Sharp Bounds on Heterogeneous *Individual* Treatment Responses

Jinhyun Lee
University of St Andrews

February 2011

Abstract

This paper discusses how to identify individual-specific causal effects of an ordered discrete endogenous variable. The counterfactual causal information is recovered by identifying the partial differences of a structural relation. The proposed nonparametric *shape* restrictions exploit the fact that the pattern of endogeneity may vary across the level of the unobserved variable. The restrictions adopted in this paper impose a sense of order to an "unordered" binary endogeneous variable. This allows for a unified structural approach to studying various "treatment" effects when self-selection is present. The usefulness of the identification results is illustrated using the data on the Vietnam-era veterans. The empirical findings reveal that when other observable characteristics are identical, military service had positive impacts for individuals with low (unobservable) earnings potential, while it had negative impacts for those with high earnings potential. This heterogeneity would *not* be detected by average effects which would underestimate the actual effects because different signs would be cancelled out. This partial identification result can be used to test homogeneity in response. When homogeneity is rejected, many parameters based on averages may deliver misleading information.

1 Introduction

Policies provide individuals with incentives to change their choices. Different people might respond to a policy change differently. If there exists heterogeneity in responses, many econometric methods based on "averages" may fail to provide correct information.[1] Policy evaluation literature typically uses the potential outcomes approach in identifying treatment responses. This paper demonstrates how *additively*

[1]See Angrist (2004) for the potential outcomes approach, and Hahn and Ridder (2011) for the structural approach.

nonseparable structural functions are used in recovering heterogeneous causality and provides a model that identifies *individual* treatment effects.[2]

Restrictions are imposed on the *shape* of the Hurwicz (1950) structure. The novel restriction exploits the fact that the pattern of endogeneity may vary across the level of the unobserved variable. The proposed model does not require differentiability of the structural functions nor continuity of observed vairiables. The model does not impose weak separability which would make it impossible to recover individuals' heterogeneous treatment responses. It can be used to recover some *partial* information on individual-level causal effects of a discrete variable by identifying the partial difference of a nonadditive structural function.

1.1 Causality, Heterogeneity, and Nonseparable Structural Relations

Suppose we are interested in the impact of a variable (Y) chosen by individuals on their outcome (W) of interest, and suppose the economic decisions on W can be described by the following relation[3]

$$W = h(Y, X, U), \tag{1}$$

where X is a vector of characteristics that are exogenously given to individuals such as age, gender, and race, and U is a normalized scalar index of unobservable (possibly) multidimensional individual characteristics. Various unobserved factors can affect the outcome and the choice, but they are assumed to do so, only through the scalar indexes taking values between 0 and 1. The structural relation may be derived from some optimization processes such as demand/supply functions. We are agnostic about this. If there is not a well-defined economic theory behind them, then the structural relations can be simply understood as how the outcome and the choice are determined by other relevant (both observable and unobservable) variables. The structural relation delivers the information on "contingent" plans of choice or outcome when different values of X and U are given. Even among the individuals

[2]Under the potential outcomes framework, individuals' treatment effects - difference between the outcomes with and without a treatment - are impossible to *measure* because only either of them is observed, not both.

[3]In contrast with (1), switching regression models with a selection equation of the following form have been widely used :

$$\begin{aligned} W_0 &= h_0(0, X, U_0) \\ W_1 &= h_1(1, X, U_1) \end{aligned} \tag{2}$$

The counterfactual outcomes are determined by distinct functional relations, h_0 and h_1, and the unobserved heterogeneity for the two counterfactual events, U_0 and U_1, are allowed to be different. The partial difference of h_0 or $h_1$would not be interpretaed as causal effects.

with the same observed characterisics we observe a distribution of the outcome due to the unobserved elements, U, and the conditional distribution of the outcome, $F_{W|YX}$, is determined by the interaction between the distribution of the unobserved elements, $F_{U|YX}$ and the structural relation, $h(\cdot,\cdot,\cdot)$.

Causal effects of a variable indicate the effects of the variable only, separated from other possible influences. This counterfactual information is contained in partial differences of the structural relation. When the outcome is determined by (1), the causal effects of changing the value of Y from y^a to y^b on W other things being equal would be measured by the partial difference of the structural function, h

$$\Delta(y^a, y^b, x, u) \equiv h(y^a, x, u) - h(y^b, x, u)$$

for some fixed values of $X = x$ and $U = u$. Individuals with different values of X and U may have different values of $\Delta(y^a, y^b, x, u)$, thus, heterogeneity can constitute of both observed and unobserved components.

When Y is binary, the ceteris paribus effect of Y can be expressed by

$$\Delta(1, 0, x, u) = h(1, x, u) - h(0, x, u).$$

Adopting the notation of the potential outcomes framework, let W_{di} denote the hypothetical outcome with $Y = d$ for the individual i whose observed and unobserved characteristics are x and u.[4] Suppose there is a binary choice decision and let $d \in \{0, 1\}$. If we can assume that W_{1i} and W_{0i} are generated by the structural relation then we can write

$$W_{1i} - W_{0i} = h(1, x, u) - h(0, x, u).$$

This way we map the problem in the potential outcomes framework into the structural approach[5]. This is the key relation that justifies the *interpretation* of $h(1, x, u) - h(0, x, u)$ as *individual-specific* treatment response.

Identification of causal effects calls for special attention if there is endogeneity or selection problem. Y is called endogenous if U and Y are not independent. The selection problem exists if the distribution of counterfactual outcomes, W_0 and W_1 counterfactual distributions are different from each other[6]. The identification problem in the potential outcomes approach (identification of the object on the left) is caused

[4]See Heckman, Florens, Meghir, and Vytlacil (2008) for average effects of continuous treatment, and Angrist and Imbens (1995), and Nekipelov (2009) for average effects of multi-valued discrete treatment.

[5]By the structural approach we mean the sort of analysis in classical simultaneous equations systems model. This should be distinguished from "structural estimation" where the underlying optimization processes such as preferences are fully specified. Rather, the structural approach I am considering simply assumes the existence of decision processes which can be expressed as relationships between variables. Further specification of the decision processes is not required.

[6]If the counterfactual distributions are distinct from each other even after controling for observable characteristics, there is selection on unobservables. Selection on unobservables is the case I am considering in this paper.

by the fact that either W_{1i} or W_{0i} is observed, but not both. Thus, the difference of the two for each individual is never observed and cannot be replaced by observed W_i if there exists the selection problem. Difficulties in identification of the structural function (identification of the object on the right) arise because observed information from the relevant variables does not necessarily guarantee the information on independent variation in each coordinate of the structural relation.

The potential outcomes approach does not utilize the information on the economic processes that generate the potential outcomes. Instead of $W_{1i} - W_{0i}$, this paper focuses on identification of $h(1, x, u) - h(0, x, u)$, by assuming the existence of economic processes and by imposing restrictions on such decision mechanisms. See the recent debate between Deaton (2009) and Imbens (2009)[7]. The proposed model can be used to identify the signs of individual treatment responses. This model would be particularly informative when the signs of individual effects vary across the population, in which case average effects would underestimate the true effects with different signs being cancelled out.

1.2 Contributions

This paper contributes to the nonparametric identification literature by providing new identification results on a non-additive structural function when an endogenous variable is discrete/*binary* by using a control function approach without relying on continuity of exogenous variables. Non-additive structural functions are used to model heterogeneity. Use of nonseparable relation is not just a generalization.[8] One of the key implications of additively nonseparable functional form is that partial differences are themselves stochastic objects that have distributions[9]. Thus, heterogeneity in

[7]We advocate the structural approach for two reasons : as Deaton (2009) and Heckman and Urzua (2009) argue econometric models guided by economic models provide clearer interpretation of data analysis. Moreover, assuming the existence of a structure derived from an economic model allows us to use restrictions that may be justified by economic arguments such as monotonicity or concavity of structural relation, which can result in identification of some parameters of interest. In contrast with Imbens (2009)'s arguments, when a specific structural feature is aimed to be recovered (not the whole structure), the structural approach helps, rather than hinders, inference of causal information from data. On the other hand, the applicability may be limited to the extent that the restrictions can be justified since the identifying power comes from such restrictions.

[8]If there exist different responses among the observationally identical agents, and if there exists endogeneity, then nonseparable structural relation should be used. In this case conditional moment conditions do not have identifying power. See Hahn and Ridder (2009).

[9]If the structural function is linear, that is, $W = a + bY + cX_1 + U$, then the partial derivative of this linear function with respect to Y is b. Thus, assuming a linear structural relation corresponds to assuming "homogenous" responses. On the other hand, an additively separable structural function, for example, $W = f(Y, X_1) + U$, allows for heterogeneity in responses, but once conditioning on the observables, there are no differences among the people with different unobserved characteristics as the ceteris paribus effect measured by the partial derivative, $\dfrac{\partial f(y,x)}{\partial y}$, is determined by observed characteristics only.

individual causal effects can be found by identifying partial differences of a non-additive structural function. However, individual-specific causal effects have not been discussed so far.

On the one hand, in the *structural* approach many studies dealing with endogeneity focus on identification of the structural function, rather than its partial differences, but identification of partial differences is not necessarily guaranteed from the knowledge of identification of structural function when it is non-additive. Existing identification results of a nonadditive structural function are not applicable to identification of the partial difference of a nonadditive function with respect to a binary endogeneous variable. Single equation IV models as in Chernozhukov and Hansen (2005) and Chesher (2010) do not guarantee identification of partial differences. Imbens and Newey (2009)'s control function approach is not applicable to discrete endogenous variables. Chesher (2005) reports identification results of partial differences with respect to an ordered discrete endogenous variable, but it is not applicable to a binary endogenous vairiable. Jun, Pinkse, and Xu (2010) is not applicable if the IV is binary.

On the other hand, individual treatment effects are not recovered from the *potential outcomes* approach since both counterfactual outcomes are never observed. Instead, usually average effects are the focus of interest. Several papers (see Imbens and Rubin (1997), Abadie (2002), and more recently, Chernozhukov, Fernandez-Val, and Melly (2010), Kitagawa (2009), for example) focus on identification of the marginal distributions of the counterfactuals whose information may be useful in recovering QTE, but the *individual* treatment effect cannot be recovered from the information on the marginal distributions of the potential outcomes.

Another distinct feature of the proposed model is that the identifying power does not come from restrictions on data. In this paper nonparametric *shape* restrictions on the structure are imposed, rather than relying on properties of observed variables. Nonparametric identification under endogeneity often relies on the characteristics of IV/exogenous variables - many results exploit continuity, rich support in exogenous variation, large support conditions or certain rank conditions. Such results therefore may have limited applicability since many microeconomic variables are discrete or show limited variation in the support. In contrast with other studies, the new resuls in this paper can be applied to a discrete, including binary, endogenous variable when the IV is binary or when the IV is weak. The proposed model does not require differentiability of the structural function and thus, can be applied to discrete outcomes. The proposed weak rank condition can be applied to examples such as regression discontinuity designs, a case with a binary endogenous variable or weak IV or a binary IV.

1.3 Related Studies

Since Roehrig (1988)'s recognition of the importance of nonparametric identification, there have been many studies that aim to clarify what can be obtained from data

without parametric restrictions (see Matzkin (2007) for a survey on nonparametric identification and the references therein). When parametric assumptions are avoided, point identification is often not possible[10] with a discrete endogenous variable. In such cases one could aim to define a set in which the parameter of interest can be located. This partial identification idea, which was pioneered by Manski (1990, 1995, 2003), has been actively used in many different setups and since it now constitutes a vast literature we only focus on policy evaluation literature.

Many authors[11] emphasize the existence of heterogeneity in individual responses in practice and the importance of the information regarding individual-specific, possibly heterogeneous causal effects of a binary endogenous variable was recognized earlier. Many interesting parameters are functionals of the distribution of individual treatment effects as Heckman, Smith, and Clements (1997) noted. In contrast with average treatment effects which are found by a linear operator, other functionals such as quantiles require the knowledge of the distribution of the individual treatment effects[12].

Some information regarding heterogeneity can be recovered by using quantiles. One particular object that has been the focus of research is the quantile treatment effect (QTE) defined by Lehman (1974) and Doksum (1974)[13]. The QTE can be found from the marginal distributions in principle. to control for possible slection issues, Abadie, Angrist, and Imbens (2002) study the QTE under the LATE-type assumptions using a linear quantile regression model, Firpo (2007) under the matching assumption, and Frolich and Melly (2009) under the regression discontinuity design. Chernozhukov and Hansen (2005)'s moment condition based on their IV-QR model provides a way to estimate QTE controlling for selection or endogeneity problem. However, QTE should not be used to identify individual-specific treatment effects.

One approach to recover individual-specific causal effects is to recover hetero-

[10]Under the "complete" system of equations as Roehrig (1988) and Matzkin (2008), identification analysis relies on differentiability and invertibility of the structural functions. However, differentiability and invertibility fail to hold with discrete endogenous variables. Another well known example is discussed by Heckman (1990) using the selection model - without parametric assumptions point identification is achieved by the identification at infinity argument, which may not hold in practice.

[11]See, for example, Heckman (2000).

[12]When the treatment effects are homogeneous the problem is trivial and the distribution of the treatment effects is degenerate. See Firpo and Ridder (2008) for more discussion.

[13]By estimating quantile treatment effects (QTE) using the Connecticut experimental data Bitler, Gelbach, and Hoynes (2006) found that welfare reforms in the ninties had heterogeneous effects on individuals as predicted by labour supply theory. They conclude that "welfare reform's effects are likely both more varied and more extensive". Average effects may miss much information and can be misleading if the signs of individual treatment effects are varying across people. However, when experimental data are not available, QTE does not have causal interpretation on *individuals* because individuals' rankings in the two marginal distributions of the potential outcomes may change. Our model could be used to determine who benefits by identifying the signs of treatment effect of individuals with different rankings of the *scalar* unobserved heterogeneity even with observational data.

geneity in treatment effects by identifying the distribution of $W_1 - W_0$ directly[14]. Heckman, Smith and Clements (1997) use the Hoeffding-Frechet bounds, and Fan and Park (2010) and Firpo and Ridder (2008) used Makarov bounds to derive information on the distribution of the treatment effects from the "*known*" marginal distributions of the potential outcomes.

Alternative to these potential outcomes setups, one could use structural approaches. By adopting a triangular structural setup, Chesher (2003,2007) studies identification of $\Delta(y^a, y^b, x, u)$ when Y is continuous, by the quantile-based control function approach (QCFA, hereafter). Chesher (2005) showed how the QCFA proposed by Chesher (2003) can be used to find the intervals that the values of the structural function lie in when the endogenous variable is ordered discrete with more than three points in the support. Jun, Pinkse, and Xu (2010) report *tighter* bounds when a different rank condition from Chesher's (2005) is used, while other restrictions on the structure in Chesher (2005) are adopted. Jun, Pinkse, and Xu (2010) does not have identifying power for a binary endogenous variable if the IV is binary. Vytalcil and Yildiz (2007) use a triangular system and report a point identification result of the average treatment effect of a dummy endogenous variable under weak separability and an exclusion restriction. Their results rely on certian characteristics of variation in exogenous variables and excluded variables to achieve point identification. Vytalcil and Yildiz (2007) does not guarantee identification of partial difference. They focus on identificaiton of the average effect, not the structural function. Manski and Pepper (2000) and Bhattacharya, Shaikh, and Vytlacil (2008) have partial identification results on average effects. They exploit different monotonicity restrictions. More discussion on these studies can be found in Lee (2011).

