TREATMENT EFFECTS

1. Potential Outcomes Framework and Average Treatment Effects

Here we have two *latent* variables Y_1 and Y_0, which are the counterfactual outcomes for an observational unit when the unit is subjected to treatment and no treatment, in an idealized experiment [17, 10]. In economic context, the treatment can be a training program or a policy intervention. These quantities are "counterfactual", because they can't be simultaneously observed, unless we have replicas of the observational units that are simultaneously subjected to different treatments.

The treatment effect is

$$Y_1 - Y_0.$$

It is generally unrealistic to uncover the treatment effect, but we can hope to estimate moments of $Y_1 - Y_0$ such as the average treatment effect (ATE):[1]

$$\delta = \mathrm{E}(Y_1 - Y_0) = \mathrm{E}Y_1 - \mathrm{E}Y_0.$$

Let D denote the treatment indicator, which takes a value of one if the observational unit participated in the treatment and zero otherwise. The observed outcome is

$$Y = DY_1 + (1 - D)Y_0.$$

Hence we observe $Y = Y_1$ if $D = 1$ and $Y = Y_0$ if $D = 0$. For example, we observe the wage outcome Y_1 after completion of a training program for a given person only if this person has completed the program $D = 1$; we do not observe the wage outcome Y_1 without completion of the training program, i.e. if $D = 0$.

So we can identify the quantities

$$\mathrm{E}[Y \mid D = 1] = \mathrm{E}[Y_1 \mid D = 1] \text{ and } \mathrm{E}[Y \mid D = 0] = \mathrm{E}[Y_0 \mid D = 0].$$

The difference of the two quantities gives us the average predictive effect (APE) of treatment status on the outcome:

$$\pi = \mathrm{E}[Y \mid D = 1] - \mathrm{E}[Y \mid D = 0].$$

[1] We consider other moments such as the average treatment effect on the treated (ATT) in the appendix.

It measures the association of the treatment status with the outcome, and this quantity π in general does not agree with the ATE δ:

$$\delta \neq \pi.$$

This phenomenon is generally caused by a selection problem.

EXAMPLE 1.(Selection Effects in Observational Data) Suppose we want to study the impact of smoking marijuana on life longevity. Suppose that smoking marijuana has no causal/treatment effect on life longevity:

$$Y = Y_0 = Y_1, \text{ so that } \delta = \mathrm{E}Y_1 - \mathrm{E}Y_0 = 0.$$

However, the observed smoking behavior, D, results not from an experimental study, but from a behavior in which smoking is associated with poor health (certain types of cancers, for example), which causes shorter life expectancy. In this case

$$\pi = \mathrm{E}[Y \mid D = 1] - \mathrm{E}[Y \mid D = 0] < 0 = \delta. \quad \square$$

The problem with the observational study like the one in this contrived example is that the "treatment status" D is determined by the individual behavior which depends on the potential outcomes, causing selection bias, namely $\pi < \delta$. Arguably, the cleanest way to break the dependency is through random assignment.

Assumption 1 (Random Assignment/Exogeneity)**.** *Suppose that the treatment status is randomly assigned, namely D is statistically independent of the potential outcomes $(Y_d)_{d\in\{0,1\}}$, which is denoted as*

$$D \perp\!\!\!\perp (Y_0, Y_1),$$

and $0 < \mathrm{P}(D = 1) < 1$.

Theorem 1 (Randomization Removes Selection Bias)**.** *Under Assumption 1,*

$$\mathrm{E}[Y \mid D = d] = \mathrm{E}[Y_d \mid D = d] = \mathrm{E}[Y_d], \text{ for } d \in \{0, 1\}.$$

Hence

$$\pi = \mathrm{E}Y_1 - \mathrm{E}Y_0 = \delta.$$

Please note that the theorem is self-contained, since the proof is contained in the statement.

Hence randomized experiments, commonly called Randomized Control Trials (RCTs), in which treatment is randomly assigned, generate a simple, practical mechanism through which we can measure the impact of a treatment on average potential outcomes.

EXAMPLE 2.(No Selection Effects in Experimental Data) Suppose instead that in the previous example we worked with data where smoking was generated by random assignment, then we would have the agreement between average predictive and treatment effects: $\pi = \delta$. Of course, it is difficult to imagine an RCT where smoking or non-smoking marijuana can be forced onto participants of the study. □

Scientists most often have to rely on observational studies to try to learn causal/treatment effects – we've been developing tools for this the entire course.[2] One commonly used assumption to eliminate selection effects is the following.

