

# The Information Basis of Matching in Treatment Effect: Understanding Propensity Scores

Esfandiar Maasoumi  
SMU, Dallas, TX

## Introduction

The literature on "treatment" effects and program/policy evaluation deals with a classic problem of unobserved outcomes by "matching" through observed covariates. Two groups whose relevant characteristics are observed are, respectively, treated and untreated/controls. One cannot observe both the treatment and no-treatment outcomes for the same subject; subjects either receive or do not receive a "treatment". In order to isolate/identify the effect of treatment alone, there is a need to control for all other factors that may significantly contribute to the *apparent* differences in outcomes.

Broadly speaking, three stages or steps may be identified with the "treatment" literature. First, relevant covariates and characteristics are chosen; second, a mechanism is devised for the way these covariates influence outcomes or choices, such as the decision to participate or probability of treatment. Lastly, "treatment effect" is *defined* and identified. In this paper we do not directly deal with the first step, the division of this covariate set into observed and unobservables, or exogenous and endogenous variables. A recent insightful survey on this is Heckman and Navaro-Lozano (2004). But we deal with how these covariates are employed in order to "characterize" subjects, through estimation and/or matching, and any transformations thereof, to determine "similarity" between members of samples or population groups. In particular, we offer representation perspectives and techniques which further expose the meaning and limitations of propensity scores (PS).

We deal only partially with the third of the stages mentioned above and consider aspects of the outcome/treatment *distributions* to be examined, the means, quantiles, or entire distributions. The existing literature is predominantly focused on "averages" or mean treatment effects, conditional on various information sets. Notable exceptions attempt to model for conditional quantiles in an attempt to avoid the "vail of ignorance" and reveal more of the reality that *outcomes are distributed* and "averages" can often mislead. See Imbens (2004) for a survey. Our alternative approach suggests natural ways of dealing with the whole distribution of covariates as well as of outcomes. We only sketch and exemplify full distribution methods such as entropy divergences, equality and/or dominance between whole distributions of *attributes and outcome*, respectively.

The intuition behind our main proposals in this paper is as follows:

Let  $x_i$  be a set of  $m$  attributes or characteristics for "individual"  $i = 1, 2, \dots, n$ . Subsets of  $x_i$  determine choices, selection, and/or outcomes of policy or treatment programs. We propose to choose a function  $S(x_i)$  to represent or summarize each individual. We believe the choice of functionals  $S(x_i)$ , perhaps as utility or welfare evaluation functions, must be dealt with explicitly. Powerful implications of welfare

axioms of “fundamentality”, “impartiality”, and “anonymity” may be invoked which partially justify the same “common representations” for everyone’s preferences, in functional forms and for “representative agent” formalisms; see Kolm (1977) and Maasoumi (1986). Simply put, if we enter into a preference function all the attributes that we think would matter (i.e., as  $m$  increases indefinitely), the need for ex post and often arbitrary heterogeneity in functional representations and other approximations would be diminished. There are philosophical and empirical/data availability problems which impinge upon this “ideal”, however, leading to practical admission of heterogeneity as well as functional approximations.

The reason we should explicitly deal with this problem of *representation* is to avoid *implicit* aggregation rules and representations which are not tenable when exposed, sometimes patently so. This happens with propensity scores. For instance, a linear function of  $x$  would imply infinite substitutability between individual characteristics and other covariates! Quadratic utility functions leading to linear decisions rules are not well supported empirically. Our proposed method follows the line of reasoning in Maasoumi (1986a) which is currently a central ingredient of practical multidimensioned approaches to welfare and poverty analysis. Any scalar functional of  $x$  is a summary measure of it, and possesses an induced “distribution”. In empirical sciences, the distribution is the only full “truth” that is open to inference with data. This summary distribution must not be incompatible with the information available on its constituent variables. Again, the full information about the latter is their “distributions”. Fortunately, it is possible to find summary/representation/utility functions of  $x$  whose distribution is “closest” to that of all its constituent attributes. Any other functionals are, thereby suboptimal and distortive. While there are no consensus criteria for “closeness” or similarity in sciences, there are some very good ones that are supported by well thought out axioms and properties. Thus, our approach will not “solve” the problem, but rather provides palusible and well founded bounds to inferences, and illuminates the meaning of other “solutions”.

One such solution to “matching” is by propensity scores. To put it in the context of our discussion above, PS first takes a convenient function (typically linear, but often augmented with polynomial terms),  $g(x)$  say, and then transforms this into the unit interval by inversion through some cumulative distribution function (CDF),  $F$  say; Normal or logistic cdf predominate in determining “treatment probability”. While we may view this last transformation as a convenient standardization that helps simplify matching, we are less comfortable with the justifications of the various elements of the PS technique. Firstly, neither  $g(x)$  nor  $F(g(x))$  has any clear justification as a representation of individuals/households/policy makers in the sense made clear earlier. Second, the parameter values, or indeed even semi-parametric estimates of these functionals, are determined with reference to optimal estimation rules which, again, have no clear connection to optimal representation or aggregation. In addition, the criticism in Heckman and Navaro-Lozano (2004), about the choice of covariates in  $x$  being based on estimation and model fit measures, applies. To the latter argument we add that, estimating conditional means or quantiles, further impinges on the plausibility of variable choices and coefficient values/factor loadings. Another discomfoting aspect of the PS type solution is that the conditioning set is almost always made of the same fundamental and frequently observed/cited  $x$  variables, whatever the experiment or treatment. In the extreme, it is therefore possible that  $g(x)$  will be called upon to

determine selection into a drug treatment program, as well as a welfare program! Given the oft made expedient assumption of "independence of  $x$  and the choice variable" it is not hard to imagine the leap of faith in some applications. Finally, matching based on propensity scores is a decision on similarity of individuals based on *estimated means* of whole distributions, not the true means. Very many different distributions of individuals, and of heterogeneous groups, may have similar or even the same means. The "control function" method favored by Heckman and Navaro-Lozano (2004) is a second moment-based approach to partially deal with this shortcoming of the PS method.

Given the welfare theoretic context here, matching by PS is seen to impose certain implicit assumptions on utility functions and the meaning of the CDF transformation of the "utility functions". Individuals with similar  $p$  scores are deemed "similar" enough to solve the counterfactual problem. Thus the  $p$  scores are deemed as suitable characterization functions for the individuals with covariates  $x$ . We show below that this is not so, or likely, and find inferences are very sensitive to both this assumption, and the estimated coefficient values, as has been pointed out elsewhere; e.g., see Heckman and Navaro-Lozano (2004). We then offer information criteria to assess divergences between two sets of distributions: The first set contains the distributions of characteristics for the treated and the counterfactuals (or other groups), and is analysed for optimal matching decisions; the second set contains the distributions of their respective outcome variables, analysed to determine treatment effects. Thus entropy measures are seen to provide guidance in both optimal matching decisions, and in evaluating outcome differences for all, not just the averages, or at certain quantiles.

In what follows, we focus on the cases in which the assumption of "selection on observables" is plausible. This partially finesses a serious problem of intera-personal comparison of welfare and utility and its attendant empirical limitations; see Blundell and Lewbel (1991), and Pollak and Wales (1979).

## Optimal Multiattribute Functions

Let  $X_{ij}$  denote a measure of attribute  $j = 1, 2, \dots, m$ , associated with individual (unit, household, country)  $i = 1, 2, \dots, n$ . Define the covariate matrix  $X = (X_{ij})$ ,  $X_i$  its  $i$ -th row,  $X^j$  its  $j$ -th column, and consider any scalar function of the matrix  $X$ . Examples of such scalar functions are inequality measures or Social Welfare Functions (SWFs), or propensity scores. It has proven difficult to develop "consensus" axioms which may characterize an ideal scalar measure of  $X$ , such as aggregators or inequality measures

### Aggregate Functions

Two approaches are relevant here. The first is based on measures of closeness and affinity which may identify either attributes that are similar in some sense, *and/or* determine a "mean-value", or aggregate, *which most closely represents the constituent attributes*. The second approach which is axiomatic lays down properties that we may agree an aggregate function should possess. This second approach was examined by Tsui (1992b) and faces difficulties in determining "consensus" properties. This is not unrelated to the difficulty of adopting a criterion of "closeness" in the first approach. But the latter difficulty has had some resolution in "information theory" which seems to suggest members of the Generalized Entropy (GE) family as ideal criteria of

“closeness” or “divergence”. Axiomatic characterization of GE is, however, beyond the scope of the present paper. The interested reader may refer to Maasoumi (1993).

Let  $S_i$  denote the aggregate or mean function for the  $i$  –  $th$  unit. It makes little difference to our approach whether  $S_i$  is interpreted as an individual’s utility evaluations or the “observer’s” or policy maker’s assessments for individual  $i$ . Let us define the following Generalized Multivariate GE measure of closeness or diversity between the  $m$  densities of the chosen  $m$  attributes:

$$D_{\beta}(S, X; \alpha) = \sum_{j=1}^m \alpha_j \left\{ \sum_{i=1}^n S_i [(S_i/X_{ij})^{\beta} - 1] / \beta(\beta + 1) \right\} \quad \#$$

where  $\alpha_j$ s are the weights attached to each attribute. Minimizing  $D_{\beta}$  with respect to  $S_i$  such that  $\sum S_i = 1$ , produces the following “optimal” aggregation functions:

$$S_i \propto \left( \sum_j^m \alpha_j X_{ij}^{-\beta} \right)^{-1/\beta}, \beta \neq 0, -1 \quad \#$$

$$S_i \propto \prod_j X_{ij}^{\alpha_j}, \beta = 0 \quad \#$$

$$S_i \propto \sum_j \alpha_j X_{ij}, \beta = -1 \quad \#$$

These are, respectively, the hyperbolic, the generalized geometric, and the weighted means of the attributes, see Maasoumi (1986a). Noting the “constant elasticity of substitution”- $\sigma = 1/(1+\beta)$ , these functional solutions include many of the well known utility functions in economics, as well as some arbitrarily proposed aggregates in empirical applications. For instance, the weighted arithmetic mean subsumes a popular “composite welfare indicator” based on the principal components of  $X$ , when  $\alpha_j$ s are the elements of the first eigen vector of the  $X'X$  matrix; see Ram (1982) and Maasoumi (1989a).

The “divergence measure”  $D_{\gamma}(\cdot)$  forces a choice of an aggregate vector  $S = (S_1, S_2, \dots, S_n)$  with a distribution that is closest to the distributions of its constituent variables. This is especially desirable when the goal of our analysis is the assessment of distributed outcomes, such as treatment effects. But it is desirable generally since we have no other “information” than the distribution of variables. Information theory establishes that any other  $S$  would be extra distortive of the objective information in the data matrix  $X$ . Elsewhere it has been argued that such a distributional criterion is desirable for justifying choices of utility, production, and cost functionals since such choices should not distort the actual market allocation signals that are in the observed data. The distribution of the data reflects the outcome of all decisions of all agents in the economy; see Maasoumi (1986b).