1.4 Organization of the Paper

The remaining part is organized as follows. Section 2 introduces the model for "ordered" discrete endogenous variables and contains the main identification results. Section 3 discusses "unordered" binary endogenous variable as a different case of discrete endogenous variable. Section 4 discusses the restrictions imposed in the model and other related studies in more detail. Section 5 illustrates the usefulness of the identification results by examining the effects of the Vietnam-era veteran status on the civilian earnings using a binary IV. Section 6 concludes.

[14]The quantiles of treatment effects recovered from the distribution of $W_{1i} - W_{0i}$ are examples of $D\Delta-$ treatment effects, while the quantile treatment effects (QTE) are examples of $\Delta D-$treatment effects discussed in Manski (1997). Neither of them is implied by the other, and they deliver different information regarding distributional consequences of any policy. As Firpo and Ridder (2008) nicely discussed, $\Delta D-$treatment effects, such as QTE can deal with the issues such as the impact of a policy on the society (population) in general, while $D\Delta-$treatment effects can be used to address issues such as policy impacts on "individuals".

2 Local Dependence and Response Match (LDRM) model - $\mathcal{M}^{LDRM}$

2.1 Restrictions of the Model $\mathcal{M}^{LDRM}$

I introduce a model that interval identifies the value of the structural function evaluated at a certain point in the presence of an endogenous discrete variable by applying the quantile-based control function approach (QCFA) in Chesher (2003). The model, $\mathcal{M}^{LDRM}$, is defined as the set of all the structures that satisfy the restrictions[15].

Restriction QCFA[16] : Scalar Unobservables Index (SIU)/Monotonicity/Exclusion

$$\begin{aligned} W &= h(Y, X, U), \\ Y &= g(Z, X, V), \\ \text{with } g(z, x, v) &= y^m,\ P^{m-1}(z, x) < v \leq P^m(z, x), \\ m &\in \{1, 2, ..., M-1\} \end{aligned}$$

The function h is weakly increasing[17] with respect to variation in scalar U. From here on other exogenous variables, X, are ignored. X can be added as a conditioning variables in any steps of discussion without changing the results.

The variable W is a discrete, continuous, or mixed discrete continuous random variable. The conditional distribution of Y given $Z = z$ is discrete with points of support $y^1 < y^2 < ... < y^M$, invariant with respect to z and with positive probability masses $\{p_m(z)\}_{m=1}^M$. Cumulative probabilities $\{P^m(z)\}_{m=1}^M$ are defined as

$$\begin{aligned} P^m(z) &\equiv \sum_{l=0}^{m} p_l(z) = F_{Y|Z}(y^m|z), \quad m \in \{1, 2, ..., M\}, \\ p_0(x) &\equiv 0. \end{aligned}$$

The latent variates are jointly continuously distributed and they are normalized uniformly distributed on $(0, 1)$ independent of Z. The value $y^m, m \in \{2, ..., M-1\}$, is an interior point of support of the distribution of Y.

[15] I adopt this definition of a model as a set of structures satisfying the restrictions imposed, following Koopmans and Reiersol (1950).

[16] Triangularity assumption enables us to avoid the issue of coherency that may be caused due to discrete endogenous variables when the outcome is discrete.

[17] If g is weakly increasing in v, then if h is weakly increasing in u and if g is weakly decreasing, h should be weakly decreasing as well. This monotonicity restriction is one of the two key restrictions in QCFA identification strategy. This enables us to use the equivariance property of quantiles. In many applications this can be justified - under certain regularity conditions many optimization frameworks predict that the equilibrium relations are monotonic in certain variables - law of demand as a typical example. See Imbens and Newey (2009) for examples that justify monotonicity.

The function g evaluated at $Z = z$, $g(z, \tau_V)$ is identified by $Q_{Y|Z}(\tau_V|z)$. The monotonicity restriction on Y is reflected in the threshold crossing structure.

Restriction RC (Rank Condition)[18] There exist instrumental values of Z, $\{z'_m, z''_m\}$, such that

$$P^m(z'_m) \leq \tau_V \leq P^m(z''_m)$$

for $m \in \{0, 1, 2, ..., M-1\}$.

Restriction C-QI (Conditional Quantile Invariance) : The value of U, $u^* \equiv Q_{U|VZ}(\tau_U|\tau_V, z)$ is invariant with $z \in \overline{z}_m \equiv \{z'_m, z''_m\}$ for $P^m(z'_m) \leq \tau_V \leq P^{m-1}(z''_m)$.

Define $\boldsymbol{V} \equiv (V_L, V_U]$, where $V_L = \max_{z \in \overline{z}_m} P^{m-1}(z)$, and $V_U = \max_{z \in \overline{z}_m} P^{m+1}(z)$.[19] Define also $\boldsymbol{U} \equiv (U_L(z), U_U(z)]$, where $U_L(z) = \min_{\tau_V \in \boldsymbol{V}} Q_{U|VZ}(\tau_U|\tau_V, z)$, and $U_U(z) = \max_{\tau_V \in \boldsymbol{V}} Q_{U|VZ}(\tau_U|\tau_V, z)$. The value, u^*, is not known, but it indicates τ_U- ranked individual's value of U in the conditional distribution of U given V and Z. The case in which $F_{U|VZ}(u^*|v, z)$ is nonincreasing in v, for $u^* \in U$ is called PD (Positive Dependence) and the other case, ND (Negative Dependence). The case in which $h(y^{m+1}, u^*) \geq h(y^m, u^*)$ is called PR (Positive Response) and the other case, NR (Negative Response).

Restriction LDRM (Local (Quantile) Dependence Response Match) : $F_{U|VZ}(u|v, z)$ is weakly monotonic in $v \in \boldsymbol{V}$ for $u \in \boldsymbol{U}$. If $F_{U|VZ}(u|v, z)$ is weakly decreasing in $v \in \boldsymbol{V}$ for $u \in \boldsymbol{U}$, then $h(y^{m+1}, u^*) \geq h(y^m, u^*)$, **(PDPR)** and if $F_{U|VZ}(u|v, z)$ is weakly increasing in $v \in \boldsymbol{V}$ for $u \in \boldsymbol{U}$, then $h(y^{m+1}, u^*) \leq h(y^m, u^*)$, **(NDNR)** for any $u^* \in \boldsymbol{U}$ for $m \in \{0, 1, 2, ..., M-1\}$. See Figure 1.

2.2 Discussion on Restrictions

2.2.1 Restriction QCFA - Scalar Index Unobservables, U and V

These are the fundamental restrictions imposed in the quantile-based control function method in Chesher (2003). Monotonicity of the structural functions in the *scalar*

[18] Restriction RC is related to the "relevance" condition for IV. If Z is a strong IV, Restriction RC is satisfied. Chesher (2005)'s rank condition is that there exist values of Z, z'_m, and z''_m such that

$$P^m(z'_m) \leq \tau_V \leq P^{m-1}(z''_m)$$

thus, if Chesher (2005)'s rank condition holds, our rank condition also holds since $P^{m-1}(z'') \leq P^m(z'')$. In this sense, Chesher (2005)'s rank condition is stronger than our rank condition. Note also that Chesher (2005)'s strong rank condition is not satisfied when the instrument is weak or when a binary endogenous variable is present.

[19] For a binary endogenous variable $V \equiv [0, 1]$.

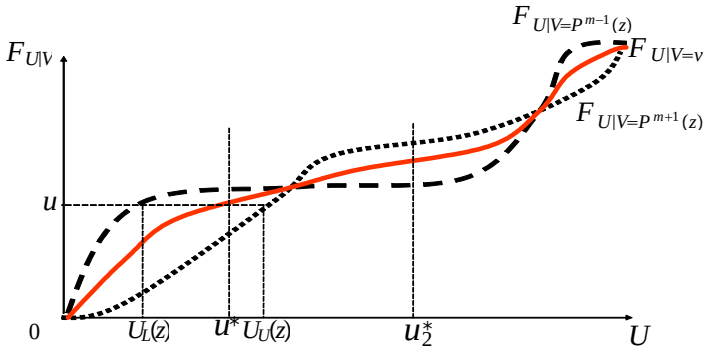

Figure 1: **"Local" nature of Restriction LDRM** : the information on endogeneity is contained in $F_{U|V}$ - if Y is exogenous, then $F_{U|V}$ is invariant with values of V.$F_{U|V}$ is drawn for different values of V by assuming monotonicity in V. The solid line is the distribution of U given $V = v$. A point in the support of U, u^* can be written as τ_U-quantile of U given $V = v$. Monotonicity of $F_{U|V}(u^*|v)$ does not have to be global in U, all that is required is monotonicity in some region U of u. In this graph, for $v' \leq v'' \leq v'''$, $F_{U|V}(u^*|v)$ is decreasing in v, while $F_{U|V}(u_2^*|v)$ is increasing in $v \in V$.

indices of unobserved factors and the existence of Z that is excluded from the outcome equation are key features together with independence between U and Z. The model admits multiple factors of unobserved heterogeneity as long as they affect the outcome though a scalar index.[20]

There is a tradeoff between using a vector and a scalar unobserved heterogeneity - allowing for a vector unobserved heterogeneity in the structural relation would be more realistic. Several studies report identification results without monotonicity restrictions (See Altonji and Matzkin (2005), Hoderlein and Mammen (2007), Imbens and Newey (2009), and Chalak, Schennach, and White (2008), and Chernozhukov, Fernandez-Val, and Newey (2009) for identification analysis without monotonicity). However, what can be identified without monotonicity is objects with the heterogeneity in responses averaged out. On the other hand, the quantile-based approaches under monotonicity can be adopted to recover heterogeneous treatment response only

[20]However, this scalar unobserved index assumption does not admit measurement error models or duration outcomes. For structures with vector unobservables that can**not** be represented by a scalar unobservable, see Chesher (2009), where examples of such case are illustrated. The vector of unobservables is called "excess heterogeneity" in Chesher (2009) - "excess" in the sense that we allow for more unobservable variables than the number of endogenous variables. The distinction of the number of endogenous variables from the number of unobservable variables stems from the analysis of classical simultaneous equations models of the Cowles Commission, and more recent studies on nonparametric identification of simultaneous equations models in Brown (1983), Roehrig (1988), Matzkin (2008), and Benkard and Berry (2006), where the number of unobservables is equal to the number of endogenous variables.

if a scalar (index) unobserved heterogeneity is assumed.

2.2.2 Local Dependence and Response Match (LDRM)

Endogeneity is often defined as the dependence between an explanatory variable and the unobserved elements in the structural relationship. They can be positively dependent or negatively dependent. "Dependence" is used instead of "correlation" to clarify the *local* information contained in Restriction LDRM. Under the triangularity of this paper the source of endogeneity is caused by the dependence between U and V. This information is contained in the conditional distribution of $F_{U|V}$.

Restriction LDRM assumes first that $F_{U|V}(u|v)$ is monotonic in v in certain ranges of U and V. Then it restricts the direction of the dependence in that range and the direction of the response - whether the response is positive or negative or zero. For example, college graduates may be different from high school graduates in terms of ability (U) when other observed characteristics are identical. Restriction LDRM is concerned with how the patterns of dependence vary with the level of the unobserved characteristic. It may be the case that individuals with very low ability are not allowed to get into college due to low test scores, on the other hand, individuals with extremely high ability may not choose to go to college if they have better options that will lead to higher income. This example shows the possibility that there is positive dependence with the low level of ability, and negative dependencewith the high level.

Restriction LDRM is regarding each point in the support of the unobserved variable U. Note that U is normalized to be uniform (0,1) and each point in (0,1) is indicated by expressing it as *quantiles*. In this sense, "local" implication of Restriction LDRM can be understood using *quantiles*.

Relaxation of LDRM What is required for the main results to hold, is less restrictive than Restriction LDRM. Even though the match does not hold (that is, either PDNR or NDPR is the case) as long as the reversal effect (different effect from stated in LDRM restriction) is small, the bounds defined by the model are still relevant and sharp. Consider the education and wage example. Suppose that education is assumed to be positively dependent with unobserved ability, in other words, more able people tend to decide to get educated. The case in which our model is not applicable is the case in which education is so deterimental that the hypothetical wage with one more year of education is smaller than without it, among those with "similar" abliity. On the other hand, LDRM restricts that wage with one more education needs to be larger or equal to that without it among the "same" level of ability if more able individuals choose to get educated more.

Restriction No interaction Reversal (NIR) : $F_{U|VZ}(u|v,z)$ is weakly monotonic in $v \in \boldsymbol{V}$ for $u \in \boldsymbol{U}$. If $F_{U|VZ}(u|v,z)$ is weakly decreasing in $v \in \boldsymbol{V}$ for $u \in \boldsymbol{U}$, then $h(y^{m+1}, \overline{u}) \geq h(y^m, u^*)$ for $\overline{u}, u^* \in \boldsymbol{U}$, with $\overline{u} \geq u^*$. Conversely, $F_{U|VZ}(u|v,z)$ is

weakly increasing in $v \in \boldsymbol{V}$ for $u \in \boldsymbol{U}$, then $h(y^{m+1}, \overline{u}) \leq h(y^m, u^*)$, $\overline{u}, u^* \in \boldsymbol{U}$, with $\overline{u} \leq u^*$ for $m \in \{0, 1, 2, ..., M-1\}$.

<Figure 2> and <Figure 3> are drawn for the case where the unobserved elements are positively dependent. Restriction LDRM specified that $h(y^{m+1}, u^*) \geq h(y^m, u^*)$, (comparison of points A and B) thus, by monotonicity of h with respect to U, $h(y^{m+1}, \overline{u}) \geq h(y^{m+1}, u^*) \geq h(y^m, u^*)$, for $\overline{u} \geq u^*$ (that is, $C \geq B \geq A$). <Figure 3> shows the case in which Restriction LDRM fails (A$\geq B$). The left panel is the case in which LDRM fails, but NIR holds, while the right panel is the case both LDRM and NIR fail.

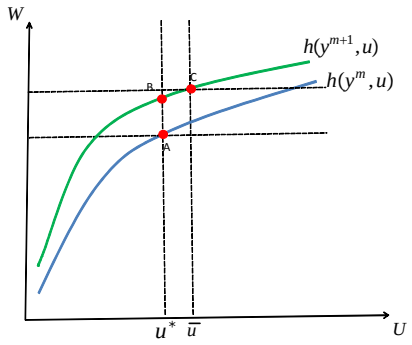

Figure 2: LDRM satisfied

2.2.3 Discrete Data

The restrictions imposed do not require continuity/differentiability of structural relations nor rely on continuity of covariates/large support condition. This makes the proposed model more useful since many variables in microeconometrics are discrete or censored.

3 Main Results

3.1 Bound on the Value of the Structural Relation

We have the following interval identification for $h(y^m, u^*)$ for $m \in \{1, 2, ..., M-1\}$, where $u^* = Q_{U|VZ}(\tau_U|\tau_V, z)$. For $m = M$, the bound in Theorem 1 is not applied[21].

