

Ok, so far we have focused our attention on mean treatment parameters.

There may be many other questions we care about that cannot be answered within the framework we have used so far.

Let (Y_0, Y_1) denote potential outcomes in state $S = 0$ and $S = 1$, respectively within a given policy regime.

Each person has a (Y_0, Y_1) pair.

We assume that (Y_0, Y_1) have finite means and can be expressed in terms of conditioning variables X in the following manner:

$$Y_0 = \mu_0(X) + U_0$$

$$Y_1 = \mu_1(X) + U_1$$

Where $E(Y_j|X) = \mu_j(X)$ and $E(U_j|X) = 0$ for $j = 0, 1$.

The gain for an individual who moves from the $S = 0$ to $S = 1$ within a policy regime is

$$\Delta = Y_1 - Y_0.$$

An evaluation problem within a policy regime arises because we do not observe the pair (Y_0, Y_1) for anybody.

This is a missing data problem. The solution to this missing data problem is to construct counterfactuals: what is the Y_0 outcome for an individual who chose $S = 1$ if she had chosen $S = 0$.

Let's think of two problems we face when we want to go beyond the mean effect of treatment question.

First, if the gains (Δ) vary across agents, there is no single number that summarizes the distribution of gains for all purposes of policy evaluation.

For each specific policy question we want to address, we must carefully define the parameter of interest.

In general, the average gain of those who are in the program (TT) is not the relevant parameter of interest.

Consider an example from the economics of education: if we want to determine the gains of a policy that reduces tuition, we need to know (a) how many entrants into education will be induced by the tuition policy, and (b) from where in the distribution of gains to schooling ($Y_1 - Y_0$) the new entrants are coming.

Given (a) and (b), we can compute aggregate gains from the tuition policy.

We can then check whether the return of the marginal entrant is above or below that of the typical person enrolled in the program.

For other problems of distributional analysis, it is of interest to determine where in an initial distribution beneficiaries come from and where they end up in the treatment outcome distribution.

The second problem is an econometric one. Once we have defined the parameter of interest, say the gain to the median voter, how can we estimate it?

If we wish to avoid special assumptions like statistical independence between Y_0 and Y_1 or perfect dependence, the solution is to recover the joint distribution of (Y_0, Y_1) .

Once we know this distribution, it is possible to calculate the distribution of $(Y_1 - Y_0)$ for any group of people we are interested in and obtain its median or any other quantile. We may compute several measures of interest from this distribution.

The proportion of people taking schooling that benefit from it in terms of gross returns $\Delta (= Y_1 - Y_0)$ is $\Pr(\Delta > 0 \mid S = 1)$. This parameter is one way to measure how widely program gains are distributed among participants.

The proportion of the total population benefiting from participating in schooling is $\Pr(\Delta > 0 \mid S = 1) \cdot \Pr(S = 1)$. It is of interest to determine how many people in society at large benefit (in the sense of $Y_1 - Y_0$ gains) from participating in schooling.

The distribution of gains from schooling for agents who are at selected base state values is $\Pr(\Delta \leq a \mid S = 1, Y_0 = y_0)$. This measure interests Rawlsian evaluators who seek to determine the impact of schooling on recipients in the lower tail of the base state distribution.

The increase in the level of outcomes above a certain threshold, say the poverty line \bar{y} , due to schooling is $\Pr(Y_1 > \bar{y} | S = 1) - \Pr(Y_0 > \bar{y} | S = 1)$. This is a parameter that describes how the distribution of the outcomes for the participants compares to the distribution of the outcomes for the same agents if they had not participated in schooling.

We can also form measures for people affected by a specific policy.

Let A and B denote two policy states, say a high tuition and a low tuition policy, respectively. The proportion of people who benefit from a policy that induces them into schooling (e.g., a reduction in tuition) is $\Pr(\Delta > 0 \mid S_A = 0, S_B = 1)$, where the measure of benefit is a gross gain measure and S^A and S^B are choice indicators under policy A and B , respectively.