The divergence criterion here is  $\alpha_j$ -weighted sum/average of pairwise GE divergences between the “distributions”  $S$  and  $X^j$ , the  $j$ -th attribute/column in  $X$ .

## A Closer Look at measures of closeness

Consider the uniform distribution over the support of any variable  $Z$ . It is the "maximum entropy", or uncertainty, distribution since all outcomes are equally likely. The Generalized Entropy (GE) divergence between the actual distribution of  $Z$  and the maximum entropy is thus a measure of similarity, concentration, equality, and certainty in  $Z$ . This versatile concept is represented by the following form:

$$I_{\beta}(Z) = \frac{1}{\beta(1 + \beta)} \int_0^{\infty} (z/\mu_z)[(z/\mu_z)^{\beta} - 1]dF, \beta \text{ real} \quad \#$$

$I_1$  is ordinally equivalent to the coefficient of variation and the Herfindahl index. The family includes the variance of logarithms and Theil's first and second inequality measures,  $I_0$  and  $I_{-1}$ , respectively. Also, up to a monotonic transformation, there is a unique member of GE corresponding to each member of the Atkinson family inequality indices.  $\nu = -\beta$  is the degree of aversion to relative inequality, or risk; the higher its absolute value the greater is the sensitivity of the measure to transfers in the tail areas of the distribution which may be the target of policy makers (e.g., chronically unemployed, chronically ill, poor performing students, etc).

The axiomatic derivation technique that identifies GE is constructive and is to be appreciated as an important breakthrough in organizing learning and knowledge in this area. It will help to set the multivariate issues in context. This axiomatic approach owes much to functional analysis first developed in "information theory", by Aczel and others; see Maasoumi (1993). In this approach one puts down an **explicit** set of requirements (axioms) which the ideal criteria must satisfy, and which may or may not be universally acceptable. Imposing these axioms as *explicit* constraints on the function space, one obtains the appropriate criterion. Shorrocks (1980) discuss the "fundamental welfare axioms" of symmetry, continuity, Principle of Transfers, and additive decomposability which identify GE as the desirable scale invariant family of **relative** "inequality" measures.

## An Application

### Data

We use the data from Dehejia and Wahba (1999) which is based on Lalonde's (1986) seminal study on the comparison between experimental and non-experimental methods for the evaluation of causal effects. The data combine the treated units from a randomized evaluation of the National Supported Work (NSW) demonstration with non-experimental comparison units drawn from PSID data. We restrict our analysis to the so-called NSW-PSID-1 subsample consisting of the male NSW treatment units, and the largest of the three PSID subsample.

The outcome of interest is real earnings in 1978; the treatment is participation in the NSW treatment group. Control covariates are age, education, real earnings in 1974 and 1975, and binary variables for black, hispanic, marital status and a nodegree. The treatment group contains 185 observations, the control group 2490 observations, for a total of 2675.

In Table 1 we report average treatment effects with various matching techniques. The first method (second column) is the standard Average Treatment on the Treated

(ATT) using propensity scores. These are replicated numbers from Daheija and Wahba (1999) for all the covariates age, age squared, education, education squared, real earnings in 1974 and its square, real earnings in 1975 and its square, dummies for black, hispanic, marital status, nodegree, and an interaction term between black and nonemployment in 1974. The additional nonlinear terms make a small difference for the DW sample and the PS method thereof.

S(i) index includes age, education, real earnings in 1974, real earnings in 1975 and dummies for black, hispanic, marital status and nodegree. In columns 3-5 we report three different aggregator functions,  $S(\cdot)$  obtained at different values of the elasticity of substitution parameter,  $\beta$ . Note that matching is done by first carrying out the same transformation on the  $S(\cdot)$  functions, namely with the Gaussian CDF,  $\Phi(\cdot)$ , as is used in the PS method (Probit). Thus, in this table, we offer a traditional use of our aggregators, and otherwise, matching by propensity scores is done in the traditional manner. For this reason, the results in Table 1 reveal the sensitivity of the PS method to the functional form of the model used, and the parameter values in these functional forms. In Columns 3-5, the " $\alpha$ " coefficients were given the values they take from the estimation of the traditional PS method. Later on we experiment further with other values, especially equal  $\alpha, s$ .

Table 1: Average Treatment Parameter Estimations with Different Indices

Specifications	Matching with Propensity Score		Matching with $\Phi(S(i))$	
	ATT (Standard Error)	$\beta=-2$	$\beta=-1$	$\beta=-1/2$
		ATT (Standard Error)		
Specification A <sup>2</sup>	1545.518 (930.926)	-8112.840 (975.704)	1545.518 (928.496)	-6929.172 (1709.479)
Specification B <sup>3</sup>	1654.566 (1042.161)	-2360.403 (982.410)	1976.884 (961.238)	-7899.357 (1096.775)

NOTES:

- (1) The coefficient estimates from the probit model specifications of A and B are used as weights for  $S(i)$
- (2) Specification A includes age, education, Black, Hispanic, married, nodegree, earnings in 1974 and earnings in 1975.
- (3) Specification B includes age, age squared, education, education squared, Black, Hispanic, married, nodegree, earnings in 1974, square of earnings in 1974, earnings in 1975, square of earnings in 1975 and an interaction term between black and zero earnings in 1974.
- (4) Standard Errors are obtained via 500 bootstrap repetitions.
- (5) Sample Size=2675, Treated Units=185 and Control Units=2490.

It is clear that inferences vary dramatically depending on which functional forms or coefficient values are used. The standard errors were obtained by simple bootstrap and are given in parentheses. Note that the case with  $\beta = -1$  is, other than for sampling variation and the intercept, identical with the PS. This is the case of "infinite substitutability" between covariates in the linear in parameters models. When less substitution is allowed, we have dramatically different results, including the reversal of the sign for the ATT! This further reinforces the criticisms offered recently in Heckman and Navaro-Lozano (2004), and Smith and Todd (2005).

In Table 2 we provide an idea of proximity of the distributions of the propensity scores between the experimental and the matched non-experimental groups. This can be used to evaluate the efficacy of the matching technique employed (here, the nearest neighbor), not as in the traditional way of assessing how close two or more matched pairs are, but how well "similar" the two samples are. This is of some importance given the recent debate regarding the sensitivity and lack of robustness to the choice of subsamples being used by various investigators; see Smith and Todd (2005) and response by Dahejia (2005). KL and other entropic measures give an idea of how far apart the two distributions are. The KL measure is defined as follows:

$$I(f, g) = \int f \log \frac{f}{g} d_x$$

$f$  and  $g$  are the distributions of the treated and control groups for a variable  $x$ , respectively. A symmetrized version of this is the well-known KL measure, but we report both of the asymmetric measures which need to be averaged to get KL. This is useful here given the common focus on the "treated". It is well known that  $I(f, g) = 0$  if and only if  $f = g$ . This measure provides for consistent tests of null hypothesis of equality of the two distributions, See Hong and White (2004). But this measure is not "metric" as it does not satisfy the triangularity rule. Thus, while it is tempting to compare the values of KL across the various cases, one must be careful in drawing strong conclusions about "distances". For proper entropic distance measures, we need

to employ other measures, such as the one recently advocated by Granger, Maasoumi and Racine (2004). This subject is beyond the scope of the present study.

**Table 2**

ATT Estimates with Different Matching Techniques				
Matching		ATT (Standard Error)	Kullback-Leibler Divergence	
			Treated/Nontreated	Nontreated/Treated
PS		1654.566 (1042.161)	0.0015	0.0014
S(i),	$\beta=-1$	2263.860 (1785.033)	3.5109e-008	3.5106e-008
S(i),	$\beta=-1/2$	606.618 (1082.455)	2.2390e-005	2.2456e-005
S(i),	$\beta=-2/3$	-557.414 (1663.907)	2.3058e-007	2.3079e-007
$\Phi(S(i))$ ,	$\beta=-1$	1545.518 (928.496)	0.0054	0.0051
$\Phi(S(i))$ ,	$\beta=-1/2$	-6929.172 (1709.479)	1.9903e-005	1.9836e-005
$\Phi(S(i))$ ,	$\beta=-2/3$	-1398.497	4.2654e-004	4.2509e-004

---

NOTES: (i)  $\Phi(S(i))$  includes the same specification as in  $S(i)$ . The  $\alpha$ -weights for  $\Phi(S(i))$  are the corresponding probability weights from the probit equation, whereas  $\alpha_j = 1/8$  was used in matching by  $S(i)$ . (ii) Standard errors are obtained via 500 bootstrap replications. (iii) Kullback-Leibler divergence measure is obtained via a Gaussian kernel with fixed bandwidth.

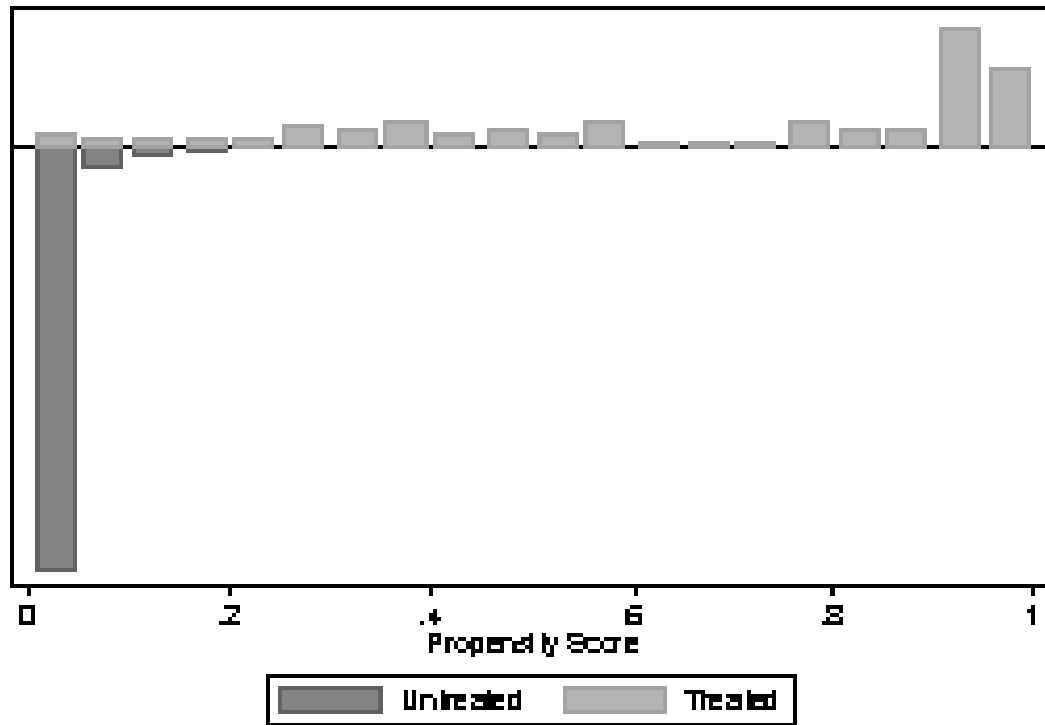
---

In Table 2 we also provide matching based on the aggregator functions,  $S(i)$ , as well as matching based on the (Normal) CDF transforms of the  $S(i)$ . The latter is done for direct comparability with the PS results, and to draw out the impact of different substitution and other parametric values. The KL measures are smaller for matching by the  $S(i)$  functions than by PS. Matching by the comparable  $\Phi(S(i))$  produces generally smaller KL values than the traditional PS. The only difference between the traditional PS (first row) and the case  $\Phi(S(i))$ ,  $\beta = -1$ , is an intercept in the former method. This accounts for the apparent difference between the two sets of KL values. The figures below shed further light on the better "distributions matches" between the two groups when  $S(i)$  is used with finite substitutability between characteristics. Figures 1-4

demonstrate the distribution of the matched pairs by each technique. Specifically, these are frequency distributions of PS for each individual, and the comparable transformation  $\Phi(S(i))$ .

