  and  $S_M = K$ . For this value of the treatment choice, the expression for the density for compliers naturally comes from the configuration of the support:

$$f_{>}(w_1, K, K - M + 1) = \frac{f_M(w_1, K) \mathcal{P}_M(K) - f_{M-1}(w_1, K) \mathcal{P}_{M-1}(K)}{\sum_{j=1}^{K-M+1} (\mathcal{P}_1(j) - \mathcal{P}_2(j))}.$$

Theorem 1 shows that the probability  $P_{>}(\cdot)$  is exactly identified from the data. Therefore, there is a unique system of equations for  $f_{>}(\cdot)$ . Given that this system has a unique solution for each  $P_{>}(\cdot)$ , the entire set of identification conditions determines a unique density  $f_{>}(w_1, k, p)$  for each suitable  $k$  and  $p$ . Note that due to the convexity assumption, the joint distribution of treatment choices on the upper boundary of the support of treatment selection rules does not provide over-identification restrictions for the distributions of interest.

## 2.5 Semiparametric moment specification

In the previous discussion, I have shown that the observed distributions identify the relevant marginal treatment outcome distribution for a set of generalized compliers. I assume that the main object of interest in this model is the Euclidean parameter  $\beta \in \mathcal{B} \subset \mathbb{R}^k$ , defined by a semiparametric moment equation

$$\varphi_p(w, x, \beta) = E\{g(w, x, \beta) \mid p = s_1 < \dots < s_M, w_2, x\} = 0. \quad (2)$$

I use this condition to define the baseline model. In Section 5 I demonstrate that this model can be used to analyze moment equations defined in terms of latent treatment outcomes.

The generalized LATE model incorporates many existing models of treatment effects as special cases. One special case where the instrument and treatment selection rule has the same support size was considered in Example 1. Two other special cases are presented in examples below.

### Example 2

This special case of the model produces a multi-dimensional generalization of the model for treatment effects under the unconfoundedness assumption. Consider the following additional assumptions:

1. The treatment selection choices  $S_k$  and instrument  $Z$  have the same support:  $K = M$
2. All observations in the sample are "compliers":  $S_k = k$ .

In this model there is going to be a unique sequence of treatment selection choices in this model:  $1 = S_1 < \dots < S_M = M$ . In this case, the model will be equivalent to the model with a single treatment  $D = Z$ , such that

$$\{Y_d\}_{d=1}^M \perp D \mid X.$$

This is the treatment effect model under the unconfoundedness and multi-valued treatments considered, for instance in Cattaneo (2007). In the equations that I used to identify the treatment effect for generalized compliers,  $f_M(w_1, p) = 0$  for  $p < M$ . Therefore,  $f(Y_p = w_1 \mid D = p) = f_p(w_1, p)$ . This suggests that the multi-valued treatment effect model for independent treatment is trivially identified under Assumptions 1.1 - 1.6 with the addition of Assumptions 1 and

2 above.

### Example 3

Consider a model where the supports of both the instrument and the treatment selection choice variable are binary. For binary support the model becomes the standard non-linear local treatment effect model with the instrument taking values  $Z = 1, 2$ . The expression for the density of the observed outcome for compliers specializes to this case as well. The obtained expressions for conditional densities of outcomes for compliers coincide with standard expressions in the literature.

The third special case of the model is the case of quantile regression with an endogenous discrete regressor. Suppose that the quantile function is defined as

$$Q_\tau(w, x, \beta) = \mathbf{1} \{w_1 \leq \beta_0 w_2 + x' \beta_1\} - \tau.$$

Defining the moment  $g(w, x, \beta)$  to be equal to the quantile function (or a vector of quantile functions if several simultaneous quantile restrictions are considered), one can estimate parameters  $\beta_0$  and  $\beta_1$  corresponding to the moment equation for generalized compliers. For the group of generalized compliers the problem of endogeneity of the discrete treatment variable effectively disappears.

These motivating examples demonstrate that the model under consideration generalizes an array of existing treatment effect models, therefore providing additional insight about these models as well.

## 3 Semiparametric Efficiency

In the previous section, I discussed identification of the non-parametric component of the model from observable distributions in the data. In this section, I provide the semiparametric efficiency bound for the estimation of a finite-dimensional parameter  $\beta$ , which is defined through the set of moment equations for compliers.

My efficiency results build on and extend the general efficiency framework of Koshevnik and Levit (1976), Begun, Hall, Huang, and Wellner (1983), Chamberlain (1987), Newey (1990), Bickel, Klaassen, Ritov, and Wellner (1993), and Severini and Tripathi (2001) for moment-based models by considering these models in the generalized LATE framework. Efficient estimation of

the effects of binary treatments has been considered in Hahn (1998), Hirano, Imbens, and Ridder (2003), Hong and Nekipelov (2007) among others. I generalize the existing semiparametric efficiency results in the treatment effect literature to endogenous multinomial choice models.

I begin by introducing the following notation. Define dummy variables indicating particular choices of treatment rules and treatment choices

$$d_m^z = \mathbf{1}(z = m), \quad \text{and} \quad d_k^{w_2} = \mathbf{1}(w_2 = k).$$

Also define the conditional probability of treatment choice as a function of the instrument value

$$\mathcal{F}(w_2, z) = \sum_{m=1}^M d_m^z \mathcal{P}_m(w_2).$$

Finally, introduce

$$\zeta_p(w_2, x) = \left( \frac{d_p^{w_2}}{\mathbf{P}_p}, \dots, \frac{d_{p+M-1}^{w_2}}{\mathbf{P}_{p+M-1}} \right)'.$$

**Theorem 2** *Under Assumptions 1.1 - 1.6, the semiparametric efficiency bound for a  $k$ -dimensional parameter  $\beta$  in moment equation (2) characterizing the subsample of generalized compliers can be expressed as:*

$$V(\hat{\beta}) = E \left( P_{>}(p)^2 E \left[ \frac{\partial \varphi_p(w_2, x, \beta)}{\partial \beta} \zeta_p(x, w_2)' \middle| x \right] \Omega^{-1} E \left[ \zeta_p(x, w_2) \frac{\partial \varphi_p(w_2, x, \beta')}{\partial \beta} \middle| x \right] \right)^{-1}.$$

In the following I use the notation  $\gamma_{z, w_2} = E[g|w_2, z, x]$  and  $\omega_{z, w_2} = V(g|w_2, z, x)$  to express the components of  $V(\hat{\beta})$ . Elements of the  $\Omega$  component in the semiparametric efficiency bound can be written explicitly. Diagonal elements of the matrix  $\Omega$  can be computed as

$$\begin{aligned} \Omega_{ii} = & \omega_{i, p+i-1} \frac{\mathcal{P}_i(p+i-1)}{\mathcal{Q}_i} + \omega_{M, p+i-1} \frac{\mathcal{P}_M(p+i-1)}{\mathcal{Q}_M} \\ & + \gamma_{M, p+i-1}^2 \mathcal{P}_M^2(p+i-1) \left[ \frac{\mathcal{Q}_M - \mathcal{P}_M(p+i-1)}{\mathcal{Q}_M^2 \mathcal{P}_M(p+i-1)} + \frac{\mathcal{Q}_i - \mathcal{P}_i(p+i-1)}{\mathcal{Q}_i^2 \mathcal{P}_i(p+i-1)} \right]. \end{aligned}$$

Off-diagonal elements have a simple form

$$\Omega_{ij} = - \frac{\mathcal{P}_M(p+i-1) \mathcal{P}_M(p+j-1) \gamma_{M, p+i-1} \gamma_{M, p+j-1}}{\mathcal{Q}_M}.$$

The optimal instrument matrix  $\mathcal{M}(x)$  is generated by the structure of the moment conditions and by the semiparametric efficient projection. This matrix takes the form

$$\mathcal{M}^*(x) = E \left[ \frac{\partial \varphi_p(w_2, x, \beta)}{\partial \beta} \zeta_p(x, w_2)' \middle| x \right] \Omega^{-1} \text{diag} \left\{ \frac{\mathcal{Q}_1}{\mathbf{P}_p}, \dots, \frac{\mathcal{Q}_{M-1}}{\mathbf{P}_{p+M-1}} \right\}.$$

This weighting matrix transforms the original over-identified conditional moment equation to the unconditional one.

## 4 Efficient Estimation

In this section I develop two classes of estimators, both of which achieve the semiparametric efficiency bound. The first one is based on the inverse probability weighting and the second one is based on the semiparametric efficient projection of conditional expectation. I also give a set of sufficient conditions to assure consistency and asymptotic normality of the suggested estimators. The structure of the inverse-probability weighted estimator is closely related to the structure of estimators for conditionally independent models in Hahn (1998) and Firpo (2006). The structure of the projection-based estimator is closely related to the estimator proposed in Imbens, Newey, and Ridder (2003) and Chen, Hong, and Tarozzi (2007) under conditional independence assumption and Frolich (2006) and Hong and Nekipelov (2007) under the binary LATE assumption.

### 4.1 Propensity score weighted estimator

I consider the problem of estimation of the finite-dimensional parameter  $\beta \in \mathcal{B} \subset R^k$  defined by conditional moment model (2) for a group of generalized compliers.

The idea of propensity score weighting is to use a set of weight functions to translate the moment condition (2) for compliers to the moment condition for the entire population. Such transformation can be made using identification conditions from Section 2.2 and using the Bayes' rule. Specifically, for the optimal choice of the instrument matrix  $\mathcal{M}(x)$  denoting  $\tilde{g} = \mathcal{M}(x)\zeta_p(w_2, x)g(w, x, \beta)$ , the conditional moment equation can be written as the uncondi-

tional one:

$$E \left( \sum_{k=p}^{p+M-1} \left[ d_k^{w_2} d_{k-p+1}^z - \frac{\mathcal{Q}_{k-p+1}}{\mathcal{Q}_M} d_k^{w_2} d_M^z \right] \tilde{g} \right) = 0. \quad (3)$$

Equation (3) can be proved in the following way. First, note that conditional moment equation (2) for compliers for each  $w_2 = p, \dots, p+M-1$  can be redefined in terms of empirically observable moments using the Bayes' rule:

$$\begin{aligned} \frac{\mathcal{P}_{k-p+1}(k,x)}{\mathcal{P}_{>(p,x)}} E [g(w, x, \beta) \mid w_2 = k, z = k - p + 1, x] \\ - \frac{\mathcal{P}_M(k,x)}{\mathcal{P}_{>(p,x)}} E [g(w, x, \beta) \mid w_2 = k, z = M, x] = 0. \end{aligned} \quad (4)$$

For a given optimal instrument matrix  $\mathcal{M}(x)$  I pick the weighting matrix  $\mathcal{A}(\cdot)$  in a particular form:

$$\mathcal{A}(w_2, x) = \mathcal{P}_{>(p)} \sum_{k=p}^{p+M-1} \frac{\mathcal{Q}_{k-p+1} d_k^{w_2}}{\mathbf{P}_k} \mathcal{M}(x) \zeta_p(w_2, x).$$

For an arbitrary non-singular weighing matrix  $\mathcal{A}(w_2, x)$ , conditional moment equation (2) implies that

$$E [\mathcal{A}(w_2, x) \varphi_p(w, x, \beta)] = 0.$$

Using this fact, I substitute the chosen weighting matrix into the (4) and defining

$$\bar{g} = \mathcal{A}(w_2, x) g(w, x, \beta),$$

I find that: multiplication of the moment function by the weighting matrix  $\mathcal{A}(\cdot)$  yields

$$\sum_{k=p}^{p+M-1} E \left( \frac{\mathcal{P}_{k-p+1}(k) d_k^{w_2}}{\mathcal{P}_{>(p)}} E [\bar{g} \mid w_2 = k, z = k - p + 1, x] - \frac{\mathcal{P}_M(k) d_k^{w_2}}{\mathcal{P}_{>(p)}} E [\bar{g} \mid w_2 = k, z = M, x] \right) = 0.$$

Substituting the expression for  $\mathcal{A}(\cdot)$  in this formula produces equation (3).

An empirical analog for equation (3) defines the inverse probability-weighted estimator. Such an estimator can be implemented in a two-step procedure which is outlined below.

**Step 1** Estimate functions  $\hat{\zeta}_p$ ,  $\hat{\mathcal{Q}}_m$  non-parametrically. Pick some non-singular constant matrix  $\overline{M}$  with the same dimensions as  $\mathcal{M}(x)$  and set up the moment equation

$$\psi_i(\hat{\mathbf{P}}, \hat{\mathcal{Q}}, \beta) = \sum_{k=p}^{p+M-1} \left[ d_k^{w_{2i}} d_{k-p+1}^{z_i} - \frac{\hat{\mathcal{Q}}_{k-p+1,i}}{\hat{\mathcal{Q}}_{M,i}} d_k^{w_{2i}} d_M^{z-i} \right] \overline{M} \hat{\zeta}_p(w_{2i}, x_i) g(w_i, x_i, \beta).$$

Find the first-stage estimate of  $\beta$  by finding a zero of the exactly identified system of moments

$$\frac{1}{N} \sum_{i=1}^N \psi_i(\hat{\mathbf{P}}, \hat{\mathcal{Q}}, \hat{\beta}^{(1)}) = 0.$$

**Step 2** Given a solution from the first stage  $\hat{\beta}^{(1)}$  evaluate the conditional variances and expectations of the moment equation  $E \left[ g(w, x, \hat{\beta}^{(1)}) | w_2, z, x \right]$  and  $\text{Var} \left( g(w, x, \hat{\beta}^{(1)}) | w_2, z, x \right)$ . Using these estimates, construct the optimal instrument matrix  $\hat{\mathcal{M}}(x)$  using the formula from Theorem 2. From matrix  $\hat{\Omega}$  construct the optimal weighting matrix and find the second stage estimate as a solution to a system of non-linear equations generated by the empirical moments:

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \psi_i(\hat{\mathbf{P}}, \hat{\mathcal{Q}}, \hat{\beta}^{(2)}) &= \frac{1}{N} \sum_{i=1}^N \sum_{k=p}^{p+M-1} \left[ d_k^{w_{2i}} d_{k-p+1}^{z_i} \right. \\ &\quad \left. - \frac{\hat{\mathcal{Q}}_{k-p+1,i}}{\hat{\mathcal{Q}}_{M,i}} d_k^{w_{2i}} d_M^{z_i} \right] \hat{\mathcal{M}}(x) \hat{\zeta}_p(w_{2i}, x_i) g(w_i, x_i, \hat{\beta}^{(2)}) = 0. \end{aligned} \tag{5}$$

Equation (6) is sample analog of the population moment condition (3) and delivers a consistent estimator which is also asymptotically normal under the regularity conditions which I provide in Section 4.3.

