Differences-in-differences, differences of quantiles and quantiles of differences

Franco Peracchi

October 13, 2006

1 A motivating example

- Economists have largely debated on the causes of health care cost increases, and a vast literature discusses the major drivers of this expenditure and the most effective policies for controlling it.
- Unfortunately, much less is known concerning the productivity of such expenditure, namely the "value for money" or what we gain in terms of improved health for each euro spent.
- Having a better knowledge of the productivity levels in the health sector is particularly important when policy makers and health care managers adopt measures aimed at controlling costs or at rationalizing the access to health care services by patients.
- Increasing the efficiency of health systems is the most promising response to pressures to contain costs while improving performance.
- Spending more is not necessarily a problem if the added benefits exceed the extra costs (OECD, 2004)
- Unfortunately, policy evaluation remains a critical issue!

The data

- Our data is based on administrative registries maintained by the Pharmaceutical Service Department of the Local Health Authority (LHA) of Treviso, a province in the North-East of Italy.
- The data has been obtained by merging three different registries containing information about daily access to public health care services by the whole local population:
  - The drug prescription database.
  - The hospitalization registry.
  - The death and transfer registry.
- We concentrate our analysis on hypertensive patients, born between 1910 and 1960, who were regularly prescribed drugs in the ACE-inhibitor class during the period 1997–2002 .

The policy changes

- Our data span three major policy changes.
- Two of them regard drug co-payment, and took place in January 2001 (abolition - Policy 1) and March 2002 (reintroduction - Policy 3) respectively.
- The other, which took place in September 30th, 2001, regards the maximum number of packages that can be purchased with a single prescription (reduction from 6 to 3 of the maximum number of packages for each prescription - Policy 2).
- These policy changes represent three "natural experiments", whose effects on medical compliance and health outcomes can be evaluated using a difference-in-difference (DiD) specification.

Figure 1: Observed and fitted hospitalization and mortality rates by gender and compliance level.

Figure 2: Average quarterly compliance for high and low compliants.

Caveats of DiD analysis

- Although interesting from the point of view of the problem investigated, a major limit of the DiD is that it focuses on the average effect.
- This can be misleading in that we ignore, say, the possibility of substantial benefits for some, little benefits for many and harm for a few.
- In other words, we are only allowed to know that treatment is "good" or "bad" on average, but we miss information on single patient responses. This is a critical point, especially for the design and evaluation of an efficient policy.

Figure 3: Quantiles of compliance by period.