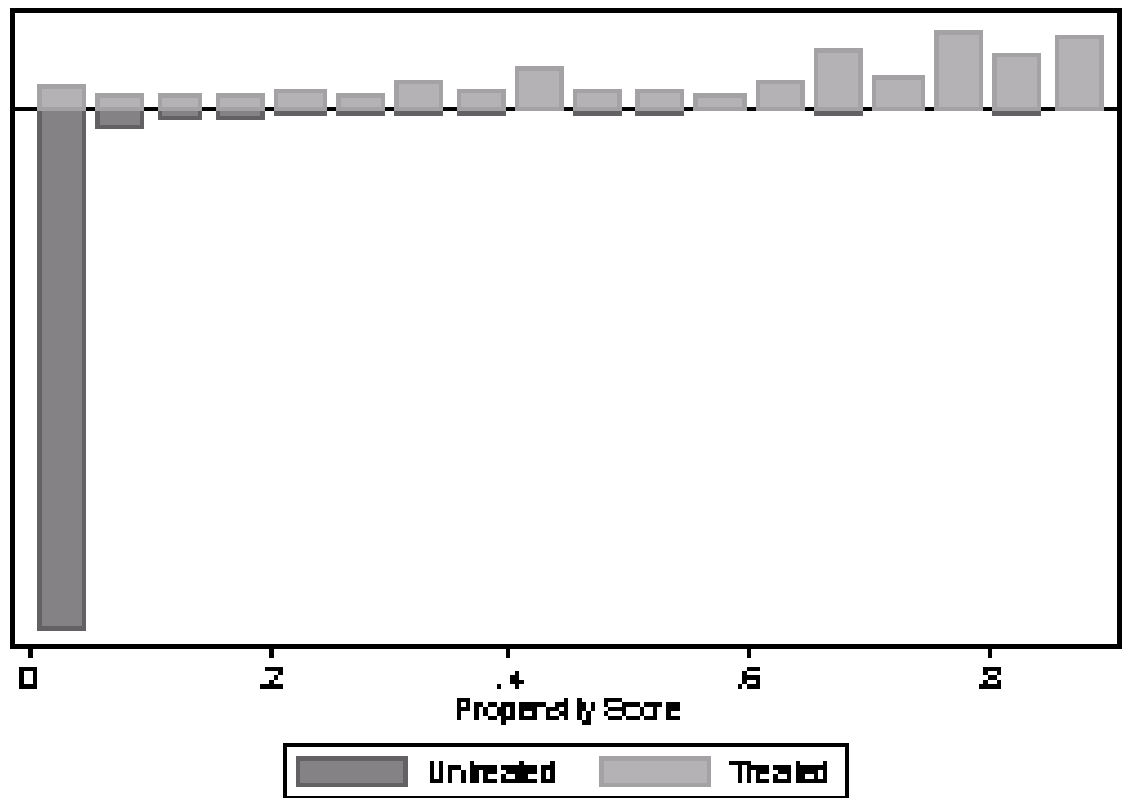
Propensity Score Distribution

Figure 1:



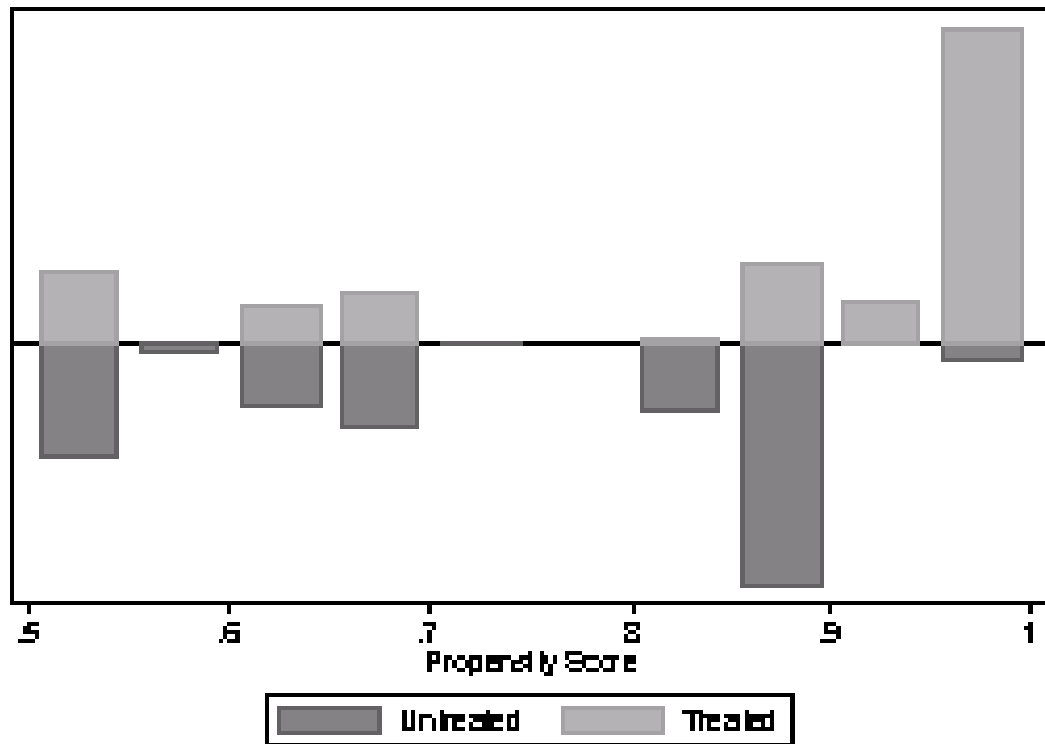
$\Phi(S(i))$  for  $\beta=-1$

Figure 2:



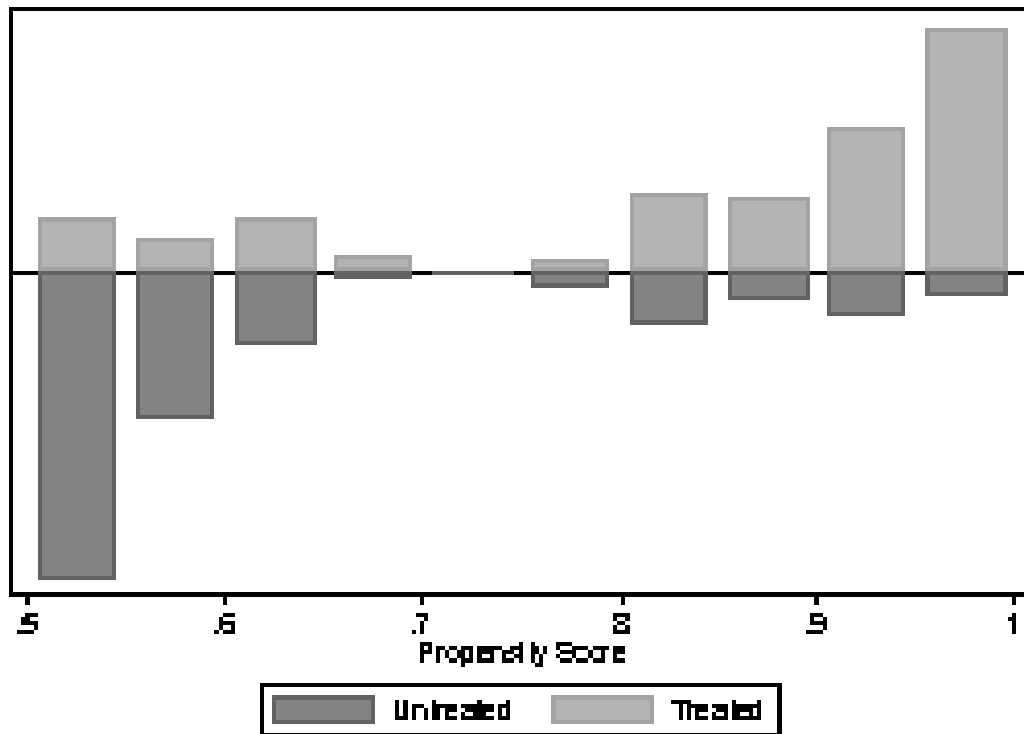
$\Phi(S(i))$  for  $\beta=-1/2$

Figure 3:



$\Phi(S(i))$  for  $\beta=-2/3$

Figure 4:



Traditional PS matching seems to correspond to very different distributions of scores for the treated and untreated. This may appear reasonable at first blush, since the untreated were not selected into the program. But these non-experimental subjects in the PSID were not considered for treatment, and should have similar scores, or be similar populations, if the PS matching theory is to work at all. It is evident that at lower levels of substitution between characteristics, treatment probability is increased for the untreated sample. Again these figures are computed using PS regression estimates for the  $\alpha$  weights in the  $S(i)$  functions. Naturally, there would be some variation in these results as we change the  $\alpha$  values. Basing these values, and the  $\beta$ , on estimation techniques is a decision to vary these graphs (PS values) to obtain statistically optimal values for treatment/selection probability. It is difficult a priori to see very many contexts in which this would be reflective of policy optimality, or in keeping with random assignment.

These results make clear the difficulties of assessing treatment effect, and expose the nature of the implicit choices being made which appear disconnected from common program objectives. Data "snooping", such as balancing, and estimation of flexible or semi-parametric functional forms for PS regressions, are attempts at changing the above distribution matches, but without the benefit of guidance by whole distribution measures of "match", and often distracted by statistical cost functions that remain to be

justified or connected to goals of matching populations and samples.

## Treatment Effects: Stochastic Dominance

Following the IT reasoning above, it is possible and desirable to assess the similarity, or any relation between the distribution of outcomes for the treated and untreated/counterfactual groups. Entropy measures such as KL can be used both to quantify differences and to test if the differences are significant. Any summary measure of this divergence/distance corresponds to a cardinal evaluation function that puts different weights on different subjects. This is true of the "average" treatment effects, and there is no way around it. Averages are particularly non-robust to outliers and special outcomes which may have nothing to do with the treatment. But averages, or specific quantiles, have the advantage of easy monetization for standard cost benefit analysis of policy outcomes. It would be useful to examine whether the *distribution* of outcomes is rankable between two scenarios, no matter what weights are attached to different subgroups. If it is desired to assess the impact of treatment on the entire "sample", and to avoid full cardinalization, such as in "average effect" or choice of any quantile, weak uniform ranking presents itself as a useful method. Such rankings are achieved by tests for stochastic dominance whereby we may be able to conclude that, the treated are better off than the untreated, whichever utility/welfare function one chooses in large classes of such functions. Note that this type of ranking does not provide quantitative values for treatment effects. For that we need to cardinalize and make the hard decisions about what weight to give to different parts of the population/sample. The recent paper of Dahejia (2005) is in the same spirit as it considers a decision theoretic context for program evaluation. We first describe the elements of ordering and statistically testing for stochastic dominance in the univariate case.