To form the empirical moment equation, distributions and conditional expectations which are not parameterized should be estimated non-parametrically. In this paper I suggest a series-based estimation procedure for non-parametric functions. Applied aspects of semiparametric estimation using sieves are considered in Chen (2007). Consider linear functional space  $\mathcal{H}_L$  generated by a set of known basis functions  $\{\eta_i(\cdot)\}_{i=1}^L$ . I assume that the sequence of functional spaces  $\mathcal{H}_L$  is dense in the functional space containing the non-parametrically specified marginal distributions in the model, and consider uniform approximations over  $x \in \mathcal{X}$ . Also consider a truncation sequence  $k_N$  for the sample of i.i.d. data of size  $N$ . This truncation sequence is

chosen adaptively such that  $k_N \rightarrow \infty$  and  $\frac{k_N}{N} \rightarrow 0$  when  $N \rightarrow \infty$ . Define a vector of basis functions and the matrix of their values on the dataset  $\{x_i\}_{i=1}^N$  as

$$\begin{aligned}\eta^{(k_N)}(x) &= (\eta_1(x), \dots, \eta_{k_N}(x)), \\ \mathcal{W} &= (\eta^{(k_N)}(x_1), \dots, \eta^{(k_N)}(x_N)).\end{aligned}$$

In the first stage of estimation procedure I project the estimated functions into linear space  $\mathcal{H}_{k_N}$  using the weighting matrix  $\mathcal{W}$ . For instance, to estimate marginal probabilities of selection of treatment selection rules and treatment choices one can use expressions

$$\widehat{\mathcal{Q}}_m = \sum_{i=1}^N d_m^{z_i} \eta^{(k_N)}(x_i) [\mathcal{W} \mathcal{W}']^{-1} \eta^{(k_N)}(x),$$

and

$$\widehat{\mathbf{P}}_k = \sum_{i=1}^N d_k^{w_{2i}} \eta^{(k_N)}(x_i) [\mathcal{W} \mathcal{W}']^{-1} \eta^{(k_N)}(x).$$

In order to construct the optimal weighting matrix one needs to estimate marginal probabilities of selection variables, obtain a preliminary estimate of parameter of interest to compute elements of the matrix  $\widehat{\Omega}$  and evaluate the Jacobi matrix. For many frequently used moment equations, such as linear, quadratic or quantile moments, the Jacobi matrix can be computed analytically. In cases where an analytic solution is not available, it would be necessary to use additional computations. If the moment function  $g(\cdot)$  is smooth under suitable support conditions the Jacobi matrix can be substituted by a moment of  $\frac{\partial g(\cdot, \beta)}{\partial \beta}$ . However, it is frequently the case that the moment function  $g(\cdot)$  itself is not differentiable. To obtain the numerical derivative for an element of the Jacobi matrix for each  $\beta \in \mathcal{B}$  evaluate the projection

$$\widehat{\varphi}(w_2 = k, x, \beta) = \left( \sum_{i=1}^N d_k^{w_{2i}} \right)^{-1} \sum_{i=1}^N d_k^{w_{2i}} g(w_{1i}, k, x_i, \beta) \eta^{(k_N)}(x_i) [\mathcal{W} \mathcal{W}']^{-1} \eta^{(k_N)}(x).$$

Then compute numerical derivatives

$$\mathcal{J}_{kl} = \frac{\widehat{\varphi}(w_2 = k, x, \beta + \delta_l) - \widehat{\varphi}(w_2 = k, x, \beta - \delta_l)}{2h_\beta},$$

where  $\delta_l = (\delta_{lk})_{k=1}^{\dim(\beta)}$  with  $\delta_{ll} = 1$  and  $\delta_{lk} = 0$  for  $k \neq l$ . The full Jacobi matrix is built from elements  $\mathcal{J}_{kl}$ .

Using expression for the optimal instrument and non-parametrically estimated probabilities, I form the empirical moment condition. The estimate for the finite-dimensional parameter  $\beta$  is obtained by solving the exactly identified system of non-linear equations defined by the optimally weighted empirical moment (5).

## 4.2 Projection-based estimator

As an alternative to the propensity weighting procedure described in the previous section, one can use direct estimation of population moment conditions from conditional expectations.

The identification condition can be explicitly used in estimation. Equation (4) demonstrates that it is possible to redefine the moment condition for compliers in terms of the population moment condition. Estimation procedure can then be organized by using this expression directly: substitution of conditional expectation with their non-parametrically estimated counterparts produces the empirical moment condition.

In the previous section it was shown that the conditional moment equation for the treatment outcome for compliers can be expressed as:

$$E \left\{ \sum_{k=p}^{p+M-1} \left( \frac{\mathcal{P}_{k-p+1}(k)d_k^{w_2}}{P_{>(p)}} E[\bar{g} | w_2 = k, z = k - p + 1, x] - \frac{\mathcal{P}_M(k)d_k^{w_2}}{P_{>(p)}} E[\bar{g} | w_2 = k, z = M, x] \right) \right\} = 0.$$

Then defining the moment function

$$\mu(\mathbf{P}, \mathcal{Q}, \beta) = \sum_{k=p}^{p+M-1} \left( \frac{\mathcal{P}_{k-p+1}(k)d_k^{w_2}}{P_{>(p)}} E[\bar{g} | w_2 = k, z = k - p + 1, x] - \frac{\mathcal{P}_M(k)d_k^{w_2}}{P_{>(p)}} E[\bar{g} | w_2 = k, z = M, x] \right),$$

for the optimal choice of weighting matrix parameter  $\beta$  can be efficiently estimated by solving a system of non-linear equations

$$\frac{1}{N} \sum_{i=1}^N \mu_i \left( \hat{\mathbf{P}}, \hat{\mathcal{Q}}, \hat{\beta}^{\text{proj}} \right) = 0 \quad (6)$$

for  $\hat{\beta}^{\text{proj}}$ . For this estimator it is necessary to use the efficient weighting matrix  $\mathcal{M}(x)$  to construct an optimally projected moment function  $\bar{g}$ . In order to compute this weighting matrix one

needs a preliminary consistent estimate of the coefficients  $\beta$ . This estimate can be obtained, for instance, from the first-stage estimation with an arbitrary non-singular weighting matrix.

Non-parametric components of the mode are estimated using the series expansion as in the case of the inverse probability-weighted estimator. To project the estimated functions into linear space  $\mathcal{H}_{k_N}$  I use the weighting matrix  $\mathcal{W}$  defined above. Then, to estimate conditional expectations of interest for a particular  $\beta$  I write

$$\widehat{E}[\bar{g} | w_2 = k, z = m, x] = \left( \widehat{\mathcal{P}}_m(k) \widehat{\mathcal{Q}}_m \right)^{-1} \sum_{i=1}^N d_k^{w_2} d_m^{z_i} \bar{g}(w_{1i}, w_{2i}, x_i, \beta) \eta^{(k_N)}(x_i) [\mathcal{W} \mathcal{W}']^{-1} \eta^{(k_N)}(x).$$

To derive the population proportion of compliers, I estimate

$$\widehat{\mathcal{P}}_m(k) = \left( \widehat{\mathcal{Q}}_m \right)^{-1} \sum_{i=1}^N d_k^{w_2} d_m^{z_i} \eta^{(k_N)}(x_i) [\mathcal{W} \mathcal{W}']^{-1} \eta^{(k_N)}(x).$$

Then, using the identification results from Section 2.1, I estimate the probability of compliers as:

$$\widehat{P}_>(p) = \sum_{k=1}^p \left( \widehat{\mathcal{P}}_1(p) - \widehat{\mathcal{P}}_2(p) \right).$$

Finally, I substitute the non-parametrically estimated quantities into the empirical moment equation. By solving the exactly identified system (6), I find the estimate of the finite-dimensional parameter  $\beta$ .

Asymptotic theory for both the propensity score-weighted estimator and for the projection based estimator is developed in the next section.

### 4.3 Consistency and asymptotic distribution

In this section I provide sufficient conditions for consistency and asymptotic normality of the semiparametric estimates provided in the previous sections. The conditions provided in this section build on general results for consistency and asymptotic normality of semiparametric M-estimators. These conditions come from applications of the theory of empirical processes, for instance, in research by Pollard (1990), Bickel, Klaassen, Ritov, and Wellner (1993), and van der Vaart and Wellner (1996).

The structure of the conditions for consistency and asymptotic normality of estimators that I propose builds on the existing literature especially in application to semiparametric M-estimation

(e.g. Andrews (1994), Newey (1994), Ai and Chen (2003), Chen, Linton, and Van Keilegom (2003)). Assumption 2 provides sufficient conditions for consistency of the semiparametric estimates, while Assumptions 3 and 4 are sufficient for asymptotic normality.

To state the regularity properties of the model I make use of the following norm definitions

$$\|h(x, \beta)\|_{\infty, \omega} = \sup_{x \in \mathcal{X}, \beta \in \mathcal{B}} \left| h(x, \beta) (1 + \|x\|^2)^{-\omega/2} \right| \quad \text{and} \quad \|h\|_{P, r} = (E_P |h|^r)^{1/r},$$

for some numbers  $\omega, r > 0$  and a probability measure  $P$  (e.g. Chen, Linton, and Van Keilegom (2003) and van der Vaart and Wellner (1996)). The class of moment functions  $\varphi_p(w, x; \beta, \kappa)$  is indexed by  $\beta \in \mathcal{B}$  and an infinite-dimensional parameter  $\kappa \in \mathcal{F}$ , the set of conditional expectations and probabilities that have to be estimated nonparametrically is such that

$$\Phi = \{\varphi_p(\cdot; \kappa, \beta), \beta \in \mathcal{B}, \kappa \in \mathcal{F}\}.$$

The true function are superscripted by 0.

**Assumption 2** *The following conditions hold:*

1. *There exists a collection of measurable functions  $F(x, w_2) \geq 0$  for  $w_2 = p, \dots, p+M-1$  such that  $|\varphi_p(w_2, x; \cdot)| \leq F(x, w_2)$  for all  $x \in \mathcal{X}$ , and  $\varphi_p(\cdot; \beta, \kappa) \in \Phi$ , and  $E[F^2(x, w_2)] < \infty$*

2. *Functions*

$$\sup_{\varphi_p \in \Phi} \left| \frac{1}{n} \sum_{i=1}^n \alpha_i \varphi_p(w_2, x_i; \beta, \kappa) \right|$$

*are measurable for all fixed sequences  $\alpha_i = \pm 1$ ,  $w_2 = p, \dots, p+M-1$  and every  $n \in \mathbb{N}$ .*

3.  *$\sup_R N(\epsilon \|F\|_{R,1}, \Phi, \|\cdot\|_{R,1}) < \infty$ , for every  $\epsilon > 0$ . Here  $R$  are probability measures and  $N(\cdot, \Phi, \|\cdot\|_{R,1})$  is the covering number for the class of functions  $\Phi$ .*

4.  *$\|\hat{\mathcal{A}}(k, \cdot) - \mathcal{A}^0(k, \cdot)\|_{\infty, \omega} \xrightarrow{P} 0$ , for  $k = p, \dots, p+M-1$  and  $\omega > 0$ .*

5. *There exists a point  $\beta^0 \in \text{int}(\mathcal{B})$  such that for all  $\epsilon > 0$*

$$\inf_{\beta: d(\beta, \beta^0) > \epsilon} E[\mathcal{A}(w_2, x) \varphi_p(w_2, x; \beta, \kappa^0)] > 0 = E[\mathcal{A}(w_2, x) \varphi_p(w_2, x; \beta^0, \kappa^0)].$$

Assumptions 2.1-2 assure that the weighted moment condition is *P-Glivenko-Cantelli*. The last condition is the identification condition for Z-estimators. Primitive conditions on the moment

function  $g(\cdot)$  implying the envelope and covering number conditions are given, for example in Chen, Linton, and Van Keilegom (2003) and require Hölder or uniform  $L_r(P)$  continuity (or both) of the moment function.

**Theorem 3** *Under Assumption 2,  $\hat{\beta} - \beta_0 = o_p(1)$ .*

Next I provide two sets of assumptions assuring asymptotic normality of the obtained estimator. The first group of assumptions assures that the estimated moment functions are asymptotically linear.

**Assumption 3** *The following conditions hold:*

1. *Using the notation of Assumption 2, assume that*

$$\int_0^1 \sqrt{\log \sup_R N(\epsilon \|F\|_{R,1}, \Phi_\delta, \|\cdot\|_{R,1})} d\epsilon < \infty,$$

*for some  $\delta > 0$  such that  $\Phi_\delta = \Phi \cap B_{\delta,\omega}$  where  $B_{\delta,\omega}$  is an  $\omega$ -weighted Hölder ball centered at  $(\beta^0, \kappa^0)$ .*

2. *Moment function  $E[\mathcal{A}(w_2, x; \kappa) \varphi_p(w_2, x; \beta, \kappa)]$  is differentiable at  $\beta^0$  uniformly in  $\kappa \in \mathcal{F}$ , differentiable in  $\kappa$  in  $B_{\omega,\delta}$  for some  $\delta > 0$  and  $L_2(P)$  continuous at  $(\beta^0, \kappa^0)$ .*
3.  *$\|\hat{\mathcal{Q}}(\cdot) - \mathcal{Q}^0(\cdot)\|_{\infty,\omega} = o_p(n^{-1/4})$ ,  $\|\hat{\mathcal{A}}(\cdot, k) - \mathcal{A}^0(\cdot, k)\|_{\infty,\omega} = o_p(n^{-1/4})$ , for  $k = p, \dots, p + M - 1$ .*

The second group of assumptions assures the normality of errors associated with the estimation of auxiliary parameters, such as marginal probabilities and weights.

In the next group of assumptions I provide conditions which assure that error associated with re-weighting the moment condition by  $\mathcal{Q}_k$  is normal. For this purpose define

$$\delta_m^0(x) = \frac{1}{\mathcal{Q}_M} E[d_k^{w_2} \tilde{g} | z = M, x] + \sum_{k=p}^{p+M-1} \frac{\mathcal{Q}_{k-p+1}}{\mathcal{Q}_M^2} E[d_k^{w_2} \tilde{g} | z = M, x].$$

Define projections of  $\delta_m^0(x)$  and  $\mathcal{Q}_m^0(x)$  on linear functional space  $\mathcal{H}_{k_N}$  for  $m = 1, \dots, M - 1$ :  $\delta^{k_N}(x)$  and  $\mathcal{Q}_m^{k_N}(x)$ . These projections are easily computed using standard formulas:

$$\mathcal{Q}_m^{k_N}(X) = q^{k_N}(X) (E[\mathcal{W}\mathcal{W}'])^{-1} E q^{k_N}(X) \mathcal{Q}_m^0(X)',$$

$$\delta_m^{k_N}(X) = q^{k_N}(X) (E[\mathcal{W}\mathcal{W}'])^{-1} E q^{k_N}(X) \delta(X)'$$

**Assumption 4** *The following conditions hold:*

$$\begin{aligned}
& NE \left[ \|\delta_m^0(X) - \delta_m^{k_N}(X)\|^2 \right] \cdot E \left[ \|\mathcal{Q}_m^0(X) - \mathcal{Q}_m^{k_N}(x)\|^2 \right] \rightarrow 0. \\
& E \left[ \|\delta_m^{k_N}(X) (\mathcal{Q}_m^0(X) - \mathcal{Q}_m^{k_N}(x))\|^2 \right] \rightarrow 0. \\
& E \delta_m^0(X) q^{k_N}(X)' \left( (\mathcal{W}'\mathcal{W}/N)^{-1} - (E\mathcal{W}'\mathcal{W}/N)^{-1} \right) \sum_{i=1}^N q^{k_N}(X_i) (d_m^{z_i} - \mathcal{Q}_m^{k_N}(X_i)) / N = o_p(1)
\end{aligned}$$

This assumption implies that the approximation error due to coarseness of space  $\mathcal{H}_{k_N}$  for sufficiently large sample sizes, does not exceed the estimation error.

**Theorem 4** *Under Assumptions 1-4 M-estimate for Euclidean parameter  $\beta$  is consistent, asymptotically normal and achieves the semiparametric efficiency bound. In other words:*

$$\sqrt{N} \left( \hat{\beta} - \beta \right) \xrightarrow{d} N \left( 0, V \left( \hat{\beta} \right) \right).$$

for  $V \left( \hat{\beta} \right)$  given in Theorem 2.

The proofs of the theorems in this section are in the Appendix A.4 and follow immediately from the assumptions.

## 5 Treatment effects in the generalized LATE framework

### 5.1 Definition of treatment effects for compliers

An important application of the moment-based model for generalized compliers is estimation of treatment effects. In this section I set up the general problem of estimation of treatment effects defined by a set of unconditional moment equations defined in terms of unobservable outcomes. My model includes as special cases average treatment effects, average treatment effects on the treated, quantile treatment effects and more general non-linear effects for compliers.

Models considered in this section are structurally different from the conditional moment model considered before. In the conditional moment setup it assumed that the moment restriction holds for all values of the conditioning variables: treatment choice and covariates. In the treatment effect models it is assumed that the moment condition is valid for the entire population, averaged over values of treatments and covariates. However, I will show further, treatment

effect models can be represented as a weighted conditional model considered in the previous sections. This will allow me to use the results that were derived for the conditional model to the models of treatment effects.