We can also measure the proportion of the total population that benefits from the policy: $\Pr(\Delta > 0 \mid S_A = 0, S_B = 1) \cdot \Pr(S_A = 0, S_B = 1)$.

So, how do we solve the problem of constructing counterfactuals? By identifying the joint distribution of (Y_1^A, Y_0^A) and the potential outcomes under policy A , conditional on S .

The possibility we are going to study is using a factor structure model. These models generalize the LISREL models of Jöreskog (1977) and the MIMIC model of Goldberger and Jöreskog (1975) to produce counterfactual distributions.

Before getting into the methodology, a brief digression on the types of policies we evaluate.

It is fruitful to distinguish between two kinds of policies: (a) those that affect potential outcomes (Y_0^A, Y_1^A) under policy regime A through price and quality effects and (b) those that affect sectorial choices (through I^A), but do not affect potential outcomes.

Tuition and educational access policies that do not produce general equilibrium effects fall into the second category of policy.

It is the second kind of policy that receives the most attention in empirical work.

How does the factor structure solve the problem of recovering joint distributions?

Let's impose assumptions then. Assume that

$$U_j = \alpha_j \theta + \varepsilon_j \text{ for } j = 0, 1,$$

where θ is a vector of mutually independent "factors". These factors θ are assumed independent of the uniquenesses $\varepsilon_0, \varepsilon_1$ which are also assumed independent of each other. Although at this point it is not required assume also that they are all independent of X .

Now, if we can somehow know the distribution of θ, ε_0 and ε_1 then

$$f(Y_0, Y_1) = \int_{\Theta} f(Y_0, Y_1 | \theta) f(\theta) d\theta$$

Notice that, conditional on θ , Y_1 and Y_0 are independent by assumption so

$$= \int_{\Theta} f(Y_0 | \theta) f(Y_1 | \theta) f(\theta) d\theta.$$

But, since we assumed we know $f(\varepsilon_0)$ and $f(\varepsilon_1)$ we know the conditional marginal distributions of Y_0 and Y_1 !

There it is, we just solved the problem of obtaining joint distributions.

Of course we had to bring some (strong!) assumptions and we haven't seen how to (if at all possible) recover any of these things but here at least is a possibility for getting joint distributions without imposing perfect ranking or independence.

Recovering Marginals

A generalized Roy model would write index I as a net utility

$$I = \mu_I(X, Z) + U_I \quad (1)$$

where the Z are observed (by the analyst) determinants of utility and U_I denotes unobserved determinants from the point of view of the analyst.

We write

$$S = 1 \quad \text{if } I \geq 0; \quad S = 0 \text{ otherwise.} \quad (2)$$

Thus if the net utility of state 1 is positive, $S = 1$ is chosen.

This model for (Y_1, Y_0, S) is sufficiently rich to serve our purposes.

Identifying Counterfactuals using Independent Factors

Ok, so let's see if we can actually recover the things we assumed known (and how flexible we can be on those).

Identifying the joint distribution of potential outcomes is a difficult problem because we do not observe both components of (Y_0, Y_1) for anyone except in special panel data situations (see Heckman and Smith, 1998).

Thus, one cannot in general directly form the joint distribution of potential outcomes (Y_0, Y_1) .

Under the assumptions that (Z, X) are statistically independent from (U_0, U_1, U_I) , $\mu_I(X, Z)$ is a nontrivial function of Z given X , and full support on $\mu_0(X)$, $\mu_1(X)$ and $\mu_I(X, Z)$, and an assumption that the elements of the pairs $(\mu_0(X), \mu_I(X, Z))$ and $(\mu_1(X), \mu_I(X, Z))$ can be varied independently of each other, then one can identify the joint distributions of $(U_0, \frac{U_I}{\sigma_I})$ and $(U_1, \frac{U_I}{\sigma_I})$ and also $\mu_0(X)$, $\mu_1(X)$, and $\frac{\mu_I(X, Z)}{\sigma_I}$.