### Definitions and Tests in the Univariate Case

Let  $X$  and  $Y$  be two variables at either two different points in time, before and after treatment, or for different regions or countries. Let  $X_1, X_2, \dots, X_n$  be  $n$  not necessarily i.i.d observations on  $X$ , and  $Y_1, Y_2, \dots, Y_m$  be similar observations on  $Y$ . Let  $U_1$  denote the class of all utility functions  $u$  such that  $u' \geq 0$ , (increasing). Also, let  $U_2$  denote the class of all utility functions in  $U_1$  for which  $u'' \leq 0$  (strict concavity), and  $U_3$  denote the subset of  $U_2$  for which  $u''' \geq 0$ . Let  $X_{(i)}$  and  $Y_{(i)}$  denote the  $i$ -th order statistics, and assume  $F(x)$  and  $G(x)$  are continuous and monotonic cumulative distribution functions (cdf,s) of  $X$  and  $Y$ , respectively.

Quantiles  $q_x(p)$  and  $q_y(p)$  are implicitly defined by, for example,  $F[X \leq q_x(p)] = p$ .

**Definition**  $X$  First Order Stochastic Dominates  $Y$ , denoted  $X$  FSD  $Y$ , if and only if any one of the following equivalent conditions holds:

$$(1) E[u(X)] \geq E[u(Y)] \quad \text{for all } u \in U_1, \text{ with strict inequality for some } u.$$

$$(2) F(x) \leq G(x) \quad \text{for all } x \text{ in the support of } X, \text{ with strict inequality for some } x.$$

$$(3) q_x(p) \geq q_y(p) \quad \text{for all } 0 \leq p \leq 1.$$

**Definition**  $X$  Second Order Stochastic Dominates  $Y$ , denoted  $X$  SSD  $Y$ , if and only if any of

the following equivalent conditions holds:

- (1)  $E[u(X)] \geq E[u(Y)]$  for all  $u \in U_2$ , with strict inequality for some  $u$ .
- (2)  $\int_{-\infty}^x F(t)dt \leq \int_{-\infty}^x G(t)dt$  for all  $x$  in the support of  $X$  and  $Y$ , with strict inequality for some  $x$ .
- (3)  $\int_0^p q_x(t)dt \geq \int_0^p q_y(t)dt$ , for all  $0 \leq p \leq 1$ , with strict inequality for some value(s)  $p$ .

The tests of FSD and SSD are based on empirical evaluations of conditions (2) or (3) in the above definitions. Mounting tests on conditions (3) typically relies on the fact that quantiles are consistently estimated by the corresponding order statistics at a finite number of sample points. Mounting tests on conditions (2) requires empirical cdfs and comparisons at a finite number of observed ordinates. Also condition (3) of SSD is equivalent to the requirement of Generalized Lorenz (GL) dominance, and lower order dominance implies higher order ones.

Whitmore (1970) introduced the concept of third order stochastic dominance (TSD) in finance. Addition of an increasing “transfer sensitivity” requirement leads to TSD ranking of distributions. This requirement is stronger than the Pigou-Dalton principle of transfers and is based on the class of welfare functions  $U_3$ . TSD is defined as follows:

**Definition** *X Third Order Stochastic Dominates Y, denoted X TSD Y, if and only if any of the following equivalent conditions holds:*

- (1)  $E[u(X)] \geq E[u(Y)]$  for all  $u \in U_3$ , with strict inequality for some  $u$ .
- (2)  $\int_{-\infty}^x \int_{-\infty}^v [F(t) - G(t)]dt dv \leq 0$ , for all  $x$  in the support, with strict inequality for some  $x$ ,

with the end-point condition:

$$\int_{-\infty}^{+\infty} [F(t) - G(t)]dt \leq 0.$$

- (3) When  $E[X] = E[Y]$ , X TSD Y iff  $\lambda_x^2(q_i) \leq \lambda_y^2(q_i)$ , for all Lorenz curve crossing points  $i = 1, 2, \dots, (n + 1)$ ; where  $\lambda_x^2(q_i)$  denotes the “cumulative variance” for incomes up to the  $i$ th crossing point. See Davies and Hoy (1996).

The McFadden-type tests require a definition of “maximal” sets, as follows:

**Definition** *Let  $\mathcal{A} = \{X_1, X_2, \dots, X_K\}$  denote a set of  $K$  distinct random variables. Let  $F_k$  denote the cdf of the  $k$ th variable. The set  $\mathcal{A}$  is first (second) order maximal if no variable in  $\mathcal{A}$  is first (second) order weakly dominated by another.*

Let  $X_{.n} = (x_{1n}, x_{2n}, \dots, x_{Kn})$ ,  $n = 1, 2, \dots, N$ , be the observed data. We assume  $X_{.n}$  is strictly stationary and  $\alpha$  – mixing. We also assume  $F_k$  is unknown and estimated by the empirical distribution function  $F_{kN}(X_k)$ . Finally, we adopt the mathematical regularity conditions pertaining to von Neumann-Morgenstern (VNM) utility functions that generally underlie the expected utility maximization paradigm. The following theorem defines our tests and the hypotheses being tested:

**Theorem** *Given the mathematical regularity conditions;*

- (a) *The variables in  $\mathcal{A}$  are first-order stochastically maximal; i.e.,*

$$d = \min_{i \neq j} \max_x [F_i(x) - F_j(x)] > 0, \quad (1)$$

if and only if for each  $i$  and  $j$ , there exists a continuous increasing function  $u$  such that  $E u(X_i) > E u(X_j)$ .

(b) The variables in  $\mathcal{A}$  are second order stochastically maximal; i.e.,

$$S = \min_{i \neq j} \max_x \int_{-\infty}^x [F_i(\mu) - F_j(\mu)] d\mu > 0, \quad (2)$$

if and only if for each  $i$  and  $j$ , there exists a continuous increasing and strictly concave function  $u$  such that  $E u(X_i) > E u(X_j)$ .

(c) Assuming the stochastic process  $X_n$ ,  $n = 1, 2, \dots$ , to be strictly stationary and  $\alpha$ -mixing with  $\alpha(j) = O(j^{-\delta})$ , for some  $\delta > 1$ , we have:

$d_{2N} \rightarrow d$ , and  $S_{2N} \rightarrow S$ , where  $d_{2N}$  and  $S_{2N}$  are the empirical test statistics defined as

1.

$$d_{2N} = \min_{i \neq j} \max_x [F_{iN}(x) - F_{jN}(x)] \quad (3)$$

and,

$$S_{2N} = \min_{i \neq j} \max_x \int_0^x [F_{iN}(\mu) - F_{jN}(\mu)] d\mu \quad (4)$$

The null hypotheses tested by these two statistics is that, respectively,  $\mathcal{A}$  is not first (second) order maximal— i.e.,  $X_i$  FSD(SSD)  $X_j$  for some  $i$  and  $j$ . We reject the null when the statistics are positive and large. Since the null hypothesis in each case is composite, power is conventionally determined in the least favorable case of identical marginals  $F_i = F_j$ .