An object of interest of the treatment studies with multiple treatments, as noted in Frolich (2004), is the effect of one treatment program relative to another. This defines the average treatment effect as

$$ATE_{mk} = E \left[ Y^m - Y^k \right].$$

Under the generalized LATE assumption it is possible to define the average treatment effect for compliers:

$$ATEC_{mk} = E \left[ Y^m - Y^k \mid p = S_1 < \dots < S_M \right].$$

Similarly, it is possible to define the treatment effect for treated compliers as:

$$ATTC_{mk} = E \left[ Y^m - Y^k \mid w_2 = l, p = S_1 < \dots < S_M \right].$$

The notion of the quantile treatment effect also admits generalization to multiple endogenous treatments. The quantile treatment effect model in case of multiple treatments has been considered in Cattaneo (2007) and defines a vector of moment restrictions on quantiles of treatment outcome variables. In the simple case the linear quantile treatment effect for compliers for quantile  $\tau$  relative to treatment program  $p$  can be defined as a solution to a system of population moment equations

$$E \left[ \mathbf{1} \{ Y_p \leq \beta_p \} - \tau \mid p = S_1 < \dots < S_M \right] = 0,$$

$$\text{and } E \left[ \mathbf{1} \{ Y_k \leq \beta_p + QTEC_{pk} \} - \tau \mid p = S_1 < \dots < S_M \right] = 0, \text{ for } k > p.$$

In this case  $QTEC_{pk}$  measures the difference between conditional quantiles of distributions of outcomes  $Y_k$  and  $Y_p$  for compliers. Similarly, one can define the quantile treatment effect for treated compliers by conditioning on a particular value of the treatment choice.

The set of treatment effects that can potentially be estimated depends on the structure of the instrument and treatment variables in the model. The reason for this limitation is that for

a particular group of compliers some treatment outcomes can be never observed. For instance, if the considered group of compliers contains  $p = S_1 < \dots < S_M = p + M - 1$ , then the outcome  $Y_{p+M}$  cannot be observed because it cannot be generated by any of the treatment choices from  $S_1$  to  $S_M$ . In this group of compliers it will be possible to estimate treatment effects only for outcomes  $Y_p$  to  $Y_{p+M-1}$ .

One notable exception where it becomes possible to identify treatment effects for all outcomes, is the case where both the instrument and treatment variables can take the same number of values. In this case there is only one group of compliers  $1 = S_1 < \dots < S_M = M$  for which all possible outcomes  $Y_1$  to  $Y_M$  are observed. An example of the estimation procedure for such a case is given in Section 6 of this paper where I study the effect of job attrition on hourly wage.

It also possible to consider treatment effects which are not specified for a particular group of compliers. For instance, the average treatment effect for compliers can be defined as

$$\text{ATEC}_{mk} = E \left[ Y^m - Y^k \mid S_1 < \dots < S_M \right].$$

This version of the average treatment effect can be computed by weighting the average treatment effect for particular groups of compliers by probabilities of observing these groups of compliers. Given that  $Y_m$  is observed only for groups of compliers where  $m - M + 1 \leq S_1 \leq m$ , this should be taken into account when one computes the total effect from effects for particular complier groups.

## 5.2 Efficient estimation of treatment effects for compliers

The types of treatment effects that I described above can be considered as a special case of a general class of separable moment models. This class contains moment equations which are separable with respect to the unobservable outcomes:

$$\varphi_p(\beta) = E \left[ \sum_{k=p}^{p+M-1} m_k(Y_k, x, \beta) \mid p = S_1 < \dots < S_M \right] = 0. \quad (7)$$

I consider the problem of estimation of the finite dimensional parameter  $\beta \in \mathcal{B}$ . This structure of the model generates extensions for conventional treatment effect models to the generalized LATE and multi-valued treatment effect setting. For instance, one can define the average treatment effect comparing all treatment selection rules to selection rule  $p$  by constructing moment

functions in (7)  $m_j(Y_j) = \left(m_j^{(1)}(Y_j), \dots, m_j^{(M)}(Y_j)\right)'$ , where superscript indicates a particular element of the moment vector, as:

$$m_p^{(i)}(Y_p) = -Y_p, \quad m_{p+k}^{(k+1)}(Y_{p+k}) = Y_{p+k} - \text{ATEC}_{p+k},$$

$$\text{and } m_{p+k}^{(j)}(Y_{p+k}) = 0, \text{ for } j \neq k+1, \text{ and } i, j = 1, \dots, M.$$

The model can also be designed to define the generalized quantile treatment effects. In this case the moment functions can be rewritten to take quantile structure into account

$$m_p^{(i)}(Y_p) = \mathbf{1}\{Y_p \leq \beta_p\} - \tau, \quad m_{p+k}^{(k+1)}(Y_{p+k}) = \mathbf{1}\{Y_{p+k} \leq \beta_p + \text{QTEC}_{pk}\} - \tau,$$

$$\text{and } m_{p+k}^{(j)}(Y_{p+k}) = 0, \text{ for } j \neq k+1.$$

I assume that moment vector (7) exactly identifies parameter vector  $\beta$  (which is the case for the multi-valued ATEC and QTEC concepts considered above). Application of Bayes's rule transforms the original conditional equation to the unconditional one in the form<sup>2</sup>:

$$E \left[ P_{>}(p) \sum_{k=p}^{p+M-1} \mathcal{Q}_{k-p+1} E \left[ m_k(Y_k, x, \beta) \mid w_2 = k, x, p = S_1 < \dots < S_M \right] \right] = 0.$$

This equation can be written in the form that contains only observable marginal probabilities and re-weighting the conditional moment equation for compliers:

$$E \left[ P_{>}(p) \sum_{k=p}^{p+M-1} \frac{\mathcal{Q}_{k-p+1} d_k^{w_2}}{\mathbf{P}_k} \times E \left[ \sum_{k=p}^{p+M-1} d_k^{w_2} m_k(w_1, x, \beta) \mid w_2, x, p = S_1 < \dots < S_M \right] \right] = 0.$$

If the original moment equation is over-identified, then one can produce an exactly identified system of moments by multiplying the original equation by a constant matrix  $\mathcal{M}$ . Denoting

$$\mathcal{A}(w_2, x) = P_{>}(p) \sum_{k=p}^{p+M-1} \frac{\mathcal{Q}_{k-p+1} d_k^{w_2}}{\mathbf{P}_k} \mathcal{M},$$

---

<sup>2</sup>Alternatively, to derive this result one can use the fact that  $Y_k \perp W_2 \mid p = S_1 < \dots < S_M, X$ . This is a straightforward feature of conditioning on the subset of compliers. In fact, if  $W_2$  moves from  $p+k_1$  to  $p+k_2$  the outcome for compliers moves from  $Y_{p+k_1}$  to  $Y_{p+k_2}$  (in the subsample of compliers one treatment outcome is generated by exactly one treatment selection rule). Therefore, effectively in the subset of compliers the treatment variable and the instrument are functionally dependent:  $W_2 = Z + p - 1$ . As  $Y_k \perp Z \mid X$ , then  $Y_k \perp W_2 \mid p = S_1 < \dots < S_M, X$ .

and

$$g(w, x, \beta) = \sum_{k=p}^{p+M-1} d_k^{w_2} m_k(w_1, x, \beta),$$

one can redefine the original problem for the unconditional moment equation in the same form that I used for the conditional moment model:

$$E[\mathcal{A}(w_2, x) E[g(w, x, \beta) \mid w_2, x, p = S_1 < \dots < S_M]] = 0.$$

Both the structure of the efficiency bound and the efficient estimation procedure in this model are similar to the case of conditional moment equation. In the following theorem I describe the structure of the semiparametric efficiency bound in this model.

**Theorem 5** *In the model given by the general moment condition (7) the efficient influence function, corresponding to finite-dimensional parameter  $\beta$  can be expressed as:*

$$\begin{aligned} \Psi = -J^{-1}\Phi(w, x, z) = & -J^{-1} \left\{ \sum_{k=p}^{p+M-1} \left( d_k^{w_2} d_{k-p+1}^z - \frac{\mathcal{Q}_{k-p+1} d_k^{w_2} d_M^z}{\mathcal{Q}_M} \right) (m_k(w_1, x, \beta) - \bar{m}_k) \right. \\ & - \sum_{k=p}^{p+M-1} \left( d_{k-p+1}^z - \frac{\mathcal{Q}_{k-p+1} d_M^z}{\mathcal{Q}_M} \right) \mathcal{P}_{k-p+1}(k) (E[m_k(w_1, x, \beta) \mid w_2 = k, z = k - p + 1] - \bar{m}_k) \\ & + P_{>}(p) \sum_{k=p}^{p+M-1} \bar{m}_k \left( d_{k-p+1}^z - \mathcal{Q}_{k-p+1} \right) + \sum_{k=p}^{p+M-1} \mathcal{Q}_{k-p+1} \bar{m}_k \sum_{j=1}^p \left( d_1^z \frac{d_j^{w_2 - \mathcal{P}_1(j)}}{\mathcal{Q}_1} - d_2^z \frac{d_j^{w_2 - \mathcal{P}_1(j)}}{\mathcal{Q}_2} \right) \\ & \left. + P_{>}(p) \sum_{k=p}^{p+M-1} \mathcal{Q}_{k-p+1} \bar{m}_k - E \left[ P_{>}(p) \sum_{k=p}^{p+M-1} \mathcal{Q}_{k-p+1} \bar{m}_k \right] \right\}. \end{aligned}$$

In the expression for the influence function I use notation

$$\bar{m}_k = E[m_k(w_1, x, \beta^0) \mid w_2 = k, x, p = S_1 < \dots < S_M].$$

The next step is the optimal choice of the weighting matrix. The Jacobi matrix for this model takes the form:

$$J = \mathcal{M} \frac{\partial \varphi_p(\beta)}{\partial \beta'}.$$

The semiparametric efficiency bound is associated with the optimal choice of the weighting matrix and can be expressed using the variance of component  $\Phi(\cdot)$  in the efficient influence function:

$$V(\hat{\beta}) = \left( \frac{\varphi_p(\beta^0)}{\partial\beta} E[\Phi(w, x, z)\Phi(w, x, z)']^{-1} \frac{\varphi_p(\beta^0)}{\partial\beta'} \right)^{-1}.$$

The optimal instrument matrix  $\mathcal{M}$  is fixed in this case and can be computed as

$$\mathcal{M}^* = \frac{\partial\varphi_p(\beta^0)}{\partial\beta} E[\Phi(w, x, z)\Phi(w, x, z)']^{-1}.$$

As matrix  $\mathcal{M}^*$  is constant, there will be no need to compute weighting matrices depending on covariates. A sample analog that defines an estimate of  $\beta$  (in the exactly identified case) is given by

$$\frac{1}{N} \sum_{i=1}^N \left[ \sum_{k=p}^{p+M-1} \frac{d_k^{w_{2i}}}{\hat{\mathbf{P}}_k} \hat{E} \left\{ \left( d_k^{w_2} d_{k-p+1}^z - d_k^{w_2} d_M^z \frac{\mathcal{Q}_{k-p+1}}{\mathcal{Q}_M} \right) m_k(w_1, x, \beta) \mid x_i \right\} \right] = 0.$$

The estimation procedure combines the projection-based approach with inverse probability weighting. First, marginal probabilities  $\mathbf{P}_k$  and  $\mathcal{Q}_k$  are evaluated. Second, re-weighted moment vectors  $m_k(\cdot)$  are projected on the covariates. Third, moments are weighted by marginal probabilities  $\mathbf{P}_k$  and averaged over the sample. In case of an overidentified moment equation, parameters can be estimated using the standard two-step GMM approach. This becomes possible because in case of unconditional moment model the optimal weighting matrix is constant. In the first step, any non-singular weighing matrix  $\mathcal{M}$  can be used to construct a re-weighted empirical moment equation. In the second step matrix  $\mathcal{M}^*$  is computed from the variance and the Jacobi matrix of the moment equation using the first-step parameter estimate.

### 5.3 Treatment effects for treated compliers

In the model with a separable moment equation it is possible to define the treatment effect for treated compliers. It will be defined by moment equation

$$\varphi_p(\beta) = E \left[ \sum_{k=p}^{p+M-1} m_k(Y_k, x, \beta) \mid w_2 = l, p = S_1 < \dots < S_M \right] = 0, \quad (8)$$

where  $p < l < p + M - 1$ . This model can be specialized to cases of the average treatment effect for treated compliers and the quantile treatment effect for treated compliers by choosing the moment functions similarly to the unconditional case above.

Equivalently to the case of the unconditional treatment effect for compliers, this equation can be redefined as a weighted conditional moment equation

$$E \left[ P_{>}(p) \mathcal{Q}_l \sum_{k=p}^{p+M-1} \frac{d_k^{w_2}}{\mathbf{P}_k} E \left[ \sum_{k=p}^{p+M-1} d_k^{w_2} m_k(w_1, x, \beta) \middle| w_2, x, p = S_1 < \dots < S_M \right] \right] = 0.$$

This makes the problem analogous to the problem of estimation of conditional moment equation, for which I offered estimation methods in Section 4. To see this, denote

$$g(w, x, \beta) = \sum_{k=p}^{p+M-1} d_k^{w_2} m_k(w_1, x, \beta),$$

and

$$\mathcal{A}(w_2, x) = P_{>}(p) \sum_{k=p}^{p+M-1} \frac{d_k^{w_2} \mathcal{Q}_l}{\mathbf{P}_k},$$

which makes the moment equation take exactly the same form as the weighted moment equation in Section 4.

## 6 Empirical Application

In this section I make use of the generalized LATE model to study the effect of job attrition on hourly wages at the primary job for past welfare applicants in Florida. The wage variable in my dataset is censored at zero which makes a linear instrumental variable model not applicable for endogeneity correction. My proposed estimator for the generalized LATE model can capture the non-linearity in the moment condition generated by data censoring and the multi-valued treatment feature of the endogenous job attrition indicator.

### 6.1 Data

The data I use came from a study between years 1994 and 2000 conducted by the Manpower Demonstration Research Corporation (MDRC) among the welfare applicants in Escambia county

in Florida, which includes the city of Pensacola. The MDRC conducted an experiment to determine the benefits of a new welfare support program - the Family Transition Program (FTP). The FTP was designed as an alternative for the existing welfare program - the Aid to Families with Dependent Children (AFDC). Unlike AFDC, the FTP program set rigid limits on the amount of time people could spend on welfare (up to 24 months within any 72 - months period). The FTP also provided a wider array of services to its applicants including training for job-related skills and additional assistance with job search. At the time of application, individuals were randomly assigned to one of the welfare programs -AFDC or FTP. The main sample includes 2,815 heads of single parent households who were randomly assigned to one of the welfare options between May 20, 1994 and February 31, 1996, 1,405 to FTP and 1,410 to AFDC.

The data include individual characteristics 4 years after application to the program. Collected individual responses from 35 minute interviews provide data about family and employment status (including education, job experience, family and dependents, housing, food security, and living conditions), details on welfare receipts and details about jobs that an individual has had during 4 years after for welfare. Administrative records provide individual incomes obtained from the states Unemployment Insurance system, AFDC payments received in the state of Florida, and Food Stamp payments received in the state of Florida.

## **6.2 Empirical model**

Ideally, a job training program associated with unemployment benefits should aim at decreasing both the average rate and the average duration of unemployment. Job attrition is an important determinant of the unemployment duration, and it is associated with the frequency of changing employment status. I use two indicators for the frequency of changing the individual employment status: the number of jobs that an individual has had and the number of moves during the period under consideration (48 months between program participation and a follow-up interview). The first indicator is a better indicator for changes in the employment status. However, one might expect that it has a higher observation error than the second one because it comes from survey data. The second one is a worse approximation for the number of changes in the employment status, but it is also more precise because it comes from the administrative records data. In fact,

a new job of an individual can be close to his or her current location and, thus, this individual does not have to move to be able to work. On the other hand, there could be other reasons for moving such as deterioration of the quality of housing, increase in the family size, etc. Therefore, the number of moves is not so closely related to the job attrition as the number of past jobs. I truncate both variables to take only four values and drop extreme outliers as they are likely coding errors.

