2021 - 2022

Contents

1 Marginal treatment effects

Instrumental variables (IV) has been used to estimate Local Average Treatment Effects (LATE) but these are only informative about the Treatment Effects (TE) on compliers, idem people whose selection decision into treatment can be shifted by the IV. In many contexts, the LATE is not the target parameter, for instance when we would like to know the TE on the whole population, in this case we have to generalize the LATE to other parts of population who don't comply with the instrument. Or when we want to know whether a given policy is effective on a specific part of population. One way to do this is by using Marginal Treatment Effects (MTE).

1.1 Basics

Set-up we study a binary treatment set-up where D is the treatment variable :

$$D \in \{0, 1\} \tag{1}$$

We have two different potential outcomes which cannot be observed at one time for a given unit :

$$Y_i = D_i Y_{i,1} + (1 - D_i) Y_{i,0} \tag{2}$$

where $Y_{i,1}$ is the potential outcome of unit i if she has taken the treatment and $Y_{i,0}$ otherwise.

Suppose we have an instrument Z that can shift in/out some units into/from treatment, in the literature they say Z affects the likelihood of choosing the treatment. You can see the instrument as a cost shifter which would influence someone decision to get treated or to not take the treatment.

The Roy model the framework for studying the Marginal Treatment Effect is The Roy model. In this model, each individual potential outcome has a component that is common to everyone in the population and a component that is idiosyncratic. We suppose separable

unobserved (heterogeneous component) and observed (common component) determinants of outcomes :

$$Y_{i,1} = \mu_1 + U_{i,1} \tag{3}$$

$$Y_{i,0} = \mu_0 + U_{i,0} \tag{4}$$

with $E(U_{i,j}) = 0$. Therefore, the individual treatment effect has a common component and an idiosyncratic component which differs by individuals :

$$\Delta_i = Y_{i,1} - Y_{i,0} = \mu_1 - \mu_0 + U_{i,1} - U_{i,0} \tag{5}$$

This means that the likelihood of taking treatment will be different between people because the benefit from the treatment differs between them. The model suppose that on average the idiosyncratic difference component is zero, but for some people this difference is positive and means they earn gains from getting treated. The Roy model allows to account for heterogeneity in treatment effects.

The selection equation like in micro theory we want to model the selection choice into treatment and we want to use the instrument Z and an idiosyncratic component as determinants in the selection equation. A way to do this is to consider someone's latent net benefit of choosing the binary treatment :

$$D^* = \alpha + \beta Z_i + V_i \tag{6}$$

α is the common gain from taking the treatment, $\alpha + \beta Z_i$ is the total gain from taking the treatment and this differs between people. Note that we can also consider random coefficient β to account for heterogeneous reactions to instrument variations. $-V_i$ is the unobserved idiosyncratic cost of taking the treatment idem the resistance to treatment. People can have different resistances to treatment because of their different preferences or different cost opportunities, etc. Note that one can add other observed determinants to the equation 6. D^* is unobserved and we can only see whether $D = 1$ or $D = 0$ and the instrument Z. We have $D = 1$ if $D^* \geq 0$ and $D = 0$ otherwise. If the gain from taking the treatment exceeds the cost then $D = 1$.

Let put all things together into a regression framework. By substituting equations 4 and 3 into equation 2 and using equation 5, we obtain the regression equation of the outcome on the dummy of treatment :

$$Y_i = \mu_0 + \Delta_i D_i + U_{i,0} \tag{7}$$

Δ_i here is not identified because we can only observe one potential outcome at a time for each individual. Instead we can know the average treatment effect $E(\Delta_i)$. Recall that $E(U_{i,j}) = 0$ so by using equation 5 we have

$$\Delta_i = E(\Delta_i) + U_{i,1} - U_{i,0} \tag{8}$$

and by substituting it into equation 7 we obtain the regression equation :

$$Y_i = \mu_0 + E(\Delta_i) D_i + \epsilon_i \tag{9}$$

such that $\epsilon_i = U_{i,0} + D_i[U_{i,1} - U_{i,0}]$. We can then identify $E(\Delta_i)$. But we have endogeneity because $cov(\epsilon, D) \neq 0$ by construction of ϵ, therefore the estimates of ATE will be biased : we talk about the selection bias. Selection bias corresponds to the difference between true ATE and the simple mean difference of observable outcomes :

$$E(Y_i/D_i = 1) - E(Y_i/D_i = 0) - E(\Delta_i) = E(U_{i,1}/D_i = 1) - E(U_{i,0}/D_i = 0) \tag{10}$$

If the treatment was randomly assigned and compliance with the assignment was a hundred percent, then the selection bias would be zero because the unobservable outcome U_i of the treated in case of treatment would be the same as the unobservable outcome of the untreated in case of no treatment.

Implications of selection bias in randomized experiments, when compliance is perfect idem $P(D_i = 1/Z_i) = 1$ (Z_i is an instrument used to assign treatment) then there is no selection bias; with imperfect compliance idem $P(D_i = 1/Z_i) < 1$ (in this case people choose or not choose to get treatment even if they receive the invitation), we can only estimate some causal effects like LATE but not the ATE. In observational studies when we don't have randomization of the treatment, we have bias unless we find a valid instrument or any other credible identification strategy. Sometimes even with credible strategy we may not estimate consistently the ATE. One step along the way to solve the problem of selection bias is Heckman selection model.