Assumption 2 (Ignorability and Overlap). *(a) Ignorability. Suppose that the treatment status D is independent of potential outcomes $(Y_d)_{d\in\{0,1\}}$ conditional on a set of covariates X, that is*

$$D \perp\!\!\!\perp (Y_0, Y_1) \mid X.$$

(b) Overlap. Suppose that the propensity score $p(X) := \mathrm{P}(D = 1|X)$, which is the probability of receiving treatment given X, is non-degenerate:

$$\mathrm{P}(0 < p(X) < 1) = 1.$$

Note that the conventional name used in econometrics for ignorability is *conditional exogeneity* or *conditional independence* assumption. Since we emphasize potential outcomes as a framework to think of causality here, we use the naming conventions of this literature.

Assumption 2 means that the treatment is as good as randomly assigned conditional on X. This assumption underlies most of the regression strategies for identifying the causal/treatment effects from observational data.

Theorem 2 (Conditioning on X Removes Selection Bias). *Under Assumption 2,*

$$\mathrm{E}[Y \mid D = d, X] = \mathrm{E}[Y_d \mid D = d, X] = \mathrm{E}[Y_d \mid X].$$

Hence the Conditional Average Predictive Effect (CAPE),

$$\pi(X) = \mathrm{E}[Y \mid D = 1, X] - \mathrm{E}[Y \mid D = 0, X]$$

is equal to the Conditional Average Treatment Effect (CATE),

$$\delta(X) = \mathrm{E}[Y_1 \mid X] - \mathrm{E}[Y_0 \mid X].$$

Hence APE and ATE also agree:

$$\delta = \mathrm{E}\delta(X) = \mathrm{E}\pi(X) = \pi$$

[2] We have used structural equation models and made various assumptions such as conditional exogeneity and other conditional moment restrictions to identify the structural parameters from the statistical parameters such as predictive effects.

Please note that the theorem is self-contained, since the proof is contained in the statement. The overlap assumption makes it possible to condition on the events $\{D = 0, X\}$ and $\{D = 1, X\}$.

Example 3.(Removing Selection Bias Conditional on Controls) In the context of smoking, it might be plausible to think of the smoking behavior as independent of potential outcomes, once we condition on observed characteristics, such as medical records, or demographic characteristics. □

One implication of this assumption deals with the linear model

$$Y_d = d\alpha + X'\beta + \epsilon_d, \quad \mathrm{E}\epsilon_d X = 0, \tag{1.1}$$

where X contains an intercept. Under Assumption 2, D is independent of ϵ_1 and ϵ_0 and hence

$$\epsilon = \epsilon_1 D + \epsilon_0(1 - D)$$

obeys

$$\mathrm{E}\epsilon D = \mathrm{E}\epsilon_1 D = 0$$

so that

$$Y = \alpha D + X'\beta + \epsilon, \quad \mathrm{E}\epsilon(D, X) = 0,$$

implying that we can identify δ from the coefficient α of this model:

$$\alpha = \delta.$$

Of course the assumption of linearity with respect to d in (1.1) is restrictive, but convenient. In particular, we can use the partialling out methods for estimating α, including cases where β is high-dimensional.

Without the linearity assumption (1.1), the projection parameter α can be shown to estimate the weighted average of CATEs,

$$\alpha = \mathrm{E}w(X)\delta(X).$$

Another approach is to drop the linear model and consider the interactive model,

$$Y_d = dX'\alpha + X'\beta + \epsilon_d, \quad \mathrm{E}\epsilon_d \mid X = 0 \tag{1.2}$$

Under ignorability ϵ_d's are independent of D and we have that

$$X'\alpha = \mathrm{E}Y_1 \mid X - \mathrm{E}Y_0 \mid X$$

and δ is identified as the average of $X'\alpha$:

$$\mathrm{E}X'\alpha = \delta.$$

The fact that conditioning on the right set of controls removes the bias has long been recognized by researchers employing regression methods. Rosenbaum and Rubin [INSERT REF] made the much more subtle point that conditioning on the propensity score suffices to remove the selection bias.