The frequency of changing employment status should have a negative impact on an individual's wage. If an individual changes jobs frequently, this slows down acquisition of job-specific skills and, therefore, his productivity will be lower. The frequency of changing employment status can be interpreted then as the lack "seniority" of an individual in a job. The frequency of changing individual employment status can also reflect the individual's job search effort. From the point of view of the job search theory, individuals stay longer at jobs which have higher wages. This implies that an indicator of job attrition will be endogenous.

Most of the demographic variables available from the FTP dataset are discrete. For this reason, I choose to use the wage level  $WAGE_i$  as a dependent variable in the following specification where dependent variables are the indicator of frequency of changing employment status during the period under consideration  $\#JOBS_i$ , and the set of binary individual characteristics  $x$ :

$$WAGE_i = \alpha \#JOBS_i + x_i' \beta + \epsilon_i.$$

Due to the effect from individual job search along with unobserved heterogeneity across individuals correlated with the error term, it cannot be assumed that in this equation  $\epsilon_i \perp \#JOBS_i \mid x_i$ . Summary statistics in the wage equation are presented in Table 1<sup>3</sup> and include an individual's age indicator, marital status, ethnicity, and basic family characteristics.

It is well known, that a censoring problem arises when one estimates the wage regression because it is not possible to observe "potential" wages for individuals who are not currently employed. To reduce heterogeneity in my sample I use only a subset of individuals who indicated that they have had at least one job after participating in the welfare program<sup>4</sup>. Figure 2 demonstrates the distribution of hourly wages in the considered subsample of individuals.

---

<sup>3</sup>Tables and graphs for this section are provided in Appendix B

<sup>4</sup>Individuals who were never employed can have non-economic reasons for not working such as family problems, temporary disability, etc.

Figure 2 shows that a large proportion of wage observations in the sample (more than 80%) is left-censored. Censoring of the dependent variable implies that potential selection will bias the coefficient on attrition if one estimates the model only for individuals who are currently employed. On the other hand, the presence of bottom-coding of the observed wage does not allow one to estimate the model for the entire sample without introducing the bias. In case of exogenous regressors and normal distribution of random disturbance, estimation of coefficients in the wage equations can be performed using the standard *tobit* model. In Table 3, I demonstrate the results of basic OLS regressions and tobit models in the indicated columns. One can see that the estimated coefficient for *#JOBS* in the tobit model is almost ten times larger than the coefficient obtained in the OLS model, indicating that the latter is substantially biased.

When one of the regressors is endogenous and the distribution of error terms is not normal, estimates from the tobit model are not consistent. To provide an estimation procedure which does not rely on the structure of distributions in this problem, I impose a quantile restriction on the error term  $\epsilon_i$  requiring that:

$$Q_{\tau}(\epsilon_i | x_i) = 0,$$

where  $Q_{\tau}(\cdot)$  is a  $\tau$ -quantile of the corresponding conditional distribution. This is the moment restriction analogous to that in the censored quantile regression model considered in Powell (1984).

The censored quantile regression will produce consistent estimates for the case where quantile restriction is valid, the regressor is exogenous, and the distribution of error terms is not specified parametrically. Estimates from the censored quantile regression model for the wage are presented in Table 3. The regressions demonstrated in Table 3 correspond to the moment generated by the 90-th quantile of the wage distribution. The choice of such a high distribution quantile is motivated by Figure 2, demonstrating that a large proportion of the sample is left-censored. Khan and Powell (2001) suggest using higher quantiles in such a case to produce consistent parameter estimates. Standard estimation methods for the censored quantile regression model, however, cannot produce consistent parameter estimates in case of endogenous regressors. In these circumstances, I can effectively use the availability of the dummy for assignment to welfare programs to obtain consistent parameter estimates.

The FTP program develops not only job-search skills but also encourages the development

of skills associated with keeping a job. By the design of the randomized experiment, assignment to the AFDC or the FTP programs was random. With regard to the model for wage this implies that for the assignment dummy  $FTP_i$ :

$$(\#JOBS_i, \epsilon_i) \perp FTP_i | x_i.$$

Table 2 demonstrates the empirical distribution of the number of jobs that an individual has had since entering into the welfare program for two values of the instrument. This table indicates that the distribution of the number of jobs for the AFDC participants first-order stochastically dominates the distribution of the number of jobs for the FTP participants.

### 6.3 The generalized LATE model for analysis of the FTP participants

In my dataset I coded the indicator of job attrition to take four values. I employ an additional assumption that the attrition indicator becomes exogenous for large values of the attrition indicator. The rationale for this assumption is that the wage level is important for transition from unemployment to the first employment, while if an individual has an experience of finding a job, future job search costs are relatively small. This allows me to introduce potential treatment selection choices  $S_k$  for  $k = 1, \dots, 4$  such that the number of values of each  $S_k$  is the same as the number of values of endogenous indicator of attrition  $\#JOBS_i \in \{1, \dots, 4\}$ . Then the outcome variable and attrition indicator can be expressed in terms of the generalized LATE model as:

$$\begin{aligned} W_{1i} &= WAGE_i, \\ W_{2i} &= \#JOBS_i, \\ Z_i &= \begin{cases} 1 + AFDC_i, & \text{if } \#JOBS_i < 3, \\ \#JOBS_i, & \text{if } \#JOBS_i \geq 3. \end{cases} \end{aligned}$$

In this expression I define  $Z$  to be an indicator of participation in the AFDC (which results in a weakly higher frequency of job changes) when the indicator of attrition  $\#JOBS$  moves from value 1 to value 2 using the assignment dummy  $AFDC_i$ . For values 3 to 4 the regressor and the instrument coincide. Therefore, random assignment to one of the welfare programs plays the role of an instrument for transition between the lowest values of the attrition indicator, while the attrition indicator becomes exogenous when it takes larger values and can be used as an

instrument as well. The generated instrumental variable and the endogenous regressor have the same support and there will be only one subset of compliers with  $1 = S_1 < \dots < S_4$ .

For the subsample of compliers the countable regressor  $w_{2i}$  becomes effectively exogenous. Therefore, specifying the moment function to be the  $\tau$ -quantile function:

$$g(w, x, \beta) = Q_\tau(w, x, \beta) = \mathbf{1}\{w_1 \leq \alpha w_2 + x'\beta\} - \tau,$$

the estimated moment equation can be written in the form:

$$E[Q_\tau(w, x, \beta) | w_2, x, 1 = s_1 < \dots < s_4] = 0. \quad (9)$$

I use the inverse probability weighting method to estimate the parameter  $\beta$  in this equation. Given the optimal choice of the weighing matrix  $\mathcal{M}$ , the empirical moment condition can be specified as:

$$\psi(\beta) = \left[ \sum_{k=1}^3 \left( d_k^{w_2} d_k^z - \frac{Q_k(x)}{Q_4(x)} d_k^{w_2} d_4^z \right) + d_4^{w_2} d_4^z - \frac{Q_4(x)}{Q_3(x)} d^{w_2} d_3^z \right] \mathcal{M} \zeta_1(w_2, x) Q_\tau(w, x, \beta).$$

This moment equation is used for estimation of parameters  $\alpha$  and  $\beta$ . Parameters  $\beta$  reflect differences in quantiles of wage distribution across individuals with different binary attributes. Parameter  $\alpha$  defines the (negative) return to an additional job reflected by the wage on the primary job.

## 6.4 Estimation results

The estimation procedure is organized in two steps. In the first step I construct the Jacobi matrix computed at the quantile regression estimates from Table 3 in lieu of  $\mathcal{M}$  to get a preliminary parameter estimate. In the second step, I use the preliminary estimate from the first stage to form the optimal weighting matrix. I use the local linear regression<sup>5</sup> with kernel smoothing to estimate the non-parametric components of the model.

I use a range of quantiles of the wage distribution to estimate the parameters of the wage equation. I present my results in two forms. Table 3 presents the estimates corresponding to the 90% quantile of the wage distribution. The coefficient for the number of moves in the corresponding model is insignificant. However, in the model where the number of past jobs is

---

<sup>5</sup>I choose the bandwidth using cross-validation arguments.

used as a regressor, the coefficient for the past number of jobs is negative. The coefficient for the number of past jobs in the generalized LATE model is significantly higher in absolute terms than the coefficient in the quantile regression. This indicates a substantial presence of endogeneity bias in estimation of the effect of this variable.

I compare the estimates of coefficients for the number of past jobs in Figures 2 and 3. Figure 2 shows the estimate of the effect of the number of past jobs on wage for quantiles from 85% to 99%. The estimate tends to be lower for higher quantiles. However, the standard error is high and confidence bands include zero for a wide range of quantiles. Figure 3 demonstrates the estimated effect of the number of jobs on wage is significantly higher in the LATE context. Asymptotic standard errors obtained from the expression for the semiparametric efficiency bound for the model are substantially lower than those for the ordinary quantile regression except for the top quantiles. The estimates show that the bias of the quantile regression is especially high close to the bottom of the support of the wage distribution in the sample.

## 7 Conclusion

In this paper I develop the generalized local average treatment effect model (LATE) where the endogenous treatment variable can take multiple discrete values, provide semiparametric efficiency bound for a finite-dimensional parameter defined by a semiparametric moment equation, and suggest efficient estimation procedures for this parameter. The generalized LATE model is characterized by an outcome variable, an endogenous discrete regressor, and a discrete-valued instrumental variable which is independent of treatment choices and outcomes given a set of covariates. I prove identification of the constructed model and provide a framework for efficient semiparametric estimation of finite-dimensional parameters of the moment equations under the generalized LATE assumption. I suggest two types of estimation procedures: an inverse probability weighting-based procedure and a conditional expectation projection-based procedure. I prove that both suggested estimators produce semiparametrically efficient estimates. I show that conventional treatment effect models with multi-valued treatments can be analyzed using the generalized LATE approach when the treatment variable is endogenous. In particular, my approach covers such important applications as the quantile treatment effect model and the average treatment effect model with an endogenous multi-valued treatment variable. I apply

my methodology to estimate the effect of job attrition on hourly wage among former applicants for state welfare support in Florida. I find a significant negative effect of job attrition on hourly wage which is underestimated when endogeneity of an attrition indicator is not taken into account, demonstrating that the proposed framework is a powerful tool for correcting endogeneity bias in non-linear models.

## References

- ABADIE, A. (2003): “Semiparametric Instrumental Variable Estimation of Treatment Response Models,” *Journal of Econometrics*, 113(2), 231–263.
- ABADIE, A., J. ANGRIST, AND G. IMBENS (2002): “Instrumental Variables Estimates of the Effects of Subsidized Training on the Quantiles of Trainee Earnings,” *Econometrica*, 70, 91–117.
- AI, C., AND X. CHEN (2003): “Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions,” *Econometrica*, 71(6), 1795–1843.
- ALTONJI, J., AND R. MATZKIN (1997): “Panel Data Estimators for Nonseparable Models with Endogenous Regressors,” Working Paper.
- ANDREWS, D. W. (1994): “Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity,” *Econometrica*, 62.
- ANDREWS, D. W. K., S. BERRY, AND P. JIA (2003): “Confidence Regions for Parameters in Discrete Games with Multiple Equilibria, with an Application to Discount Chain Store Location,” unpublished manuscript, Yale University.
- ANGRIST, J. (2004): “Treatment Effect Heterogeneity in Theory and Practice,” *Economic Journal*, 114(494), 83.
- ANGRIST, J., G. IMBENS, AND D. RUBIN (1996): “Identification of Causal Effects Using Instrumental Variables.,” *Journal of the American Statistical Association*, 91(434).
- BEGUN, J., W. HALL, W. HUANG, AND J. WELLNER (1983): “Information and Asymptotic Efficiency in Parametric-Nonparametric Models,” *The Annals of Statistics*, 11(2), 432–452.
- BERESTEANU, A., AND F. MOLINARI (2006): “Asymptotic Properties for a Class of Partially Identified Models,” *Department of Economics, Cornell University*.
- BICKEL, P. J., C. A. KLAASSEN, Y. RITOV, AND J. A. WELLNER (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag New York, Inc.

- CATTANEO, M. (2007): “Efficient Semiparametric Estimation of Multi-valued Treatment Effects,” *Department of Economics, UC Berkeley, PhD thesis*.
- CHAMBERLAIN, G. (1987): “Asymptotic Efficiency in Estimation with Conditional Moment Restrictions,” *Journal of Econometrics*, 34, 305–334.
- CHEN, X. (2007): “Large Sample Sieve Estimation of Semi-Nonparametric Models,” *Handbook of Econometrics*, 6.
- CHEN, X., H. HONG, AND E. TAMER (2005): “Measurement Error Models with Auxiliary Data,” *Review of Economic Studies*, 72(2), 343–366.
- CHEN, X., H. HONG, AND A. TAROZZI (2007): “Semiparametric efficiency in GMM models with auxiliary data,” *Annals of Statistics Forthcoming*.
- CHEN, X., O. LINTON, AND I. VAN KEILEGOM (2003): “Estimation of Semiparametric Models when the Criterion Function Is Not Smooth,” *Econometrica*, 71(5), 1591–1608.
- CHERNOZHUKOV, V., AND C. HANSEN (2005): “An IV Model of Quantile Treatment Effects,” *Econometrica*, 73(1), 245–261.
- CHERNOZHUKOV, V., H. HONG, AND E. TAMER (2007): “Parameter Set Inference in a Class of Econometric Models,” forthcoming, *Econometrica*.
- CHESHER, A. (2003): “Identification in Nonseparable Models,” *Econometrica*, 71, 1405–1441.
- DUNFORD, N., AND J. T. SCHWARZ (1958): *Linear Operators. Part I: General Theory*. Wiley.
- FIRPO, S. (2006): “Efficient Semiparametric Estimation of Quantile Treatment Effects,” *Econometrica*.
- FROLICH, M. (2004): “Programme Evaluation with Multiple Treatments,” *Journal of Economic Surveys*, 18(2), 181–224.
- FROLICH, M. (2006): “Nonparametric IV estimation of local average treatment effects with covariates,” *Journal of Econometrics*, 139(1), 35–75.
- HAHN, J. (1998): “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,” *Econometrica*, 66(2), 315–331.
- HIRANO, K., AND G. IMBENS (2004): “The Propensity Score with Continuous Treatments,” *Missing Data and Bayesian Methods in Practice*.
- HIRANO, K., G. IMBENS, AND G. RIDDER (2003): “Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score,” *Econometrica*, 71(4), 1161–1189.