## SD rankings for DW data

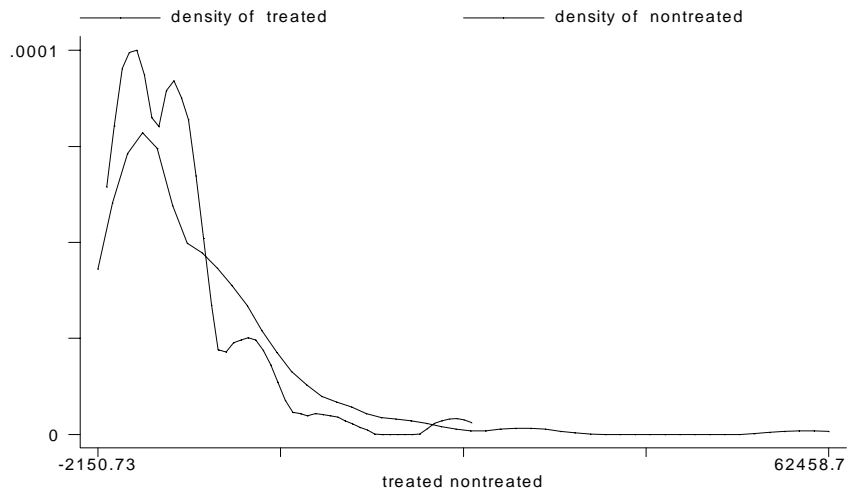


Fig. 5

pdf of PS

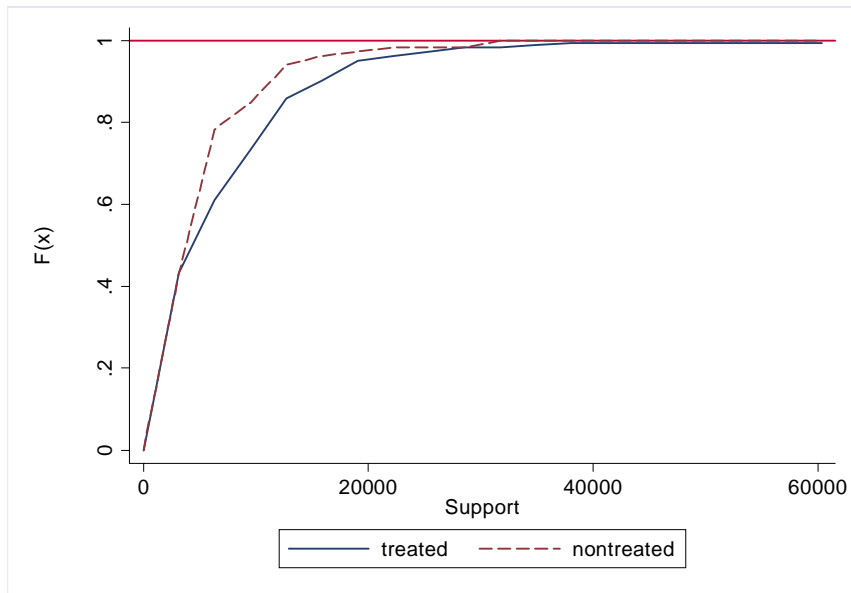


Fig.6

CDFs of PS

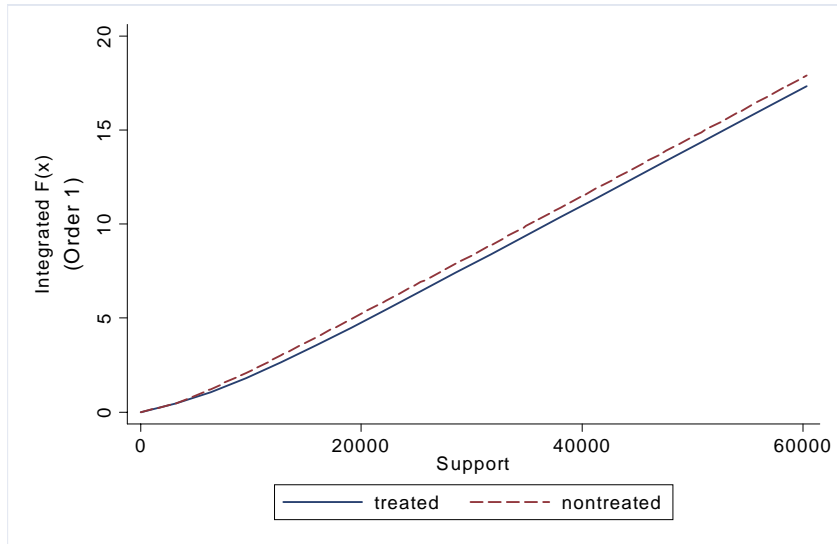


Fig.7

Integrated CDFs of PS

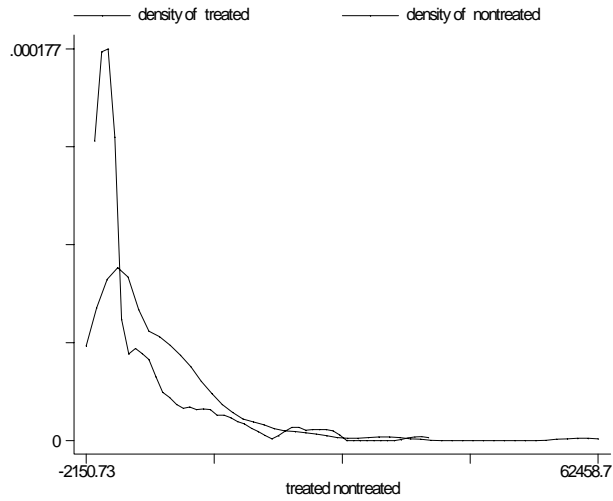


Fig.8 pdf of

$S(i); \beta = -1$

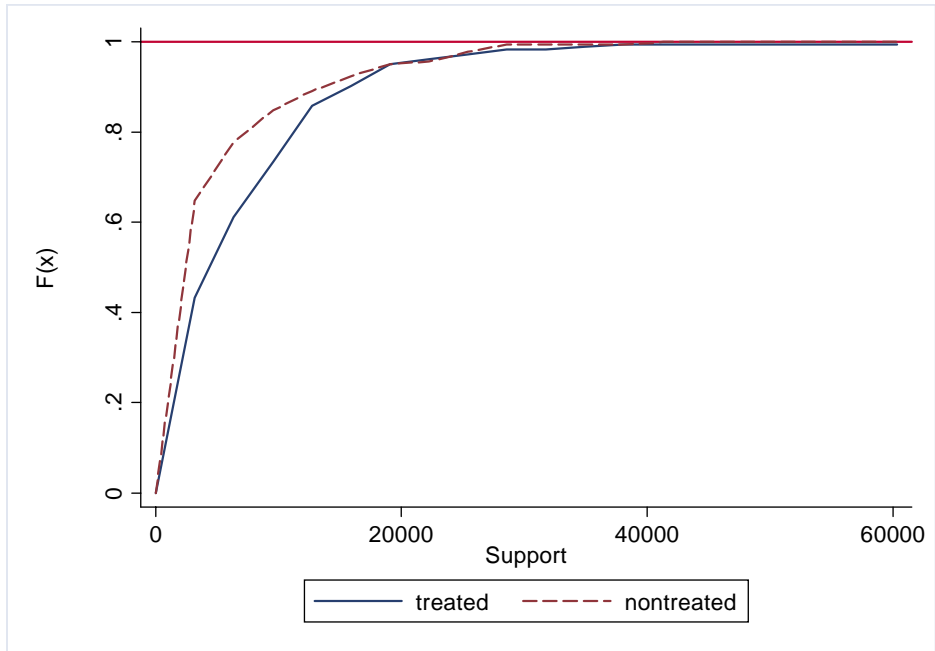


Fig.9 CDFs

of  $S(i)$ ;  $\beta = -1$

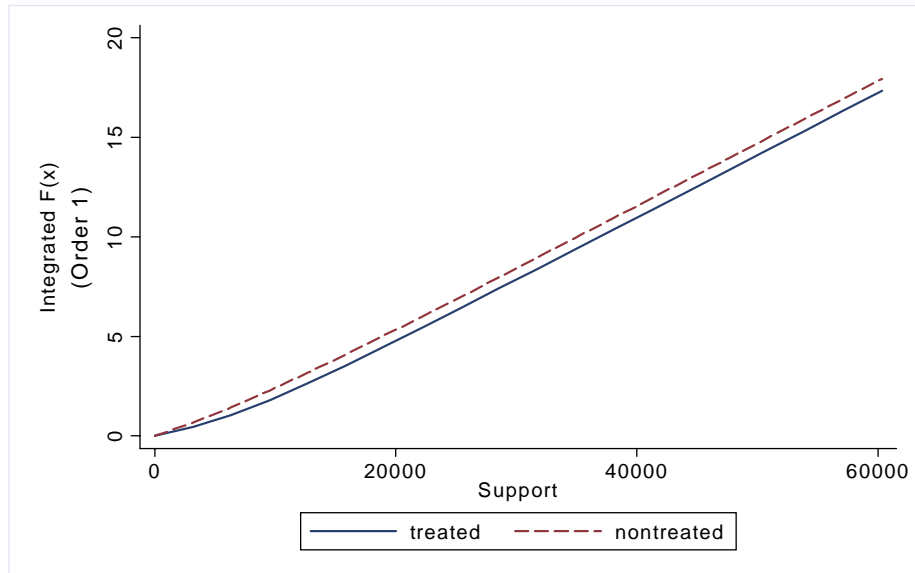


Fig.10 Integrated

CDFs  $S(i)$ ;  $\beta = -1$

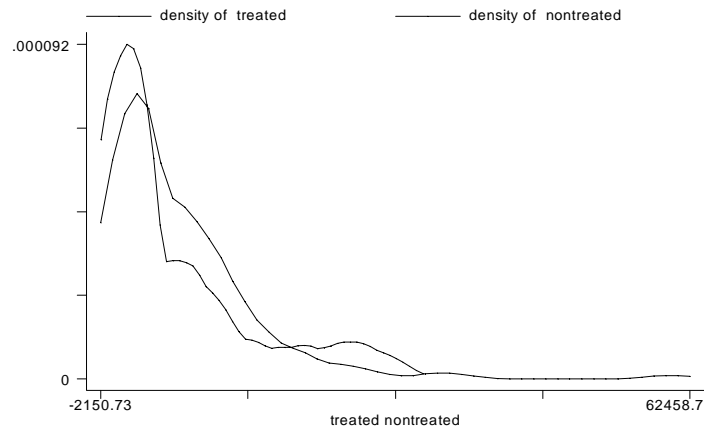


Fig. 11 PDFs of

$S(i)$  for  $\beta=-1/2$

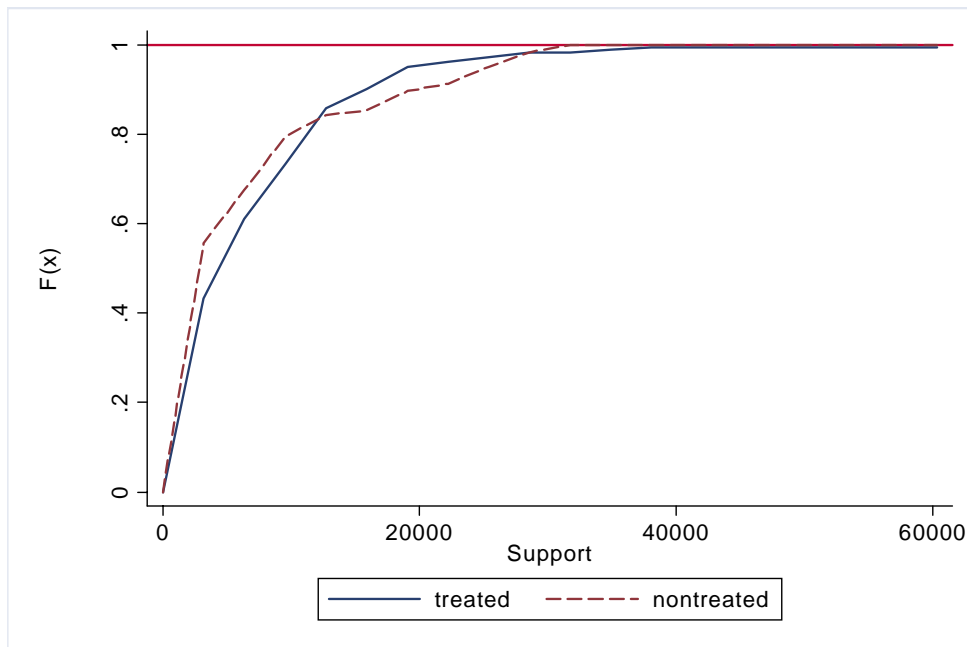


Fig.12 CDFs of

$S(i)$  for  $\beta=-1/2$

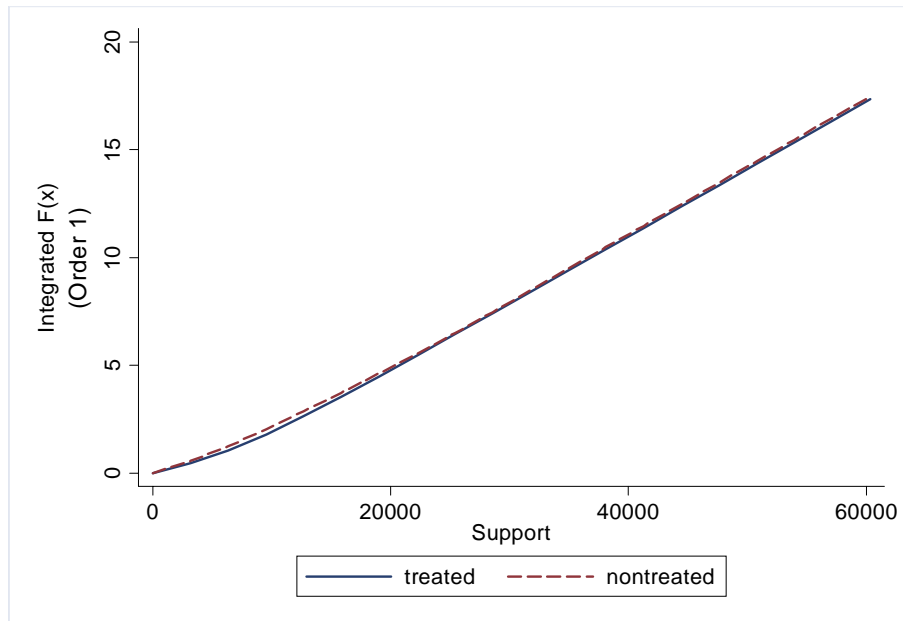


Fig.13 Integrated CDFs of

$S(i)$  for  $\beta=-1/2$  (3rd order SD)

This is an interesting case in which 2nd order SD is not evident. Even a risk/inequality averse policy maker could not use either the average treatment effect, or any quantiles, to assess the program outcome. Only a function that reflects both aversion to inequality and increasing sensitivity to transfers to the poorer of the wage earners could quantify the outcome and not hide any important details.