[21] The bounds cannot be applied to $m = M$. This restricts the identification results when $M = 2$, as we will see in the next section.

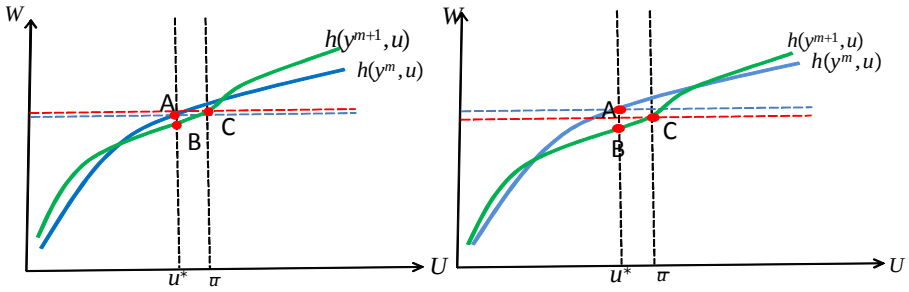

Figure 3: Failure of LDRM (A>B) for the case of positive dependence - the main results still hold for for the left panel (A<C), while the main result does not hole for the right panel (A>C).

Theorem 1 Under Restriction QCFA,C-QI,RC,and LDRM, there are the inequalities for $m \in \{0, 1, 2, ..., M-1\}$ and $\tau \equiv \{\tau_U, \tau_V\}$

$$\begin{aligned}
q_m^L(\tau, y^m, \overline{z}_m) &\leq h(y^m, u^*) \leq q_m^U(\tau, y^m, \overline{z}_m) \\
\text{where } u^* &= Q_{U|VZ}(\tau_U|\tau_V, z), \\
\text{for some } \tau_U &\in (0,1) \text{ and } \tau_V \in [P^m(z'_m), P^m(z''_m)], \\
z &\in \overline{z}_m = \{z'_m, z''_m\}, \\
q_m^L(\tau, y^m, \overline{z}_m) &= \min\{Q_{W|YZ}(\tau_U|y^m, z'_m), Q_{W|YZ}(\tau_U|y^{m+1}, z''_m)\}, \\
q_m^U(\tau, y^m, \overline{z}_m) &= \max\{Q_{W|YZ}(\tau_U|y^m, z'_m), Q_{W|YZ}(\tau_U|y^{m+1}, z''_m)\}.
\end{aligned}$$

The interval is **not** empty.
Proof. See the Appendix. ■

To identify all the values of the structural function, say, $h(y^1, u^*), h(y^2, u^*), ..., h(y^{M-1}, u^*)$, for fixed u^*, we need to guarantee the rank condition holds for all $m \in \{1, 2, ..., M-1\}$.There should exist two values of Z, $\{z'_m, z''_m\}$ for each m, such that $P^m(z'_m) \leq \tau_V \leq P^m(z''_m)$. Therefore, how closely Y and Z are related and whether we have enough variation in Z are key to the identification of the whole function.

3.2 Sharpness

Suppose a set identifies the value of the structural feature. Then *all* distinct "*admitted*" structures that are "*observationally equivalent*" to the true structure should produce values of the structural feature that are contained in the identified set. All such structures that generate a point in the set, are indistinguishable by data. A sharp identified set contains *all and only* such values that are generated by admitted and observationally equivalent structures.

If the identified set is *not* sharp, some of the points in the set are not possible candidates for the value of the structural feature, which would make the identified set less informative.

Different points in a sharp identified set may have been generated by different structures, but the distinct structures (i) should all satisfy the restrictions of the model (consistent with the model), (ii) should be observationally equivalent (consistent with the data), and (iii) any point in the interval should be considered to be the possible value of the structural feature. (See Lee (2011) for the definition of sharpness).

Common support restriction is imposed for sharpness.

Restriction CSupp (Common Support) The support of the conditional distribution of W given Y and Z has support that is invariant across the values of Y and Z.

Theorem 2 Under Restrictions CSUPP, QCFA,C-QI,RC,and LDRM, the bound $I(\tau, y^m, \overline{z}) \equiv [q_m^L(\tau, y^m, \overline{z}_m), q_m^U(\tau, y^m, \overline{z}_m)]$, specified in **Theorem 5.1** for each $m = 0, 1, 2, .., M-1$ and for some $\tau \equiv \{\tau_U, \tau_V\}$, is sharp.
Proof. Use Lemma 1 in Section 2.2. See the Appendix. ■

Sharpness of an identified set is a logically essential property for inference. When the identified set is not sharp, the confidence region is constructed "conservatively". However, if an identified set is not shown to be sharp, then the inference based on the non-sharp identified set can be *meaningless* - it could be the case that one never rejects a hypothesis (regarding the structural feature) even if the hypothesis were not true. For example, consider one is interested in $H_0 : \theta(S) = 0$. Without the information on the outer region, confidence region constructed on the non-sharp identified set might not be informative, since even if zero lied in the confidence region, one never knows whether zero is in the outer region or not. If zero lied in the outer region of the identified set, then the inference would fail with the power of the test being zero.

3.3 Many Instrumental Values, Overidentification, and Refutability

If there are many pairs of values of Z that satisfy Restriction RC (overidentification), then each pair defines the causal effect for a *different* subpopulation defined by each pair. Taking intersection of each identified set **cannot** be a sharp identified set as is discussed Lee (2011). To use all the information available from data and to justify taking intersection of each set defined by distinct pairs of values of Z in producing *a sharp identified set* in this case, **a different restriction** is imposed.

Let $SUPP(Z)$ be the support of Z. Define $\mathbf{V}_m \equiv [P^m(z'_m), P^m(z''_m)]$ for the pair, $\{z'_m, z''_m\}$ that satisfies Restriction RC. Each pair defines different subpopulation over which a causal interpretation is given. Define $\mathcal{Z}_m$ as the set of pairs of $\{z'_m, z''_m\}$ that satisfies Restriction RC, $\mathcal{Z}_m \equiv \{\overline{z}_m : P^m(z'_m) \leq \tau_V \leq P^m(z''_m)$, with $\overline{z}_m = \{z'_m, z''_m\}\}$. Let $\mathbf{V}_m(\mathcal{Z}_m) \equiv \{\mathbf{V}_m(\overline{z}_m) : \overline{z}_m \in \mathcal{Z}_m\}$ be a class of the set defined by $\mathcal{Z}_m$. Denote $\mathbb{V} \equiv \cap V_m(\mathcal{Z}_m)$.

Restriction C-QI^M (Conditional Quantile Invariance with Many Instrumental Values) : The value of U, $u^* \equiv Q_{U|VZ}(\tau_U|\tau_V, z)$ is invariant with all $z \in \overline{z}_m (\in \mathcal{Z}_m)$.

Corollary 1 Under Restriction QCFA,C-QI^M, RC,and LDRM, there are the inequalities for $m \in \{0, 1, 2, ..., M-1\}$, $\tau \equiv \{\tau_U, \tau_V\}$,

$$\begin{aligned}
Q^L_m(\tau, y^m, \mathcal{Z}_m) &\leq h(y^m, u^*) \leq Q^U_m(\tau, y^m, \mathcal{Z}_m) \\
\text{where } u^* &= Q_{U|VZ}(\tau_U|\tau_V, z), \\
\text{for some } \tau_U &\in (0,1) \text{ and } \tau_V \in \mathbb{V} \equiv \cap_m V_m(\overline{z}_m) \\
Q^L_m(\tau, y^m, \mathcal{Z}_m) &= \max_{\overline{z}_m} \; q^L_m(\tau, y^m, \overline{z}_m), \; \overline{z}_m \in \mathcal{Z}_m \\
Q^U_m(\tau, y^m, \mathcal{Z}_m) &= \min_{\overline{z}_m} \; q^U_m(\tau, y^m, \overline{z}_m), \; \overline{z}_m \in \mathcal{Z}_m \\
q^L_m(\tau, y^m, \overline{z}_m) &= \min\{Q_{W|YZ}(\tau_U|y^m, z'_m), Q_{W|YZ}(\tau_U|y^{m+1}, z''_m)\} \\
q^U_m(\tau, y^m, \overline{z}_m) &= \max\{Q_{W|YZ}(\tau_U|y^m, z'_m), Q_{W|YZ}(\tau_U|y^{m+1}, z''_m)\}
\end{aligned}$$

This intersection interval is sharp and **can be empty**.

Proof. Idenitified intervals for each pair $\overline{z}_m \in \mathcal{Z}_m$, are shown in **Theorem 1**. The bound in this corollary is found by taking intersection of all such identified intervals. This intersection bound is sharp. The same sharpness proof of **Thorem 2** applies with some modification in (S2) constructed in the proof in Appendix. When there exist many instrumental values that satisfy the rank condition, RC, the partition,

$\{P^l\}_{l=1}^M$ defined in the proof of **Theorem 2** can be re-defined as the following :

$$\begin{aligned}
P^l &= \left\{ \begin{array}{lll} \min_{z\in SUPP(Z)}\{P^l(z)\}, & if & l < m-1 \\ \max_{z\in SUPP(Z)}\{P^l(z)\}, & if & l > m \end{array} \right\} \\
P^{m-1} &= \min_{z\in\overline{z}_L}\{P^m(z)\} \\
P^m &= \max_{z\in\overline{z}_U}\{P^m(z)\}, \\
\text{where } \overline{z}_L &\equiv \{z_L : z_L = \min \overline{z}_m, \overline{z}_m \in \mathcal{Z}_m\} \\
\overline{z}_U &\equiv \{z_U : z_U = \max \overline{z}_m, \overline{z}_m \in \mathcal{Z}_m\} \\
\mathcal{Z}_m &\equiv \{\overline{z}_m : P^m(z'_m) \leq \tau_V \leq P^m(z''_m), \text{with } \overline{z}_m = \{z'_m, z''_m\}\}.
\end{aligned}$$

$\overline{z}_L(\overline{z}_U)$ is the set of smaller (larger) values of $\overline{z}_m = \{z'_m, z''_m\} \in \mathcal{Z}_m$. The partition of the support of V, $(0,1)$, is constructed such that $P^1 < P^2 < ... < P^M$. ■

Intersection of identified sets may be empty, and even if it is not empty, the causal interpretation of the intersection bound needs to be given to a different subpopulation.

Suppose that the intersection, $\mathbb{V} \neq \varnothing$. Then the bound defined by **Corollary 1** should be interpreted as causal effects for the subpopulation defined by $\mathbb{V}$. If $\mathbb{V} = \varnothing$, no causal interpretation would be possible, even though the intersection bound may not be empty since the subpopulation that is affected by the change in the values of Z does not exist. If $\mathbb{V} \neq \varnothing$, but the intersection bound is empty, then this means that some of the restrictions in the model are not satisfied[22]. However, which restrictions are misspecified is not known by the fact that the identified set is empty. This way one can falsify the econometric model, rather than a specific restriction.

4 Binary Endogenous Variable

Although in many empirical studies, the distribution of the treatment effects can deliever valuable information for any policy design, quantiles of the distribution of differences of potential outcomes, $W_1 - W_0$, have been considered to be difficult to point identify without strong assumptions.[23] In this section I apply the LDRM model to a binary endogenous variable and identify the ceteris paribus impact of the binary variable, or treatment effects. As Chesher (2005) noted, models for an ordered discrete endogenous variable can not directly be applied to binary endogenous variables due to the "unordered" nature of binary variables, however, Restriction LDRM imposes a sense of order to a binary endogenous variable, which enables the model to identify the partial differences. The number of points in the support of Y restricts the identification result.

[22] I am grateful to Pierre Debois, and Brendon McConell for this point.

[23] Note that in general, quantiles of treatment effects, $Q_{W_1-W_0|X}(\tau|x) \neq Q_{W_1|X}(\tau|x) - Q_{W_0|X}(\tau|x)$, where the right hand side is the QTE.

4.1 Bound on the Value of the Structural Relation

The model interval identifies $h(1,u^*)$ and $h(0,u^*)$ as is shown in the following corollary.

Corollary 2 Under Restriction QCFA,C-QI,RC,and LDRM there are the inequalities for $y \in \{0,1\}$, $z \in \overline{z} = \{z', z''\}$, and $\tau \equiv \{\tau_U, \tau_V\}$,

$$\begin{aligned} q^L(\tau,y,\overline{z}) &\leq h(y,u^*) \leq q^U(\tau,y,\overline{z}) \\ \text{where } u^* &= Q_{U|VZ}(\tau_U|\tau_V,z), \\ \text{for some } \tau_U &\in (0,1) \text{ and } \tau_V \in [P(z'),P(z'')], \\ q^L(\tau,y,\overline{z}) &= \min\{Q_{W|YZ}(\tau_U|0,z'), Q_{W|YZ}(\tau_U|1,z'')\} \\ q^U(\tau,y,\overline{z}) &= \max\{Q_{W|YZ}(\tau_U|0,z'), Q_{W|YZ}(\tau_U|1,z'')\} \end{aligned}$$

The bound is sharp.
Proof. See the Appendix ■

The identified intervals for $h(1,u^*)$ and $h(0,u^*)$ are the same. Nevertheless, this is still informative in the sense that the identified interval restricts the possible range that the values $h(1,u^*)$ and $h(0,u^*)$ lie in, and that under Restriction LDRM either the upper bound or the lower bound on $h(1,u^*) - h(0,u^*)$ should be zero.

Lemma 3 Under Restriction QCFA,C-QI,RC,and LDRM,

$$\begin{aligned} PDPR \text{ implies } Q_{W|YZ}(\tau_U|y^{m+1},z''_m) &\geq Q_{W|YZ}(\tau_U|y^m,z'_m), \text{ and} \\ NDNR \text{ implies } Q_{W|YZ}(\tau_U|y^{m+1},z''_m) &\leq Q_{W|YZ}(\tau_U|y^m,z'_m). \end{aligned}$$

Proof. See the Appendix. ■

Corollary 2 and **Lemma 3** are used to recover heterogeneous treatment responses. **Theorem 3** states the partial identification result of heterogeneous treatment effects.