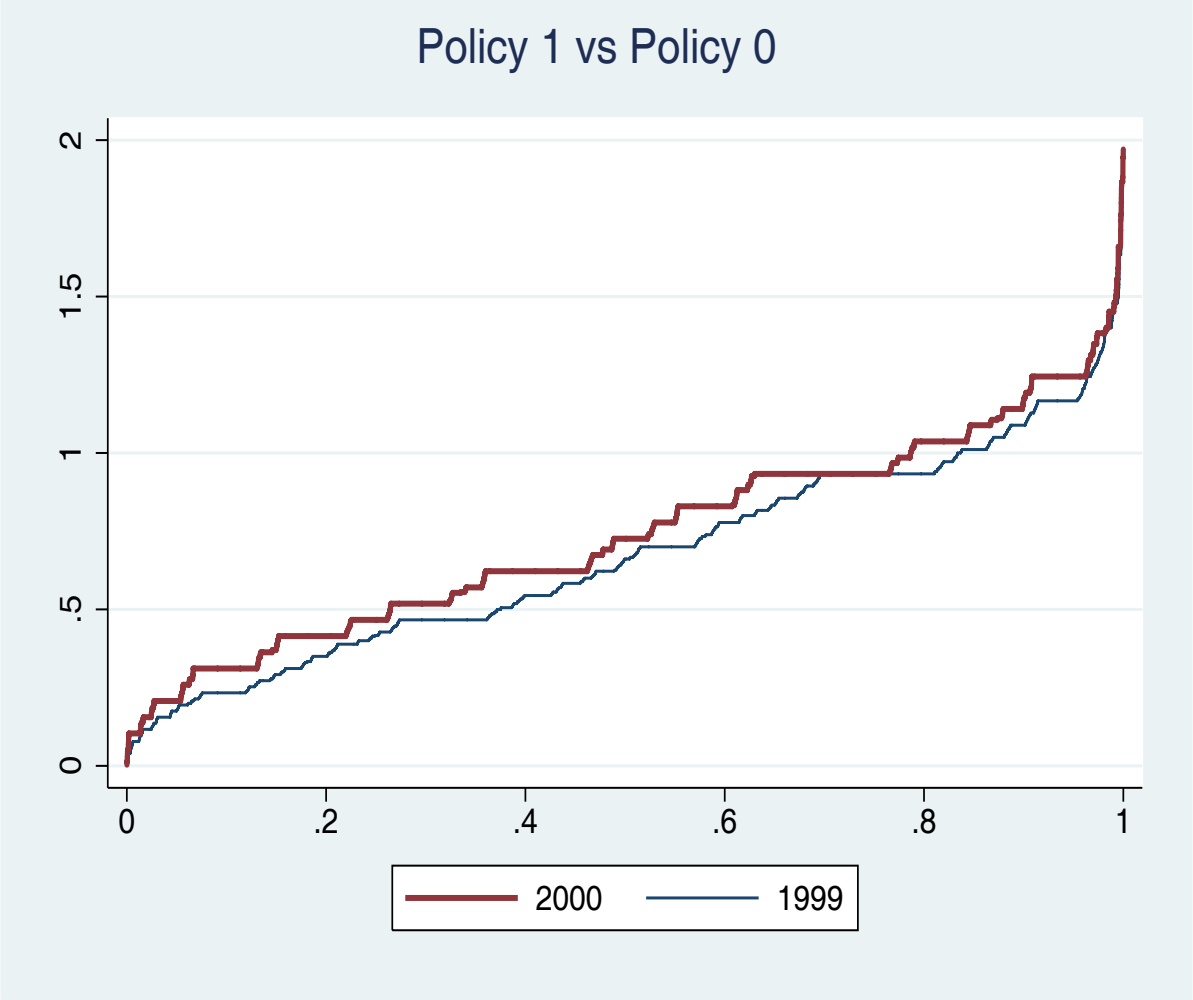

Figure 4: Quantiles of differences in compliance and difference of quantiles of compliance by period.

2 The statistical problem

Consider a population where, at time τ, someone receives a treatment but someone does not. Let D be the treatment indicator and write the observed outcome in period t as

$$Y_t = (1 - D)Y_t^0 + DY_t^1, \tag{1}$$

where Y_t^0 and Y_t^1 are latent rv's representing the two "potential outcomes " of not receiving and receiving the treatment. Although these potential outcomes are well defined for anyone in the population, only one is ever observed. Because they are never both observed, the individual treatment effect $Z_t = Y_t^1 - Y_t^0$ is unobservable.

The statistical problem is how to use data on pre- and post-treatment outcomes to estimate the differential impact of the treatment on a heterogeneous population. In general terms, this amounts to estimating the distribution of Z_t, or some parameter of this distribution. Two commonly studied parameters are:

- the average treatment effect in the population $\mathrm{E}\, Z_t$,
- the average treatment effect in the subpopulation of the treated $\mathrm{E}(Z_t \mid D = 1)$, or average treatment on the treated effect.

3 Differences-in-differences

The difference-in-difference (DiD) approach is a popular way of estimating the average treatment on the treated effect without the need of relying on instrumental variables (IV) or specifying a fully structural choice model.

Suppose that one has available a balanced panel data covering the treatment date τ. Represent counterfactual outcomes for the ith population unit at time t as

$$Y_{it}^d = \alpha_t^d + U_{it}^d, \qquad d = 0, 1, \qquad i = 1, \ldots, n, \tag{2}$$

where the $U_{it}^d = Y_{it}^d - \alpha_t^d$ are regression errors with zero mean. In principle, α_t may depend on a vector of exogenous variables X_i, but we ignore this to simplify the presentation.