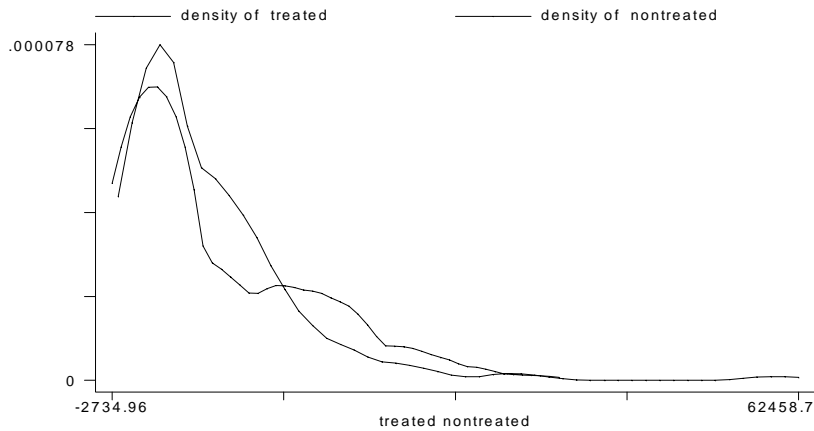


Fig.14 pdf  $S(i)$  ;

$$\beta = -2/3$$

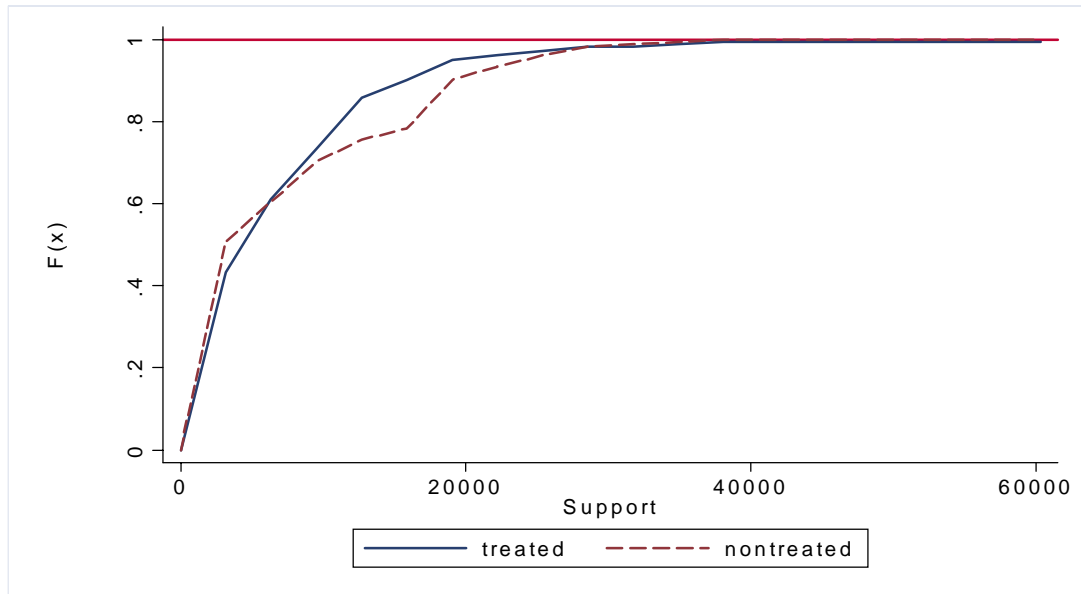


Fig. 15 CDFs of

$S(i); \beta = -2/3$

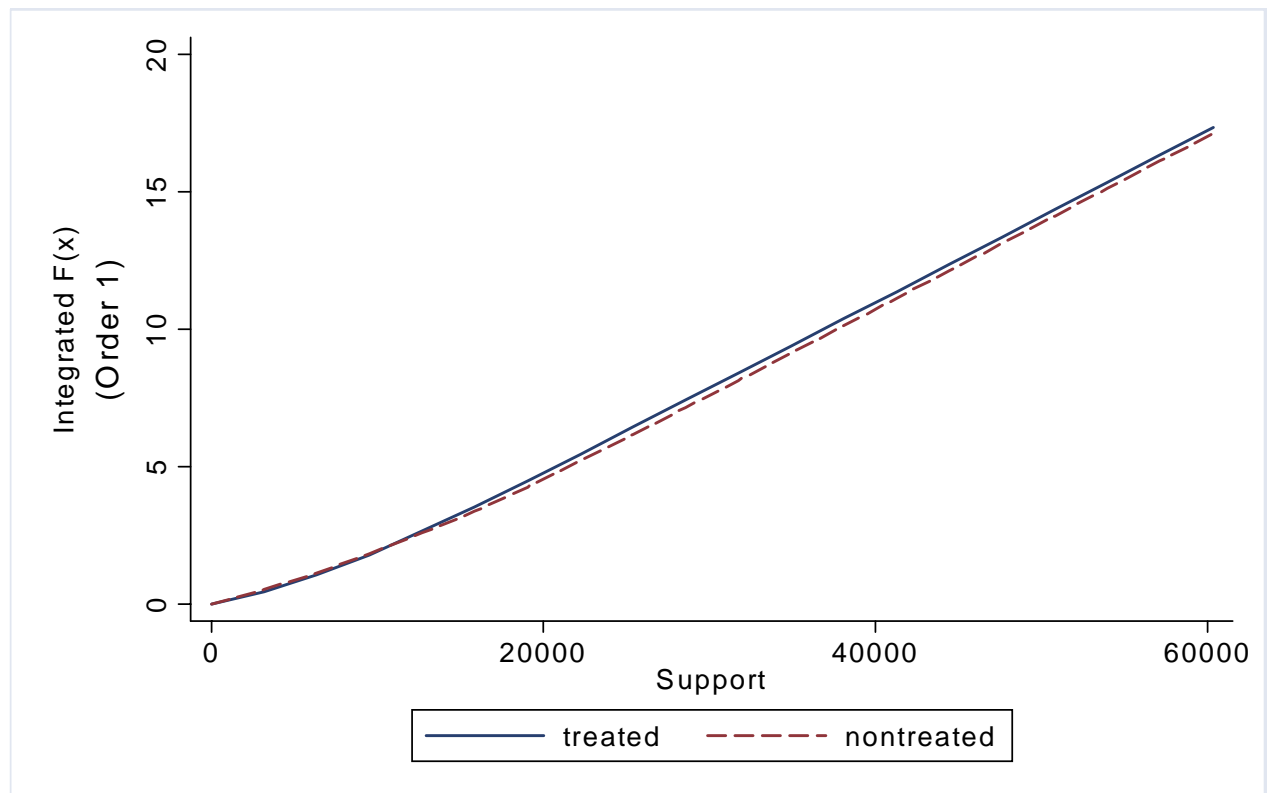


Fig.16 Integrated CDFs

$S(i); \beta = -2/3$  (3rd order SD)