4.2 Bound on Partial Difference of the Structural Relation

Theorem 3 Under Restriction QCFA,C-QI,RC,and LDRM, $h(1,u^*) - h(0,u^*)$ is identified by the following interval:

$$\begin{aligned} B^L &\leq h(1,u^*) - h(0,u^*) \leq B^U \\ B^U &= \max\{0, Q^{\Delta}_{10}(\tau_U)\} \\ B^L &= \min\{0, Q^{\Delta}_{10}(\tau_U)\}, \\ \text{where } Q^{\Delta}_{10}(\tau_U) &\equiv Q_{W|YZ}(\tau_U|1,z'') - Q_{W|YZ}(\tau_U|0,z') \end{aligned}$$

Proof. Suppose $Q_{W|YZ}(\tau_U|1,z'') \geq Q_{W|YZ}(\tau_U|0,z')$. From **Corollary 2** we have

$$\begin{aligned} Q_{W|YZ}(\tau_U|0,z') &\leq h(1,u^*) \leq Q_{W|YZ}(\tau_U|1,z'') \\ Q_{W|YZ}(\tau_U|0,z') &\leq h(0,u^*) \leq Q_{W|YZ}(\tau_U|1,z'') \end{aligned}$$

then we have

$$\begin{aligned} -(Q_{W|YZ}(\tau_U|1,z'') - Q_{W|YZ}(\tau_U|0,z')) &\leq h(1,u^*) - h(0,u^*) \\ &\leq Q_{W|YZ}(\tau_U|1,z'') - Q_{W|YZ}(\tau_U|0,z'). \end{aligned} \tag{3}$$

By **Lemma 3**, if $Q_{W|YZ}(\tau_U|1,z'') \geq Q_{W|YZ}(\tau_U|0,z')$, we should have

$$h(1,u^*) - h(0,u^*) \geq 0$$

applying this to (3) yields the result. The case when $Q_{W|YZ}(\tau_U|1,z'') \leq Q_{W|YZ}(\tau_U|0,z')$ can be shown similarly. ■

Whether the treatment effect is positive or negative can be determined by data from the sign of $Q^{\Delta}_{10}(\tau_U) \equiv Q_{W|YZ}(\tau_U|1,z'') - Q_{W|YZ}(\tau_U|0,z')$ based on **Theorem 3**. If $Q^{\Delta}_{10}(\tau_U) > 0$, then

$$0 \leq h(1,u^*) - h(0,u^*) \leq Q^{\Delta}_{10}(\tau_U),$$

and if $Q^{\Delta}_{10}(\tau_U) < 0$, then

$$Q^{\Delta}_{10}(\tau_U) \leq h(1,u^*) - h(0,u^*) \leq 0.$$

If $Q^{\Delta}_{10}(\tau_U) = 0$, then $h(1,u^*) - h(0,u^*)$ is point identified as zero. Either the upper bound or the lower bound is always zero.

If Restriction LDRM were true about the underlying structure, then from this restriction we could infer whether the dependence between the two unobservables is positive or negative locally in a certain range by Lemma 3. If economic arguments can justify the nature of the dependence pattern found from data, then this model can be credibly applicable.

4.3 Discussion

4.3.1 Heterogeneous Causality Measured by Partial Differences

The major object of interest in this paper is the partial difference of the structural quantile function, $h(1,u^*) - h(0,u^*)$. The value u^* is unknown, but is assumed to be $u^* = Q_{U|VZ}(\tau_U|\tau_V,z)$ for some $\tau_U, \tau_V \in (0,1)$. $h(1,u^*) - h(0,u^*)$ is interpreted as a ceteris paribus impact of Y. When the value of Y changes from 1 to 0, the value of U would change as well if there exists endogeneity. This is in contrast with other identification results in additively nonseparable models. Other studies identify the values of a *nonadditive* structural function, but their results do not guarantee identification of partial differences.

4.3.2 Rank Condition and Causal Interpretation

The rank condition restricts the group for whom the identification of causal impacts is justifiable into those who are ranked between $P(z')$ and $P(z'')$, where $P(z) = \Pr(Y = 0|Z = z)$. $h(1, u^*) - h(0, u^*)$ would be understood as the treatment effects of the τ_U-ranked individuals in the subpopulation whose $V-$ ranking is between $P(z')$ and $P(z'')$. When the value of Z changes from z' to z'', their treatment status changes from $y = 1$ to $y = 0$. We call this group "compliers" following the potential outcomes framework.

4.3.3 Applicability to Regression Discontinuity Designs (RDD) and Randomised Trials

Recently, many studies (see Lee and Lemieux (2009), for a survey) adopted regression discontinuity design (RDD) to measure causal effects. Under this design if the continuity condition at the threshold point of the "forcing variable" holds, the causal effects of individuals with the forcing variable just above and below the threshold point are shown to be identified. When the RDD is available, our rank condition[24] is guaranteed to hold, thus, as long as Restriction LDRM is applicable in the context of interest, the proposed model can be applicable to an RD design even when all other variables are not continuous in the treatment - determining variable at the threshold.[25]

5 Further Comments

5.1 Control Function Methods and Discrete Endogenous Variables in *Non-additive* Structural Relations

Control function approaches are understood as a way to correct endogeneity or the selection problem by conditioning on the residuals obtained from the reduced form equations for the endogenous variables in a triangular simultaneous equations system. Control function methods (see Blundell and Powell (2003) for a survey) are not considered to be applicable when the structural function is *non-additive* and the endogenous variable is *discrete*. If the structural relation is additively separable, the control function method can be applied to a case with a discrete endogenous variable. (See Heckman and Robb (1986)).

Imbens and Newey's (2009) control function method under non-additive structural relation is conditioning on the conditional distribution of the endogenous variable

[24] Suppose a threshold point t_0 of a variable T is known by a policy design such that the treatment status (Y) is partly determined by this vairiable. Then we can construct a binary variable Z such that $Z = 1(T > t_0)$. In such a case, our rank condition holds.

[25] For example, age or date of birth, which are used for eligibility criteria, are often only available at a monthly, quarterly, or annual frequency level.

given other covariates as an extra regressor for the outcome equation. Chesher (2003) used the QCFA. This uses the same information as Imbens and Newey (2009), but instead of conditioning on the conditional distributions of the endogenous variable given other covariates, the QCFA can be understood as conditioning on a quantile of the conditional distribution. Imbens and Newey (2009) show that the two control function approaches are equivalent when the endogenous variable is continuous.

When the endogenous variable is discrete, Imbens and Newey (2009)'s approach does not have identifying power.[26] Chesher (2003)'s QCFA fails to produce point identification since the one-to-one mapping between the endogenous variable and the unobserved variable that exists with a continuous endogenous variable does not exist any more with a discrete endogenous variable. Rather, with a discrete endogenous variable, a specific value of the endogenous vairiable maps into a set of values of the unobservable variable, (called **V-set)**, thus, the QCFA with a discrete variable could be roughly described as conditioning on v-quantiles of the conditional distribution of the endogenous variable given covariates, where $v \in$ **V-set**. The smaller the V-set is, the smaller the identified set is. Without imposing further restrictions, a sharp bound cannot be defined. Chesher (2005) suggested to impose monotonicity of $F_{U|V}(u|v)$ in v to define a bound on the value of the structural funciton. This monotonicity restriction is adopted in this paper and Jun, Pinkse, and Xu (2010).

5.2 Nonparametric *Shape* Restrictictions

The identifying power of an eceonometric model comes from restrictions imposed by the model. The restrictions can be categorized into two : those imposed on the structure, and those on data. One could impose restrictions on data - existence of a variable exhibiting certain patterns such as large support condition, rank conditions, or completeness conditions.

Alternatively, one could adopt restrictions on the structure. Apart from Chesher (2005) and Jun, Pinkse, and Xu (2010)'s monotonicity imposed on the distribution of the unobservables, which is mentioned earlier, Manski and Pepper (2000)[27] and Bhattacharya, Shaikh and Vytlacil (2008) adopt certain monotonicity restrictions in the structural relations. Under the MTS (Monotone Treatment Selection) - MTR (Monotone Treatment Reponse) restriction Manski and Pepper (2000) estimated the upper bounds on the returns to schooling. With monotonicity in response, the lower bound is always zero.

Manski and Pepper (2000) develop their arguments by assuming that both selection and response are increasing, but assuming that both are decreasing also leads

[26] Imbens and Newey (2009) defines a bound, but this is for the case in which the common support assumption fails, not for a discrete endogenous variable.

[27] Okumura and Usui (2009) impose concavity to Manski and Pepper (2000) framework and show that the identified interval can be shortened. However, when the endogenous variable is binary, the Okumura and Usui (2009) bounds would be the same as those of Manski and Pepper (2000).

to identification of average effects. However, with the LDRM restriction, weakly increasing response should be matched with weakly increasing selection and vice versa. MTR is equivalent to monotone response assumption in our model, and MTS holds if $F_{U|V}(u|v)$ is weakly decreasing in v over the whole support of U. The LDRM allows the direction (either PDPR or NDNR) of the match to vary over the support of U. On the other hand, the MTR-MTS should be matched for the mean - either positive response with positive selection or negative response with negative selection. Roughly speaking, the LDRM restriction can be described as a *local (quantile)*[28] version of MTR-MTS. Manski and Pepper (2000) identifies average treatment effects, thus the heterogeneity in treatment effects can be found for the subpopulation defined by the observed characteristics, while LDRM model can recover heterogeneity in treatment effects *even* among observationally identical individuals.

Bhattacharya, Shaikh and Vytlacil (2008) compare Shaikh and Vytlacil (2005) bounds with Manski and Pepper (2000)[29] by applying them to a binary outcome - binary endogenous variable case. Bhattacharya, Shaikh and Vytlacil (2008)'s bounds are found under the restriction that the binary endogenous variable is determined by an IV monotonically. When IV, Z, and Y are binary, their monotonicity is equivalent to the monotonicity here. Note also that when Y is binary, we can always reorder 0 and 1 due to the "unordered nature" of a binary variable. In contrast with their claim, when Manski and Pepper (2000) is applied to a binary case, the direction of the monotonicity of response and selection does not have to be determined a priori[30]. Data will inform about the direction of the monotonicity, however, the direction of MTR and MTS should be matched in a certain way[31].

[28]Restriction MTR-MTS is regarding the *mean*, while Restriction LDRM is regarding each point (locally) in the support of the unobserved variable, U. Every point in the support of U can be expressed as quantiles of the distribution of U.

[29]In fact, what they consider is MTR-MIV in Manski and Pepper (2000) with the upper bound of the outcome 1 and the lower bound 0 when the outcome is binary.

[30]When the endogenous variable is ordered discrete with more than two points in the support, the direction should be assumed a priori to find the bounds.

[31]Following the notation of Manski and Pepper (2000) if data show that $E(y|z=0) \leq E(y|z=1)$, then this is the case where non-decreasing MTR and non-decreasing MTS are matched because

$$\begin{aligned} E(y|z &= 0) = E(y(0)|z=0) \overset{MTR}{\leq} E(y(1)|z=0) \\ \overset{MTS}{\leq} E(y(1)|z &= 1) = E(y|z=1). \end{aligned}$$

Whereas if the data show that $E(y|z=0) \leq E(y|z=1)$, then this is the case where non-increasing MTR matched with non-increasing MTS as follows :

$$\begin{aligned} E(y|z &= 0) = E(y(0)|z=0) \overset{MTR}{\geq} E(y(1)|z=0) \\ \overset{MTS}{\geq} E(y(1)|z &= 1) = E(y|z=1). \end{aligned}$$

The counterfactural $E(y(1)|z=0)$ can be bounded by $E(y|z=0)$ and $E(y|z=1)$, and the data will inform us of which is the upper/lower bound - the direction of the match will be determined by

The advantage of the LDRM assumption is that it allows the match to vary across the level of the unobserved characteristic unlike MTS-MTR in Manski and Pepper (2000) or Bhattacharya, Shaikh and Vytlacil (2008). The LDRM model would be useful when the direction of the dependence is likely to be different across different values of the unobserved characteristic. However, LDRM may not be very informative when the outcome is binary in practice, since the values that the partial difference can take are -1,0, and 1, although it is still legitimate to apply the model to binary outcomes in principle.

6 Empirical Illustration - Heterogeneous *Individual* Treatment Responses

By heterogeneous treatment responses I mean idiosyncratic treatment effects even after accounting for observed characteristics[32]. Several studies[33] allowed for individual heterogeneity in response. However, identification is achieved by integrating out the heterogeneity[34] in these studies. By identifying average responses, much information regarding the distributional consequences of a policy - heterogeneity in response - would be lost.

I demonstrate how "partial" information (the signs and the bounds of treatment effects, not the exact size of them) regarding who benefits (individual heterogeneous response) can be recovered from data by using quantiles rather than averages when "who" is indicated by individual observed characeristics and the ranking in the distribution of the unobserved characteristic[35]. This is "illustrated" by examining the effects of the Vietnam-era veteran status on the civilian earnings using the data used in Abadie (2002)[36] - a sample of 11,637 white men, born in 1950-1953, from the March Current Population Surveys of 1979 and 1981-1985. Annual labour earnings are used

data.

[32]This is called "essential heterogeneity" by Heckman, Urzua, and Vytlacil (2006).

[33]The standard linear IV model cannot identify heterogeneus treatment effects. See Heckman and Navarro (2004) and Heckman and Urzua (2009).

For identification under heterogeneous responses see Heckman, Urzua, and Vytlacil (2006) for binary endogenous variable, and Florens, Heckman, Meghir, and Vytlacil (2008), Athey and Imbens (2006), Imbens and Newey (2009),Chernozhukov, Fernandez-Val, and Newey (2009), Hoderlein and White (2009), among others. There is another line of research using random coefficient models to recover the distribution of the response, see Card (2001) and Heckman and Vytlacil (1999) for example.

[34]The averaged objects however can exhibit a certain degree of heterogeneity by allowing for treatment heterogeneity.

[35]Most welfare programs are designed to support certain groups of people. If "who benefits" from such programs could be recovered from data, this would be informative in judging whether the groups targeted by the policy actually benefit from it.

[36]The data are obtainable in Angrist Data Archive :
http://econ-www.mit.edu/faculty/angrist/data1/data

as an outcome, and the veteran status is the binary endogenous variable of concern.

Veterans have been provided with various forms of benefits in terms of insurance, education, etc. How serious the impact of military service on veterans' labour market outcomes, or whether they are compensated for their service enough has been an important political issue and there has not been any consensus on this matter. Angrist (1990) reports negative average impact of veteran status on earnings later in life, which shows that on average military service had a negative impact on earnings possibly due to the loss of labour market experience.

6.1 Bounds on Individual-specific Causal Effects of Vietnam-era Veteran Status on Earnings

By applying his identification results of the marginal distribution of the potential outcomes for compliers, Abadie (2002) reports that military service during the Vietnam era reduces lower quantiles of the earnings distribution, leaving higher quantiles unaffected. The information from the marginal distribution of the potential outcomes (for compliers) may be used to recover QTE, however, it does not reveal any information on individual-specific impact on earnings of Vietnam-era veteran experience.

Let W be annual labour earnings, Y be the veteran status, and Z be the binary variable determined by the draft lottery. Age, race, and gender are controlled so that the subgroup considered is observationally homogenous. The unobserved variables U and V indicate scalar indexes for "earnings potential" and "participation preference" or "aptitude for the army" each. Note that there can be many factors that determine these indexes, but we assume that these multi-dimensional elements can be collapsed into a "scalar" index.

6.1.1 Selection on Unobservables

Enrollment for military service during the Vietnam era may have been determined by the factors which may have been associated with the unobserved earnings potential. This concern about selection on unobservables is caused by several aspects of decision processes both of the military and of those cohorts to be drafted. On the one hand, the military enlistment process selects soldiers on the basis of factors related to earnings potential. For example, the military prefers high school graduates and screens out those with low test scores, or poor health. As a consequence, men with very low earnings potential are unlikely to end up in the army. On the other hand, for some volunteers military service could be a better option because they expected that their careers in the civilian labour market would not be successful, while others with high earnings potential probably found it worthwhile to escape the draft. This shows that the direction of selection could vary with where each individual is located in the distribution of the earnings potential.