Substituting into the definition of Y_t and rearranging gives the random coefficient model

$$Y_{it} = Y_t^0 + (Y_t^1 - Y_t^0)D = \alpha_t^0 + Z_{it}\, D_i + U_{it}^0, \tag{3}$$

where the rv

$$Z_{it} = Y_{it}^1 - Y_{it}^0 = (\alpha_t^1 - \alpha_t^0) + (U_{it}^1 - U_{it}^0)$$

is the individual treatment effect (ITE).

The OLS estimator

The average treatment effect (ATE) is

$$\beta_t = \mathrm{E}\, Z_{it} = \alpha_t^1 - \alpha_t^0.$$

Substituting in (3) gives

$$Y_{it} = \alpha_t^0 + \beta_t D_i + V_{it},$$

where

$$V_{it} = U_{it}^0 + D_i(U_{it}^1 - U_{it}^0).$$

There is selection on unobservables if D_i is correlated with V_{it}, that is, with U_{it}^0 or U_{it}^1.

If there is no selection on unobservables, then β_t is consistently estimated by the slope coefficient $\hat{\beta}_t$ from an OLS regression of Y_{it} on a constant and D_i. This estimator corresponds to the difference $\bar{Y}_t^1 - \bar{Y}_t^0$ between the sample means of Y_{it} for the two treatment groups.

If there is selection on unobservables, then

$$\hat{\beta}_t \xrightarrow{\mathrm{P}} \beta_t + \mathrm{E}(U_t^1 \mid D = 1) - \mathrm{E}(U_t^0 \mid D = 0),$$

and so the OLS estimator $\hat{\beta}_t$ is inconsistent for β_t.

The DiD estimator

Consider the following set of assumptions:

A.1. $U_{it}^0 = \phi_i + \theta_t + \epsilon_{it}$ (3-error compoentn structure of U_{it}^d).

A.2. ϵ_{it} is mean independent of D_i.

A.3. $Y_t^0 = Y_t^1 = Y_t$ for $t < \tau$ (outcomes in the pre-treatment periods $t < \tau$ are independent of treatment).

From (3)

$$\mathrm{E}(Y_{it} \,|\, D_i) = \alpha_t^0 + \mathrm{E}(Z_{it}\, D_i \,|\, D_i) + \mathrm{E}(U_{it}^0 \,|\, D_i),$$

where, under A.1–A.2,

$$\mathrm{E}(U_{it}^0 \,|\, D_i) = \mathrm{E}(\phi_i \,|\, D_i) + \theta_t.$$

Therefore, under A.1–A.3,

$$\mu_t^0 = \mathrm{E}(Y_{it} \,|\, D_i = 0) = \alpha_t^0 + \mathrm{E}(\phi_i \,|\, D_i = 0) + \theta_t,$$

for all t, while

$$\mu_t^1 = \mathrm{E}(Y_{it} \,|\, D_i = 1) = \begin{cases} \alpha_t^0 + \delta_t + \mathrm{E}(\phi_i \,|\, D_i = 1) + \theta_t, & \text{if } t \geq \tau, \\ \alpha_t^0 + \mathrm{E}(\phi_i \,|\, D_i = 1) + \theta_t, & \text{if } t < \tau, \end{cases}$$

where $\delta_t = \mathrm{E}(Z_{it} \,|\, D_i = 1)$ is the average treatment on the treated effect (TTE).

Thus, for any $t_0 < \tau < t_1$, we have

$$\delta_{t_1} = (\mu_{t_1}^1 - \mu_{t_1}^0) - (\mu_{t_0}^1 - \mu_{t_0}^0) = (\mu_{t_1}^1 - \mu_{t_0}^1) - (\mu_{t_1}^0 - \mu_{t_0}^0).$$

A consistent estimator of δ_{t_1} is therefore the DiD estimator

$$\hat{\delta}_{t_1} = (\bar{Y}_{t_1}^1 - \bar{Y}_{t_1}^0) - (\bar{Y}_{t_0}^1 - \bar{Y}_{t_0}^0) = (\bar{Y}_{t_1}^1 - \bar{Y}_{t_0}^1) - (\bar{Y}_{t_1}^0 - \bar{Y}_{t_0}^0),$$

which is also the slope coefficient in an OLS regression of $\Delta Y_i = Y_{it_1} - Y_{it_0}$ on a constant and D_i.