We defined above the propensity score as $p(X) := \mathrm{P}(D = 1|X)$, which is the probability of receiving treatment given X. Under the previous assumption we can represent the treatment selection rule statistically as:

$$D = 1\{U \leq p(X)\}, \quad U \mid X \sim U(0, 1),$$

and

$$U \perp\!\!\!\perp (Y_0, Y_1) \mid X.$$

Note that

$$\mathrm{E}[Y \mid D = 1, p(X)] = \mathrm{E}[Y_1 \mid U \leq p(X), p(X)] = \mathrm{E}[Y_1 \mid p(X)],$$

and similarly

$$\mathrm{E}[Y \mid D = 0, p(X)] = \mathrm{E}[Y_0 \mid U \geq p(X), p(X)] = \mathrm{E}[Y_0 \mid p(X)].$$

Hence we obtain the theorem of Rosenbaum and Rubin [16].

Theorem 3 (Conditioning on The Propensity Score Removes Selection Bias)**.** *Under Assumption 2,*

$$\mathrm{E}[Y \mid D = d, p(X)] = \mathrm{E}[Y_d \mid p(X)].$$

Hence the Conditional (on the propensity score) Average Predictive Effect,

$$\pi(X) = \mathrm{E}[Y \mid D = 1, p(X)] - \mathrm{E}[Y \mid D = 0, p(X)]$$

is equal to the Conditional (on the propensity score) Average Treatment Effect (CATE),

$$\delta(X) = \mathrm{E}[Y_1 \mid p(X)] - \mathrm{E}[Y_0 \mid p(X)].$$

Hence APE and ATE also agree:

$$\delta = \mathrm{E}\delta(X) = \mathrm{E}\pi(X) = \pi.$$

Note that the propensity score is the minimal "sufficient" statistic in the sense that conditioning on the propensity score generally removes the bias under ignorability. This is a useful strategy when X is high dimensional and $p(X)$ is available or can be approximated accurately, as happens in the recent work [2]. In this case, we can simply use $p(X)$ as control.

Another critical point, known as the Horvitz-Thompson method, uses propensity score reweighting to recover averages of potential outcomes. Indeed,

$$\mathrm{E}[DY/p(X)] = \mathrm{E}[DY_1/p(X)] = \mathrm{E}[p(X)\mathrm{E}[Y_1 \mid X]/p(X)] = \mathrm{E}[Y_1]$$

and similarly

$$\mathrm{E}[(1-D)Y/(1-p(X))] = \mathrm{E}[Y_0]$$

Hence we obtain the following result.

Theorem 4 (Propensity Score Reweighting Removes Bias)**.** *Under Assumption 2,*

$$\gamma = \mathrm{E}[DY/p(X)] - \mathrm{E}[(1-D)Y/(1-p(X))] = \delta.$$

2. Estimation and Inference on ATE

Given a strategy outlined above we can proceed with GMM estimation and inference using the methods we have developed previously. Define

$$\mu_d(X) = \mathrm{E}[Y \mid D = d, X], \quad p_d(X) = \mathrm{P}(D = d \mid X).$$

Then denoting $\theta_d = \mathrm{E}Y_d$ we have

$$\theta = (\theta_d)_{d\in\{0,1\}}.$$

The identification argument of Theorem 2 leads to the regression-based moment conditions:

$$\mathrm{E}[\mu_d(X)] - \theta_d = 0, \quad d \in \{0,1\},$$

(where $\mu_d(X)$ can also be replaced for $\mathrm{E}[Y \mid D = d, p(X)]$ by Theorem 3)[CHECK]. Given the parameteriziation

$$\mu_d(X) = \mu_d(X, \beta_d),$$

we can proceed with GMM approach to estimating θ using this moment equation approach, stacking the score

$$g_1(Z, \theta, \beta) = (\mu_d(X, \beta_d) - \theta_d)_{d\in\{0,1\}}$$

with the score that corresponds to the estimation of $\beta = (\beta_0', \beta_1')'$ [7]. Such an approach is well-suited if β is low or moderately-low dimensional.

The identification argument of Theorem 4 leads to the propensity-score based moment conditions,

$$\mathrm{E}[1(D = d)Y/p_d(X)] - \theta_d = 0,$$

Given the parameteriziation

$$p_d(X) = p_d(X, \ell_d),$$

where we can use the binary response models in L6, we can proceed with GMM approach to estimating θ using this moment equation approach, stacking the score

$$g_2(Z, \theta, \ell) = (1(D = d)Y/p_d(X, \ell_d) - \theta_d)_{d\in\{0,1\}}$$

with the score that corresponds to the estimation of $\ell = (\ell_0', \ell_1')'$ [8]. Such an approach is well-suited if ℓ is low or moderately-low dimensional.