6.1.2 Draft Lottery as an Instrument - Exclusion, Rank Condition, and Independence

As in Angrist (1990) the Vietnam era draft lottery is used as an instrument to identify the effects of veteran status on earnings. The lottery was conducted every year between 1970 and 1974. The lottery assigned numbers from 1 to 365 to dates of birth in the cohorts being drafted. Men with the lowest numbers were called to serve up to a ceiling[37]. The ceiling was unknown in advance. I construct a binary IV based on the lottery number the threshold point being chosen as 100 following Abadie(2002).

It would be natural to believe that this IV is not a determinant of earnings, and the unobserved scalar indexes are independent of draft eligibility[38].

To apply the identification results in **Theorem 3,** I investigate first whether the data satisfy Restriction RC in the model. The participation rate[39] among the draft-non-eligible ($Z = 0$) is about 0.14 and the participation rate among the eligible is 0.22.

$$P(Y = 0|Z = 1, X = x) = 0.78 < P(Y = 0|Z = 0, X = x) = 0.86 \qquad \text{(RC)}$$

Thus, $z' = 1$ and $z'' = 0$ in this example. The compliers (or draftees) are defined as those whose V-ranking is between 78% and 86%. Note that the V- ranking is never observed, so we cannot tell whether an individual is a complier or not.

6.1.3 The Result and Causal Interpretation

The bounds on the partial differences, $Q_{W|YZ}(\tau_U|1, z'') - Q_{W|YZ}(\tau_U|0, z')$, are found by the differences in the quantiles of earnings for the veterans who were not eligible and those of non-veterans who were draft-eligible.

LATE can be found by the model in Imbens and Angrist (1994). LATE is found for compliers by integrating out the heterogeneity, therefore, hiding possibly useful information regarding heterogeneity. While Angrist (1990) report negative impact on earnings on average, our quantile based analysis reveals that when age, gender, and race are controlled the veteran status had positive causal impacts for individuals with low earnings potential, but negative causal impacts for individuals with high earnings potential(see Figure 4).

The costs of military service may be larger than the benefits provided by the government for those with high earnings potential, while the benefits provided may

[37]The draft eligibility ceilings were 195 for men born in 1950, 125 for men born in 1951, and 95 for men born in 1952. The eligibility ceiling is determined by the Department of Defense depending on the needs in the year.

[38]There has been some discussion on individuals' draft lottery number caused behavior : some men therefore volunteered in the hope of serving under better terms and gaining some control over the timing of their service. If those who change their behavior according to their draft lottery number show certain patterns in their unobserved factors, then the quantile invariance restriction may be violated.

[39]Note that $P(z)$ is not the usual propensity score, and $1 - P(z)$ is the propensity score.

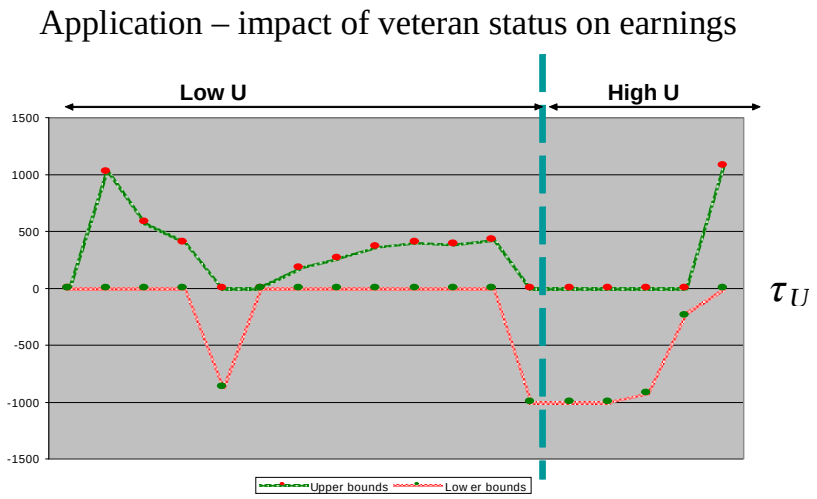

Figure 4: LDRM bounds on heterogeneous treatment effects of Vietnam era veteran status among the observationally similar individuals

be sufficient for those with low earnings potential. Considering the fact that benefits in the form of insurance, pension, or education opportunities should be targeted at people with less potential, the findings indicate that the compensation was enough for this group. However, the Vietnam-era military service may have higher opportunity costs for individuals with high earnings potential. This may be used against conscription.

The results in Figure 4 are interpreted as the causal effects for those who change their participation decision as the value of Z changes. To the extent that we believe the implication from Restriction LDRM on the distribution of the unobservable the bounds would be considered to be informative regarding the population.

7 Conclusion

The presence of endogeneity and discreteness of the endogenous variable causes the loss of the identifying power of the quantile-based control function approach (QCFA) in the sense that the model based on the QCFA does not produce point identification. I propose a model that set identifies the structural features when one of the regressors is ordered discrete. I then apply the model to a binary endogenous variable. This structural approach turns out to be useful in defining the bounds on the heterogeneous individual treatment effects, which have not been studied so far under the structural framework without distributional assumptions.

The set identification result of this paper is applied to recover heterogeneous impacts of the Vietnam-era military service on earnings later in life. As we can

see in this example, average effects may miss much information in some cases. Even though the proposed model can give only partial information on the individual causal effect, this may be useful in some economic contexts, especially when the sign of the effects may be varying across individuals with different characteristics. The causal interpretation is justified on the group of compliers defined by the pair of instrumental values that satisfy the rank condition. Different pairs define different "compliers". Heterogeneity in responses is recovered for different earnings potentials. If there exist heterogeneity in responses between draftees and volunteers, then our findings cannot be extrapolated onto volunteers.

In conclusion, by using nonparametric shape restrictions that can be argued in each economic context, the proposed model provides partial information regarding individual causal effects. This information can be more credible than parametric restrictions to the extent they are justifiable by economic logic. The information on the signs of individual treatment effects is crucial if they vary across the population, since in such a case the average effect would be smaller with different effects with different signs canceled out. This would lead to a misleading conclusion. The model can also be used for robustness checks in data analysis for whether there exists any heterogeneity in causal responses.

References

[1] Abadie, A. (2002), "Bootstrap tests for distributional treatment effects in instrumental variable models," *Journal of the America Statistical Association*, 97, 284-292.

[2] Abadie,A., J. Angrist, and G. Imbens (2002), " Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earning," *Econometrica*, 70(1), 91-117.

[3] Angrist, J. (1990), "Lifetime earnings and the Vietnam era draft lottery : evidence from the social security administrative records," *American Economic Review*, 80, 313-336.

[4] Angrist, J. and G. Imbens (1995), "Two-stage least squares estimation of average causal effects in models wth variable treatment intennsity," *Journal of the American Statistical Association*, 90 (430), 431-442.

[5] Angrist, J. (2004), "Treatment effect heterogeneity in theory and practice," *The Economics Journal*, 114, C52-C83.

[6] Bhattacharya, J., A. Shaikh, and E. Vytlacil (2008), "Treatment effects bounds under monotonicity assumptions : an application to Swan - Ganz Catheterization," *American Economic Review*, 98, 351-346.

[7] Bitler, M., J. Gelbach, and H. Hoynes (2006), "What mean impacts miss : distributional effects of welfare reform experiments," *American Economic Review*, 96, 988-1012.

[8] Blundell, R. and J. Powell (2003), "Endogeneity in nonparametric and semiparametric regression models, " in M. Drewatripont, L.P. Hansen and S.J. Turnovsky (eds.) Advances in Economics and Econometrics : Theory and Applications, Eighth World Congress, Vol II (Cambridge : Cambridge University Press).

[9] Blundell, R. and J. Powell (2004), "Endogeneity in semiparametric binary response models," *Review of Economic Studies*, 71, 655-679.

[10] Chalak, K., S. Schennach, and H. White, "Estimating average marginal effects in nonseparable structural systems," mimeo.

[11] Chernozhukov, V. and C. Hansen (2005), "An IV Model of quantile treatment effects," *Econometrica*, Vol.73, No. 1, 245-261.

[12] Chernozhukov, V., H. Hong, and E. Tamer, (2007), "Estimation and confidence regions for parameter sets in econometric models, " *Econometrica*, 75. 1243-1284.

[13] Chernozhukov, V., I. Fernadez-Val, and W. Newey (2009), "Quantile and average effects in nonseparable panel model," mimeo.

[14] Chernozhukov, V., I. Fernandez-Val, and B. Melly (2010), "Inference on counterfactual distributions in nonseparable Models," mimeo.

[15] Chesher, A. (2003), "Identification in nonseparable models," *Econometrica*, 71, 1405-1441.

[16] Chesher, A. (2005), "Nonparametric identification under discrete variation," *Econometrica*, 73(5), 1525-1550.

[17] Chesher, A. (2007), "Instrumental values," *Journal of Econometrics*, 139, 15-34.

[18] Chesher, A. (2009), "Excess heterogeneity, endogeneity, and index restrictions," *Journal of Econometrics*, 152, 35-47.

[19] Deaton, A. (2009), "Instruments of development : randomization in the Tropics, and the search for the elusive keys to economic development," NBER working paper no. 14690.

[20] Doksum, K. (1974) "Empirical probability plots and statistical inference for nonlinear models in the two sample case," *The annals of Statistics*, 2. 267-277.

[21] Fan, Y. and S. Park (2010), "Sharp bounds on the distribution of treatment effects and their statistical inference," *Econometric Theory*, 26, 931-951.

[22] Firpo, S. (2007), "Efficient semiparametric estimation of quantile treatment effects," *Econometrica*, 75. 259-276.

[23] Firpo. S. and G. Ridder (2008), "Bounds on functionals of the distribution of treatment effects," IEPR working paper 08.09.

[24] Florens, J., J. Heckman, C. Meghir, and E. Vytlacil (2008), "Identification of treatment effects using control functions in models with continuous endogenous treatment and heterogeneous treatment effects," *Econometrica*, 76. 1191-1206.

[25] Frolich, M. and B. Melly (2008), "Quantile treatment effects in the regression discontinuity design," mimeo.

[26] Hahn, J. and G. Ridder (2011), "Conditional moment restrictions and triangular simulatenous equations," *Review of Economics and Statistics*, forthcoming.

[27] Heckman, J., J. Smith, and N. Clements (1997), "Making the most out of program evaluations and social experiments accounting for heterogeneity in program impacts," *Review of Economic Studies*, 64, 487-535.

[28] Heckman, J., S. Urzua, and E. Vytalcil (2006), "Understanding instrumental variables in models with essential heterogeneity," *The Review of Economics and Statistics*, 88, 389-432.

[29] Heckman, J. and E. Vytalcil (2001), "Instrumental variables, selection models, and tight bounds on the average treatment effect," in M. Lechner and F. Pfeiffer (Eds.), *Econometric evaluation of labour market policies*, New York.

[30] Heckman, J. and S. Urzua (2009), "Comparing IV with structural models : what simple IV can and cannot identify," NBER working paper. no. 14760.

[31] Hoderlein, S. and E. Mammen (2007), "Identification of marginal effects in nonseparable models without monotonicity," *Econometrica*, 75. 1513-1518.

[32] Horowitz, J. and C. Manski (2000), "Nonparametric analysis of randomized experiments with missing covariate and outcome data," *Journal of the American Statistical Association*, 95. 77-84.

[33] Hurwicz, L. (1950) "Generalization of the concept of identification," in Statistical Inference in Dynamic Economic Models, Cowles Commision Monograph 10. Wiley, New York.

[34] Imbens, G. (2009), "Better LATE than Nothing : some comments on Deaton (2009) and Heckman and Urzua (2009), " mimeo.

[35] Imbens, G. and J. Angrist (1994), "Identification and estimation of Local average treatment effects, " *Econometrica*, 62, 467-476.

[36] Imbens, G. and D. Rubin (1997), Estimating outcome distributions for compliers in instrumental variable models," *Review of Economic Studies*, 64, 555-574.

[37] Imbens, G. and W. Newey (2009), "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77(5),1481 - 1512.

[38] Jun, S., J. Pinkse, and H. Xu (2010), "Tighter bounds in triangular system," mimeo.

[39] Kitagawa, T. (2009), "Identification region of the potential outcome distributions under instrument independence," cemmap working paper.

[40] Koenker, R. (2005), *Quantile Regression, Econometric Society Monographs*, Cambridge University Press.

[41] Lee, D. and T. Lemieux (2009), "Regression discontinuity designs in economics," NBER working paper.

[42] Lee, J. (2011) "Beyond averages : a revisit of Angrist (1990) - what do QTE, IVQR, and partial differences of a structural relation measure?," mimeo.

[43] Lee, J. (2011) "Falsifiability of economie assumptions under partial identification," mimeo.

[44] Lehman, E. (1974), Nonparametric Statistical Methods Based on Ranks, San Francisco, Holden-Day.

[45] Manski, C. (2003), Partial identification of probability distribution, Springer-Verlag.

[46] Manski, C. and J. Pepper (2000), "Monotone instrumental variables : with an application to the returns to schooling," *Econometrica*, 68(4), 997-1010.

[47] Matzkin, R. (2003), " Nonparametric estimation of nonadditive random functions," *Econometrica*, 71, 1332-1375.

[48] Nekipelov, D. (2009), "Endogenous multi-valued treatment effect model under monotonicity," mimeo.

[49] Okumura, T. and E. Usui (2009), "Concave-monotone treatment response and monotone treatment selection : with returns to schooling application," mimeo.

[50] Roehrig, C (1988), "Conditions for identification in nonparametric and parametric models," *Econometrica*, Vol. 56, No. 2, pp. 433-447.

[51] Vytlacil, E. and N. Yildiz (2007), "Dummy endogenous variables in weakly separable models," Econometrica, 75(3), 757-779.

Appendix - proofs

A.1 Proof of Theorem 1

Proof. We adopt Lemma 2 in Appendix in Chesher (2005).

Recall that $\mathbf{V} \equiv (V_L, V_U]$, where $V_L = \max_{z \in \overline{z}_m} P^{m-1}(z)$, and $V_U = \max_{z \in \overline{z}_m} P^{m+1}(z)$, and where $\overline{z}_m$ is the set of values of the values of Z that satisfy the rank condition.