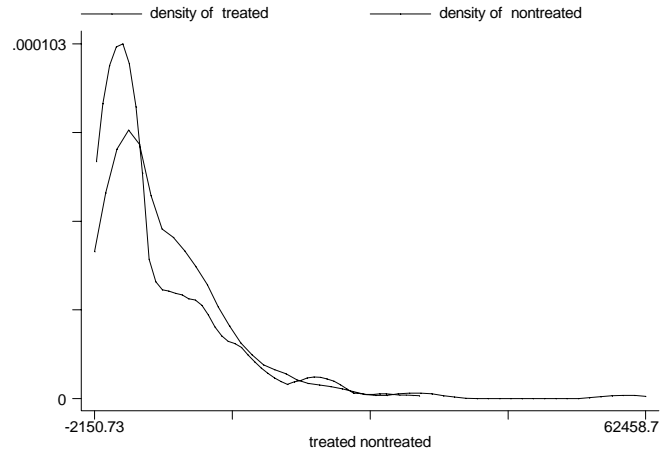


Fig.17 pdfs

of  $\Phi(S(i))$  for  $\beta=-1$

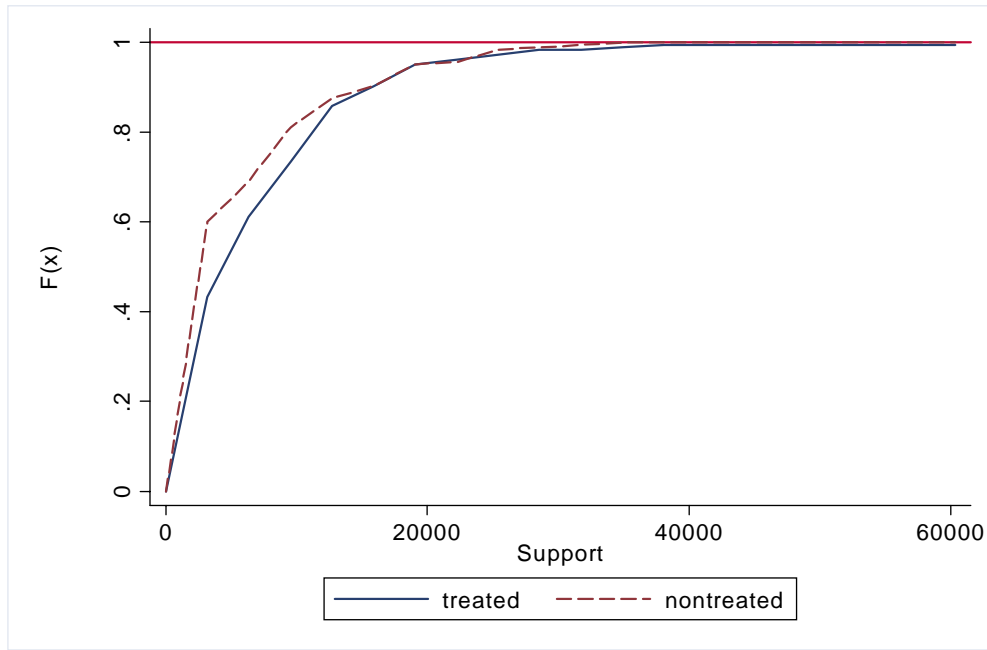


Fig.18 CDFs

of  $\Phi(S(i))$  for  $\beta=-1$

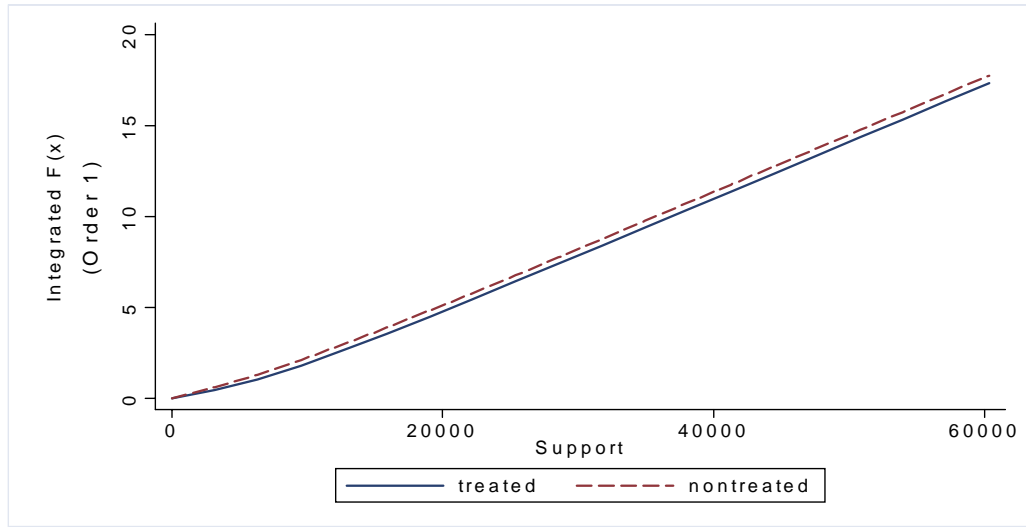


Fig. 19 Integrated CDFs

of  $\Phi(S(i))$  for  $\beta=-1$

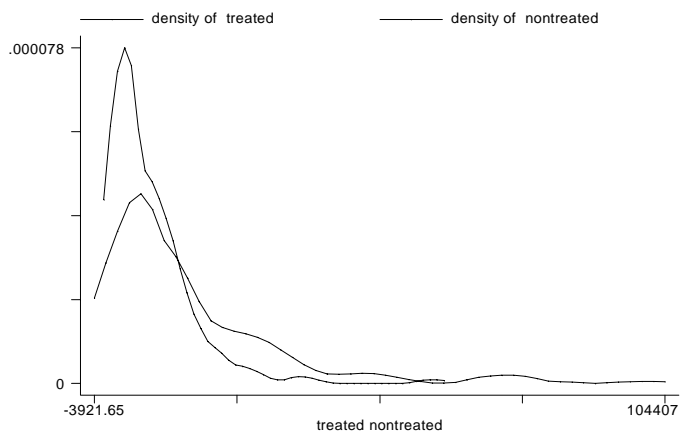


Fig. 20 pdfs

of  $\Phi(S(i))$  for  $\beta=-1/2$

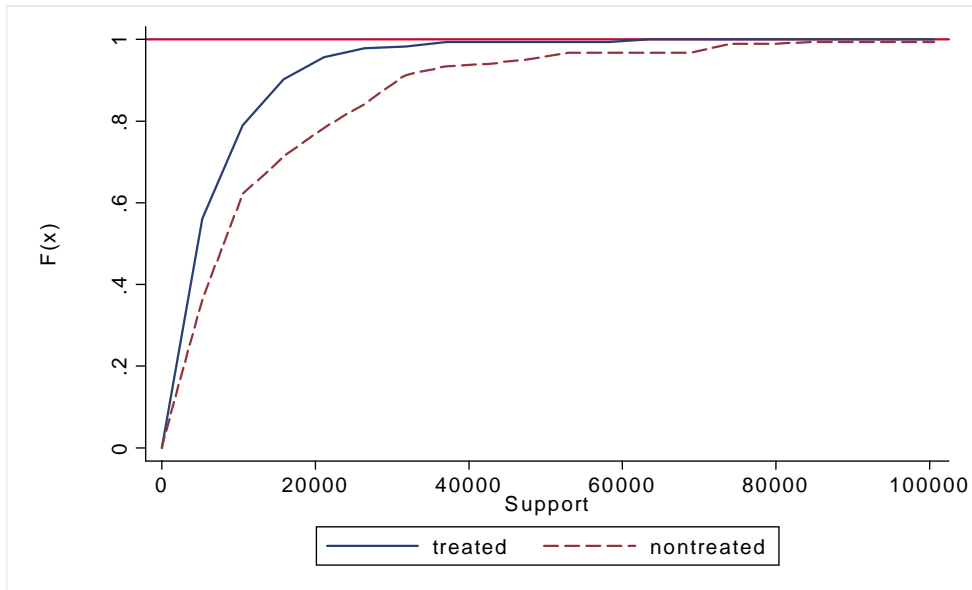


Fig.21 CDFs of

$\Phi(S(i))$  for  $\beta=-1/2$

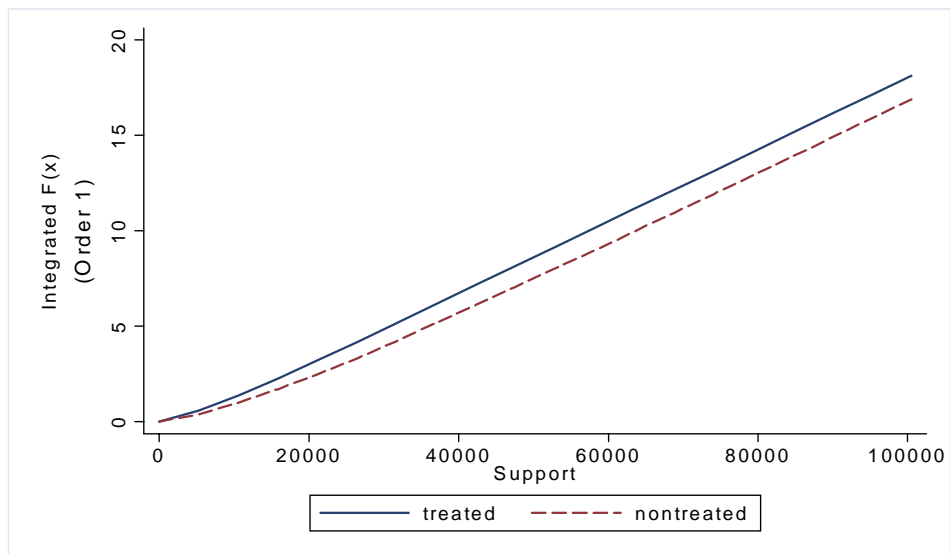


Fig.22 Integrated

CDFs of  $\Phi(S(i))$  for  $\beta=-1/2$

## Multivariate Rankings and Comparisons

Atkinson and Bourguignon (1982, 1987, 1989) have developed some useful conditions for ranking multidimensional distributions. SWFs are taken to be individualistic and (for convenience) separable. But anonymity may be dropped in recognition of the fact that households (individuals) must be distinguished according to some distinct characteristic. It is desired to rank the distributions of outcomes in two states conditional on a given distribution of discrete characteristics such as gender or family composition.

Let there be  $G$  groups which are characterized in terms of two characteristics, gender and wages, say. All members of a group  $g \in G$  have the same valuations and marginal valuations of the wages. If there were no wage transfers between groups as a result of the treatment, the necessary and sufficient conditions for FSD and SSD given above would have to hold for *all* groups for FSD and SSD to hold. If there is any transfer between groups, however, one must deal with each group's evaluation as well as *between group valuations* of the trade-offs between wages and the discrete distinguishing covariate, "gender", say. Thus interpersonal comparisons of well-being are inevitable whenever heterogenous populations are involved. In this general situation Shorrocks (1995) provides an "impossibility" result for unambiguous or consensus rankings. But, "majority" rankings are possible with plausible restrictions. To see this it is worthwhile to formally describe the conditions of Atkinson and Bourguignon (1987) here as they combine the desirable elements of "decomposability" alluded to earlier, and partial ordering which, although it avoids full cardinalization, shows the directions in which an analyst may wish to make increasingly normative assumptions to *approach cardinality*; see Basu (1980).

Let  $u(X, nd)$  denote private valuations of wage  $X$  and "nd",  $w(u(X, nd))$  or just  $w(X, nd)$  represent the social welfare (or decision) function, and  $p_g$ ,  $g=1,2,\dots,G$ , the marginal frequency in class  $g$ . The cumulative function is  $P_g = \sum_{j=1}^g p_j$ ,  $P_G=1$ . The social valuation of wage received by household  $g$  is  $U^g(X)$  which is assumed continuously differentiable as needed. The  $U_1$  and  $U_2$  classes are as defined earlier with the partial derivatives  $U_X^g \geq 0$ , as well as  $U_{XX}^g \leq 0$  for  $U_2$ .

If no assumptions are made about how  $U^g$  varies with  $g$  the conditions of FSD and SSD must hold for all groups  $g$  for FSD and SSD to hold. These are strong conditions. Among other things, they require that the mean wages of all groups must be no lower in the dominant distribution. This would rule out equalizing treatments between groups with different "needs". To resolve this situation one must specify some aspects of the trade-off between incomes and needs.

The traditional univariate/homogenous analysis is implicitly based on the extreme assumption that  $U_X^g(X) = U_X(X)$ ,  $\forall g$ . The *level* of welfare can vary with "nd", but no more. This assumption is sufficient to allow a consideration only of the marginal distribution of wages. But suppose one follows Sen in weakening the "Equity Axiom" by assuming that groups can be ranked by their *marginal valuation of wages*. For instance, if the group with smallest "nd" has the highest marginal valuation of wages, the next higher "nd" group has the second highest marginal valuation, and so on, then the necessary and sufficient condition for FSD of  $F_1$  over  $F_2$  is:

$$\sum_{g=1}^j p_g [F_1^g - F_2^g] \leq 0, \text{ for all } X \text{ and all } j = 1, \dots, G \quad \#$$

where superscript indicates the wage distribution for  $g$ -th group. Note that the final condition here is the FSD of the entire marginal distribution of wages. As was shown by Sen (1973b), it is possible for a utilitarian SWF to violate this type of "weak equity axiom". But as Atkinson and Bourguignon (1987) point out, marginal valuation by *society* can take into account the *level* of individual welfare. Therefore it is possible that the assumed negativity of  $u_{Xh}$  may be offset by sufficient degree of concavity ( $-w''/w'$ )

of the additive social valuation function  $w(\cdot)$ . Thus the ranking of groups assumed by Atkinson and Bourguignon (1987) coincides with the ranking of levels of welfare where higher needs increase marginal valuations of wages, or the social welfare function has a sufficiently large degree of concavity.

The above FSD condition may be weakened further for SSD if we are willing to assume that “the differences in the social marginal valuation of wages between groups become smaller as we move to higher wage levels”; see Atkinson and Bourguignon (1987). That is  $-U_{XX}$  decreases for smaller “nd” groups, reflecting less social concern with differences in “nd” for higher wage groups. If this assumption is adopted, a necessary and sufficient condition for SSD is:

$$\sum_{g=1}^j p_g \left[ \int_0^x (F_1^g - F_2^g) dX \right] \leq 0 \text{ for all } x, \text{ and } j = 1, \dots, G \quad \#$$

This includes the usual SSD condition for the marginal distribution of wages.