1.2 Heckman's normal selection model (1976)

In this model we have a regression equation that includes the outcome that we observe Y_i, the treatment D_i, an exogenous variable X_i and an unobserved difference in the outcome U_i that is neither explained by the treatment nor by the exogenous term :

$$Y_i = \mu + \gamma X_i + \Delta D_i + U_i \tag{11}$$

μ is the mean outcome. At first and to simplify let consider that the treatment effect is the same across individuals and then we can relax this assumption in future models. However we consider heterogeneity in the model which comes from the selection equation :

$$D^* = \alpha + \beta Z_i + V_i \tag{12}$$

we can interpret α as the common gain from treatment and $-\beta Z_i - V_i$ as the cost of it.

$$D_i = 1(D_i^* \geq 0) \tag{13}$$

This means that if the difference between gains and costs is greater then zero then the individual chooses the treatment. We assume that the idiosyncratic cost V_i and idiosyncratic determinant of outcome U_i are mean zero and we have correlation between between U_i and V_i ($cov(U_i, V_i) \neq 0$) so that the treatment is not randomly assigned and there are systematic differences between people who choose to get treatment and who don't and these differences affect the way each one sees his outcome after/without treatment and this leads to heterogeneous self selection decisions into treatment (idem endogenous selection into treatment).

Switching Regression model we can split the regression equation into two regimes. Regime 1 is when $D_i = 1$, in this case $V_i \geq -\alpha - \beta Z_i$ (gain exceeds cost) and $Y_i = \mu + \gamma X_i + \Delta + U_i$. Regime 0 is when $D_i = 0$, in this case $V_i \leq -\alpha - \beta Z_i$ (cost exceeds gain) and $Y_i = \mu + \gamma X_i + U_i$. Remember that U_i may differ between people under these regimes and the difference may be systematic between the two groups because $cov(U_i, V_i) \neq 0$ so that $cov(U_i, D_i) \neq 0$ and this leads to having a selection bias while estimating the ATE by the simple mean difference of outcomes (see equation 10 and do not consider a heterogeneous TE) unless the assignment of treatment is random and compliance is perfect (which is not the case here). Heckman (1978) solved this selection bias problem by estimating the two expectations $E(U_i/D_i = 1)$ and $E(U_i/D_i = 0)$. The key ingredient of his method is the truncated normal distributions. He made parametric assumptions about unobserved determinants of outcomes and then calculated their conditional expectations using the assumptions. The assumption is that U_i and V_i are jointly normally distributed.

Truncated normal distributions we recall that for a standard normal distribution Y whose support is R, the mean of the truncated distribution X defined on $[a, +\infty[$ (X is the truncated distribution taken from Y) is

$$E(X) = 0 + 1.\frac{f(a) - f(+\infty)}{F(+\infty) - F(a)} = \frac{f(a)}{1 - F(a)} \tag{14}$$

Let assume that U and V are jointly normally distributed with mean zero, standard deviations σ_U, σ_V and co variance σ_{UV}. Let $f_U(.)$ be the normal density of the distribution of U and $F_U(.)$ its CDF. We present properties of the truncated normal distributions:

$$E(\frac{U}{\sigma_U}|\frac{U}{\sigma_U} > a) = \frac{f_U(a)}{1 - F_U(a)} \tag{15}$$

$$E(\frac{U}{\sigma_U}|\frac{U}{\sigma_U} < b) = -\frac{f_U(b)}{F_U(b)} \tag{16}$$

$$E(\frac{U}{\sigma_U}|a < \frac{U}{\sigma_U} < b) = \frac{f_U(a) - f_U(b)}{F_U(b) - F_U(a)} \tag{17}$$

These ratios are called the inverse Mill's ratios. We also present properties of truncated joint normal distributions:

$$E(\frac{U}{\sigma_U}|\frac{V}{\sigma_V} > a) = \sigma_{UV}\frac{f_U(a)}{1 - F_U(a)} \tag{18}$$

$$E(\frac{U}{\sigma_U}|\frac{V}{\sigma_V} < b) = \sigma_{UV}\frac{f_U(b)}{F_U(b)} \tag{19}$$

These expectations are interpreted as the mean of unobserved determinants of outcomes U in the control and treatment groups. They will capture selectivity.

Heckman's two step procedure in the first step, he estimated the coefficients of the selection equation α and β by using Probit :

$$D^* = \alpha + \beta Z_i + V_i \tag{20}$$

and then computed the inverse Mill's ratios (IMRs) in treatment and control groups:

$$\lambda_{1,i} = \frac{f(\hat{\alpha} + \hat{\beta} Z_i)}{F(\hat{\alpha} + \hat{\beta} Z_i)} \tag{21}$$

$$\lambda_{0,i} = \frac{f(\hat{\alpha} + \hat{\beta} Z_i)}{1 - F(\hat{\alpha} + \hat{\beta} Z_i)} \tag{22}$$