- HONG, H., AND D. NEKIPELOV (2007): “Semiparametric Efficiency in Nonlinear LATE Models,” *Working paper, Stanford University*.
- HONG, H., AND E. TAMER (2003): “Inference in Censored Models with Endogenous Regressors,” *Econometrica*, 71(3), 905–932.
- IMBENS, G., AND W. NEWEY (2002): “Identification and Estimation of Triangular Simultaneous Equations Models without Additivity,” working paper, MIT.
- IMBENS, G., W. NEWEY, AND G. RIDDER (2003): “Mean-squared-error Calculations for Average Treatment Effects,” *Department of Economics, UC Berkeley, unpublished manuscript*.
- IMBENS, G., AND D. RUBIN (1997): “Estimating Outcome Distributions for Compliers in Instrumental Variables Models,” *The Review of Economic Studies*, 64(4), 555–574.
- KHAN, S., AND J. POWELL (2001): “Two-Step Estimation of Semiparametric Censored Regression Models,” *Journal of Econometrics*, 103(1-2), 73–110.
- KOSHEVNIK, Y., AND B. LEVIT (1976): “On a Non-Parametric Analogue of the Information Matrix,” *Theory of Probability and its Applications*, 21, 738–753.
- NEWEY, W. (1990): “Efficient Instrumental Variables Estimation of Nonlinear Models,” *Econometrica*, 58(4), 809–837.
- (1994): “The Asymptotic Variance of Semiparametric Estimators,” *Econometrica*, 62, 1349–82.
- PAKES, A., J. PORTER, K. HO, AND J. ISHII (2005): “Moment Inequalities and Their Application,” Working paper, Harvard University.
- POLLARD, D. (1990): *Empirical Processes: Theory and Applications*. Hayward, CA: Institute of Mathematical Statistics.
- POWELL, J. (1984): “Least absolute deviations estimation for the censored regression model,” *Journal of Econometrics*, 25(3), 303–25.
- RUBIN, D. (1974): “Estimating causal effects of treatments in randomized and nonrandomized studies,” *Journal of Educational Psychology*, 66(5), 688–701.
- SEVERINI, T., AND G. TRIPATHI (2001): “A simplified approach to computing efficiency bounds in semiparametric models,” *Journal of Econometrics*, 102, 23–66.
- VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak convergence and empirical processes*. Springer-Verlag, New York.

# Appendix

## A Proofs

### A.1 Proof of Theorem 1

The logic of the proof is the following. Consider the full structure of the data imposing the weak monotonicity assumption on the treatment selection rules. I characterize the structure of identifying equations and extract the set of linearly independent components in these equations. Then I show that the set of these linearly independent components is isomorphic to the set of identifying equations 1 under convexity.

Consider the following assumption as an alternative assumption to Assumption 1.4.

*1.4' (Monotonicity) Selection rules  $S_k$  for  $k = 1, \dots, M$  are contained in weakly increasing sequences of treatment selection choices, moreover this ranking is preserved almost everywhere in  $\mathcal{X}$ :*

$$\Pr \{S_1 \leq \dots \leq S_M | X = x\} = 1.$$

This assumption is certainly weaker than the convexity assumption. For instance it allows non-zero probabilities for such sequences of treatment selection choices as  $p = S_1 < \dots < S_{k-1} = S_k < \dots < S_M$ , which are not convex. The following equation relates the joint probability of the sequence of latent treatment selection choices to the probability of observable outcomes:

$$\sum_{j_{k-1}=p-1}^p \dots \sum_{j_1=j_2-1}^{j_2} \sum_{j_{k+1}=p}^{p+1} \dots \sum_{j_M=j_{M-1}}^{j_{M-1}+1} P_x(j_1, \dots, j_{k-1}, p, j_{k+1}, \dots, j_M) = \mathcal{P}_k(p, x), \quad (10)$$

I collect all non-zero components of the joint distribution of treatment selection choices into a single vector. In this vector the elements are arranged by sorting first by treatment selection choice  $S_M$ , then  $S_{M-1}$ , etc. Introduce the following notations:

$$a = (P(1, \dots, 1), P(1, \dots, 1, 2), P(1, \dots, 1, 2, 2), \dots, P(2, \dots, 2), P(1, \dots, 1, 3), \dots, P(1, \dots, 1, K), \dots, P(K, \dots, K)),$$

$$\alpha = (\mathcal{P}_M(1), \dots, \mathcal{P}_1(1), \dots, \mathcal{P}_M(K), \dots, \mathcal{P}_1(K)).$$

In these new notations I substitute sparse arrays of conditional probabilities by vectors containing only potentially non-trivial components. In this case system (10) is a usual system of linear equations (defined for each  $x \in \mathcal{X}$ ):

$$Aa = \alpha.$$

Note that identification properties of this system of equation depend on the rank properties of matrix  $A$ .

Consider matrix  $A$ . It is a sparse rectangular matrix with all non-zero entries equal to 1. The first dimension of this matrix is equal to  $MK$  while the second dimension is equal to the number of all weakly monotone paths on the support  $K \times M$ . Except for the binary case the number of monotone sequences of treatment selection choices is significantly larger than  $K M^6$ . Denote this number  $\mathcal{N}_{MK}$ . This clearly suggests that the model with only monotonicity assumption in place is not identified. Next I will describe the "maximal" identified system that can be constructed from identification equations under monotonicity. The approach is to extract the set of basis columns from the matrix of the system of identifying equations.

I perform a set of elementary transformations (non-singular linear transformations) with rows and columns of matrix  $A$ . First, I perform transformations with rows of  $A$  and then with its columns. I index the row of  $A$  corresponding to equation with  $\mathcal{P}_m(k)$  on the right-hand side by  $mk$  with  $m = M, \dots, 1$  (in the descending order) and  $k = 1, \dots, K$ . In the first set of transformations subtract each row  $Mk$  from the subsequent  $M$  rows. Then subtract each row  $1k$  from row  $M(k+1)$  for  $k = 1, \dots, K-1$ . Finally, subtract each row  $(m-1)k$  from row  $mk$  for  $m = 2, \dots, M$ . Note that because these transformations are non-singular, they are represented by a non-singular square  $KM \times KM$  transformation matrix  $\Sigma$ .

I index columns of  $A$   $j_1, \dots, j_M$  by the order of treatment selection rules. Then I use the following iterative procedure. For steps  $m = 2, \dots, M$  in step  $m$  subtract column  $j_1 = p-1, \dots, j_{m-1} = p-1, j_m = p, \dots, j_M = p$  from columns  $j_1 = p-1, \dots, j_{m-1} = p-1, j_m = p, p+k, \dots, p+k$  for  $k = 1, \dots, K-p$ . This transformation is represented by a non-singular square matrix  $\tilde{\Sigma}$  with  $\mathcal{N}_{MK}$  columns. The obtained matrix  $\tilde{A} = \tilde{\Sigma} A \Sigma$  has only  $M(K-1)$  non-zero entries. In row  $mk$  the non-zero entry is indexed  $j_1 = \dots = j_{m-1} = k-1, j_m = \dots = j_M = k$ . Note that the  $M(K-1) \times M(K-1)$  submatrix containing the columns with non-zero elements is the identify matrix. Therefore, these columns represent the basis of interest. This result also suggests that we can identify at most  $M(K-1)$  elements of matrix  $A$ . In fact, as transformations  $\tilde{\Sigma}$  and  $\Sigma$  are non-singular, any vector of probabilities with non-zero entries not corresponding to basis columns of  $A$  will be observationally equivalent to the vector with corresponding zero entries.

Lastly, note that the system of identifying equations for the probabilities under Assumption 1.4 becomes a triangular matrix with  $K-1$  main diagonals and zeros everywhere else. By a system of elementary transformations such matrix can be reduced to a non-singular diagonal matrix with  $(K-1)M \times (K-1)M$ . Therefore the matrix for the system of identifying equations under convexity is isomorphic to the system of basis columns under monotonicity.

---

<sup>6</sup>The number of this paths is determined by multinomial sum of Bernoulli numbers determined by dimensions  $K$  and  $M$  and for  $K, M > 1$  grows at a factorial rate as support increases.

## A.2 Set Estimation with Partial Identification

In cases where the convexity and the separability Assumptions 1.4 and 1.5 are not attractive, one can use a weaker monotonicity assumption 1.4' given in Appendix A.1 to provide a set estimator for the treatment effect parameter.

As is shown in Appendix A.1, in this case neither the joint distribution of treatment selection choices nor the distribution of treatment outcomes is identified. However, one can use set inference to provide bounds on parameters of interest. In this section I briefly discuss the construction of sets of distributions satisfying the identification assumptions. General set estimation and inference methods have been developed in Pakes, Porter, Ho, and Ishii (2005), Andrews, Berry, and Jia (2003), Beresteanu and Molinari (2006), and Chernozhukov, Hong, and Tamer (2007). These methods can be used to translate the mechanism for constructing a set of complier distributions provided in this section to the set of parameters satisfying identification assumptions.

Appendix A.1 provides the details for constructing the system of identifying equations for the joint distribution of the treatment choices. Using notations

$$a = (P(1, \dots, 1), P(1, \dots, 1, 2), P(1, \dots, 1, 2, 2), \dots, P(2, \dots, 2), P(1, \dots, 1, 3), \dots, P(1, \dots, 1, K), \dots, P(K, \dots, K)),$$

$$\alpha = (\mathcal{P}_M(1), \dots, \mathcal{P}_1(1), \dots, \mathcal{P}_M(K), \dots, \mathcal{P}_1(K)),$$

I represent the system of identifying equations (10) in a parsimonious form:

$$Aa = \alpha.$$

Matrix  $A$  has dimensions  $M(K-1) \times \mathcal{N}_{MK}$ , where  $\mathcal{N}_{MK}$  is the number of monotone sequences of treatment choices which contain a particular treatment choice  $S_m = k$ . Similarly, I form a system of identifying equations for the densities of the treatment outcomes for compliers. To proceed I fix the continuous variables  $w_1$  and  $x$ . Denote:

$$b_p = (f(y_p|1, \dots, 1), f(y_p|1, \dots, 1, 2), f(y_p|1, \dots, 1, 2, 2), \dots, f(y_p|2, \dots, 2), P(1, \dots, 1, 3), \dots, f(y_p|1, \dots, 1, K), \dots, f(y_p|K, \dots, K)),$$

$$\beta = (f_M(w_1, 1), \dots, f_1(w_1, 1), \dots, f_M(w_1, K), \dots, f_1(w_1, K)).$$

Then the density of interest solves the system of linear equations

$$A_k b_k * a = (\beta * \alpha)_k, \quad k = 1, \dots, K,$$

where  $A_k$  is the selection rows of matrix  $A$  corresponding to elements  $\mathcal{P}_m(w_1, k)$  for  $m = 1, \dots, M$ ,  $*$  denotes the Hadamard product of vectors, and  $(\beta * \alpha)_k$  is the subvector of elements corresponding to elements  $\mathcal{P}_m(w_1, k)$  for  $m = 1, \dots, M$ .

I will first characterize the set of solutions of systems of equations for  $a$  and  $b_k$  and then impose regularity conditions which will assure that the solutions are proper conditional probabilities and densities. First of all, note that considered systems of equations have at least one solution satisfying Assumptions 1. To describe all other solutions one can use the fact that any solution of linear system can be represented as a particular solution to the non-homogenous system and a linear combination of basis solutions of the homogenous system. As a particular solution is already known, I only need to describe the linear space defined by

$$\left\{ \xi \mid A \xi = 0, \xi \neq 0 \right\}.$$

Matrix  $A$  is sparse with entries equal either zero or one and the structure of its elements has been described in Section 2.3. To characterize solutions of the homogeneous system it is possible to find a sequence of elementary transformation of this matrix to a sparse matrix containing an identity submatrix (an example of such procedure can be found in Appendix A.1). The obtained transformed matrix  $\tilde{A}$  has only  $M(K-1)$  non-zero entries in an  $M(K-1) \times M(K-1)$  identity submatrix.

Transformations of matrix  $A$  to a sparse matrix  $\tilde{A}$  are elementary transformations in which columns and rows of  $A$  are substituted by their non-trivial linear combinations. These transformations can be expressed as:

$$\tilde{A} = \left( \prod_{k=1}^{MK} S_k^1 \right) A \left( \prod_{k=1}^{\mathcal{N}_{MK}} S_k^2 \right),$$

where each matrix  $S_k^i$  is responsible for a single substitution of a row or a column by its non-trivial linear combination with other rows or columns. In this expression  $S_k^1 = I - E^k$  where  $I$  is the identify matrix and  $E^k$  is the matrix with zero entries except for one off-diagonal element, equal to 1. This matrix defines transformation where a particular column of  $A$  is subtracted from a column with a higher index. Matrix  $S_k^2 = I + G^k$ . The structure of  $G^k$  depends on whether  $k$  is even or odd and its non-zero entries are equal to  $-1$  if  $k$  is odd and  $1$  when it is even. As a result,  $S_k^2$  defines either addition or subtraction between two rows of  $A$ .

Solutions to the system of equations defined by  $\tilde{A}$  can be written as a vector  $\tilde{\xi}$  where  $\tilde{\xi}_1 = \xi_{M(K-1)} = 0$  while the rest of the elements can be arbitrary real numbers. Solutions to the transformed uniform linear system (where the right-hand-side is substituted with a zero column) formed by matrix  $\tilde{A}$  are contained in a linear space with basis vectors forming a matrix  $I - H$ , where  $H$  is an upper triangular sparse matrix with non-zero entries equal to 1 and  $-1$ . The general solution to the original uniform system of equations (generated by matrix  $A$ ) can be then written as  $\left( \prod_{k=1}^{\mathcal{N}_{MK}} S_k^2 \right)^{-1} (I - H) \eta$ , where  $\eta \in \mathbb{R}^{\tilde{K}}$  and  $\tilde{K} = M(K-1)$ .

Suppose that  $\tilde{\zeta}^1$  is a vector representing the solution to the exactly identified system of equations for the joint probabilities of treatment choices under Assumptions 1,  $\tilde{\zeta}^2$  is the Hadamard product of  $\tilde{\zeta}^1$  and the vector of conditional densities computed under separability assumption, and  $\tilde{\zeta}^{2k}$  is a vector where  $\tilde{\zeta}_k^{2k} = \tilde{\zeta}_k^2$  and all other entries are zeros. Then the general solution to the system of equations under consideration can be written as:

$$\begin{aligned} a &= \tilde{\zeta}^1 + \left( \prod_{k=1}^{\mathcal{N}_{MK}} S_k^2 \right)^{-1} (I - H) \eta^1, \\ b_k &= \tilde{\zeta}^{2k} + \left( \prod_{k=1}^{\mathcal{N}_{MK}} S_k^2 \right)^{-1} (I - H) \eta^2, \end{aligned} \tag{11}$$

with  $\eta^1, \eta^2 \in \mathbb{R}^{\tilde{K}}$ . Expressions (11) completely characterize the linear space of solutions to the system under consideration.

The last step in characterizing the solutions of the considered linear system is to insure that the obtained vectors are proper distributions. I impose an additional restriction  $\sum_{i=1}^{\tilde{K}} a_i = 1$ , and  $a, b_k \geq 0$ . If  $P_a$  is a projection into simplex in  $\mathbb{R}^{\tilde{K}}$  and  $P_b$  projects into a positive subspace, then solutions can be characterized by:

$$\begin{aligned} a &= \tilde{\zeta}^1 + P_a \left( \prod_{k=1}^{\mathcal{N}_{MK}} S_k^2 \right)^{-1} (I - H) \eta^1, \\ b_k &= \tilde{\zeta}^{2k} + P_b \left( \prod_{k=1}^{\mathcal{N}_{MK}} S_k^2 \right)^{-1} (I - H) \eta^2. \end{aligned} \tag{12}$$

My proposed estimation procedure employs the structure of solution (12). First, estimate a vector of distributions corresponding to the exactly identified case when Assumptions 1 are satisfied. Second, use the structure of matrix  $A$  to define all possible distributions satisfying monotonicity.

At each point of the support of  $w_1$  and  $x$  the density of the outcome distribution for compliers in the set-identified case is equal to the sum of the density obtained under convexity and separability assumptions and a set of "shifts" in the density which do not violate the monotonicity assumption. The latter is computed using a linear projection. Then, for instance, using a grid over the outcome variable one can compute an additional contribution of shifts in the density to the moment equation. This computation has the same order of complexity as computing a one-dimensional integral of an explicitly defined function. The set of solutions to permuted moment equations will generate the estimate of the identified set for finite-dimensional parameter  $\beta$ . When additional restrictions on joint probabilities of treatment selection choices are available, they can be taken into account in the structure of elementary transformations  $S_k^i$ .