Thus, one can identify the joint distributions of (Y_0, I^*) and (Y_1, I^*) given X and Z where $I^* = I/\sigma_I$. One cannot recover the conditional (on X, Z) joint distribution of (Y_0, Y_1) or (Y_0, Y_1, I^*) without further assumptions.

The thrust of the argument is that under the stated conditions, we can identify the distribution of I up to a factor of proportionality, σ_I .

We can also identify $F(Y_0, I|I < 0, X, Z) = F(Y_0|D = 0, X, Z) \Pr(D = 0|X, Z)$ and $F(Y_1, I|I \geq 0, X, Z) = \Pr(Y_1|D = 1, X, Z) \Pr(D = 1|X, Z)$.

By varying X, Z we can trace out the full distributions of $F(Y_0, I)$ and $F(Y_1, I)$ respectively.

With these distributions in hand, we can perform conventional factor analysis on (Y_0, I^*) and (Y_1, I^*) because, effectively, we observe these two distributions.

For simplicity focus on the case where θ is a scalar.

First case

Assume we can also write

$$U_I = \alpha_I^* \theta + \varepsilon_I$$

with θ independent of ε_I and ε_i independent of $\varepsilon_0, \varepsilon_1$.

We just “described” how one can identify $F(U_0, U_{I^*})$ and $F(U_1, U_{I^*})$.

From these distributions one can identify the left hand sides of the following two equations:

$$\begin{aligned}Cov(U_0, U_{I^*}) &= \alpha_0 \alpha_{I^*} \sigma_\theta^2 \\Cov(U_1, U_{I^*}) &= \alpha_1 \alpha_{I^*} \sigma_\theta^2.\end{aligned}$$

As previously noted, the scale of the unobserved I is normalized, a standard condition for discrete choice models.

A second normalization that we need to impose is that $\sigma_\theta^2 = 1$. This is required since the factor is not observed and we must set its scale. That is, since $\alpha\theta = k\alpha\frac{\theta}{k}$ for any constant k , we need to set the scale by, say, normalizing the variance of θ .

We could alternatively normalize some α_0 or α_1 to one.

Finally, if we set $\alpha_{I^*} = 1$ (something we can relax, as noted below and in the next section), then we identify α_1 and α_0 from the known covariances above. Since

$$Cov(U_1, U_0) = \alpha_1 \alpha_0 \sigma_\theta^2$$

we can identify the covariance between Y_1 and Y_0 even though we do not observe both Y_0 and Y_1 for anyone.

We then use the variances $Var(U_1), Var(U_0)$ and the normalization $Var(U_{I^*}) = 1$ to recover the variances of the uniquenesses $\sigma_{\varepsilon_0}^2, \sigma_{\varepsilon_1}^2, \sigma_{\varepsilon_{I^*}}^2$.

The fact that we needed to normalize both $\sigma_\theta^2 = 1$ and $\alpha_{I^*} = 1$ is a consequence of our assumption that we have only one observation for Y_1 or Y_0 for each person.

If we have access to more observations (say from panel data) or to more equations that depend on the factor (as in the next section), we can relax the normalizations, say $\sigma_\theta^2 = 1$, since then we could form for a panel of length T , the left hand sides of the following equations:

$$\frac{Cov(Y_{1,t'}, I^*)}{Cov(Y_{1,t'}, Y_{1,t})} = \alpha_{1,t}, \quad t = 1, \dots, T$$

$$\frac{Cov(Y_{0,t'}, I^*)}{Cov(Y_{0,t'}, Y_{0,t})} = \alpha_{0,t}, \quad t = 1, \dots, T$$

and recover σ_θ^2 from, say, $Cov(Y_{1,t}, I^*) = \alpha_{1,t}\sigma_\theta^2$, given the normalization $\sigma_{I^*}^2 = 1$. The variances of the uniquenesses follow as before.