One important assumption that we have to make is that $cov(U_i, Z_i) = 0$ so that the instrument doesn't influence directly the outcome but only through the selection into treatment.
The second step is done by including the IMRs into equation 11 for each group :

$$Y_i = \mu + \gamma X_i + \Delta + \rho_1 \lambda_{1,i} + v_i \tag{23}$$

$$Y_i = \mu + \gamma X_i + \rho_0 \lambda_{0,i} + v_i \tag{24}$$

ρ_1 and ρ_0 are estimators for $\sigma_U \sigma_{UV}$ and $-\sigma_U \sigma_{UV}$ respectively. We can estimate the treatment effect by comparing the two equations. By introducing λ_1 and λ_0 we control for selectivity and account for the differences between control and treatment groups (that the treatment group would have a higher outcome that pushed its individuals to take the treatment). Now days researchers don't use this approach anymore but as we will see later even the more modern approaches of marginal treatment effects (MTE) are very much based on the idea of using parametric assumptions and truncated normal distributions to solve the problem of selection bias and the treatment being more beneficial for treatment group then for control group.

We've looked at the Heckman's two-step selection model and we have seen that there is a way to solve the problem of selection bias by accounting for selectivity into treatment by using parametric assumption and truncated normal distributions. However the Heckman's model was based on homogeneous treatment effects which means that the gain from treatment will be the same for anyone who took the treatment but there was heterogeneity in the costs V_i of people that choose to take or not the treatment. The homogeneous TE assumption is not realistic because the treatment effects themselves are heterogeneous across people. Now we have a lot of heterogeneity to deal with and in order to see the impact of a public policy for example, we have to take into account both types of heterogeneity.

1.3 Heterogeneous treatment effects and target parameters

In this subsection we will continue our path towards understanding marginal treatment effects. A key ingredient to understand why at first place researchers thought about MTE is the heterogeneity in treatment effects. Bjorklund and Moffitt (1987) incorporate heterogeneous treatment effects into a selection model :

$$Y_i = X_i \beta + \alpha_i D_i + \epsilon_i \tag{25}$$

$$D_i^* = \alpha_i - \phi_i \tag{26}$$

$$\alpha_i = Z_i\delta + u_i \tag{27}$$

$$\phi_i = W_i\eta + v_i \tag{28}$$

$$E(\epsilon_i) = E(u_i) = E(v_i) = 0 \tag{29}$$

This model is an extension of the Heckman's two step selection model. In equation 25, Y_i is function of observable determinants X and unobserved ones ϵ. D_i is again correlated with ϵ_i so that some people choice to take the treatment is influenced by unobserved determinants of outcome and $D_i = 1(D_i^* > 0)$. In the selection equation 26, α_i is the heterogeneous gain from treatment and the treatment effect whereas ϕ_i is the heterogeneous cost of taking the treatment.α_i varies across people and affects the outcome differently and this is the extension to the Heckman's two step selection model that Bjorklund and Moffitt (1987) offered in a richer model. The reduced form of this model is given by

$$Y_i = X_i\beta + Z_i\delta + \epsilon_i + u_i \tag{30}$$

when $D_i = 0$,

$$Y_i = X_i\beta + \epsilon_i \tag{31}$$

when $D_i = 0$. And

$$D_i^* = Z_i\delta - W_i\eta + u_i - v_i \tag{32}$$

We will look at the distributions of u_i and v_i and see how we can express treatment effects using them. If we suppose that both of them are normally distributed then the difference $u_i - v_i$ is also normally distributed. Remember that this difference is is the difference between gains and costs of taking the treatment and the decision rule is based on the comparison between $u_i - v_i$ and s_i defined as :

$$s_i = \frac{-Z_i\delta + W_i\eta}{\sigma_{U-V}} \tag{33}$$

since $u_i - v_i$ is normally distributed we can use as before the truncated normal distribution trick and express the conditional mean of $u_i - v_i$ as follows :

$$\lambda_i = \frac{f(s_i)}{1 - F(s_i)} \tag{34}$$

and this inverse Mill's ration can be used for estimating the treatment effects and other economically interesting effects, for example the expected gain of from treatment for those who select into treatment :

$$E(\alpha_i|D_i = 1, Z_i\delta, W_i\eta) = Z_i\delta + E(u_i|u_i - v_i > -Z_i + W_i\eta) = Z_i\delta + (\sigma_{U,U-V}/\sigma_{U-V})\lambda_i \tag{35}$$

the average gain from taking the treatment depends on the sum of the average gain that comes from the instrument and of a term that accounts for people whom net gain from taking the treatment is high. We can go further to see how the average gain will change by changing the cost of choosing the treatment :

$$\frac{\partial E(\alpha_i|D_i = 1, Z_i\delta, W_i\eta)}{\partial W_i\eta} = [(\sigma_{U,U-V}/\sigma_{U-V}^2)]\lambda_i(\lambda_i - s_i) > 0 \tag{36}$$

The proof of equation 36 is given in the paper Bjorklund and Moffitt (1987). If we make the treatment cheaper to get we will see people taking the treatment but getting lower gains on average. This is because this change on cost will shift people that were indifferent between taking or not taking the treatment into treatment but the gain from treatment for these people is not high which makes the average gain lower. Therefore estimating the MTE is possible using parametric assumptions and the estimated effects depends on who the instrument (part of cost captured by instrument) shifts into or out of treatment.