### A.3 Proof of Theorem 2

To derive the efficiency bound for the model of interest I first characterize the structure of the tangent set. Consider a particular likelihood decomposition and a particular parametrization path  $\theta$ . Note that conditioning on compliers restricts relevant treatments to  $w_2 \in [p, p + M - 1]$ , while the estimated functions use the data starting from  $w_2 = 1$ . This provides a verification argument for the derived semiparametric score. If  $\phi_\theta(x)$  is the Radon-Nykodym density of  $x$  with the support on  $\mathcal{X}$ , the likelihood function for the data can be written as:

$$f_\theta(w, z, x) = f_\theta(w_1 | w_2, z, x) \prod_{k=1}^K \mathcal{F}_\theta^{d_k^{w_2}}(k, z) \prod_{m=1}^M \mathcal{Q}_{\theta m}^{d_m^z} \phi_\theta(x).$$

For this parametrization it is possible to characterize the score and the tangent set of the model which will be ultimately associated with the Fréchet derivative of the non-parametric likelihood. I take pathwise derivative along the parametrization path. Then the score can be written as:

$$\begin{aligned} S_\theta(w, z, x) &= \sum_{k=1}^K s_\theta(w_1 | w_2 = k, z, x) + \sum_{m=1}^M \sum_{k=1}^{K-1} \dot{\mathcal{P}}_m(k) d_m^z \left[ \frac{d_k^{w_2}}{\mathcal{F}(k, z)} - \frac{d_K^{w_2}}{\mathcal{F}(K, z)} \right] \\ &+ \sum_{m=1}^{M-1} \dot{\mathcal{Q}}_m \left[ \frac{d_m^z}{\mathcal{Q}_m} - \frac{d_M^z}{\mathcal{Q}_M} \right] + s_\theta(x), \end{aligned}$$

where  $s(\cdot | w_2, z, x)$  is the score of observed conditional distribution of  $w_1$ ,  $\dot{\mathcal{Q}}_m(\cdot)$  and  $\dot{\mathcal{P}}_k(\cdot)$  are probabilities for discrete-valued treatment choices and instrument variable  $z$ ,  $s_\theta(x)$  is the score corresponding to  $\phi_\theta(x)$ . The structure of the pathwise derivative demonstrates that the score of the model splits into four uncorrelated components. The expression for the tangent set of the model for conditional distribution moments is given by

$$\begin{aligned} \mathcal{T} &= \left\{ \sum_{k=1}^K s_\theta(w_1 | w_2 = k, z, x) + \sum_{m=1}^M \sum_{k=1}^{K-1} \zeta_{mk}(x) d_m^z \left[ \frac{d_k^{w_2}}{\mathcal{F}(k, z)} - \frac{d_K^{w_2}}{\mathcal{F}(K, z)} \right] \right. \\ &\quad \left. + \sum_{m=1}^{M-1} \xi_m(x) \left[ \frac{d_m^z}{\mathcal{Q}_m} - \frac{d_M^z}{\mathcal{Q}_M} \right] + t(x) \right\}, \end{aligned}$$

where  $E_\theta[s_\theta(w_1 | w_2 = k, z, x) | w_2, z, x] = 0$ ,  $E\{t(x)\} = 0$ , and  $\zeta_{km}(\cdot)$  and  $\xi_k(\cdot)$  are square - integrable functions. In a special case where the support of regressors is infinite, I add one more requirement, that partial sums of series with elements  $E[\xi_m^2(x)] \mathcal{Q}_m$  and  $E[\zeta_{km}^2(x)] \mathcal{P}_m(k)$  converge to finite limits.

Considered conditional moment equation remains valid if it is multiplied by a non-singular matrix depending on  $w_2$  and  $x$ . Therefore, one can define a linear functional  $A$  which maps the conditional moment equation into  $\mathbb{R}^k$  transforming the conditional moment equation to an exactly identified system

of unconditional moments. Such functional can be represented by function  $\mathcal{A}(x, w_2) : \mathcal{X} \times \{0, 1\} \mapsto \mathbb{R}^k$  such that for function  $f(x, w_2)$ :

$$A \circ f = E[\mathcal{A}(x, w_2) f(x, w_2)]. \quad (13)$$

In fact, assuming that  $\varphi(w_2, x, \beta)$  and  $\frac{\partial \varphi(w_2, x, \beta)}{\partial \beta}$  are elements of  $\mathbf{L}_2(\mathcal{X} \times \{0, 1\}, \mathfrak{F}, \mu)$  for each  $\beta \in \mathcal{B}$ , where  $\mathfrak{F}$  is a Borel  $\sigma$ -algebra in the product space  $\mathbb{Z} \times \mathcal{X}$  and  $\mu$  is the corresponding probability measure. In this case the dual space to  $\mathbf{L}_2(\mathbb{Z} \times \mathcal{X}, \mathfrak{F}, \mu)$  will also be  $\mathbf{L}_2(\mathbb{Z} \times \mathcal{X}, \mathfrak{F}, \mu)$  (see Dunford and Schwarz (1958)). As a result, considering weighting functions  $\mathcal{A}(\cdot)$  in  $\mathbf{L}_2(\mathbb{Z} \times \mathcal{X}, \mathfrak{F}, \mu)$  allows one to find the optimal structure of the unconditional moment vector corresponding to the conditional moment equation. In fact, as it is shown in Dunford and Schwarz (1958) any continuous linear functional on  $\mathbf{L}_p(T, \mathfrak{F}, \mu)$  can be represented as (13) for  $\mathcal{A} \in \mathbf{L}_q(T, \mathfrak{F}, \mu)$  for  $p^{-1} + q^{-1} = 1$  and  $1 < p < \infty$ .

For an arbitrary choice of the weighting matrix  $\mathcal{A}(w_2, x)$  moment equation generates an unconditional moment

$$E[E\{\mathcal{A}(w_2, x)g(w, x, \beta) \mid p = s_1 < \dots < s_M, w_2, x\}] = 0.$$

Introduce  $\dim(\beta) \times M$  matrix  $\mathcal{M}(x)$  and use  $\zeta_p(w_2, x)$  to define

$$A(w_2, x) = \mathcal{M}(x)\zeta_p(w_2, x),$$

and similarly to Hong and Nekipelov (2007) choose

$$\mathcal{A}(w_2, x) = P_{>}(p) \sum_{k=p}^{p+M-1} \frac{\mathcal{Q}_{k-p+1} d_k^{w_2}}{\mathbf{P}_k} A(w_2, x).$$

Next, I use parametrization of the likelihood in the moment equation transformed with the weighting matrix  $\mathcal{A}(\cdot)$ . Note that matrix  $\mathcal{A}(\cdot)$  contains a non-parametric component and, thus, will have a non-zero derivative along the parametrization path. The transformed moment condition takes the form :

$$E_\theta[\mathcal{A}_\theta(w_2, x) E_\theta[g(w, x, \beta_\theta) \mid p = s_1 < \dots < s_M, w_2, x]] = 0.$$

Define the Jacobi matrix:

$$J = E \left[ \mathcal{A}(w_2, x) \frac{\partial \varphi_p(w_2, x, \beta)}{\partial \beta'} \right].$$

Following Newey (1990) one can obtain the expression for the directional derivative of  $\beta$  solving a linear system of equations:

$$J \frac{\partial \beta_\theta}{\partial \theta} = -\frac{\partial}{\partial \theta} E_\theta[\mathcal{A}_\theta(w_2, x) E_\theta[g(w, x, \beta_\theta) \mid p = s_1 < \dots < s_M, w_2, x]]. \quad (14)$$

The right-hand side component of equation (14) can be written as:

$$\begin{aligned}
& E \left[ \mathcal{A}(w_2, x) \int g(w, x, \beta) s_{>}(w_1, w_2, p) f_{>}(w_1, w_2, p) dw_1 \right] \\
& + E \left[ \mathcal{A}(w_2, x) s_{\theta}(w_2, x) \int g(w, x, \beta) f_{>}(w_1, w_2, p) dw_1 \right] \\
& + E \left[ \frac{\partial \mathcal{A}_{\theta}(w_2, x)}{\partial \theta} \int g(w, x, \beta) f_{>}(w_1, w_2, p) dw_1 \right].
\end{aligned} \tag{15}$$

In this expression  $s_{>}(\cdot)$  is the score corresponding to the conditional distribution of treatment outcomes for compliers, and  $s_{\theta}(w_2, x)$  is the score corresponding to the joint distribution of regressors in the model. For conditional moment based model, weighting of the moment equation by a function of  $w_2$  and  $x$  keeps the equation valid. Therefore, the second and the third components in the expression for the directional derivative are equal to zero. Denote

$$\tilde{g} = -J^{-1} A(w_2, x) g(w, x, \beta), \quad \text{and} \quad \bar{g} = P_{>}(p) \sum_{k=p}^{p+M-1} \frac{\mathcal{Q}_{k-p+1} d_k^{w_2}}{\mathbf{P}_k} \tilde{g}.$$

This transforms the expression for the directional derivative to

$$\frac{\partial \beta_{\theta}}{\partial \theta} = E \left[ \int \bar{g} s_{>}(w_1, w_2, p) f_{>}(w_1, w_2, p) dw_1 \right].$$

The results for  $K = M = 2$  case are provided in Hong and Nekipelov (2007) and here I assume that  $K > M$ . To derive the score  $s_{>}(\cdot)$  consider the following four cases.

**Case 1:**  $w_2 = 1, p = 1, M = 2$ . In this case, the score takes the form

$$\begin{aligned}
s_{>}(w_1, 1, 1) f_{>}(w_1, 1, 1) &= \frac{\mathcal{P}_1(1)}{\mathcal{P}_{>}(1)} f_1(w_1, 1) s(w_1 | w_2 = 1, z = 1) - \frac{\mathcal{P}_2(1)}{\mathcal{P}_{>}(1)} f_2(w_1, 1) s(w_1 | w_2 = 1, z = 2) \\
&+ \frac{\dot{\mathcal{P}}_{1\theta}(1) \mathcal{P}_{2\theta}(1) - \dot{\mathcal{P}}_{2\theta}(1) \mathcal{P}_{1\theta}(1)}{\mathcal{P}_{>}(1)^2} (f_2(w_1, 1) - f_1(w_1, 1)).
\end{aligned}$$

**Case 2:**  $w_2 = 1, p = 1, M > 2$ . In this case the score is

$$\begin{aligned}
s_{>}(w_1, 1, 1) f_{>}(w_1, 1, 1) &= \frac{\mathcal{P}_1(1)}{\mathcal{P}_{>}(1)} f_1(w_1, 1) s(w_1 | w_2 = 1, z = 1) - \frac{\mathcal{P}_M(1)}{\mathcal{P}_{>}(1)} f_M(w_1, 1) s(w_1 | w_2 = 1, z = M) \\
&+ \frac{\dot{\mathcal{P}}_{1\theta}(1)(f_M(w_1, 1) \mathcal{P}_M(1) - f_1(w_1, 1) \mathcal{P}_2(1)) + \dot{\mathcal{P}}_{2\theta}(1)(f_M(w_1, 1) \mathcal{P}_M(1) - f_1(w_1, 1) \mathcal{P}_1(1))}{\mathcal{P}_{>}(1)^2} - \frac{\dot{\mathcal{P}}_{M\theta} f_M(w_1, 1)}{\mathcal{P}_{>}(1)}.
\end{aligned}$$

**Case 3:**  $w_2 = k, 1 < k < K, M > 2$ . In this case the score is

$$\begin{aligned}
s_{>}(w_1, k, p) f_{>}(w_1, k, p) &= \frac{\mathcal{P}_{k-p+1}(k)}{\mathcal{P}_{>}(p)} f_{k-p+1}(w_1, k) s(w_1 | w_2 = k, z = k - p + 1) \\
&- \frac{\mathcal{P}_M(k)}{\mathcal{P}_{>}(p)} f_M(w_1, k) s(w_1 | w_2 = k, z = M) \\
&+ \frac{\dot{\mathcal{P}}_{k-p+1}(k) f_{k-p+1}(w_1, k) - \dot{\mathcal{P}}_M(k) f_M(w_1, k)}{\mathcal{P}_{>}(p)} - \frac{\mathcal{P}_{k-p+1}(k) f_{k-p+1}(w_1, k) - \mathcal{P}_M(k) f_M(w_1, k)}{(\mathcal{P}_{>}(p))^2} \sum_{j=1}^p (\dot{\mathcal{P}}_1(j) - \dot{\mathcal{P}}_2(j)).
\end{aligned}$$

**Case 4:**  $w_2 = K$ ,  $M > 2$ . In this case the score is

$$\begin{aligned}
s_{>}(w_1, K, K - M + 1) f_{>}(w_1, K, K - M + 1) &= \frac{\mathcal{P}_M(K)}{\mathcal{P}_{>}(K-M+1)} f_M(w_1, K) s(w_1|w_2 = K, z = M) \\
&\quad - \frac{\mathcal{P}_1(K)}{\mathcal{P}_{>}(K-M+1)} f_1(w_1, K) s(w_1|w_2 = K, z = 1) \\
&\quad + \frac{\dot{\mathcal{P}}_M(K) f_M(w_1, K) - \dot{\mathcal{P}}_1(K) f_1(w_1, K)}{\mathcal{P}_{>}(K-M+1)} - \frac{\mathcal{P}_{K-M+1}(k) f_{K-M+1}(w_1, K) - \mathcal{P}_1(K) f_1(w_1, K)}{(\mathcal{P}_{>}(K-M+1))^2} \sum_{j=1}^{K-M+1} \left( \dot{\mathcal{P}}_1(j) - \dot{\mathcal{P}}_2(j) \right).
\end{aligned}$$

Case 1 gives an expression for the score corresponding to the density of outcome distribution for generalized compliers which coincides with the expression for the score in a binary case as in Hong and Nekipelov (2007). This assures consistency of the model considered in this paper with the model for the binary case. Expressions for the score from Cases 1-4 can be summarized as

$$s_{>}(w_1, w_2, p) = \sum_{k=p}^{p+M-1} d_k^{w_2} s_{>}(w_1, k, p).$$

Given these expressions for the score, I look for the efficient influence function as a solution to the functional equation

$$\frac{\partial \beta_\theta}{\partial \theta} = E[\Psi(w, z, x) S_\theta(w, x, z)].$$

Applying the standard projection technique, the efficient influence function is derived in the form:

$$\begin{aligned}
\Psi(w, z, x) &= \sum_{k=p}^{p+M} \left\{ \frac{\mathbf{P}_k d_k^{w_2} d_{k-p+1}^z}{\mathcal{P}_{>}(p) \mathcal{Q}_{k-p+1}} (\bar{g} - E[\bar{g}|w_2 = k, z = k - p + 1]) \right. \\
&\quad \left. - \frac{\mathbf{P}_k d_k^{w_2} d_M^z}{\mathcal{P}_{>}(p) \mathcal{Q}_M} (\bar{g} - E[\bar{g}|w_2 = k, z = M]) \right\} \\
&\quad + \sum_{m=1}^M \sum_{k=1}^{K-1} a_{mk} dz_m \left[ \frac{d_k^{w_2}}{\mathcal{P}_m(k)} - \frac{d_K^{w_2}}{\mathcal{P}_m(K)} \right].
\end{aligned}$$

To find coefficients  $a_{mk}$  in this representation, consider covariance between this expression for the influence function and the semiparametric score of the model. Then, I solve for coefficients that make this expression the same as expression for the directional derivative of finite-dimensional parameter. For instance, suppose that in the expression for the directional derivative, the coefficient for  $\dot{\mathcal{P}}_m(k)$  is  $G_{mk}$ . Define vector  $a^{(m)} = (a_{m1}, \dots, a_{m(K-1)})$  and  $\mathcal{P}^{(m)} = (\mathcal{P}_m(1), \dots, \mathcal{P}_m(K-1))'$ , and  $\gamma^{(m)} = (0, \dots, 0, G_{mk}, 0, \dots, 0)'$ . Then the vector of coefficients solves

$$\left( I + \frac{\mathcal{P}^{(m)} \mathbf{1}'}{\mathcal{P}_m(K)} \right) a^{(m)} = \gamma^{(m)}.$$