The above strategies are not well-behaved when the nuisance parameters β and ℓ are high-dimensional and we have to employ penalization/regularization to estimate them. Instead, we shall rely on the following moment condition that identifies θ_d:

$$\mathrm{E}[1(D = d)(Y - \mu_d(X))/p_d(X) + \mu_d(X)] - \theta_d = 0.$$

Using this moment condition for identification and inference is called the "doubly-robust approach" (Robins and Rotnizky, [15]).[3] Given the previous parameterizations for $\mu_d(X)$ and $p_d(X)$, this takes us to the score function:

$$g(Z, \theta, \eta) = (1(D = d)(Y - \mu_d(X, \beta_d))/p_d(X, \ell_d) + \mu_d(X, \beta_d) - \theta_d)_{d\in\{0,1\}},$$

where

$$\eta = (\beta, \ell).$$

This score can be derived by "optimally combining the scores" $g_1(Z, \theta, \beta)$ and $g_2(Z, \theta, \ell)$ using the GMM framework (or its extensions), and using the explicit calculation of the optimal weighting matrix.

The score function g has the following orthogonality/local-robustness property:

$$\partial_\eta \mathrm{E} g(Z, \theta_0, \eta) \mid_{\eta=\eta_0} = 0,$$

where η_0 and θ_0 denote the true values of the parameters. This makes it very robust, in particular, we can use penalized estimators of η in the high-dimensional settings.

Because of this property, errors occuring in the estimation of nuisance parameters wash out, and we don't have the first order impact on the asymptotic behavior of the GMM estimator. Specifically the leading term in the linear approximation to the GMM estimator obeys:

$$\frac{1}{n^{1/2}} \sum_{i=1}^{n} g(Z, \theta_0, \hat{\eta}) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} g(Z, \theta_0, \eta_0) + o_P(1)$$

This is the reason why in this case we do not need to stack $g(Z, \theta_0, \eta)$ with the scores corresponding to the estimation of η in the GMM approach. We can estimate θ_0 and η_0 separately.

[3] The naming refers to the fact that even if either μ_d or p_d are misspecified (but not both), then θ_d is still correctly recovered.

Note that it is easy to show that the scores g_1 and g_2 do not have the orthogonality/local-robustness property, making them unsuitable for uses in very high-dimensional settings.

3. Distribution and Quantile Treatment Effects

The marginal distributions of the potential outcomes, F_{Y_0} and F_{Y_1}, are identified under either Assumption 1 or 2. This can be seen directly from Theorems 1–4 replacing Y by the indicators $1(Y \leq y)$ with $y \in \mathcal{Y}$ for a finite set $\mathcal{Y} \subset \mathbb{R}$. Thus, under Assumption 2, by the same argument as in Theorem 2

$$F_{Y_d}(y) = \mathrm{E}[1(Y_d \leq y)] = \mathrm{E}\{\mathrm{E}[1(Y \leq y) \mid X, D = d]\}, \;\; d \in \{0,1\},$$

or, by the same argument as in Theorem 4

$$F_{Y_d}(y) = \mathrm{E}[1(Y_d \leq y)] = \mathrm{E}[1(D = d)1(Y \leq y)/p_d(X)], \;\; d \in \{0,1\}.$$

The τ-quantiles of the potential outcomes are the (left)-inverse of the distributions at τ:

$$Q_{Y_d}(\tau) = F^{\leftarrow}_{Y_d}(\tau) = \inf\{y \in \mathcal{Y} : F_{Y_d}(y) \leq \tau\}, \;\; d \in \{0,1\}, \tau \in (0,1).$$

The difference between the τ-quantiles of the potential outcomes Y_1 and Y_0 yields the τ-quantile treatment effect (τ-QTE)

$$\delta_\tau = Q_{Y_1}(\tau) - Q_{Y_0}(\tau), \;\; \tau \in (0,1).$$

In the training program example, $\delta_{1/2}$ measures the difference in the median wage between the situation where everyone participates in the program and the situation where none participates in the program. Looking at the QTE function $\tau \mapsto \delta_\tau$ we can determine if the treatment effect is heterogenous across the distribution.