Roy model with all of this in mind we can go back now to the Roy model which is not the same as Bjorklund and Moffitt model but most of the modern approaches around MTE are based on it. Let recall that the individual TE in Roy model was :

$$\Delta_i = Y_{i,1} - Y_{i,0} = \mu_1 - \mu_0 + U_{i,1} - U_{i,0} \tag{37}$$

which is a sum of common gain and idiosyncratic gain. Now using this model we will see what parameter does our estimator actually identify. Let's state the possible parameters. the first one is the average treatment effect (ATE)

$$ATE(x) = E(Y_1 - Y_0/X = x) = \mu_1 - \mu_0 \tag{38}$$

ATT is the average treatment effect on the treated

$$ATT(x) = E(Y_1 - Y_0/X = x, D = 1) = \mu_1 - \mu_0 + E(U_1 - U_0|X = x, D = 1) \tag{39}$$

ATU is the average treatment effect on the untreated

$$ATT(x) = E(Y_1 - Y_0/X = x, D = 0) = \mu_1 - \mu_0 + E(U_1 - U_0|X = x, D = 0) \tag{40}$$

LATE is the local average treatment effect

$$LATE(x) = E(Y_1 - Y_0/X = x, D_1 > D_0) = \mu_1 - \mu_0 + E(U_1 - U_0|X = x, D_1 > D_0) \tag{41}$$

$D_1 > D_0$ means we select people whose behavior has been changed and shifted into treatment due to a change in the instrument, the people that were indifferent at first and their choices changed at the margin. In many cases ATE, ATT, ATU may not be the target parameters or the policy relevant parameter. For this reason, Heckman and Vytlacil (2001) introduced a Policy-Relevant Treatment Effects (PRTE) by considering a policy change that affects the propensity score $P(X_i, Z_i)$ without affecting potential outcomes $(Y_{i,0}, Y_{i,1})$ or unobserved selection V_i. Denote by D_i the treatment choice under baseline policy and by $\tilde{D}_i$ the treatment choice under the alternative policy that changes who selects into treatment. The PRTE which is the mean effect of going from a baseline to an alternative policy per net person shifted is given by

$$\begin{aligned} PRTE(x) &= \frac{E(Y_i|X_i = x, alternativepolicy) - E(Y_i|X_i = x, baselinepolicy)}{E(D_i|X_i = x, alternativepolicy) - E(D_i|X_i = x, baselinepolicy)} \\ &= \mu_1 - \mu_0 + \frac{E(U_{1,i} - U_{0,i}|X_i = x, \tilde{D}_i = 1)E(\tilde{D}_i|X_i = x) - E(U_{1,i} - U_{0,i}|X_i = x, D_i = 1)E(D_i|X_i =}{E(\tilde{D}_i|X_i = x) - E(D_i|X_i = x)} \end{aligned} \tag{42}$$

LATE and other parameters if treatments effects were homogeneous across people then LATE coincides with ATE,ATT,ATU and PRTE but this is not the usual case. Instead the TE are heterogeneous, in this case the parameters are not the same unless we have randomized assignment of treatment and full compliance to the assignment but in most cases people choose to get treatment based on their gains and costs of treatment.

1.4 Marginal treatments effects curve and weighted estimators

What we have seen till now is that LATE can estimate the TE for people who comply with instrument and changed their treatment choice when instrument changes or people who comply with the treatment assignment. The MTE can be useful to extrapolate the TE to the whole population and estimate the target parameter. MTE models explicitly models the selection into treatment and the analysis is based on potential outcomes like in Roy model where potential outcomes are functions of observed and unobserved determinants. The marginal treatment effect is defined as :

$$MTE(x, v) = E(Y_1 - Y_0 | V = v, X = x) \tag{43}$$

MTE(x,v) is the treatment effect given the observable characteristics x and the likelihood of taking the treatment v which is an unobserved characteristic that accounts for the cost of taking the treatment. MTE(x,v) gives the treatment effect of people who are indifferent to taking the treatment at the level cost v. MTE(v) is a decreasing function because treatment effect is low for people with high cost of treatment and people who select into treatment are the ones with the largest gains from treatment. Here is an example from Carneiro et al. (2011) where they are interested in the wage return to one more year of education. The MTE curve is the final result of MTE analysis and we show it here for illustration.

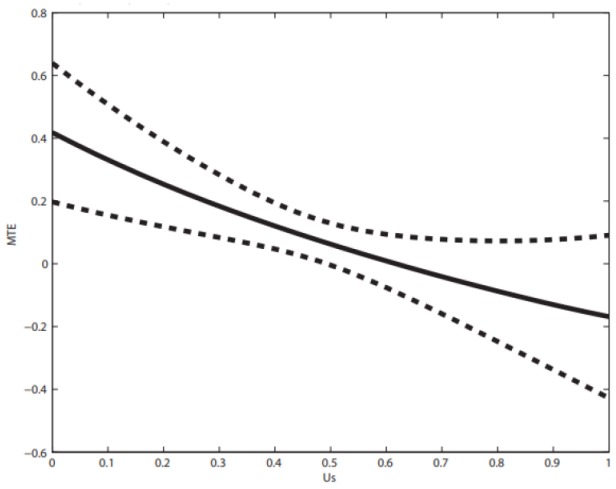

Figure 1: MTE curve of returns to education from Carneiro et al. (2011)