Suppose that $Q_{U|VZ}(\tau_U|\tau_V, z)$ is weakly increasing in $\tau_V \in \mathbf{V}$. Then by Lemma 2 in Chesher (2005) we have for $Y = y^m$,

$$\begin{aligned} h(y^m, Q_{U|VZ}(\tau_U|V_L, z'_m)) &\leq Q_{W|YZ}(\tau_U|y^m, z'_m) && \text{(A-1)} \\ &\leq h(y^m, Q_{U|VZ}(\tau_U|P^m(z'_m), z'_m)) \\ h(y^m, Q_{U|VZ}(\tau_U|V_L, z''_m)) &\leq Q_{W|YZ}(\tau_U|y^m, z''_m) && \text{(A-2)} \\ &\leq h(y^m, Q_{U|VZ}(\tau_U|P^m(z''_m), z''_m)) \end{aligned}$$

and for $Y = y^{m+1}$

$$\begin{aligned} h(y^{m+1}, Q_{U|VZ}(\tau_U|P^m(z'_m), z'_m)) &\leq Q_{W|YZ}(\tau_U|y^{m+1}, z'_m) && \text{(A-3)} \\ &\leq h(y^{m+1}, Q_{U|VZ}(\tau_U|V_U, z'_m)) \\ h(y^{m+1}, Q_{U|VZ}(\tau_U|P^m(z''_m), z''_m)) &\leq Q_{W|YZ}(\tau_U|y^{m+1}, z''_m) && \text{(A-4)} \\ &\leq h(y^{m+1}, Q_{U|VZ}(\tau_U|V_U, z''_m)) \end{aligned}$$

Under Restriction RC, $P^m(z'_m) \leq \tau_V \leq P^m(z''_m)$, when $Q_{U|VZ}(\tau_U|\tau_V, z)$ is weakly increasing in v, then :

$$\begin{aligned} Q_{U|VZ}(\tau_U|\tau_V, z''_m) &\leq Q_{U|VZ}(\tau_U|P^m(z''_m), z''_m) \\ Q_{U|VZ}(\tau_U|P^m(z'_m), z'_m) &\leq Q_{U|VZ}(\tau_U|\tau_V, z'_m) \end{aligned}$$

and because h is weakly increasing in U,

$$\begin{aligned} h(y^m, Q_{U|VZ}(\tau_U|\tau_V, z''_m)) &\leq h(y^m, Q_{U|VZ}(\tau_U|P^m(z''_m), z''_m)) && \text{(B-1)} \\ h(y^m, Q_{U|VZ}(\tau_U|P^m(z'_m), z'_m)) &\leq h(y^m, Q_{U|VZ}(\tau_U|\tau_V, z'_m)). && \text{(B-2)} \end{aligned}$$

Combining (A-4) and (B-1) we can find the upper bound on $h(y^m, Q_{U|VZ}(\tau_U|\tau_V, z''_m))$

$$\begin{aligned} h(y^m, Q_{U|VZ}(\tau_U|\tau_V, z''_m)) &\leq h(y^m, Q_{U|VZ}(\tau_U|P^m(z''_m), z''_m)) \\ &\leq h(y^{m+1}, Q_{U|VZ}(\tau_U|P^m(z''_m), z''_m)) \\ &\leq Q_{W|YZ}(\tau_U|y^{m+1}, z''_m) \end{aligned}$$

The first inequality is due to (B-1) and the second inequality is due to Restriction LDRM, and the third inequality is due to (A-4).

The lower bound on $h(y^m, Q_{U|VZ}(\tau_U|\tau_V, z'_m))$ can be found by (A-3) and (B-2) :

$$Q_{W|YZ}(\tau_U|y^m, z'_m) \leq h(y^m, Q_{U|VZ}(\tau_U|P^m(z'_m), z'_m)) \leq h(y^m, Q_{U|VZ}(\tau_U|\tau_V, z'_m)).$$

The first inequality is due to (A-3), the second is due to (B-2).

Finally, under the conditional quantile invariance (C-QI) and exclusion Restrictions (QCFA), there is for $z \in \{z'_m, z''_m\}$ for $u^* = Q_{U|VZ}(\tau_U|\tau_V, z)$,

$$Q_{W|YZ}(\tau_U|y^m, z'_m) \leq h(y^m, u^*) \leq Q_{W|YZ}(\tau_U|y^{m+1}, z''_m)$$

Similarly, when $Q_{U|VZ}(\tau_U|\tau_V, z)$ is weakly decreasing in $\tau_V \in \mathbf{V}$, we have

$$Q_{W|YZ}(\tau_U|y^{m+1}, z''_m) \leq h(y^m, u^*) \leq Q_{W|YZ}(\tau_U|y^m, z'_m)$$

■

A.2 Proof of Theorem 2 : Sharpness[40]

Define

$$h^{-1}(y^m, w) \equiv \sup_u \{u : h(y^m, u) \leq w\}. \qquad (*)$$

This implies

$$h(y^m, h^{-1}(y^m, w)) \leq w \qquad (**)$$

with equality holding when $h(y^m, u)$ is strictly increasing in u.

Recall that the structural feature of interest, is the value of the structural function evluated at $Y = y^m$ and $U = u^*$, where $u^* = Q_{U|VZ}(\tau_U|\tau_V, z)$ for $\tau_V \in [P^m(z'_m), P^m(z''_m)]$.

What is required to show sharpness[41] (following Lemma 1 in Section 2), is to construct a structure (S^a) such that (A) for any value, $w^* \in I(\tau, y^m, \overline{z}_m)$, $w^* = h^a(y^m, u^*)$, (B) is

[40] A more detailed version of proof can be found in the author's webpage : http://www.homepages.ucl.ac.uk/~uctpjil/

[41] In contrast with sharpness proofs in the potential outcomes approach (see for example, Firpo and Ridder (2008), Heckman and Vytlacil (2001)), to show whether the points in the identified set are consistent with the model we need to construct the underlying structural relation and the distribution of the unobservables since the model is characterized by the restrictions imposed on them.

Consider a binary endogenous variable case. In the potential outcomes framework $F_{W_1W_0|X}$ is the hidden data generating process to be recovered from the observed data, $F_{W|X}$, thus, sharpness proof involves construction of $F_{W_1W_0|X}$ using $F_{W|X}$ that is consistent with the model and data. In the structural approach the underlying economic data generating process is $\{h(1, u^*), h(0, u^*), F_{U|V}\}$ for given u^* in the triangular structure when ignoring other covariates, X. Therefore, sharpness proof involves the construction of $\{h(1, u^*), h(0, u^*), F_{U|V}\}$ using $F_{W|YX}$ and consistency with the model and with the data should be shown.

admitted by LDRM model ($S^a \in M^{LDRM}$) and (C) the constructed structure is observationally equivalent to the true structure ($F^a_{W|YZ} = F^0_{W|YZ}$). In Part 1 we construct a structure $S^a \equiv \{h_a, F^a_{U|VZ}(u|v,z)\}$ and in Part 2 we show (A),(B), and (C)

Part 1 - Construction of an admitted and observationally equivalent (o.e.) structure $S^a \equiv \{h_a, F^a_{U|VZ}\}$, such that $\theta(S^a) = a$

The candidate structure is constructed such that all the values of h and $F_{U|VZ}$ can be determined. Some of the restrictions such as Restriction LDRM imposed in the model are regarding local properties of the structure, while some of the restrictions such as monotonicity of h in u or whether the constructed distribution of the unobservables is weakly increasing should be shown for all the points in the support of U. To show such restrictions all the values in the support of the arguments of the structural function and the distribution of the unobservables need to be determined by the construction. Note also that there can be other ways of construction. The distribution of observables, $F_{W|YZ}$, is used in the construction of h_a and $F^a_{U|VZ}(u|v,z)$, such that by the interaction of h_a and $F^a_{U|VZ}(u|v,z)$, $F_{W|YZ}$ can be generated.

1-A Construction of a structural function[42]

Let $I(\tau, y^m, \overline{z}_m)$ denote the identified interval, say, $[Q_{W|YZ}(\tau_U|y^m, z'_m), Q_{W|YZ}(\tau_U|y^{m+1}, z''_m)]$. The structural function is constructed as

$$\begin{aligned} h_a(y^m, u^*) &\equiv Q^0_{W|YZ}(\overline{\tau}_m|y^m, z) \text{ for some } \overline{\tau}_m \text{ and } \overline{v}_m \\ \text{where } u^* &\equiv Q^a_{U|VZ}(\tau_U|\tau_V, z) \\ &= Q^a_{U|VZ}(\overline{\tau}_m|\overline{v}_m, z) \\ &= Q^0_{U|YZ}(\overline{\tau}_m|y^m, z) \\ \text{for } m &= 1, 2, ..., M. \end{aligned} \tag{S1}$$

For given $V = \tau_V$ and $Y = y^m$, by varying $\tau_U \in (0,1)$, the whole values of $h_a(y^m, u^*)$ is determined by (S1). On the other hand, the whole values of the structural function for given τ_U can be defined.

1-B Construction of the distribution of the unobservables.

For a given structural relation, h_a, and given the values of $Y = y^m$ and an arbitrary value of $U = u \in (0,1)$ can be written as

$$h_a^{-1}(y^m, w^\#)$$

for some $w^\#$ by (C*). Then we can find $w^1, w^2, ..., w^M$ for the fixed value u such that

$$w^l = h_a(y^l, u), \text{ for } l = 1, 2, ..., M$$

[42]Unlike other bounds studies we need to construct the structural relation since we exploited restrictions imposed on the structure such as monotonicity of the structural function in the unobservable and restriction LDRM.

so that

$$h_a^{-1}(y^m, w^{\#}) = h_a^{-1}(y^1, w^1) = h_a^{-1}(y^2, w^2) = \cdots = h_a^{-1}(y^M, w^M)$$

for continuous W .

Let $SUPP(Z)$ be the support of Z. For an arbitrary value $u \in (0,1)$, u is expressed as $u = h_a^{-1}(y^m, w^{\#})$, for some $w^{\#}$. For a given $z \in SUPP(Z)$, for any $u, v \in (0,1) \times (0,1)$, $F^a_{U|VZ}(u|v,z)$ is constructed as follows :

$$\begin{aligned}
& F^a_{U|VZ}(u|v,z) \\
= \; & F^a_{U|VZ}(\underbrace{h_a^{-1}(y^m, w^{\#})}_{u}|v,z) \\
\equiv \; & \left(\begin{array}{cccc}
F^0_{W|YZ}(w^1|y^1,z), & & \text{if} & 0 < v \le P^1 \\
F^0_{W|YZ}(w^2|y^2,z), & & \text{if} & P^1 < v \le P^2 \\
& \cdots & & \\
F^0_{W|YZ}(\underset{\uparrow\uparrow}{w^{\#}}|y^{m-1},z), & (*) & \text{if} & P^{m-2} < v \le P^{m-1} \\
F^0_{W|YZ}(\underset{\uparrow\uparrow}{w^{\#}}|y^m,z), & (*) & \text{if} & P^{m-1} < v \le P^m \\
F^0_{W|YZ}(w^{m+1}|y^{m+1},z), & & \text{if} & P^m < v \le P^{m+1} \\
& \cdots & & \\
F^0_{W|YZ}(w^M|y^M,z), & & \text{if} & P^{M-1} < v \le 1
\end{array} \right) \qquad \text{(S2)}
\end{aligned}$$

where $w^1, w^2, ..., w^M$ are found such that

$$\begin{aligned}
w^l &= h_a(y^l, u), \text{ and} \\
P^l &= \max_{z \in SUPP(Z)} \{P^l(z)\}, \; l \neq m-1, m \\
P^{m-1} &= \min_{z \in \overline{z}_m} \{P^m(z)\} \text{ and } P^m = \max_{z \in \overline{z}_m} \{P^m(z)\} \\
l &= 1, 2, ..., M
\end{aligned}$$

Remarks

- For any given value v, if $v \in (P^{l-1}, P^l]$, uses $Y = y^l$, as the conditioning value.
- If u is expressed as $h_a^{-1}(y^m, w^{\#})$ for some $w^{\#}$, in the identified interval, and $v \in (P^{l-1}, P^l]$, where $l \neq m-1$ and m, then find the value, w^l such that

  $$w^l = h_a(y^l, u)$$

  then assign the value $F^a_{U|VZ}(u|v,z) \equiv F^0_{W|YZ}(w^l|y^l,z)$.
- In (*) in (S2) if $u = h_a^{-1}(y^m, w^{\#})$ and $v \in (P^{m-2}, P^{m-1}]$, then assign the value $F^a_{U|VZ}(u|v,z) \equiv F^0_{W|YZ}(w^{\#}|y^{m-1},z)$. Note the value, w, (indicated by $\uparrow\uparrow$) is assigned.
- In (*) in (S2) if $u = h_a^{-1}(y^m, w^{\#})$ and $v \in (P^{m-1}, P^m]$, then assign the value $F^a_{U|VZ}(u|v,z) \equiv F^0_{W|YZ}(w^{\#}|y^m,z)$.

- $\{P^l\}_{l=1}^M$ is a weakly increasing sequence. The partition of the support of $V, (0,1)$, by $\{P^l\}_{l=1}^M$ is determined once a variable Z is given.
- $P^{m-1} = \min_{z \in \overline{z}_m}\{P^m(z)\}$ and $P^m = \max_{z \in \overline{z}_m}\{P^m(z)\}$ is chosen to guarantee conditional quantile invariance restriction, which locally holds for τ_U quantile of U given V and Z, for the range of V specified by the rank condition.
- If W is discrete, $F^a_{U|VZ}(u|v,z)$ should be a step function in u as well as in v. For notational simplicity, we assume that W is continuous. Other parts in the proof are not affected when W is discrete, but in each part of the proof extra complication of notation occurs.

Proof of proper distribution

It is required to check whether the constructed distribution is proper : since each $F^0_{W|YZ}(w|y^m,z)$, for all $m \in \{1,2,...,M\}$ is a proper distribution, $F^0_{W|YZ}(w|y^m,z)$ lies between zero and one, and weakly increasing in w. Thus, the constructed distribution $F^a_{U|VZ}(u|v,z)$ lies between zero and one, but to guarantee nondecreasing property of $F^a_{U|VZ}(u|v,z)$ in u, we need to show that as w increases, $u = h_a^{-1}(y,w)$ increases for given v and z. This can be shown by Lemma A.1.

Lemma A1 For given v and z, $F^a_{U|VZ}(u|v,z)$ weakly increases in u.