In fact, constructing $\hat{\delta}_{t-1}$ does not require panel data but only two repeated cross-sections, one before and one after τ. In this case, however, we must rule out systematic composition changes within each group.

Failures of A.1–A.3 may cause inconsistency of the DiD estimator:

- A.1 may fail if time-effects have a differential impact on the two treatment groups (Bell, Blundell, Van Reenen 1999).
- A.2 may fail if unobserved temporary and individual-specific effects influence the participation decision (Ashenfelter 1978). In this case,

$$\hat{\delta}_{t_1} \xrightarrow{\mathrm{p}} \delta_{t_1} + \mathrm{E}(\epsilon_{it_1} - \epsilon_{it_0} \,|\, D_i = 1) - \mathrm{E}(\epsilon_{it_1} - \epsilon_{it_0} \,|\, D_i = 0).$$

- A.3 may fail if the treatment is anticipated.

How can we use our knowledge of the TTE parameter? Suppose that we know $\{(\mu_t^0, \mu_t^1)\}$ and $\pi = \Pr\{D_i = 1\}$, and assume that:

- social welfare W_t is the average of individual welfare levels,
- the welfare level of individual i at time t is equal to Y_{it}.

Then social welfare at time t is $W_t = (1-\pi)\mu_t^0 + \pi\mu_t^1$, and the welfare difference between t_0 and t_1 is

$$W_{t_1} - W_{t_0} = (\mu_{t_1}^0 - \mu_{t_0}^0) + \pi[(\mu_{t_1}^1 - \mu_{t_0}^1) - (\mu_{t_1}^0 - \mu_{t_0}^0)] = (\mu_{t_1}^0 - \mu_{t_0}^0) + \pi\,\delta_{t_1},$$

where $\pi\,\delta_{t_1}$ measures the change in welfare due purely to treatment.

4 Troubles with averages

First, sample means and estimated OLS coefficients may be non-robust. Second, the DiD coefficient only tells us whether a treatment is effective relative to the alternative of no-treatment. Whether the treatment is also desirable is a completely different story.

Example 1 Suppose that $\mu^1_{t_1} > \mu^1_{t_0} \geq \mu^0_{t_0} = \mu^0_{t_0}$. In this case, $\delta_{t_1} = \mu^1_{t_1} - \mu^1_{t_0} > 0$. Although the treatment would be Pareto improving, between-group variance in the outcome would be higher after the treatment. If there is no change in within-group variance, total variance in the population would be higher. Would this be acceptable to a decision maker?

Heckman, Smith and Clements (1997) argue that "using the mean impact to evaluate a program rests on two key assumptions: (a) that increases in total output increases welfare; and (b) that undesirable distributional aspects of programmes are either unimportant or are offset by transfers governed by a social welfare function [...] Both of these assumptions are strong. Many programmes produce output that cannot easily be redistributed [...] Programme outputs cannot always be valued and summed to produce a measure of total welfare. Appeal to a mythical social welfare function begs fundamental questions of political economy."

They also argue that "many interesting evaluation questions require knowledge of feature of the distribution of programme gains other than the mean." Besides the mean of Z_t, they list:

- the proportion of people taking the program who benefit from it, $\Pr\{Z_t > 0 \,|\, D = 1\}$;
- the proportion of the population benefiting from the program, $\Pr\{Z_t > 0 \,|\, D = 1)\ \Pr\{D = 1\}$;
- selected quantiles of the impact distribution, $\inf_z\{z : F_t(z \,|\, D = 1) > p\}$, where F_t is the df of Z_t;
- the distribution of gains at selected base state values, $\Pr\{Z_t \leq z \,|\, D = 1, Y^0_t = y_0\}$.

5 Varieties of quantile treatment effects

Abadie, Angrist and Imbens (2002)

Abadie, Angrist and Imbens (2002) advocate the use of quantile treatment effects in order to properly take into account "the effects of policy variables on distributional outcomes beyond simple averages".

They define the quantile-treatment effect (QTE) as the difference in the quantiles of Y^1 and Y^0 for the subpopulation of "compliers", namely those whose potential outcomes are independent of treatment status.