U_s is the resistance to treatment or cost of treatment. For people in the upper range of U_s, the treatment is the most costly, the higher U_s the lower the marginal gain will be from treatment. The MTE curve tells us that people with the lowest cost of treatment are people who get the highest gain from treatment or people who get the highest gain from treatment

are the most likely to select into treatment. Remember that what an IV estimate will give us is the treatment effect of a small range of U_s which corresponds to people who comply with the instrument and by using MTE we can extrapolate the TE to other people. The MTE curve is very useful because it will allow to estimate any target TE parameter and express them as a weighted average of all values of MTE in the curve. The weights are probabilities of the distribution of U_s idem the weight of each group in the population. The MTE curve is very useful in the sense that it allows to estimate the TE of a public policy depending on who take the treatment and the range of people aimed by the policy, and then estimate the TE when we consider an expansion of the policy to other people. In a review paper by Mogstad and Torgovitsky (2018), they splited the MTE in what they called several Marginal Treatment Responses (MTR) for the treated and the control group :

$$m_0(v, x) = E(Y_0|V = v, X = x) \tag{44}$$

$$m_1(v, x) = E(Y_1|V = v, X = x) \tag{45}$$

m_0 and m_1 are counterfactual outcomes for people with the same resistance to treatment or the likelihood of taking the treatment. The MTE is just $m_1 - m_0$. Mogstad and Torgovitsky (2018) showed that many interesting parameters can be written as weighted averages of the MTRs m_0 and m_1. The application we can find in Mogstad and Torgovitsky (2018) is studying the effect of mosquito nets (D) on malaria (Y). They used a subsidy Z as an instrument $Z \in \{1, 2, 3, 4\}$ and they observed these propensity scores $P(Z) \in \{0.12, 0.29, 0.48, 0.78\}$ which means that people who got a high subsidy are the most likely to buy a mosquito net. Mogstad and Torgovitsky used these propensity scores and observed outcomes to construct the curves of MTE, MTR of the treated and MTR curve of the same units if they had not been treated (the counterfactual m_0).

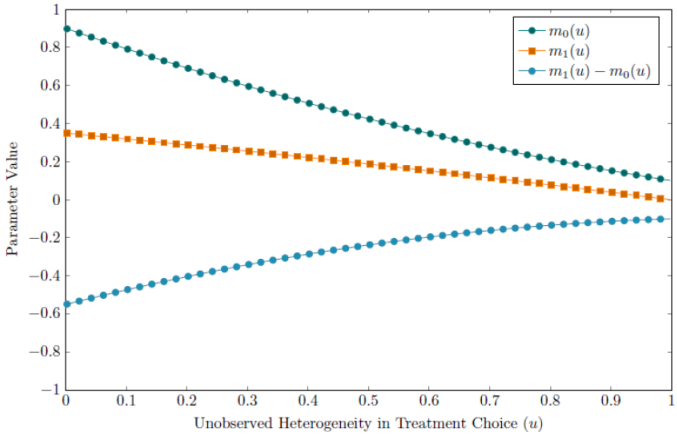

Figure 2: MTE and MTR curves, Mogstad and Torgovitsky (2018)

Let take the type of people whose cost of treatment is zero, these guys are the most likely to take the treatment. The blue curve tells us that for this type of people the MTE of taking the treatment is almost -0.6 which means that for people who are indifferent between taking or not taking the treatment at the zero level cost, the effect of buying mosquito net and using

it will be negative on the malaria. The orange and green curves show the MTR. The orange curve states that if people of type zero took the treatment they would have a level of outcome almost equal to 0.4 which is lower then the level of outcome if the same type of people had not take the treatment. By extrapolating the MTE and MTRs we can see that the average benefit from treatment becomes lower when people with the highest cost shift into treatment, this means that the efficiency of a policy depends heavily on who the policy shifts in or out of the treatment. Mogstad and Torgovitsky (2018) wrote the target parameter β^* as a function of MTRs : as a weighted average of MTRs such that the weights functions $w_1^*(v,x,z)$ and $w_0^*(v,x,z)$ depend on which target parameter we aim to estimate.

$$\beta^* = E\left[\int_0^1 m_0(v,x)w_0^*(v,x,z)dv\right] - E\left[\int_0^1 m_1(v,x)w_1^*(v,x,z)dv\right] \tag{46}$$

Mogstad and Torgovitsky (2018) gave how different possible target parameters can be expressed as function of MTRs.