Proof. Consider two distinct values u' and u''. We express u' and u'' using h_a^{-1}, for given $Y = y^m$ as the following

$$\begin{aligned} u' &= h_a^{-1}(y^m, w') \\ u'' &= h_a^{-1}(y^m, w'') \end{aligned}$$

Fix $V = v$ and $Z = z$ and suppose that $V = v$ and $Z = z$ corresponds to $Y = y^l$, $l = 1,2,...,M$. Then by (S2) we have for some τ', τ'', w'_l, and w''_l

$$\begin{aligned} \tau' &= F^a_{U|VZ}(u'|v,z) \\ &\overset{(S2)}{=} \left\{ \begin{matrix} F^0_{W|YZ}(w'_l|y^l,z), & \text{if } l \neq m-1, m \\ F^0_{W|YZ}(w'|y^l,z), & \text{if } l = m-1,\ m \end{matrix} \right\} \\ \text{where } u' &= h_a^{-1}(y^m,w') = h_a^{-1}(y^l,w'_l) \end{aligned} \tag{1-1}$$

and

$$\begin{aligned} \tau'' &= F^a_{U|VZ}(u''|v,z) \\ &\overset{(S2)}{=} \left\{ \begin{matrix} F^0_{W|YZ}(w''_l|y^l,z), & \text{if } l \neq m-1, m \\ F^0_{W|YZ}(w''|y^l,z), & \text{if } l = m-1,\ m \end{matrix} \right\} \\ \text{where } u'' &= h_a^{-1}(y^m,w'') = h_a^{-1}(y^l,w''_l). \end{aligned} \tag{1-2}$$

If we can show that $w_l'' \geq w_l'$, when $u'' \geq u'$, then the proof is done because then the assigend value following (S2) for $F^a_{U|VZ}(u''|v,z)$ is larger than $F^a_{U|VZ}(u'|v,z)$. Suppose

$$\begin{aligned} u'' &= Q^a_{U|VZ}(\tau''|v,z) \\ &\geq Q^a_{U|VZ}(\tau'|v,z) = u' \\ \text{for } \tau'' &\geq \tau'. \end{aligned}$$

Then $w_l'' \geq w_l'$, since from (1-1) and (1-2)

$$w_l'' = Q^0_{W|YZ}(\tau''|y^l,z) \geq Q^0_{W|YZ}(\tau'|y^l,z) = w_l'$$

whenever $u'' = Q^a_{U|VZ}(\tau''|v,z) \geq Q^a_{U|VZ}(\tau'|v,z) = u'$, that is, whenever $\tau'' \geq \tau'$. ■

Part 2

Part 2 - A :

Note that under Restriction Common Support, any point in the identified interval, $w^* \in I(\tau, y^m, \overline{z}_m)$ can be written as (see <Figure 5>)[43]

$$w^* = Q^0_{W|YZ}(\overline{\tau}_m|y^m, z_m') \text{ for some } \overline{\tau}_m \geq \tau_U.$$

That is,

$$\overline{\tau}_m = F^0_{W|YZ}(w^*|y^m, z_m') \text{ for some } \overline{\tau}_m \geq \tau_U$$

Note also that for any $v \in (P^{m-1}, P^m]$ by construction from (S2)

$$F^a_{U|VZ}(\underbrace{h_a^{-1}(y^m,w^*)}_{\overline{\tau}_m\text{-quantile of } F^a_{U|VZ}} |v, z_m') \overset{(S2)}{=} F^0_{W|YZ}(w^*|y^m,z_m') = \overline{\tau}_m,$$

thus, by definition of quantiles,

$$h_a^{-1}(y^m,w^*) = Q^a_{U|VZ}(\overline{\tau}_m|v,z_m') \text{ for } \textbf{some } v \in (P^{m-1},P^m] \qquad (h_a - a)$$

For a given value, w^*, in the indentified interval, $\overline{\tau}_m(\geq \tau_U)$ is determined by w^*. Then $(h_a - a)$ holds for a range of values of $v \in (P^{m-1}, P^m]$. Now we choose $\overline{v}_m \in (P^{m-1}, P^m]$ such that

$$\begin{aligned} u^* &\equiv Q^a_{U|VZ}(\tau_U|\tau_V, z_m') \qquad (h_a - b) \\ &= Q^a_{U|VZ}(\overline{\tau}_m|\overline{v}_m, z_m'). \end{aligned}$$

Then by inverting h_a^{-1} in $(h_a - a)$, for given w^* and $\overline{\tau}_m(\geq \tau_U)$, we have

$$\begin{aligned} w^* &= h_a(y^m, Q^a_{U|VZ}(\overline{\tau}_m|\overline{v}_m, z_m')) \\ &= h_a(y^m, u^*). \end{aligned}$$

[43] Alternatively, one can find $\overline{\tau}_m$ such that $w^* = Q^0_{W|YZ}(\overline{\tau}_m|y^m, z_m'')$ for some $\overline{\tau}_m \leq \tau_U$

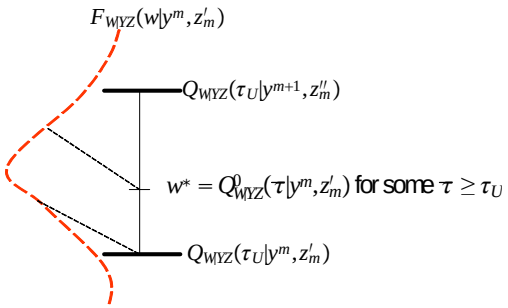

Figure 5: Any point in the interval, $w^* \in I(\tau, m, \overline{z}_m)$, can be expressed using the quantiles of $F_{W|YZ}(w|y^m, z'_m)$ under the common support restriction.

Thus, this construction guarantees that the constructed structural function crosses the arbitrary value in the identified interval

$$w^* = h_a(y^m, u^*),$$

that is, there exists a structural relation (that satisfies all the restrictions imposed by the model, which will be shown in the next section) which crosses an arbitrary point, w^*, in the identified interval.

Part 2 - B : Observational equivalence[44] $(F^a_{W|YZ} = F^0_{W|YZ})$

We need to show that $F^a_{W|YZ} = F^0_{W|YZ}$, for $S^a = \{h_a, F^a_{U|VZ}\}$ constructed as in Part 1 : for $p^a_m = P^m - P^{m-1}$, for all $m \in \{1, 2, ..., M\}$,

$$\begin{aligned} F^a_{W|YZ}(w|y^m, z) &= \frac{1}{p^a_m}\int_{P^{m-1}}^{P^m} F^a_{U|VZ}(h_a^{-1}(y^m, w)|s, z)ds \\ &= \frac{1}{p^a_m}\int_{P^{m-1}}^{P^m} F^0_{W|YZ}(w|y^m, z)ds \\ &= F^0_{W|YZ}(w|y^m, z) \end{aligned}$$

the first equality is due to Lemma 1 in Chesher (2005), the second equality is due to construction in (S2), that is, $F^a_{U|VZ}(h_a^{-1}(y^m, w)|v, z) = F^0_{W|YZ}(w|y^m, z)$, for $v \in (P^{m-1}, P^m]$ and the last equality is due to integration over the constant and the definition of p^a_m.

Part 2 - C : Admissibility by the model $S^a \in M^{LDRM}$

0. Rank condition : this can be shown using data. I suppose this restriction is satisfied.

1. Monotonicity of $h_a(y^m, u)$ **in** u

[44] That is, the data distribution that is generated by the structure constructed in part 1 is actually what we observe. Note that this can be shown because we have constructed the structure using the observed distribution.

I consider whether $h_a(y,u)$ is nondecreasing in u. Recall that

$$h_a(y^m,u) = h_a(y^m, Q^a_{U|VZ}(\overline{\tau}_m|\overline{v}_m,z)) \overset{\text{by } (S1)}{\equiv} Q^0_{W|YZ}(\overline{\tau}_m|y^m,z)$$

by choosing $\overline{v}_m$ such that $u = Q^a_{U|VZ}(\tau_U|\tau_V,z) = Q^a_{U|VZ}(\overline{\tau}_m|\overline{v}_m,z) = Q_{U|YZ}(\overline{\tau}_m|y^m,z)$, for $\forall \tau_U,\tau_V,\overline{\tau}_m \in (0,1)$ and $\overline{v}_m \in (P^{m-1},P^m]$.

- First, fix $\overline{v}_m$, then $h_a(y^m,u)$ is weakly increasing in u since higher $\overline{\tau}_m$ implies higher $u = Q_{U|VZ}(\overline{\tau}_m|\overline{v}_m,z)$, as well as higher $Q_{U|YZ}(\overline{\tau}_m|y^m,z)$.
- Next fix $\overline{\tau}_m$, if we observe higher u, then it is because of higher $\overline{v}_m$ if $F_{U|V}(u|\overline{v}_m,z)$ is non-increasing in $\overline{v}_m$ and lower $\overline{v}_m$ if $F_{U|VZ}(u|\overline{v}_m,z)$ is nondecreasing in $\overline{v}_m \in (P^{m-1},P^m]$. However, regardless of the direction of monotonicity, for $\overline{v}_m \in (P^{m-1},P^m]$, $Y = y^m$. Thus, the value of $\overline{v}_m$ does not affect the value of h_a as long as Y is fixed at $Y = y^m$. That is, for fixed $\overline{\tau}_m$, and Y, $h_a(y,u)$ is constant as u increases due to change in $\overline{v}_m$.

2. Conditional Quantile Invariance : $u^* \equiv Q^a_{U|VZ}(\tau_U|\tau_V,z)$ is invariant with respect to $z \in \overline{z}_m \equiv \{z'_m, z''_m\}$, for $\tau_V \in [P^m(z'_m), P^m(z''_m)]$. Note that there should exist a true structure, $S^0 = \{h_0, F^0_{U|VZ}\} \in \mathcal{M}^{LDRM} \cap \Omega_0$, that generates the data we observe. The distinction of the true structure, S^0 from the constructed structure, S^a, should be noted in this proof. For $u^* = h_a^{-1}(y^m,w^*)$

$$\begin{aligned}
\tau_U &\equiv F^a_{U|VZ}(u^*|\tau_V,z'_m) \\
&= F^0_{W|YZ}(w^*|y^m,z'_m) \\
&= \frac{1}{p_m(z'_m)}\int_{P^{m-1}(z'_m)}^{P^m(z'_m)} F^0_{U|VZ}(h_0^{-1}(y^m,w^*)|s,z'_m)ds \\
&= \frac{\Pr(U \le h_0^{-1}(y^m,w^*) \cap P^{m-1}(z'_m) \le V \le P^m(z'_m))}{p_m(z'_m)} \\
&= F^0_{U|V}(h_0^{-1}(y^m,w^*)|V \in (P^{m-1}(z'_m), P^m(z'_m)]) \\
&= F^0_{U|Y}(h_0^{-1}(y^m,w^*)|y^m) \\
&= F^0_{U|Y}(u^*|y^m)
\end{aligned}$$

the first equality is by construction in (S2), the second equality is due to Lemma 1 in Chesher (2005), and the third equality follows by integration. The fourth equality is by definition of the conditional probability, the fifth equality is due to how the value of Y is determined. Similarly for $Z = z''_m$,

$$
\begin{aligned}
\tau_U &\equiv F^a_{U|VZ}(u^*|\tau_V, z''_m) \\
&= F^0_{W|YZ}(w^*|y^m, z''_m) \\
&= \frac{1}{p_m(z''_m)} \int_{P^{m-1}(z''_m)}^{P^m(z''_m)} F^0_{U|VZ}(h_0^{-1}(y^m, w^*)|s, z''_m) ds \\
&= \frac{\Pr(U \le h_0^{-1}(y^m, w^*) \cap P^{m-1}(z''_m) \le V \le P^m(z''_m))}{p_m(z''_m)} \\
&= F^0_{U|V}(h_0^{-1}(y^m, w^*)|V \in (P^{m-1}(z''_m), P^m(z''_m)]) \\
&= F^0_{U|Y}(h_0^{-1}(y^m, w^*)|y^m) \\
&= F^0_{U|Y}(u^*|y^m)
\end{aligned}
$$

yielding $u^* = Q^a_{U|VZ}(\tau_U|\tau_V, z'_m) = Q^a_{U|VZ}(\tau_U|\tau_V, z''_m) = Q^0_{U|Y}(\tau_U|y^m)$, invariant with respect to $z \in \overline{z}_m$.

3. LDRM :

(1) First, it is noted that $F^a_{U|VZ}(u|v, z)$ is monotonic in v, for $u \in \mathbf{U}, v \in \mathbf{V}$, where $\mathbf{U}$ and $\mathbf{V}$ are defined previously in Restrction LDRM. This is so since $F^a_{U|VZ}(u|v, z)$ is defined as a step function, for the range of $V = (P^{m-1}(z), P^{m+1}(z)]$ only two constants ($F^0_{W|YZ}(w^*|y^m, z)$, and $F^0_{W|YZ}(w^{m+1}|y^{m+1}, z)$) should be considered, and with two constants, monotonicity always holds.

(2) Now we check whether the constructed $S^a = \{h_a, F^a_{U|VZ}\}$ satisfies the specified match. Suppose for some $\tau''_m, \tau''_{m+1}, P^m(z''_m)$ and $P^{m+1}(z''_m)$,

$$
\begin{aligned}
u^* &\equiv Q^a_{U|VZ}(\tau_U|\tau_V, z) \qquad (3\text{-}1) \\
&= Q^a_{U|VZ}(\tau''_m|P^m(z''_m), z''_m) = Q^a_{U|VZ}(\tau''_{m+1}|P^{m+1}(z''_m), z''_m),
\end{aligned}
$$

This can be shown by observing the sign of

$$
\begin{aligned}
& h_a(y^m, u^*) - h_a(y^{m+1}, u^*) \qquad (3\text{-}2) \\
= \; & h_a(y^m, Q^a_{U|VZ}(\tau''_m|P^m(z''_m), z''_m)) - h_a(y^{m+1}, Q^a_{U|VZ}(\tau''_{m+1}|P^{m+1}(z''_m), z''_m)) \\
= \; & Q^0_{W|YZ}(\tau''_m|y^m, z''_m) - Q^0_{W|YZ}(\tau''_{m+1}|y^{m+1}, z''_m),
\end{aligned}
$$

where the first equality follows by (3-1), and the second equality is by construction in (S1).

To determine the sign of $h_a(y^m, u^*) - h_a(y^{m+1}, u^*)$, it is required to determine the sign of $Q^0_{W|YZ}(\tau''_m|y^m, z''_m) - Q^0_{W|YZ}(\tau''_{m+1}|y^{m+1}, z''_m)$. We first fix $U = u^*$, and vary the value of V. Then use the monotonicity of $F_{U|VZ}$ in v in a certain range specified in the restriction and see if the match holds.

(3-3)-(3-5) link the distribution of the unobservables with the distribution of the observables, and they are found by expressing u^* using h_a^{-1} and the construction in Part 1.