Standard inversion technique for matrices of this special kind gives the solution:

$$a_{mk} = \frac{\mathcal{P}_m(k)}{\mathcal{Q}_m} G_{mk} (1 - \mathcal{P}_m(k)),$$

$$a_{mj} = -\frac{\mathcal{P}_m(k)}{\mathcal{Q}_m} G_{mk} \mathcal{P}_m(j).$$

Using separability of the solution and rearranging terms, one can write down the final expression for the efficient influence function in the form:

$$\begin{aligned} \Psi(w, z, x) = & \sum_{k=p}^{p+M-1} \left\{ \frac{\mathbf{P}_k d_k^{w_2} d_{k-p+1}^z}{P_{>}(p) \mathcal{Q}_{k-p+1}} (\bar{g} - E[\bar{g}|w_2 = k, z = k - p + 1]) \right. \\ & - \frac{\mathbf{P}_k d_k^{w_2} d_M^z}{P_{>}(p) \mathcal{Q}_M} (\bar{g} - E[\bar{g}|w_2 = k, z = M]) \\ & + \frac{E[\bar{g}|w_2=k, z=k-p+1] d_{k-p+1}^z (d_k^{w_2} - \mathcal{P}_m(k))}{\mathcal{Q}_m P_{>}(p)} - \frac{E[\bar{g}|w_2=k, z=M] d_M^z (d_k^{w_2} - \mathcal{P}_M(k))}{\mathcal{Q}_M P_{>}(p)} \\ & - \frac{\mathcal{P}_{k-p+1}(k) E[\bar{g}|w_2=k, z=k-p+1] - \mathcal{P}_M(k) E[\bar{g}|w_2=k, z=M]}{\mathcal{Q}_1 (P_{>}(p))^2} d_1^z \sum_{j=1}^p (d_j^{w_2} - \mathcal{P}_1(j)) \\ & \left. + \frac{\mathcal{P}_{k-p+1}(k) E[\bar{g}|w_2=k, z=k-p+1] - \mathcal{P}_M(k) E[\bar{g}|w_2=k, z=M]}{\mathcal{Q}_2 (P_{>}(p))^2} d_2^z \sum_{j=1}^p (d_j^{w_2} - \mathcal{P}_2(j)) \right\}. \end{aligned}$$

This expression can be compressed using identities for conditional expectation:

$$E[\bar{g}|w_2 = k, z = k - p + 1] \mathcal{P}_{k-p+1}(k) = E[\bar{g}|w_2 = k, z = M] \mathcal{P}_M(k),$$

and expression for the weighting matrix  $\mathcal{A}(w_2, x)$ . Then the efficient influence function can be rewritten as

$$\Psi(w, z, x) = \sum_{k=p}^{p+M-1} \left( d_k^{w_2} d_{k-p+1}^z - \frac{\mathcal{Q}_{k-p+1} d_k^{w_2} d_M^z}{\mathcal{Q}_M} \right) \tilde{g} - \sum_{k=p}^{p+M-1} \left( \frac{d_{k-p+1}^z}{\mathcal{Q}_{k-p+1}} - \frac{d_M}{\mathcal{Q}_M} \right) E[d_k^{w_2} d_{k-p+1}^z \tilde{g}].$$

The semiparametric efficiency bound is

$$V(\hat{\beta}) = E[\Psi \Psi'].$$

For further manipulations I denote the first component of the efficient influence function by  $\Psi_1(\cdot)$  and the second one  $\Psi_2(\cdot)$ . Both components have mean zero. The last step of efficiency calculations is to find the optimal weighting matrix  $\mathcal{M}(x)$ . Compute the transformed Jacobi matrix. Denoting

$$m(x) = E \left[ \zeta_p(w_2, x) \frac{\partial \varphi_p(w, x, \beta_0)}{\partial \beta'} \Big| x \right] \quad \text{and} \quad \mathbf{D} = \text{diag} \left\{ \frac{\mathcal{Q}_1}{\mathbf{P}_p}, \dots, \frac{\mathcal{Q}_M}{\mathbf{P}_{p+M-1}} \right\}$$

I obtain that

$$J = E[P_{>}(p) \mathcal{M}(x) \mathbf{D} m(x)].$$

Denote  $\omega_{z,w_2} = V(g | w_2, z, x)$  and  $\gamma_{z,w_2} = E(g | w_2, z, x)$ . To derive the efficiency bound note that

$$E[\Psi_1 \Psi_1'] = J^{-1} E[\mathcal{M}(x) \mathbf{D} \Sigma_1 \mathbf{D} \mathcal{M}(x)'] J^{-1'}$$

where

$$\Sigma_1 = \text{diag} \left\{ \omega_{k-p+1,k} \frac{\mathcal{P}_{k-p+1}(k)}{\mathcal{Q}_{k-p+1}} + \omega_{M,k} \frac{\mathcal{P}_M(k)}{\mathcal{Q}_M} + \frac{\gamma_{k,M}^2 \mathcal{P}_M(k) [\mathcal{P}_{k-p+1}(k) \mathcal{Q}_{k-p+1} + \mathcal{P}_M(k) \mathcal{Q}_M]}{\mathcal{P}_{k-p+1}(k) \mathcal{Q}_M \mathcal{Q}_{k-p+1}} \right\}_{k=1}^{p+M-1}.$$

For the second component of the influence function

$$\Psi_2(w, x, z) = -J^{-1} \mathcal{M}(x) \mathbf{D} \begin{pmatrix} \left( \frac{d_1^z}{\mathcal{Q}_1} - \frac{d_M^z}{\mathcal{Q}_M} \right) \gamma_{M,p} \mathcal{P}_M(p) \\ \vdots \\ \left( \frac{d_{M-1}^z}{\mathcal{Q}_{M-1}} - \frac{d_M^z}{\mathcal{Q}_M} \right) \gamma_{M,p+M-1} \mathcal{P}_M(p+M-1) \end{pmatrix}.$$

To find the variance I introduce a diagonal matrix and a vector such that

$$\Sigma_2 = \text{diag} \left( \frac{\gamma_{M,k}^2 \mathcal{P}_M^2(k)}{\mathcal{Q}_{k-p+1}} \right)_{k=p}^{p+M-1} \text{ and } \xi = (\gamma_{M,p} \mathcal{P}_M(p), \dots, \gamma_{M,p+M-1} \mathcal{P}_M(p+M-1))'.$$

Then the variance of the second part of the score

$$E[\Psi_2 \Psi_2'] = J^{-1} E \left[ \mathcal{M}(x) \mathbf{D} \left( \Sigma_2 + \frac{\xi \xi'}{\mathcal{Q}_M} \right) \mathbf{D} \mathcal{M}(x)' \right] J^{-1'}$$

Computing the covariance between two components of the efficient influence function, note that

$$\text{cov}(\Psi_1, \Psi_2) = -E[\Psi_2 \Psi_2'].$$

Then the formula for the semiparametric efficiency bound can be written as

$$E[\Psi \Psi'] = J^{-1} E[\mathcal{M}(x) \mathbf{D} \Omega \mathbf{D} \mathcal{M}(x)'] J^{-1'}$$

where

$$\Omega = \Sigma_1 - \Sigma_2 - \frac{\xi \xi'}{\mathcal{Q}_M}.$$

Recalling the structure of the Jacobi matrix, the semiparametric efficiency bound for the model with the optimal choice of  $\mathcal{M}(x)$  will take the form

$$V(\hat{\beta}) = E[P_{>}(p)^2 m(x)' \Omega^{-1} m(x)]^{-1}.$$

## A.4 Proof of Theorems 2 and 3

Proof of Theorem 2 immediately follows from assumptions. In fact, Assumptions 2.1-2.3 are standard assumptions implying that the moment function is P-Glivenko-Cantelli. Assumption 2.4 assures that projection matrix  $\mathcal{A}(\cdot)$  does not destroy the properties of the moment function. Lastly, Assumption 2.5 provides condition for local uniqueness of solution to the empirical moment equation.

The fact that the model under consideration is constructed from a conditional moment equation, assures that estimation of the weighting matrix  $\mathcal{M}(x)$  does not have impact on the asymptotic distribution. Consider first the estimation problem driven by inverse probability weighting. I use projection results in Newey (1994), to represent the moment equation asymptotically as:

$$\begin{aligned} \frac{1}{\sqrt{N}} \sum_{i=1}^N \widehat{\psi}_i(\widehat{\mathbf{P}}, \widehat{\mathcal{Q}}, \beta) &= \frac{1}{\sqrt{N}} \sum_{i=1}^N \mathcal{M}(x_i) \sum_{k=p}^{p+M-1} \left\{ \chi_{ik}(\beta) + E \left[ \frac{\partial \chi_{ik}(\beta)}{\partial \mathcal{Q}_{k-p+1}} \middle| x_i \right] \left( d_{k-p+1}^{z_i} - \mathcal{Q}_{k-p+1} \right) \right. \\ &\quad \left. + E \left[ \frac{\partial \chi_{ik}(\beta)}{\partial \mathcal{Q}_M} \middle| x_i \right] \left( d_M^{z_i} - \mathcal{Q}_M \right) \right\} + o_p(1). \end{aligned}$$

Next, I will show that the variance of the moment equation with such asymptotic representation coincides with the semiparametric efficiency bound. I use notations for two components of the efficient influence function that were introduced in the proof of Theorem 2:  $\Psi_1$  and  $\Psi_2$ . Note that there is direct correspondence between the first component of the efficient influence function and the first component of the semiparametric moment equation such that:

$$\mathcal{M}(x) \sum_{k=p}^{p+M-1} \chi_k(\beta) = J \Psi_1(w, x, z),$$

where index  $i$  for individual observations was omitted, and  $J$  is the Jacobi matrix which is described in the proof of Theorem 2.

The second component of the asymptotic representation reflect estimation error for marginal probability of instrument  $Z$ :

$$\begin{aligned} \mathcal{M}(x) \sum_{k=p}^{p+M-1} \left[ E \left[ \frac{\partial \chi_{ik}(\beta)}{\partial \mathcal{Q}_{k-p+1}} \middle| x_i \right] \left( d_{k-p+1}^{z_i} - \mathcal{Q}_{k-p+1} \right) + E \left[ \frac{\partial \chi_{ik}(\beta)}{\partial \mathcal{Q}_M} \middle| x_i \right] \left( d_M^{z_i} - \mathcal{Q}_M \right) \right] \\ = \mathcal{M}(x) \left( \frac{\mathcal{P}_M(k) \gamma_{Mk}}{\mathbf{P}_k} d_{k-p+1}^z - \frac{\mathcal{Q}_{k-p+1} \mathcal{P}_M(k) \gamma_{Mk}}{\mathbf{P}_k \mathcal{Q}_M} d_M^z \right)_{k=p}^{p+M-1} = -J \Psi_2. \end{aligned}$$

This means in particular that:

$$\widehat{\psi}(\beta) = J [\Psi_1(w, x, z) - \Psi_2(w, x, z)] = J \Psi(w, x, z),$$

so that the influence function is a scaled efficient influence function. Therefore, the variance of this estimator can be represented as

$$V\left(\widehat{\beta}\right) = J^{-1}V\left(J\Psi\left(w, x, z\right)\right)J^{-1} = V\left(\Psi\left(w, x, z\right)\right).$$

Thus, the estimator achieves the semiparametric efficiency bound.

Similar variance calculations are valid for the conditional expectation projection estimator. In particular using similar asymptotic representation technique I will demonstrate that the empirical moment function is asymptotically equivalent to the efficient influence function. Note, that due to validity of the conditional moment equation, estimation noise in the weighting matrix does not affect the asymptotic behavior of the moment equation. I will demonstrate the asymptotic representation for one component of the sum in the moment equation, and for other components derivations will be identical. Using the asymptotic representation theorem in Newey (1994) expand the first component as:

$$\begin{aligned} & \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{\widehat{P}_{k-p+1}(k) d_k^{w_2}}{P_{>}(p)} \widehat{E}[\bar{g} | w_2 = k, z = k - p + 1, x] \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^N \left\{ \frac{\mathbf{P}_k d_k^{w_2} d_{k-p+1}^z \bar{g}}{P_{>}(p) \mathcal{Q}_k} - \frac{\mathbf{P}_k P_{k-p+1}(k) E(\bar{g} | w_2 = k, z_i = k - p + 1, x_i)}{P_{>}(p) \mathcal{Q}_{k-p+1}} \left( d_{k-p+1}^{z_i} - \mathcal{Q}_{k-p+1} \right) \right. \\ & \quad \left. + \frac{d_k^{w_2} - \mathbf{P}_k}{P_{>}(p)} E\left(d_k^{w_2} \bar{g} | z_i = k - p + 1, x_i\right) \right\} + o_p(1). \end{aligned}$$

Other terms in the expression for empirical moment equation (6) can be expanded similarly. Obtained expansions should be summed taking into account the identity

$$E\left(d_k^{w_2} \bar{g} | z = k - p + 1, x\right) = E\left(d_k^{w_2} \bar{g} | z = M, x\right).$$

Analysis of the resulting expression shows that it coincides with the expression for the efficient influence function.

Assumption 3.1 is the standard P-Donsker class condition. Assumption 3.2 assures local regularity of the semiparametric moment condition, and Assumption 3.3 assures that regularity holds after multiplication by a weighting matrix. Assumption 4 basically provides the validity of local linear representation

$$E\left[\delta_m^0(X) \left(\widehat{\mathcal{Q}}_m(X) - \mathcal{Q}_m^0(X)\right)\right] = \frac{1}{\sqrt{N}} \sum_{i=1}^N \delta_m^0(X_i) \left(d_m^{z_i} - \mathcal{Q}_m^0(X_i)\right) + o_p(1).$$

Therefore, conditions (A.7) to (A.10) for Theorem 6.1 of Newey (1994) are satisfied.