Target Parameter	Expression	Weights $\omega_0^*(u,x,z)$	Weights $\omega_1^*(u,x,z)$
Average Untreated Outcome	$E[Y_0]$	1	0
Average Treated Outcome	$E[Y_1]$	0	1
Average Treatment Effect (ATE)	$E[Y_1 - Y_0]$	-1	1
ATE given $X = \overline{x}$ where $P[X = \overline{x}] > 0$	$E[Y_1 - Y_0 \mid X = \overline{x}]$	$-\omega_1^*(u,x,z)$	$\frac{\mathbb{1}[x = \overline{x}]}{P[X = \overline{x}]}$
Average Treatment on the Treated (ATT)	$E[Y_1 - Y_0 \mid D = 1]$	$-\omega_1^*(u,x,z)$	$\frac{\mathbb{1}[u \leq p(x,z)]}{P[D=1]}$
Average Treatment on the Untreated (ATU)	$E[Y_1 - Y_0 \mid D = 0]$	$-\omega_1^*(u,x,z)$	$\frac{\mathbb{1}[u > p(x,z)]}{P[D=0]}$
Local Average Treatment Effect (LATE) for $z \to z'$ given $X = x$, where $p(x,z') > p(x,z)$	$E[Y_1 - Y_0 \mid p(x,z) < U \leq p(x,z'), X = x]$	$-\omega_1^*(u,x,z)$	$\frac{\mathbb{1}[p(x,z) < u \leq p(x,z')]}{p(x,z') - p(x,z)}$

Figure 3: Weights of a variety of target parameters, Mogstad and Torgovitsky (2018)

Mogstad and Torgovitsky (2018) have also provided estimations of weights using the example of effect of mosquito nets on malaria (see figure 4). In the horizontal axis we have the cost of treatment or the inverse of the likelihood of taking the treatment. In the vertical axis we have the average weights w_1 one can plug in the equation 46 to estimate the target parameter.The weights w_1 are normalized to one and gives to each unit of the sample the same weight. For the average treatment effect on the treated (ATT) the distribution of weights is not uniform but more concentrated on the lower side of the cost, which means that people with the lowest cost of treatment are the most likely to be selected into treatment. The average treatment effect on the untreated (ATU) is the opposite, meaning that people with the highest cost are the most unlikely to take the treatment. So people with the highest likelihood to take the treatment get the highest weights in ATT but the lowest weights in ATU and vice versa. Regarding

the LATE, remember that it estimates only the treatment effect of people who comply with instrument and whose treatment choice changes by changing the instrument thus the support of weights doesn't cover all possible values of u. For example, let discuss weights that lead to the estimation of $LATE_{1\to2}$.

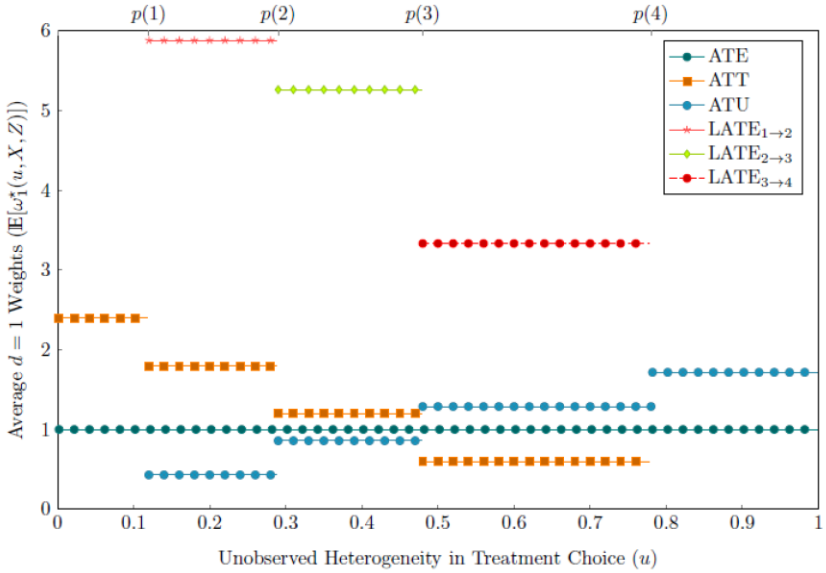

Figure 4: Weights of a variety of target parameters, Mogstad and Torgovitsky (2018)

With only subsidy $Z = 1$, only people whose cost is lower then 0.1 are likely to take the treatment and then by changing the subsidy to $Z = 2$ the range of people taking the treatment will expand and new units will shift into treatment, the $LATE_{1\to2}$ provides estimation of effect for these people who shifted into treatment and high weights are giving for them in the formula 46.
To conclude this subsection, it is important to know which target parameter we aim to estimate. MTE is a useful tool that allows to estimate a target parameter as weighted average of MTRs and accounts for heterogeneity of likelihood of taking the treatment across units thanks to the weights that attribute different level of importance to each group of people depending on their costs/gains of treatment, how they react to instrument and what target parameter we want to estimate. In the next subsection, we will see how we can use observed data to construct MTE and estimate weights.