Atkinson and Bourguignon (1987) consider weaker SSD conditions by exploring further assumptions toward cardinality. One such assumption allows further comparability between the *differences* of  $U_X$  and  $U_{XX}$ . Thus, if the rate of decline of social marginal valuation of wages across groups is positive *and declines with g*, and the same property holds for the degree of concavity ( $-U_{XX}$ ), the necessary and sufficient condition for SSD is given as follows:

$$\sum_{j=1}^k \sum_{g=1}^j p_g \left[ \int_0^x (F_1^g - F_2^g) dX \right] \leq 0 \text{ for all } x \text{ and } k=1, \dots, G-1$$

and

$$\sum_{g=1}^G p_g \left[ \int_0^x (F_1^g - F_2^g) dX \right] \leq 0 \text{ for all } x \quad \#$$

It is worth noting that all the above conditions are testable using the tests outlined above.

Consistent with a philosophy of “partial comparability” developed by Sen (1970b), Atkinson and Bourguignon (1987) have therefore shown that nihilism is avoidable if certain plausible assumptions are made about the trade offs between wages/outcomes and “needs”, and at different levels of needs, should we agree that groups can be ranked by such “other” characteristics as “needs”. These additional assumptions lead to the development of empirically implementable tests for stochastic dominance which are somewhat more general but less demanding than those described earlier.

## Multiple Outcomes

Suppose several outcomes are thought to flow from a treatment and are summarized by our proposed S aggregates. An example of this is experiments in which both earnings and probability of future employment are germane to policy evaluation. A multivariable assessment is then required of several treatment effects. Let policy evaluation, or a Social Welfare Functions (SWFs) be “equality preferring”, or “risk

averse". Such SWFs are typically non-decreasing, symmetric, quasi-concave, and thus also Schur-concave. We consider a composite measures like  $S_i$  which also reflect some trade offs between outcomes. But  $S_i$  are not subject of treatment policy, their components, the distinct  $Y_{ij,s}$  are. The following Proposition establishes a Principle of Transfers property of the multidimensional welfare functions that is useful for changes in the matrix  $Y = (Y_{ij})$ :

Let  $B$  be a bistochastic matrix such that  $b_{ij} \geq 0$ ,  $\sum_i b_{ij} = 1 = \sum_j b_{ij} \forall i \text{ and } j$ . Such matrices perform mean preserving spreads or "equalizing" transformations. Also, let  $H$  denote the set of all positive, real valued concave (concave increasing or concave non-decreasing) functions  $h(\cdot)$ .

**Proposition** *Let  $\tilde{Y} = BY$ , where  $B$  is a bistochastic matrix. Then  $W(\tilde{S}) \geq W(S)$  for all Schur-concave  $W(\cdot)$ , and  $h \in H$  such that  $\tilde{S}_i = h(\tilde{Y}_i)$  and  $S_i = h(Y_i)$ , where  $Y_i$  denotes the vector of  $M$  outcomes for the  $i$ th subject.*

Proof: See Kolm (1977, Th.6). In fact the converse also holds.

Maasoumi (1986a, Proposition 3) was an attempt at deriving a similarly strong result for rankings by Schur-convex inequality measures. This can be done for only a limited range of  $S$  functions, however.

This result and other multivariate majorization techniques can be used for multivariable outcome rankings. The next section, however, offers a more direct, nonparametric method of ranking for both univariate and multivariate situations.

## Redundant Covariates

It is worthwhile to first note a possible difficulty with potential "double counting" of the same *characteristics* by inclusion of measurements on *seemingly* distinct attributes. Put differently, two apparently distinct attributes may offer almost identical amounts of "information" to the information set inevitably utilized by any statistical measure. This issue is studied by Hirschberg, Maasoumi, and Slottje (1991) for international data. The basic idea is to detect "clusters" of attributes which are statistically similar. Once this is accomplished, a "representative" aggregator attribute is chosen for each cluster. These representative or composite attributes are then included in the desired but lower dimensional multivariate welfare analysis. The approach of Hirschberg et al (1991) is based on statistical clustering techniques as well as a new entropy based criterion. In Hirschberg et al. (1991) 24 attributes of well being were analyzed for 120 countries. These included such attributes as GNP and related concepts, literacy and mortality rates, labor force participation rates, basic amenities (*e.g.*, radios and roads), militarization indices, health status, infrastructure indicators, political freedom and civil liberty measurements. Interesting and quite plausible "clusters" were identified based on several criteria of similarity. The authors then proceeded to compute aggregate measures of well being on the basis of the "representative" attributes for the five clusters. This type of study also allows an investigation of the robustness of inferences, for example, with respect to levels of aggregation (clustering), weighting factors, and degrees of "aversion" (parameter  $\beta$  of GE).

## References

- Abadie, A., D. Drukker, J.L. Herr, and G.W. Imbens (2004), "Implementing Matching Estimators for Average Treatment Effects in STATA," *STATA Journal*, 4, 290-311
- bibitem Atkinson, A., and F. Bourguignon (1982), "The comparison of multi-dimensioned distributions of economic status," **Review of Economic Studies**, 49, 183-201.
- atkinson and Bourguignon 2 \_\_\_\_\_ (1987), "Income Distribution and Differences in Needs," in G.R. Feiwel (ed.), **Arrow and the Foundations of the Theory of Economic Policy**, Macmillan: London.
- ATKINSON, BOURGUIGNON 89 \_\_\_\_\_ (1989), "The design of Direct Taxation and Family Benefits," **Journal of Public Economics**, 41, 3-29.
- atkinson et al Atkinson, A., F., Bourguignon, F., and C. Morrison (1992). **Empirical studies of earnings mobility**, *Fundamentals of Pure and Applied Economics*, vol. 52, Distribution Section, Harwood Academic.
- Basu1980 Basu, K. (1980) **Revealed preference of government**, Cambridge: Cambridge University Press.
- blundell and Lewbel 91 Blundell, R. and A. Lewbel (1991), "The information content of equivalence scales," *Journal of Econometrics*, (Annals issue edited by E. Maasoumi, cited below), 50, No. 1/2, 49-68.6, 335-346.
- Dahejia 2005 Dahejia, R.H. (2005), "Program evaluation as a Decision Problem", *Journal of Econometrics*, 125, 141-173.
- dah-Wah 99 Dahejia, R., and S. Wahba (1999), "Causal Effects in Non-Experimental Studies: Reevaluating the evaluation of training programs", *Journal of American Statistical Association*, 94, 1053-1062.
- dahejia-Wahba Dehejia, R. and S. Wahba (2002), "Propensity Score-Matching Methods for Nonexperimental Causal Studies," *Review of Economics & Statistics*, 84, 151-161.
- GMR<sub>04</sub> Granger, C., E. Maasoumi and J. S. Racine (2004), "A Dependence Metric for Possibly Nonlinear Time Series", *Journal of Time Series Analysis*, vol. 25, 5, pages 649-669.
- heckman 04 Heckman, J. and S. Navarro-Lozano (2004), "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models," *Review of Economics & Statistics*, 86, 30-57
- hirschberg, maasoumi, slottje 91 Hirschberg, J.G., E. Maasoumi, and D.J. Slottje (1991), "Cluster analysis for measuring welfare and quality of life across countries," *Journal of Econometrics*, Annals issue, 50, No. 1/2, 131-150.
- imbens04 Imbens, G.W. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics & Statistics*, 86, 4-29.
- Kolm77 Kolm, S.C. (1977), "Multidimensional egalitarianism," *Quarterly Journal of Economics*, **91**, 1-13.
- LMW05 Linton, O., E. Maasoumi, and Y. J. Whang (2005), "Consistent Testing for Stochastic Dominance under general Sampling Schemes", *Review of*

- Economic Studies*, forthcoming.
- Maasoumi79 Maasoumi, E. (1979), "An entropy index of multivariable inequality", Mimeograph, Department of Economics, University of Southern California.
- MAASOUMI 86a Maasoumi, E.(1986a), "The Measurement and Decomposition of Multi-Dimensional Inequality," *Econometrica*, 54, 991-97.
- maasoumi1986b \_\_\_\_\_(1986b), "Unknown regression functions and efficient functional forms: An interpretation," *Advances in Econometrics*, 5, 301-9.
- MAASOUMI 89A Maasoumi, E.(1989a), "Composite Indices of Income and Other Developmental Indicators : A General Approach," *Research on Economic Inequality*, Vol. 1, 269-286.
- MAASOUMI 93 \_\_\_\_\_. (1993), "A Compendium to Information Theory in Economics and Econometrics," *Econometric Reviews*, Vol 12, 3, 1-49.
- MAASOUMI, NICKELSBURG 88 Maasoumi, E., and G. Nickelsburg (1988), "Multivariate Measures of Well Being and an Analysis of Inequality in the Michigan Data," *Journal of Business and Economic Statistics*, 6, 3, 327-334.
- MCFADDEN 89 McFadden, D. (1989), "Testing for stochastic dominance," in Part II of T. Fomby and T.K. Seo (eds.) *Studies in the Economics of Uncertainty* (in honor of J. Hadar), Springer-Verlag.
- POLLAK 91 Pollak, R.(1991), "Welfare comparisons and situation comparisons," *Journal of Econometrics*, (Annals issue edited by E. Maasoumi, op.cited), 50, No. 1/2, 31-48.
- POLLAK, WALES 79 Pollak, R., and T. Wales (1979), "Welfare comparisons and equivalence scales," *American Economic Review*, 69, 216-221.
- RAM 82 Ram, R.(1982), "Composite Indices of Physical Quality of Life, Basic needs Fulfillment, and Income : A •Principal Component• representation," *Journal of Development Economics*, 11, 227-47.
- SEN 70 Sen, A. (1970a) **Collective choice and social welfare**, Holden Day: San Francisco, (reprinted, North-Holland, 1979).
- sen1970b \_\_\_\_\_(1970b), "Degrees of cardinality and aggregate partial orderings," *Econometrica*, 43, 393-409.
- SEN 77 \_\_\_\_\_(1977) , "On Weights and Measures: Informational constraints in social welfare analysis," *Econometrica*, 45, 7, 1539-1572.
- SHORROCKS 80 \_\_\_\_\_(1980), "The class of additively decomposable inequality measures," *Econometrica*, 48, 613-625.
- shorrocks95 \_\_\_\_\_(1995), "Inequality and welfare evaluation of heterogeneous income distributions," Unpublished paper, University of Essex, UK.
- Smith-Todd 2005 Smith, J. A. and P.E. Todd (2005), "Does Matching overcome LaLonde's critique of nonexperimental estimators", *Journal of Econometrics*, 125, 305-353.
- TSUI, K. Y. 92B Tsui, K – Y. (1992), "Composite indices of well-being: Axiomatic foundations," Chinese Univ. of Hong Kong, Dept. of Economics, Mimeo., September.
- XU, FISHER, WILSON 95 Xu, K., G. Fisher, and D. Wilson (1995), "New distribution-free tests for stochastic dominance,"