For $u^* = h_a^{-1}(y^m, w^*)$ and $v = P^m(z''_m)$, let τ''_m be

$$\begin{aligned}\tau''_m &\equiv F^a_{U|VZ}(u^*|P^m(z''_m), z''_m) \\ &= F^a_{U|VZ}(\underbrace{h_a^{-1}(y^m, w^*)}_{\boldsymbol{\tau}''_m-\text{ quantile of } F^a_{U|VZ}} |P^m(z''_m), z''_m) \\ &= F^0_{W|YZ}(\underbrace{w^*}_{\boldsymbol{\tau}''_m-\text{ quantile of } F^0_{W|YZ}} |y^m, z''_m)\end{aligned} \tag{3-3}$$

Note that for $u^* = h_a^{-1}(y^{m+1}, w^{m+1})$ and $v = P^{m+1}(z''_m)$, let τ''_{m+1} be:

$$\begin{aligned}\tau''_{m+1} &\equiv F^a_{U|VZ}(u^*|P^{m+1}(z''_m), z''_m) \\ &= F^a_{U|VZ}(\underbrace{h_a^{-1}(y^{m+1}, w^{m+1})}_{\boldsymbol{\tau}''_{m+1}-\text{ quantile of } F^a_{U|VZ}} |P^{m+1}(z''_m), z''_m) \\ &= F^0_{W|YZ}(\underbrace{w^{m+1}}_{\boldsymbol{\tau}''_{m+1}-\text{ quantile of } F^0_{W|YZ}} |y^{m+1}, z''_m)\end{aligned} \tag{3-4}$$

Also, for $P^m(z'_m) < \overline{v} < P^m(z''_m)$, we have[45]

$$\begin{aligned}\overline{\tau} &\equiv F^a_{U|VZ}(u^*|v, z''_m) \\ &= F^a_{U|VZ}(\underbrace{h_a^{-1}(y^{m+1}, w^{m+1})}_{\overline{\tau}-\text{ quantile of } F^a_{U|VZ}}|\overline{v}, z''_m) \\ &= F^0_{W|YZ}(\underbrace{w^{m+1}}_{\overline{\tau}-\text{ quantile of } F^0_{W|YZ}} |y^m, z''_m)\end{aligned} \tag{3-5}$$

Step 2 : Order of (3-3)-(3-5) :
Note $P^m(z'_m) \leq P^m(z''_m) \leq P^{m+1}(z''_m)$. Then PD implies that

$$\tau''_{m+1} \leq \tau''_m \leq \overline{\tau} \tag{*PD}$$

since we are comparing the values of the three conditional distributions evaluated at the same value u^*. And ND implies that

$$\tau''_{m+1} \geq \tau''_m \geq \overline{\tau} \tag{*ND}$$

[45] This is for $P^{m-1}(z'') \leq P^m(z')$. Other cases can be shown similarly.

$$\begin{aligned}\overline{\tau} &\equiv F^a_{U|VZ}(r|v, z'') \\ &= F^a_{U|VZ}(h_a^{-1}(y^{m+1}, w^{m+1})|v, z'') \\ &= \begin{pmatrix} F^0_{W|YZ}(w^{m+1}|y^m, z'') & \text{if} \quad P^{m-1}(z'') \leq P^m(z') \\ F^0_{W|YZ}(w^{m+1}|y^{m+1}, z') & \text{if} \quad P^m(z'') \leq P^{m+1}(z') \end{pmatrix}\end{aligned} \tag{3-5$'$}$$

Step 3 : Quantile expressions for w and u^*

Now we express u^* and w^* and w^{m+1} as quantiles of the distributions so that we can find the order of the two, $h_a(y^m, u^*)$ and $h_a(y^{m+1}, u^*)$ using (*PD) and (*ND). (4-2)-(4-5) imply (4-6) and (4-7) under continuity of W and U :

$$\begin{aligned} u^* &= Q^a_{U|VZ}(\tau''_m|P^m(z''_m), z''_m) \\ &= Q^a_{U|VZ}(\tau''_{m+1}|P^{m+1}(z''_m), z''_m) \\ &= Q^a_{U|VZ}(\tau'_{m+1}|P^{m+1}(z'_m), z'_m) \\ &= Q^a_{U|VZ}(\overline{\tau}|\overline{v}, z''_m), \quad \text{for } P^m(z'_m) < \overline{v} < P^m(z''_m) \end{aligned} \tag{3-6}$$

$$\begin{aligned} w^* &\overset{(a)}{=} Q^0_{W|YZ}(\tau''_m|y^m, z''_m) = Q^0_{W|YZ}(\tau''_m|y^m, z''_m) \\ w^{m+1} &\overset{(c)}{=} Q^0_{W|YZ}(\tau''_{m+1}|y^{m+1}, z''_m) \end{aligned} \tag{3-7}$$

(a) follows from (3-3), (b) from (3-5) and (c) is by (3-4).

Step 4 : Match?

Finally we use the construction of the structural function using (3-6). Then we can determine the direction of the response : we have from (3-2)[46]

$$\begin{aligned} & h_a(y^m, u^*) - h_a(y^{m+1}, u^*) \\ = \; & h_a(y^m, Q^a_{U|VZ}(\tau''_m|P^m(z''_m), z''_m)) - h_a(y^{m+1}, Q^a_{U|VZ}(\tau''_{m+1}|P^{m+1}(z''_m), z''_m)) \\ = \; & Q^0_{W|YZ}(\tau''_m|y^m, z''_m) - Q^0_{W|YZ}(\tau''_{m+1}|y^{m+1}, z''_m) \\ = \; & Q^0_{W|YZ}(\tau''_m|y^m, z''_m) - Q^0_{W|YZ}(\overline{\tau}|y^m, z''_m) \\ & \begin{pmatrix} \leq 0 \text{ if PD} \\ \geq 0 \text{ if ND} \end{pmatrix} \end{aligned}$$

the third equality is by (c) in (3-7). Then the inequality follows because $\tau''_m \leq \overline{\tau}$ (*PD) and $\tau''_m \geq \overline{\tau}$ (*ND), and the property of quantiles.

A.3 Proof of Corollary 2

Proof. We adopt Lemma 2 in Chesher (2005) when $m = 1$ with $P^0(z) = 0$ and $P^1(z) = P(z)$, where $P(z) = \Pr(Y = 1|Z = z)$ and when $m = 2$ with $P^2(z) = 1$ and $P^1(z) = P(z)$.

[46]Recall that this is the case for $P^{m-1}(z'') \leq P^m(z')$. The other case can be shown similarly.

Suppose that $Q_{U|VZ}(\tau_U|\tau_V, z)$ is weakly increasing in v. Then we have

$$\begin{aligned}
h(0, Q_{U|VZ}(\tau_U|0, z')) &\leq Q_{W|YZ}(\tau_U|0, z') && \text{(A-1)}\\
&\leq h(0, Q_{U|VZ}(\tau_U|P(z'), z'))\\
h(0, Q_{U|VZ}(\tau_U|0, z'')) &\leq Q_{W|YZ}(\tau_U|0, z'') && \text{(A-2)}\\
&\leq h(0, Q_{U|VZ}(\tau_U|P(z''), z''))\\
h(1, Q_{U|VZ}(\tau_U|P(z'), z')) &\leq Q_{W|YZ}(\tau_U|1, z') && \text{(A-3)}\\
&\leq h(1, Q_{U|VZ}(\tau_U|1, z'))\\
h(1, Q_{U|VZ}(\tau_U|P(z''), z'')) &\leq Q_{W|YZ}(\tau_U|1, z'') && \text{(A-4)}\\
&\leq h(1, Q_{U|VZ}(\tau_U|1, z''))
\end{aligned}$$

We use (A-1) and (A-4).

$$\begin{aligned}
Q_{W|YZ}(\tau_U|0, z') &\leq h(0, Q_{U|VZ}(\tau_U|P(z'), z')) && \text{(A-1)}\\
h(1, Q_{U|VZ}(\tau_U|P(z''), z'')) &\leq Q_{W|YZ}(\tau_U|1, z'') && \text{(A-4)}
\end{aligned}$$

Under Restriction RC, $P(z') \leq \tau_V \leq P(z'')$, when $Q_{U|VZ}(\tau_U|\tau_V, z)$ is weakly increasing in v, then :

$$\begin{aligned}
Q_{U|VZ}(\tau_U|\tau_V, z'') &\leq Q_{U|VZ}(\tau_U|P(z''), z'')\\
Q_{U|VZ}(\tau_U|P(z'), z') &\leq Q_{U|VZ}(\tau_U|\tau_V, z')
\end{aligned}$$

and because h is monotonic in u and weakly increasing,

$$\begin{aligned}
h(1, Q_{U|VZ}(\tau_U|\tau_V, z'')) &\leq h(1, Q_{U|VZ}(\tau_U|P(z''), z'')) && \text{(B-1)}\\
h(1, Q_{U|VZ}(\tau_U|P(z'), z')) &\leq h(1, Q_{U|VZ}(\tau_U|\tau_V, z')). && \text{(B-2)}
\end{aligned}$$

Combining (A-4) and (B-1) we can find the upper bound for $h(1, Q_{U|VZ}(\tau_U|\tau_V, z''))$

$$h(1, Q_{U|VZ}(\tau_U|\tau_V, z'')) \leq h(1, Q_{U|VZ}(\tau_U|P(z''), z'')) \leq Q_{W|YZ}(\tau_U|1, z'')$$

Use the Restriction LDRM : $h(1, u) \geq h(0, u)$, for all values of z and u in the support of Z and U. Applying Restriction LDRM to (B-2)

$$h(0, Q_{U|VZ}(\tau_U|P(z'), z')) \leq h(1, Q_{U|VZ}(\tau_U|P(z'), z')) \leq h(1, Q_{U|VZ}(\tau_U|\tau_V, z')). \quad \text{(C)}$$

Applying (A-1) to (C), we have the lower bound for $h(1, Q_{U|VZ}(\tau_U|\tau_V, z'))$

$$Q_{W|YZ}(u|0, z') \leq h(1, Q_{U|VZ}(\tau_U|\tau_V, z')).$$

Finally, under the conditional quantile independence restriction and exclusion Restriction C-QI and QCFA, there is for $z \in \{z', z''\}$ for $u^* = Q_{U|VZ}(\tau_U|\tau_V, z)$

$$Q_{W|YZ}(\tau_U|0, z') \leq h(1, u^*) \leq Q_{W|YZ}(\tau_U|1, z'') \qquad \text{(D-1)}$$

Consider next the identification of $h(0, u^*)$.

Under Restriction RC, $P(z') \leq \tau_V \leq P(z'')$, when $Q_{U|VZ}(\tau_U|\tau_V, z)$ is weakly increasing in v, then :

$$\begin{aligned} Q_{U|VZ}(\tau_U|\tau_V, z'') &\leq Q_{U|VZ}(\tau_U|P(z''), z'') \\ Q_{U|VZ}(\tau_U|P(z'), z') &\leq Q_{U|VZ}(\tau_U|\tau_V, z') \end{aligned}$$

and because h is monotonic in U and weakly increasing,

$$h(0, Q_{U|VZ}(\tau_U|\tau_V, z'')) \leq h(0, Q_{U|VZ}(\tau_U|P(z''), z'')) \quad \text{(B-3)}$$

$$h(0, Q_{U|VZ}(\tau_U|P(z'), z')) \leq h(0, Q_{U|VZ}(\tau_U|\tau_V, z')). \quad \text{(B-4)}$$

using (A-4) and (B-3), and Restriction LDRM we can find the upper bound for $h(0, Q_{U|VZ}(\tau_U|\tau_V, z''))$

$$\begin{aligned} & h(0, Q_{U|VZ}(\tau_U|\tau_V, z'')) \overset{(a)}{\leq} h(0, Q_{U|VZ}(\tau_U|P(z''), z'')) \\ & \overset{(b)}{\leq} h(1, Q_{U|VZ}(\tau_U|P(z''), z'')) \\ & \overset{(c)}{\leq} Q_{W|YZ}(\tau_U|1, z'') \end{aligned}$$

(a) is due to (B-3), (b) follows from Restriction LDRM, and (c) is from (A-4).

Applying (A-1) to (B-4) we have

$$Q_{W|YZ}(\tau_U|0, z') \overset{(a)}{\leq} h(0, Q_{U|VZ}(\tau_U|P(z'), z')) \overset{(b)}{\leq} h(0, Q_{U|VZ}(\tau_U|\tau_V, z')).$$

(a) follows from (A-4) and (b) is from (B-4). Thus, the lower bound for $h(0, Q_{U|VZ}(\tau_U|\tau_V, z'))$

$$Q_{W|YZ}(\tau_U|0, z') \leq h(0, Q_{U|VZ}(\tau_U|\tau_V, z')).$$

Finally, by Restriction C-QI and QCFA, there is for $z \in \{z', z''\}$

$$Q_{W|YZ}(\tau_U|0, z') \leq h(0, u^*) \leq Q_{W|YZ}(\tau_U|1, z'')$$

Note that the identified intervals for $h(0, u^*)$ and $h(1, u^*)$ are the same as we see in (D-1) and (D-2). ■

A.4 Proof of Lemma 3

Proof. We show the case in which $Q_{W|YZ}(\tau_U|y^m, z'_m) \leq Q_{W|YZ}(\tau_U|y^{m+1}, z''_m)$. The other case can be shown similarly. We need to show that PDPR implies $Q_{W|YZ}(\tau_U|y^m, z'_m) \leq Q_{W|YZ}(\tau_U|y^{m+1}, z''_m)$.

Let Q''_{m+1} and Q'_m indicate the values of τ_U- quantiles, $Q'_m \equiv Q_{W|YZ}(\tau_U|y^m, z'_m)$ and $Q''_{m+1} \equiv Q_{W|YZ}(\tau_U|y^{m+1}, z''_m)$. Then by definition of quantiles we have

$$\begin{aligned} \tau_U &= F_{W|YZ}(Q'_m|Y=y^m, Z=z'_m) \\ &= \Pr(W \le Q'_m|Y=y^m, Z=z'_m) \\ &= \Pr(h(y^m,U) \le Q'_m|Y=y^m, Z=z'_m) \\ &= \Pr(U \le h^{-1}(y^m, Q'_m)|Y=y^m, Z=z'_m) \end{aligned} \tag{A}$$

similarly for Q''_{m+1}, we have

$$\tau_U = \Pr(U \le h^{-1}(y^{m+1}, Q''_{m+1})|Y=y^{m+1}, Z=z''_m) \tag{B}$$

where h^{-1} is defined as (C*) in Appendix C. Suppose PDPR. Then we have

$$\begin{aligned} \tau_U &= \Pr(U \le h^{-1}(y^{m+1}, Q''_{m+1})|Y=y^{m+1}, Z=z''_m) && \text{(C-1)} \\ &= \Pr(U \le h^{-1}(y^{m+1}, Q''_{m+1})|V \in (P^m(z''_m), P^{m+1}(z''_m)] \\ &\le \Pr(U \le h^{-1}(y^{m+1}, Q''_{m+1})|V \in (P^{m-1}(z''_m), P^m(z''_m)] \\ &= \Pr(U \le h^{-1}(y^{m+1}, Q''_{m+1})|Y=y^m, Z=z''_m) \\ &= \Pr(U \le h^{-1}(y^{m+1}, Q''_{m+1})|Y=y^m, Z=z'_m) \equiv \widetilde{u} && \text{(C-2)} \end{aligned}$$

where the first equality is by (B), the second equality follows from that the event $\{V \in (P^m(z''_m), P^{m+1}(z''_m)]\}$ is equivalent to the event $\{Y=y^{m+1}, Z=z''_m\}$. The first inequality is due to PD ($F_{U|VZ}(u|v,z)$ is non-increasing in $v \in V$), and the third equality results from the same logic as in the second equality. The last equality is due to Restriction C-QI. Then $\tau_U \le \widetilde{u}$.

From (A) and (C-2), we have

$$\tau_U = \Pr(U \le \underbrace{h^{-1}(y^m, Q'_m)}_{u^*}|Y=y^m, Z=z'_m) \le \Pr(U \le \underbrace{h^{-1}(y^{m+1}, Q''_{m+1})}_{u^{**}}|Y=y^m, Z=z'_m) \equiv \widetilde{u}$$

since $\tau_U \le \widetilde{u}$, which implies that

$$u^* \equiv h^{-1}(y^m, Q'_m) \le h^{-1}(y^{m+1}, Q''_{m+1}) \equiv u^{**}$$

by the nondecreasing property of distribution function, i.e., if $a \le a'$, $F_{A|B}(a|b) \le F_{A|B}(a'|b)$. Then we have

$$\begin{aligned} Q'_m &= h(y^m, u^*) \\ Q''_{m+1} &= h(y^{m+1}, u^{**}) \end{aligned}$$

By PDPR and monotonicity of h in u, we have

$$\begin{aligned} Q'_m &= h(y^m, u^*) \le h(y^{m+1}, u^*) \\ &\le h(y^{m+1}, u^{**}) = Q''_{m+1} \end{aligned}$$

where the first inequality is due to PDPR and the second inequality is due to monotonicity of h in u. Thus, we have shown that $Q'_m \equiv Q_{W|YZ}(\tau_U|y^m, z'_m) \le Q''_{m+1} \equiv Q_{W|YZ}(\tau_U|y^{m+1}, z''_m)$. The other case can be shown similarly. ■