1.5 Marginal treatment effects estimation

In this subsection, we will walk through how we can estimate a marginal treatment effects model. In classic Roy model we have the potential outcomes as function of observable and unobserved determinants, and we have a selection equation whereby the treatment D is a function of an instrument Z and an idiosyncratic resistance to treatment V_i

$$D = f(Z_i, V_i) \tag{47}$$

We assume that the instrument is independent of the policy we want to evaluate, the unobserved determinants of the outcomes and the specific cost parameter of each individual V_i

$$(U_0, U_1, V) \perp\!\!\!\perp Z/X \tag{48}$$

This means that conditional on some observed characteristics, the instrument should be randomly assigned. We also make the standard LATE strong first stage and monotonicity assumption which states that people who are encouraged by the instrument to shift into the treatment should have higher likelihood to take the treatment then people who don't comply with instrument and the effect of instrument on the likelihood to take the treatment has the same sign across all people. We don't need to assume neither the independence of idiosyncratic resistance to treatment and unobserved determinants of outcomes nor the independence of observed and unobserved determinants of outcomes. Let consider the selection equation

$$D^* = Z\delta - V = \mu_D - V \tag{49}$$

such that Z is the instrument and V is the resistance to treatment.μ_D is defined as above, and denote F_V is the CDF of V. An important variable we will consider is the propensity score $p(z)$ which gives the part of people taking the treatment given a level of instrument Z.

$$p(z) = P(D = 1|Z = z, X = x) = F_V(\mu_D(X, Z)) \tag{50}$$

In order to simplify the graphical visualizations and analysis, we adopt the convention

$$U_D = F_V(V) \tag{51}$$

U_D is the quantile of the resistance V at a certain level of V. With this convention we define the propensity score as the quantile of the resistance distribution. This transformation allows to look at the uniform distribution U_D instead of the distribution of V which is very unlikely to be uniform. With all of this we get :

$$D = 1 \Leftrightarrow \mu_D > V \Leftrightarrow p(z) > U_D \tag{52}$$

To arrive to the MTE estimator and to MTE curve, we make a step back to LATE and Wald estimator.

LATE and binary instrument for a given value X=x and a binary instrument, the Wald estimator is

$$Wald(x) = \frac{E(Y|Z = 1, X = x) - E(Y|Z = 0, X = x)}{E(D|Z = 1, X = x) - E(D|Z = 0, X = x)} = \frac{E(Y|Z = 1, X = x) - E(Y|Z = 0, X = x)}{p(1) - p(0)} \tag{53}$$

This estimator tells us to what extent does the average outcome change between people who have been encouraged (Z=1) to take the treatment and those who have not, scaled up by the first stage (idem by the share of people who complied with the instrument). The LATE is

$$LATE(x) = E(Y_1 - Y_0|D_1 > D_0, X = x) = \mu_1 - \mu_0 + E(U_1 - U_0|D_1 > D_0, X = x) \tag{54}$$

the second term incorporate the heterogeneity in unobserved returns from treatment across people who complied with the instrument.

LATE with a continuous instrument now we push the idea of LATE and Wald estimator to the more advanced case of continuous instrument and this is part of the road to the construction of MTE curve, we want to know what is the treatment effect of not only people who complied with instrument but of all population so we need an instrument that varies more and can make more people comply with it and this is why having a continuous instrument is important. Let's consider a pair values of instrument z and z'. The pairwise Wald estimator becomes

$$Wald(z, z', x) = \frac{E(Y_i|Z_i = z, X_i = x) - E(Y_i|Z_i = z', X_i = x)}{E(D_i|Z_i = z, X_i = x) - E(D_i|Z_i = z', X_i = x)} \tag{55}$$

Again the pairwise Wald estimator here estimates the treatment effect of the specific group of people who complied with the change of instrument. In Cornelissen et al. (2016) they used the distance to college as a continuous instrument.

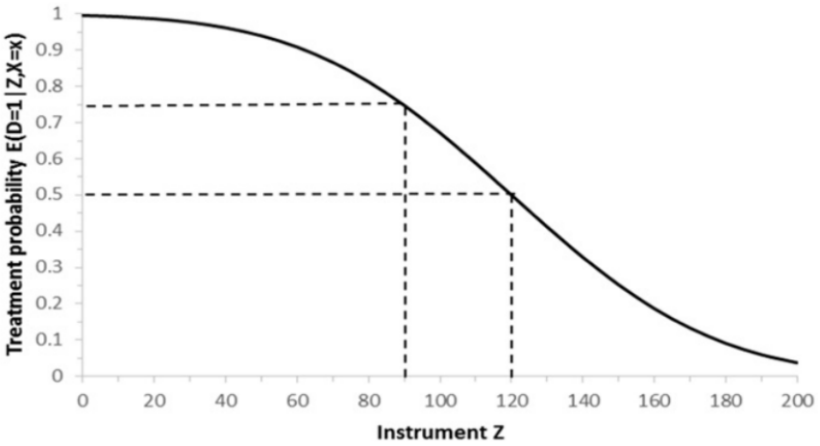

Figure 5: Compliers with distance between birth place and college, Cornelissen et al. (2016)

The first stage which is about how people react to changes in instrument is visualised in the curve of $p(z)$. By decreasing the distance to college from 120 to 90, we see that $p(z)$ went from 0.5 to 0.8 thus people whose U_v (the resistance to treatment) is between 0.5 and 0.75 (and who didn't take the treatment with the ex value of Z) are now shifted into treatment (remember that a person takes the treatment if $p(z) > U_v$, see proposition 52). The second stage consists on estimating the LATE at a given X using the Wald estimator and this is simply done by regressing the outcome on the propensity score for each value of the instrument.