## A.5 Proof of Theorem 5

I use the same decomposition of the likelihood function as in the proof of Theorem 2. Define

$$\hat{g} = \mathcal{M}\left(g(w, x, \beta) - E\left[g(w, x, \beta^0) | w_2, x, p = S_1 < \dots < S_M\right]\right),$$

and  $\hat{q} = \mathcal{M} E [g(w, x, \beta^0) | w_2, x, p = S_1 < \dots < S_M]$ . Then function  $\hat{g}$  satisfies the unconditional moment equation, and, therefore has the same structure of the efficient influence function as the influence function in the unconditional model. Define this part of the influence function  $\hat{\Psi}_1$ . Then:

$$-J \hat{\Psi}_1 = \sum_{k=p}^{p+M-1} \left( d_k^{w_2} d_{k-p+1}^z - \frac{\mathcal{Q}_{k-p+1} d_k^{w_2} d_M^z}{\mathcal{Q}_M} \right) \hat{g} - \sum_{k=p}^{p+M-1} \left( \frac{d_{k-p+1}^z}{\mathcal{Q}_{k-p+1}} - \frac{d_M^z}{\mathcal{Q}_M} \right) E [d_k^{w_2} d_{k-p+1}^z \hat{g}]$$

Note that utilized decomposition of the moment equation has transformed it to:

$$E_\theta [\mathcal{A}_\theta(w_2, x) E_\theta [\hat{g}(w, x, \beta_\theta) | w_2, x, p = S_1 < \dots < S_M]] + E_\theta [\mathcal{A}_\theta(w_2, x) \hat{q}] = 0.$$

The component for the influence function corresponding to the first part of this expression is given by  $\hat{\Psi}_1$ . To find the influence function corresponding to the second component, note that the directional derivative along the parametrization path can be written as:

$$E [\mathcal{A}(w_2, x) s_\theta(w_2, x) \hat{q}] + E \left[ \frac{\partial \mathcal{A}_\theta(w_2, x)}{\partial \theta} \hat{q} \right].$$

The expression for the directional derivative in the expectation can be written as:

$$\begin{aligned} P_{>}(p) \sum_{k=p}^{p+M-1} \dot{\mathcal{Q}}_{k-p+1} \hat{q}(k, x) + \sum_{j=1}^p \left( \dot{P}_1(j) - \dot{P}_2(j) \right) \sum_{k=p}^{p+M-1} \mathcal{Q}_{k-p+1} \hat{q}(k, x) \\ + s_\theta(x) P_{>}(p) \sum_{k=p}^{p+M-1} \mathcal{Q}_{k-p+1} \hat{q}(k, x) \end{aligned}$$

To find coefficients of the influence function corresponding to the term  $G_l \dot{\mathcal{Q}}_l$  in the expression for the directional derivative, I look for solution in the form

$$\sum_{m=1}^{M-1} a_m^l \left[ \frac{d_m^z}{\mathcal{Q}_m} - \frac{d_M^z}{\mathcal{Q}_M} \right].$$

Denote  $a^{(l)} = (a_1^l, \dots, a_{M-1}^l)'$ ,  $\mathcal{Q} = (\mathcal{Q}_1, \dots, \mathcal{Q}_{M-1})'$  and  $\gamma^{(l)} = (\gamma_m^l)_{m=1}^{M-1}$ , where  $\gamma_l^l = G_l$  and  $\gamma_k^l = 0$  for  $k \neq l$ . Then the system of equations for coefficients of interest is

$$\left( I + \frac{\mathcal{Q} \mathbf{1}'}{\mathcal{Q}_M} \right) a^{(l)} = \mathcal{Q}_l \gamma^{(l)}.$$

The solution to this system gives

$$\begin{aligned} a_l^l &= \mathcal{Q}_l (1 - \mathcal{Q}_l) G_l, \\ a_k^l &= -\mathcal{Q}_l \mathcal{Q}_k G_l. \end{aligned}$$

Using results for the efficient influence function for the unconditional model, the second component of the efficient influence function can be written as

$$\begin{aligned}
-J\widehat{\Psi}_2 &= P_{>}(p) \sum_{k=p}^{p+M-1} \widehat{q}(k, x) \left( d_{k-p+1}^z - \mathcal{Q}_{k-p+1} \right) \\
&+ \sum_{k=p}^{p+M-1} \mathcal{Q}_{k-p+1} \widehat{q}(k, x) \sum_{j=1}^p \left( d_1^z \frac{d_j^{w_2 - \mathcal{P}_1(j)}}{\mathcal{Q}_1} - d_2^z \frac{d_j^{w_2 - \mathcal{P}_1(j)}}{\mathcal{Q}_2} \right) \\
&+ P_{>}(p) \sum_{k=p}^{p+M-1} \mathcal{Q}_{k-p+1} \widehat{q}(k, x) - E \left[ P_{>}(p) \sum_{k=p}^{p+M-1} \mathcal{Q}_{k-p+1} \widehat{q}(k, x) \right].
\end{aligned}$$

Denote  $\bar{m}_k = E [m_k(w_1, x, \beta^0) | w_2 = k, x, p = S_1 < \dots < S_M]$ , then the efficient influence function can be written as:

$$\begin{aligned}
-J\Psi &= \sum_{k=p}^{p+M-1} \left( d_k^{w_2} d_{k-p+1}^z - \frac{\mathcal{Q}_{k-p+1} d_k^{w_2} d_M^z}{\mathcal{Q}_M} \right) (m_k(w_1, x, \beta) - \bar{m}_k) \\
&- \sum_{k=p}^{p+M-1} \left( d_{k-p+1}^z - \frac{\mathcal{Q}_{k-p+1} d_M}{\mathcal{Q}_M} \right) \mathcal{P}_{k-p+1}(k) (E [m_k(w_1, x, \beta) | w_2 = k, z = k - p + 1] - \bar{m}_k) \\
&+ P_{>}(p) \sum_{k=p}^{p+M-1} \bar{m}_k \left( d_{k-p+1}^z - \mathcal{Q}_{k-p+1} \right) + \sum_{k=p}^{p+M-1} \mathcal{Q}_{k-p+1} \bar{m}_k \sum_{j=1}^p \left( d_1^z \frac{d_j^{w_2 - \mathcal{P}_1(j)}}{\mathcal{Q}_1} - d_2^z \frac{d_j^{w_2 - \mathcal{P}_1(j)}}{\mathcal{Q}_2} \right) \\
&\quad + P_{>}(p) \sum_{k=p}^{p+M-1} \mathcal{Q}_{k-p+1} \bar{m}_k - E \left[ P_{>}(p) \sum_{k=p}^{p+M-1} \mathcal{Q}_{k-p+1} \bar{m}_k \right].
\end{aligned}$$

## B Tables and Graphs

Table 1: Summary statistics for combined dataset

| Variable                     | N.obs. | Mean   | Std. dev. | min    | max     |
|------------------------------|--------|--------|-----------|--------|---------|
| Hourly wage (\$)             | 404    | 4.8684 | 2.0599    | 0.0410 | 18.6047 |
| Employment status            | 2724   | 0.8363 | 0.3701    | 0      | 1       |
| Number of jobs taken         | 2815   | 3.0252 | 1.9033    | 1      | 4       |
| Number of moves per year     | 2736   | 1.6833 | 1.8101    | 0      | 3       |
| Age <20                      | 2815   | 0.0710 | 0.2570    | 0      | 1       |
| 20 ≤ Age <24                 | 2815   | 0.2515 | 0.4340    | 0      | 1       |
| 25 ≤ Age <34                 | 2815   | 0.4458 | 0.4971    | 0      | 1       |
| 35 ≤ Age <44                 | 2815   | 0.1989 | 0.3993    | 0      | 1       |
| 45 ≤ Age                     | 2815   | 0.0327 | 0.1778    | 0      | 1       |
| Female dummy                 | 2732   | 0.9714 | 0.1666    | 0      | 1       |
| Black dummy                  | 2721   | 0.5182 | 0.4998    | 0      | 1       |
| Hispanic dummy               | 2721   | 0.0114 | 0.1061    | 0      | 1       |
| Married, live together       | 2729   | 0.0092 | 0.0953    | 0      | 1       |
| No high school degree or GED | 2734   | 0.3936 | 0.4886    | 0      | 1       |
| No children                  | 2815   | 0.0469 | 0.2114    | 0      | 1       |
| FTP dummy                    | 2815   | 0.4991 | 0.5001    | 0      | 1       |

Figure 1: Empirical distribution of hourly wages

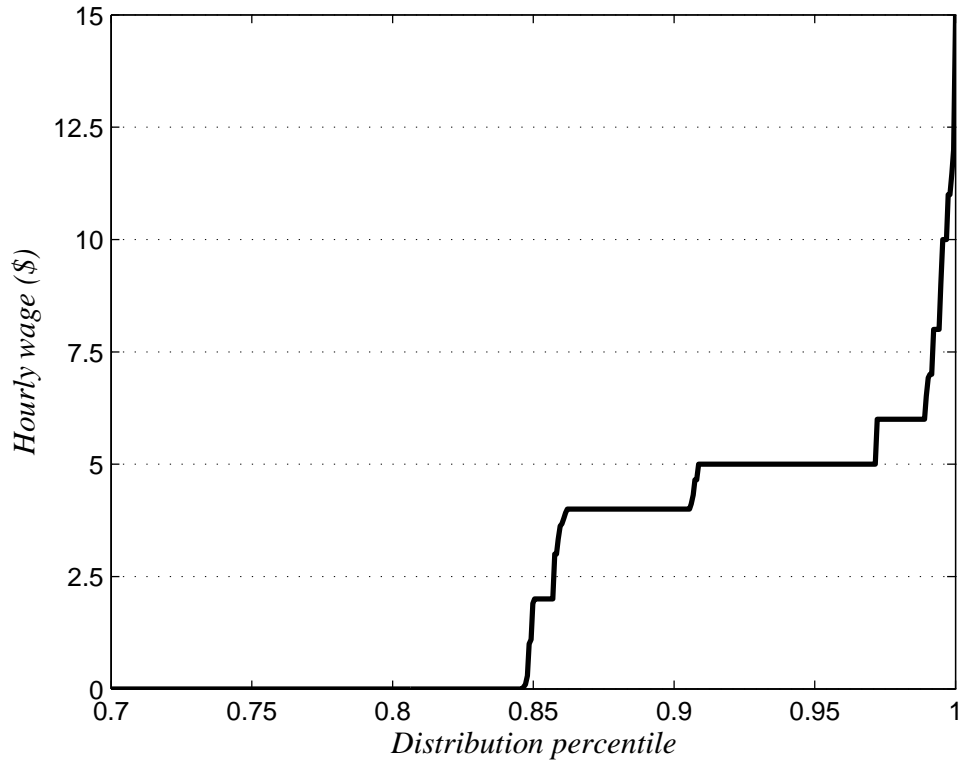


Table 2: Tabulation of empirical distribution of the number of taken jobs numbers in (%)

| #JOBS    | $Z = 0$ | $Z = 1$ |
|----------|---------|---------|
| 1        | 33.55   | 35.30   |
| 2        | 35.25   | 37.37   |
| 3        | 35.25   | 37.51   |
| 4        | 35.25   | 37.51   |
| $\geq 5$ | 100     | 100     |

Figure 2: Coefficient for # JOBS in quantile regression for WAGE (with 10% confidence bounds)

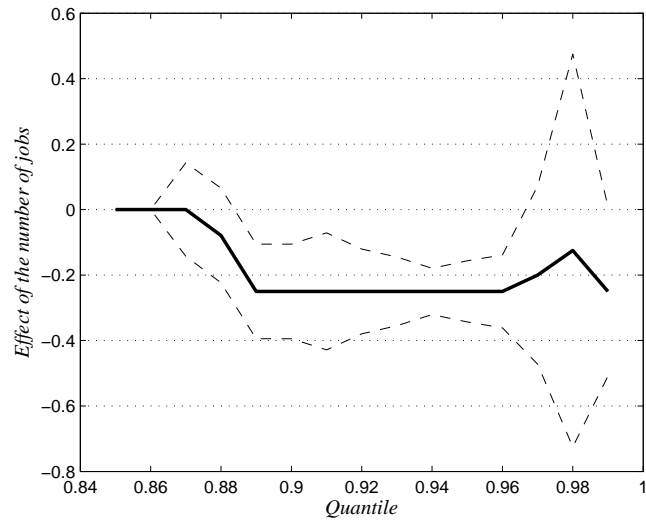


Figure 3: Coefficient for # JOBS in the generalized LATE model for WAGE (with 10% confidence bounds)

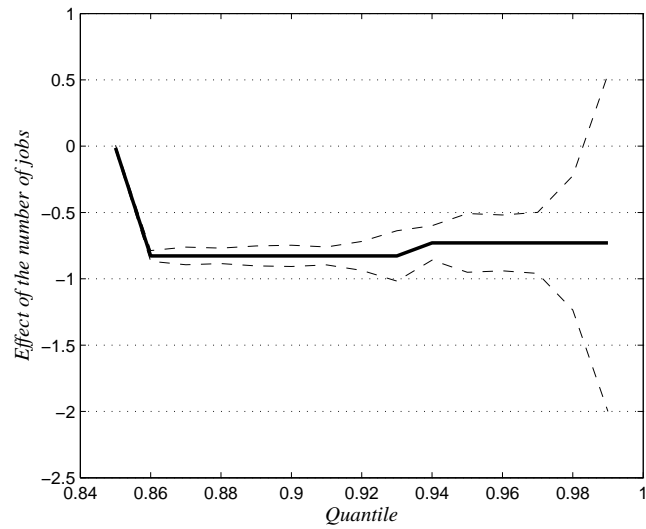


Table 3: Dependence of wage from explanatory factors<sup>a</sup>

|                           | OLS                     | Tobit                    | Quantile                | LATE                    | OLS                     | Tobit                    | Quantile                 | LATE                     |
|---------------------------|-------------------------|--------------------------|-------------------------|-------------------------|-------------------------|--------------------------|--------------------------|--------------------------|
| # Moves                   | -0.0443<br>( 0.0333 )   | -0.2764<br>( 0.2118 )    | 9.50E-16<br>( 0.1066 )  | -0.0541<br>( 0.1031 )   |                         |                          |                          |                          |
| # Past jobs               |                         |                          |                         |                         | -0.0660<br>(0.0236)***  | -0.4147<br>( 0.1655 )**  | -0.2500<br>( 0.0879 )*** | -0.8274<br>( .2235 )***  |
| Age<20                    | -0.2122<br>( 0.3361 )   | -1.6769<br>( 2.1614 )    | -1.0000<br>( 1.1191 )   | -2.0989<br>( 1.7781 )   | -0.2386<br>( 0.3408 )   | -1.9997<br>( 2.1386 )    | -2.0000<br>( 1.0113 )**  | -2.7618<br>( 1.1108 )**  |
| 20 ≤ Age<24               | -0.2852<br>( 0.3198 )   | -2.0147<br>( 1.9059 )    | -1.0000<br>( 1.0146 )   | -2.0710<br>( 1.8860 )   | -0.3268<br>( 0.3249 )   | -2.4008<br>( 1.8854 )    | -2.0000<br>( 0.8840 )**  | -0.5887<br>( 1.1245 )    |
| 25 ≤ Age<34               | -0.0405<br>( 0.3202 )   | -0.5444<br>( 1.8446 )    | 0.0000<br>( 0.9983 )    | -2.6539<br>( 2.9877 )   | -0.0712<br>( 0.3244 )   | -0.8599<br>( 1.8284 )    | -1.0000<br>( 0.8565 )    | -2.1195<br>( 1.9805 )    |
| 35 ≤ Age<44               | 0.0145<br>( 0.3353 )    | -0.2673<br>( 1.9067 )    | 0.0000<br>( 1.0405 )    | -0.0011<br>( 0.0173 )   | -0.0123<br>( 0.3392 )   | -0.5688<br>( 1.8924 )    | -1.0000<br>( 0.9062 )    | -1.7511<br>( 0.9098 )    |
| Black                     | -0.1254<br>( 0.1001 )   | -0.6312<br>( 0.6036 )    | 0.0000<br>( 0.3012 )    | 0.0000<br>( 0.0697 )    | -0.0984<br>( 0.0976 )   | -0.4670<br>( 0.5905 )    | 0.0000<br>( 0.3227 )     | 0.0069<br>( 0.0104 )     |
| Has a child               | 0.0851<br>( 0.2449 )    | 0.2300<br>( 2.1215 )     | 0.0000<br>( 0.3620 )    | -0.0704<br>( 0.1797 )   | 0.0969<br>( 0.2450 )    | 0.3447<br>( 2.1203 )     | 0.0000<br>( 0.7339 )     | 0.0000<br>( 1.5689 )     |
| Has a High Scholl diploma | 0.4986<br>( 0.0913 )*** | 3.2226<br>( 0.6710 )***  | 4.0000<br>( 0.3395 )*** | 4.0000<br>( 0.9867 )*** | 0.4788<br>( 0.0916 )*** | 3.1338<br>( 0.6698 )***  | 3.0000<br>( 0.3565 )***  | 2.3020<br>( 1.0001 )***  |
| Constant                  | 0.6072<br>( 0.3966 )    | -8.3693<br>( 2.8561 )*** | 1.0000<br>( 1.0997 )    | 0.5102<br>( 0.1162 )*** | 0.7181<br>( 0.3967 )    | -7.6037<br>( 2.8532 )*** | 3.2500<br>( 1.1569 )***  | -5.4435<br>( 1.2253 )*** |
| (pseudo) R <sup>2</sup>   | 0.0247                  | 0.0151                   | 0.0565                  | –                       | 0.0278                  | 0.0169                   | 0.0612                   | –                        |

<sup>a</sup>Standard errors are presented in the parenthesis. \*\*\* indicates significance on 1% level, \*\* - 5% level, and \* - 10% level. Asymptotic standard errors are used for quantile regression and the generalized LATE estimates.