Specifically, write the treatment indicator as

$$D = (1 - W)D^0 + WD^1,$$

where the latent binary indicators D^1 and D^0 are two potential treatment states, corresponding respectively to applying or not applying the binary 0-1 instrument W. The "compliers" are those for whom $D^1 > D^0$ (those whose treatment status can be manipulated by the instrument).

Let $Q^D(p \mid D^1 > D^0)$ denote the pth quantile of observed outcome conditional on D for "compliers". Their key assumption is that

$$Q^D(p \mid D^1 > D^0) = \alpha_p D + \beta_p X.$$

This assumption simply defines the QTE as the difference

$$\text{QTE}(p) = Q^1(p \mid D^1 > D^0) - Q^0(p \mid D^1 > D^0) = \alpha_p.$$

If we could identify the subpopulation of "compliers", then an estimator of (α_p, β_p) may be obtained by standard asymmetric LAD. Although the "compliers" are not identifiable, they show that a $\sqrt{n}$-consistent and asymptotically normal estimator of (α_p, β_p) may be obtained by a weighted asymmetric LAD problem, with weights that must be estimated in a preliminary step.

Chernozhukov and Hansen (2005)

They define QTE, more directly, as $\text{QTE}(p) = Q^1(p) - Q^0(p)$, where Q^d is the quantile function of potential outcome Y^d, $d = 0, 1$. The link with ATE is immediate for, if Z has finite mean, then

$$\text{E}\, Z = \text{E}\, Y^1 - \text{E}\, Y^0 = \int_0^1 [Q^1(p) - Q^0(p)]\, dp = \int_0^1 \text{QTE}(p)\, dp.$$

In the absence of selection, their definition and that in Abadie, Angrist and Imbens (2002) coincide.

The origin of this definition is Doksum (1974). He considers a rv X, with a continuous df F, and a treatment that shifts X by a random amount to a new rv Y, with df G. He shows that this shift may be characterized through the "shift function"

$$\Delta(x) = G^{-1}(F(x)) - x = Q_Y(F(x)) - Q_X(F(x)),$$

where $Q_Y = G^{-1}$ and $Q_X = F^{-1}$ are the quantile functions of Y and X respectively. Thus $\Delta(x)$ is just the QTE at the quantile $p = F(x)$. This relates the QTE to the concept of stochastic dominance, for $\text{QTE}(p) > 0$ for all p if G first-order stochastically dominates F.

Chernozhukov and Hansen (2005) provide conditions under which the quantiles of potential outcomes can be identified and estimated through the contional moment restriction

$$0 = p - \Pr\{Y \le Q^D(p) \,|\, W\} = p - \text{E}\, 1\{Y \le Q^D(p) \,|\, W\},$$

where W is an instrument that affects D but is independent of Y^d.

Their paper makes extensive use of the representation of a rv as the quantile-transform of a uniform rv

$$Y^d = Q^d(U^d), \qquad U^d \sim \mathcal{U}(0, 1).$$

They refer to U^d as the "rank variable", as it determines the relative ranking in terms of potential outcomes.

Their result depends crucially on a condition that, in its simplest form requires $U^d = U$ for all d. This "rank invariance" condition implies that a single unobserved factor U determines the ranking of an individual across all treatment states: "people who are strong (highly ranked) earners without a training program, remain strong earners having done the training".

Firpo (2006)

The setup is the same as in Chernozhukov and Hansen(2005), but selection is assumed to depend only on exogeneous covariates X (exogeneous selection).

He argues that, although the covariates X are important in order to control for selection, interest is typically on unconditional quantities. For mean like objects, such as ATE, this is not a problem because

$$\text{ATE} = \int \text{ATE}(x)\, dH(x),$$

where $\text{ATE}(x)$ is the ATE conditional on $X = x$ and H is the df of X. Unfortunately, integrating over the distribution of covariates, does not recover the marginal QTE from the conditional QTE.

He shows how to directly estimate the marginal QTE, and the marginal quantile treatment effect on the treated (QTT)

$$\text{QTT}(p) = Q^1(p \mid D = 1) - Q^0(p \mid D = 1),$$

without estimating the corresponding conditional quantiles.

Differences of quantiles or quantiles of differences?

Are the above sensible ways of evaluating the effect of a treatment?