The crucial idea motivating this identification strategy is that even though we never observe (Y_0, Y_1) as a pair, both Y_0 and Y_1 are linked to S through the choice equation.

From information on choice S we can recover I^* from a standard identification argument in econometrics. Thus, we essentially observe (Y_0, I^*) and (Y_1, I^*) .

The common low dimensional dependence of Y_0 and Y_1 on I^* secures identification of the joint distribution of Y_0, Y_1, I^* .

Second Case

Alternative strategy based on the same idea where in addition to a choice equation, we have a measurement equation observed for all observations whether or not Y_1 or Y_0 is observed. This plays the role of I^* and in certain respects identification with a measurement is more transparent and more traditional.

In educational statistics, a test score is often used to proxy ability. Suppose that the analyst has access to one ability test M for each person. Measured ability M is

$$M = \mu_M(X) + U_M.$$

We can estimate $\mu_M(X)$ by standard methods and we can form the residual from this equation.

Assume that the residual has a factor structure

$$U_M = \alpha_M \theta + \varepsilon_M,$$

where ε_M is mutually independent from $(\varepsilon_0, \varepsilon_1, \varepsilon_I)$ and θ .

We can, in addition to the covariances presented before, determine the left-hand sides

$$\begin{aligned} \text{Cov}(U_M, U_0) &= \alpha_M \alpha_0 \sigma_\theta^2 \\ \text{Cov}(U_M, U_1) &= \alpha_M \alpha_1 \sigma_\theta^2 \\ \text{Cov}(U_M, U_{I^*}) &= \alpha_M \alpha_{I^*} \sigma_\theta^2. \end{aligned}$$

These are obtained, respectively, from correlations of the residuals of U_M with the residuals from (selection corrected) Y_0 :

$$Y_0 - \mu_0(X) = U_0,$$

from correlations of U_M with the residuals from (selection corrected) Y_1 :

$$Y_1 - \mu_1(X) = U_1,$$

as well as the residuals of U_M with the residuals of I^* using discrete choice analysis.

If we impose the normalization $\alpha_M = 1$, which can be interpreted as requiring that higher ability signals a higher level of factor θ , we can form the ratio

$$\frac{Cov(U_0, U_{I^*})}{Cov(U_M, U_{I^*})} = \alpha_0$$

and identify α_0 . In a similar fashion,

$$\frac{Cov(U_1, U_{I^*})}{Cov(U_M, U_{I^*})} = \alpha_1$$

and we recover α_1 . Now, from

$$\text{Cov}(U_M, U_0) = \alpha_0 \sigma_\theta^2,$$

we can obtain σ_θ^2 . Finally, we can identify α_{I^*} based on information from

$$\text{Cov}(U_M, U_{I^*}) = \alpha_{I^*} \sigma_\theta^2,$$

so we can obtain α_{I^*} up to scale.

Thus, with one measurement, one choice equation and two outcomes we can identify σ_θ^2 and α_{I^*} up to scale. We can use the identified variances $\text{Var}(U_0)$, $\text{Var}(U_1)$, $\text{Var}(U_{I^*}) = 1$, and $\text{Var}(U_M)$ to recover the variance of the uniquenesses $\sigma_{\varepsilon_0}^2$, $\sigma_{\varepsilon_1}^2$, $\sigma_{\varepsilon_{I^*}}^2$, and $\sigma_{\varepsilon_M}^2$. Thus, having access to a measurement (M) and choice data, allows us to estimate the covariances among the outcomes across the two counterfactual states.

The measurements can replace the choice equation provided that the analyst surmounts the selection problem that Y_0 is observed only if $S = 0$ and Y_1 is observed only if $S = 1$.

Remaining is the problem of identifying the distributions of the unobservables. Traditional factor analysis assumes normality. We present a more general nonparametric analysis.