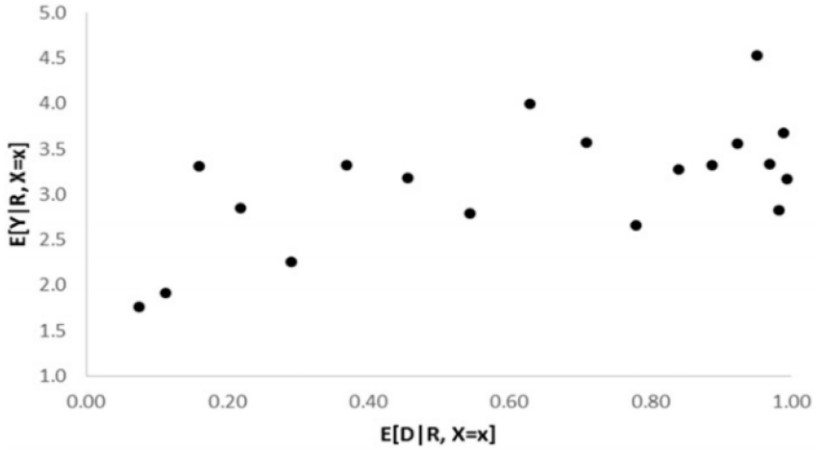

Figure 6: 2SLS estimation of LATE(x) using a continuous instrument, Cornelissen et al. (2016)

Now to have the overall IV estimator of TE, we estimate LATE(x) at all values of x and average them

$$IV = \sum_{x \in X} w(x) LATE(x) \tag{56}$$

The weights $w(x)$ are equal to the contribution of units with $X = x$ to the first stage, the higher the variation of treatment within a group, the higher the weight will be.

Back to the marginal treatment effect By replacing the cost V by its quantiles U_D in equation 43 we get

$$MTE(x, u_D) = E(Y_1 - Y_0 | X = x, U_D = u_D) = \mu_1 - \mu_2 + E(U_1 - U_0 | X = x, U_D = u_D) \tag{57}$$

and this gives the treatment effect of units with observed characteristics x and who are indifferent to treatment at the u_D^{th} quantile of resistance to treatment. The MTE gives what is the change in outcome if the change in instrument (or equivalently in the propensity score) shifts units in/out treatment. So for a given x and p(Z), the MTE is

$$MTE(X = x, U_D = p) = \frac{\partial E(Y | X = x, P(Z) = p)}{\partial p} \tag{58}$$

How to estimate MTE the estimation of MTE is challenging because we don't know the unobserved determinants of outcome and the unobserved cost of treatment. To solve this issue many solutions were proposed

- Solution 1 : impose parametric assumptions on the distribution of (U_1, U_0, V) (cf. Bjorklund and Moffitt 1987) this assumption may be very restrictive and strong, this is why Heckman and Vytlacil proposed a non parametric solution.
- Solution 2 : Heckman and Vytlacil (1999, 2001, 2005) used a fully non parametric approach that looks like the so many pairwise comparisons we saw in LATE with a continuous instrument, but for this we need an instrument that varies a lot to do all the

comparisons, which means an instrument that has all the levels that can shift all people in or out of the treatment, so this solution is very demanding on data.

- Solution 3 : the solution that the literature seems to have converged to is by the mean of using shape restrictions (Cornelissen et al., 2016). We have to assume linear separability in potential outcomes $Y_j = X_j\beta_j + U_j$ and independency of the MTE curve and X.

In our work and in the remaining parts we will focus on Solution 2 in a non parametric setting, we will review Heckman and Vytlacil (1999, 2001, 2005) and talk about its limitations. Then we will propose a non parametric method to solve for the problem of estimating TE when instrument variation is limited and $p(z)$ doesn't have a support defined on $[0, 1]$.

1.6 Heckman and Vytlacil (1999, 2001, 2005)

We will use most of the notions and ideas we have seen till now in this subsection. The model that Heckman and Vytlacil have developed doesn't rely on parametric assumptions like we've seen earlier and other functional form assumptions, for example the linear form $\mu(X) = X\beta$. They considered a nonparametric selection model with binary treatments. With no assumption on the support of outcome and by relaxing the linear functional form, they considered a non linear and nonseparable outcome model :

$$Y_1 = \mu_1(X, U_1) \tag{59}$$

$$Y_0 = \mu_0(X, U_0) \tag{60}$$

X are observable and U are unobserved determinants of outcomes. Denote by Δ the ceteris paribus causal effect of shifting a person into/out of treatment: $Y_1 - Y_0 = \Delta$. They characterized the selection equation by an index model:

$$D^* = \mu_D(Z) - U_D, D = 1(D^* \geq 0) \tag{61}$$

Z is observed and U_D is unobserved. The assumptions are the following :

- The term $\mu_D(Z)$ is a nondegenerate random variable conditional on X, which means $P(\mu_D(Z)/X) < 1$
- (U_1, U_D) and (U_0, U_D) are independent of Z conditional on X.
- The distribution of U_D is absolutely continuous with respect to Lebesgue measure
- The values of $E|Y_1|$ and $E|Y_0|$ are finite
- 1>P(D=1|X)>0
- $X_1 = X_0$

We impose on X to be exogenous to treatment in order to capture the full effect of D on Y, but X can be endogenous in the outcome model since it can be correlated with unobserved error terms ...