Abadie, Angrist and Imbens (2002) argue that quantiles of $Z = Y^1 - Y^0$ may also be of interest, but "we focus on the marginal distributions of potential outcomes because identification of the distribution of $Y_1 - Y_0$ [$Y^1 - Y^0$ in our notation] requires much stronger assumptions and because economists making social welfare comparisons typically use differences in distributions and not the distribution of differences for this purposes".

The second argument is not very convincing. For example, Cowell (2000) argues that if one is unable to observe pre- and post-treatment outcomes for the same individual, then the best that one can do is simply to compare distributions before and after the treatment. This may no longer be true if panel data are available.

As for the first argument, Heckman, Smith and Clements (1997) state that "from the two marginal distributions for participants and nonparticipants, it generally not possible to estimate the joint distribution of outcomes and so it is generally not possible to estimate the distribution of impacts or its median".

Drop the t suffix for simplicity, and assume that the df F of Z is continuous and strictly increasing. Then, the pth quantile of Z, with $0 < p < 1$, is the unique root $Q(p)$ of the equation $F(z) = p$. If Z has finite mean, then $\mathrm{E}\, Z = \int_0^1 Q(p)\, dp$.

In general, $\Delta(q) = Q^1(p) - Q^0(p) \neq Q(p)$ although, if Y^0 and Y^1 both have finite mean, then

$$\int_0^1 \Delta(p)\, dp = \int_0^1 Q(p)\, dp,$$

Under what conditions do we have that $\Delta(p) = Q(p)$ or, at least, $\Delta(p) > 0$ implies $Q(p) > 0$?

Example 2 Supppose that

$$Y^1 = \mu + \sigma Y^0, \qquad 0 < \sigma < \infty.$$

Then

$$\Delta(p) = \mu + (\sigma - 1)Q^0(p) = Q(p).$$

This condition is very strong, for it implies a deterministic relationship between Y^1 and Y^0.

Special cases:

- if $\sigma = 1$, that is $Y^1 = \mu + Y^0$ (homogeneous treatment effects model), then $Q(p) = \mu$ for all p;
- if $\mu = 0$, that is $Y^1 = \sigma Y^0$ (increasing treatment effects model or homogeneous treatment effects model on the log scale), then $Q(p) = (\sigma - 1)Q^0(p)$.

Example 3 If we confine attention to medians, then another example is the general bivariate normal distribution. Because the difference $Y^1 - Y^0$ is also normal, we have that $\Delta(.5) = Q(.5)$. This result does not carry over to quantiles different from the median. □

6 A characterization

Lee (2000) provides necessary and sufficient conditions.

Consider the following two properties:

M0: The difference of the medians is equal to the median of the difference.

M1: A positive difference in the medians implies that the median of the difference is positive.

Let $\xi^d = Q^d(.5)$ be the median of Y^d $(d = 0, 1)$, let $W^d = Y^d - \xi^d$ be the deviation of Y^d from its median, and let $\Delta = \xi^1 - \xi^0$ be the difference of the two medians.

Theorem 1 *Suppose that the joint density of* (W^0, W^1) *is continuous, and is positive in an open neighborhood of the origin. Then:*

(a) Property M0 holds if and only if "equal probability of octants" holds, that is,

$$\Pr\{W^0 < W^1 < 0\} = \Pr\{0 < W^1 < W^0\}$$

or, equivalently,

$$\Pr\{W^1 < W^0 < 0\} = \Pr\{0 < W^0 < W^1\}.$$

(b) Property M1 holds if and only if "weak monotonicity" holds, that is,

$$\Pr\{W^0 - \Delta < W^1 < 0\} > \Pr\{0 < W^1 < W^0 - \Delta\}$$

or, equivalently,

$$\Pr\{W^1 < W^0 - \Delta,\, W^0 < 0\} > \Pr\{0 < W^0,\, W^0 - \Delta < W^1\}.$$

Example 4 In addition to the homogeneous and the increasing treatment effects models, three other examples when properties M0 and M1 hold are:

- $\Pr\{Z \geq 0\} = 1$ (monotone treatment effects model, Manski 1997),
- the joint density of W^0 and W^1 is symmetric, that is, $f(w_0, w_1) = f(-w_0, -w_1)$;
- W^0 and W^1 are exchangeable, that is, $f(w_0, w_1) = f(w_1, w_0)$. □

Now consider, more generally, the following two properties:

Q0: $\Delta(q) = Q(p)$ or, equivalently, $\Pr\{Z > 0\} \leq 1 - p$.