Let's start stating a theorem (without proof) due to Kotlarski (1967):

Suppose that we have two random variables T_1 and T_2 that satisfy:

$$T_1 = \eta + v_1$$

$$T_2 = \eta + v_2$$

with η, v_1, v_2 mutually statistically independent, $E(\eta) < \infty$, $E(v_1) = E(v_2) = 0$; that the conditions for Fubini's theorem are satisfied for each random variable, and the random variables possess nonvanishing characteristic functions, then the densities $f(\eta), f(v_1), f(v_2)$ are identified (nonparametrically).

Applied to the current problem, we have a choice equation, two outcome equations and a measurement equation.

Assume that we normalize $\alpha_M = 1$. As a consequence of this assumption and the analysis of the preceding subsection, all factor loadings, factor variance and the variances of the uniquenesses are known.

The system is

$$I^* = \mu_{I^*}(X, Z) + \alpha_{I^*}\theta + \varepsilon_{I^*}$$

$$Y_0 = \mu_0(X) + \alpha_0\theta + \varepsilon_0$$

$$Y_1 = \mu_1(X) + \alpha_1\theta + \varepsilon_1$$

$$M = \mu_M(X) + \theta + \varepsilon_M.$$

This system can be rewritten as

$$\begin{aligned}\frac{I^* - \mu_{I^*}(X, Z)}{\alpha_{I^*}} &= \theta + \frac{\varepsilon_{I^*}}{\alpha_{I^*}} \\ \frac{Y_0 - \mu_0(X)}{\alpha_0} &= \theta + \frac{\varepsilon_0}{\alpha_0} \\ \frac{Y_1 - \mu_1(X)}{\alpha_1} &= \theta + \frac{\varepsilon_1}{\alpha_1} \\ M - \mu_M(X) &= \theta + \varepsilon_M.\end{aligned}$$

Applying Kotlarski's theorem to any pair of equations, we conclude that we can identify the densities of $\theta, \frac{\varepsilon_{I^*}}{\alpha_{I^*}}, \frac{\varepsilon_0}{\alpha_0}, \frac{\varepsilon_1}{\alpha_1}, \varepsilon_M$.

Since we know α_{I^*}, α_0 and α_1 we can identify the densities of $\theta, \varepsilon_{I^*}, \varepsilon_0, \varepsilon_1, \varepsilon_M$.*

*Recall that U_I is only known up to scale σ_I .

Thus, we can identify the distributions of all of the error terms.

Finally, to recover the joint distribution of (Y_1, Y_0) given X , denoted $F(Y_1, Y_0 | X)$, note that

$$F(Y_1, Y_0 | X) = \int F(Y_1, Y_0 | \theta, X) dF(\theta),$$

where $F(\theta)$ is the distribution of θ . From Kotlarski's theorem, $F(\theta)$ is known.

Because of the factor structure, Y_1, Y_0 and S are independent once we condition on θ . So

$$F(Y_1, Y_0 | \theta, X) = F(Y_1 | \theta, X) F(Y_0 | \theta, X).$$

But $F(Y_1 | \theta, X)$ and $F(Y_0 | \theta, X)$ are identified once we condition on the factors since

$$\begin{aligned} F(Y_1 | \theta, X, S = 1) &= F(Y_1 | \theta, X) \\ F(Y_0 | \theta, X, S = 0) &= F(Y_0 | \theta, X). \end{aligned}$$

Note further that if θ were known to the analyst, this procedure would be matching on θ and X . Our method generalizes matching by allowing the variables that would produce the conditional independence assumed in matching to be unobserved by the analyst.

Over the support of $\mu_I(X, Z)$, $\mu_1(X)$ and $\mu_0(X)$, we can evaluate policies that change Z for each X . We can evaluate new policies that can be expressed as some value of (X, Z) in the historical support. We can extrapolate to new supports by making functional form assumptions *e.g.*, $\mu_1(X) = X\beta_1$, $\mu_0(X) = X\beta_0$ and $\mu_I(X, Z) = (X, Z)\beta_I$.