Q1: $\Delta(p) > 0$ implies $Q(p) > 0$ or equivalently $\Pr\{Z > 0\} > 1 - p$.

Theorem 2 *Let $W_p^d = Y^d - Q^d(p)$, $d = 0, 1$, $0 < p < 1$, and suppose that the joint density of (W_p^0, W_p^1) is continuous, and is positive in an open neighborhood of the origin. Then*

(a) Property Q0 holds if and only if

$$\Pr\{W_p^0 < W_p^1 < 0\} = \Pr\{0 < W_p^1 < W_p^0\}$$

or, equivalently,

$$\Pr\{W_p^1 < W_p^0 < 0\} = \Pr\{0 < W_p^0 < W_p^1\} + 2p - 1.$$

(b) Property Q1 holds if and only if

$$\Pr\{W_p^0 - \Delta(p) < W_p^1 < 0\} > \Pr\{0 < W_p^1 < W_p^0 - \Delta(p)\}$$

or, equivalently,

$$\Pr\{W_p^1 < W_p^0 - \Delta(p),\, W_p^0 < 0\} < \Pr\{0 < W_p^0,\, W_p^0 - \Delta(p) < W_p^1\} + 2p - 1.$$

(c) $\Delta(p) = 0$ implies $Q(1-p) = 0$ if and only if

$$\Pr\{W_p^1 < W_p^0 < 0\} = \Pr\{0 < W_p^0 < W_p^1\}.$$

(d) $\Delta(p) > 0$ implies $Q(1-p) < 0$ if and only if

$$\Pr\{W_p^1 < W_p^0 - \Delta(p),\, W_p^0 < 0\} > \Pr\{0 < W_p^0,\, W_p^0 - \Delta(p) < W_p^1\}.$$

These conditions are not easy to interpret. Lee (2000) adds a third result, with the bivariate distribution centered at the bivariate median (ξ^0, ξ^1) instead of $(Q^0(p), Q^1(p))$, but concludes that it "does not seem to be easier to interpret than theorem 2 is".

7 Conclusions

- Evaluating programs using only mean impacts may be misleading.
- Comparisons of marginal distributions of program gains may also be misleading.
- If panel data are available, one should focus on quantiles (or distribution functions) of differences.

References

Abadie A., Angrist J., Imbens G. (2002), "Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings", *Econometrica*, 70, 91–117.

Angrist J.D. (2004), "Treatment effect heterogeneity in theory and practice", *Economic Journal*, 114: C52–C83.

Ashenfelter O. (1978), "Estimating the effects of training programs on earnings", *Review of Economics and Statistics*, 60: 47–57.

Bell B., Blundell R., Van Reenen J. (1999), "Getting the unemployed back to work: An evaluation of the New Deal proposals", *International Tax and Public Finance*, 6: 339–360.

Blundell R., Costa Dias M. (2002), "Alternative approaches to evaluation in empirical microeconomics", CEMMAP Working Paper CWP 10/02.

Chernozhukov V., Hansen C. (2005), "An IV model of quantile treatment effects", *Econometrica*, 73: 245–261.

Cowell F. (2000), "Measurement of inequality", in A.B. Atkinson and F. Bourguignon (eds.), *Handbook of Income Distribution*, Elsevier, North Holland, 87–166.

Doksum, K. (1974), "Empirical probability plots and statistical inference for nonlinear models in the two sample case", *The Annals of Statistics*, 2: 267–277.

Firpo S. (2006), "Efficient semiparametric estimation of quantile treatment effects", *Econometrica*, fortcoming.

Heckman J., Smith J., Clements N. (1997), Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts, *Review of Economic Studies*, 64: 487–535.

Lee M. (2000), "Median treatment effect in randomized trials", *Journal of the Royal Statistical Society, Series B*, 62: 595–604.

Manski (1997), "Monotone treatment response", *Econometrica*, 65: 1311–1334.