We can make inference on the QTE function using the generic methods of L7. These methods convert estimates and confidence bands for distributions functions into estimates and confidence bands for quantile and quantile effects function. To estimate the distributions, we use a GMM approach similar to the previous section. We focus here on the doubly-robust approach. Define

$$p_d(X) := \mathrm{P}(D = d \mid X), \text{ and } \mu_{d,y}(X) := \mathrm{E}[1(Y \leq y) \mid D = d, X], \;\; d \in \{0,1\}, y \in \mathcal{Y}.$$

Let $\theta_{d,y} := F_{Y_d}(y)$. The moment conditions of the doubly robust approach are

$$\mathrm{E}[1(D = d)(1(Y \leq y) - \mu_{d,y}(X))/p_d(X) + \mu_{d,y}(X)] - \theta_{d,y} = 0, \;\; d \in \{0,1\}, y \in \mathcal{Y}.$$

Given the parametrizations

$$p_d(X) = p_d(X, \ell_d), \quad \mu_{d,y} = \mu_{d,y}(X, \beta_{d,y}),$$

the moment function for the GMM approach is

$$g(Z, \theta, \eta) = [1(D = d)(1(Y \leq y) - \mu_{d,y}(X, \beta_{d,y}))/p_d(X, \ell_d) + \mu_{d,y}(X, \beta_{d,y}) - \theta_{d,y}]_{d \in \{0,1\}, y \in \mathcal{Y}},$$

where

$$\theta = [(\theta_{0,y})_{y \in \mathcal{Y}}, (\theta_{1,y})_{y \in \mathcal{Y}}], \quad \eta = [(\beta_{0,y})_{y \in \mathcal{Y}}, (\beta_{1,y})_{y \in \mathcal{Y}}, \ell_0, \ell_1].$$

4. Treatment Effects with Endogeneity

In many observational studies the ignorability assumption is not plausible. Even in RCTs the observational units might decide not to comply with their treatment assignments for reasons related to potential outcomes. For example, individuals randomly assignment to a training program might decide not to participate if they expect that the training is not beneficial to them. In these cases we can still identify average treatment effects in the presence of a variable related to the treatment that is randomly assigned or randomly assigned conditional on covariates. This variable is called the instrument, Z. We focus on the leading case of binary Z. This covers the leading case where Z is a random offer to participate in the treatment. Without ignorability of the treatment, the ATE is generally not identified. We can still identify the average effect for a subpopulation under some conditions on the instrument. To state these conditions it is useful to introduce the latent variables D_0 and D_1, which correspond to the counterfactual treatment assignments when the unit is subjected to offer and no offer of treatment.

Assumption 3 (LATE)**.** *(a) Ignorability. Suppose that the instrument Z is independent of potential outcomes $(Y_d)_{d\in\{0,1\}}$ and potential treatments $(D_z)_{z\in\{0,1\}}$ conditional on a set of covariates X, that is*

$$Z \perp\!\!\!\perp (Y_0, Y_1, D_0, D_1) \mid X.$$

(b) Overlap. Suppose that the propensity score $p(X) := \mathrm{P}(Z = 1|X)$, which is the probability of receiving the offer of treatment given X, is non-degenerate:

$$\mathrm{P}(0 < p(X) < 1) = 1.$$

(c) Monotonicity:

$$\mathrm{P}(D_1 \geq D_0) = 1.$$

(d) First stage:

$$\mathrm{P}(D_1 > D_0) > 0.$$

Part (a) and (b) are conditional exogeneity and overlap conditions on the instrument similar to Assumption 2. Part (c) imposes that the instrument affects the treatment in the same direction for all the units. In other words, it rules out *defiers*, units that take the treatments only when they do not receive the offer. Part (d) is a relevance condition that guarantees a positive mass of *compliers*, units that comply with the treatment offer Z because $D_1 > D_0$. Assumption 3 covers the case where Z is randomly assigned by setting $X = 1$.

Under Assumption 3, the average treatment effect is identified only for the compliers. These are the units for which $D_1 > D_0$, i.e. the treatment status can be manipulated with the instrument. The ATE for the compliers is called the *local average treatment effect* or *LATE*:

$$\delta^c := \mathrm{E}[Y_1 - Y_0 \mid D_1 > D_0].$$

It is local in that it is the ATE for the subpopulation of compliers, which is not observable. [9] and [1] showed that the LATE can be expressed as the ratio of the average treatment effect of Z on Y to the average treatment effect of Z on D. Since the instrument is ignorable with respect to the outcome and treatment, by Theorem 2 the LATE is identified by

$$\delta^c = \frac{\mathrm{E}[\mathrm{E}(Y \mid Z=1,X) - \mathrm{E}(Y \mid Z=0,X)]}{\mathrm{E}[\mathrm{E}(D \mid Z=1,X) - \mathrm{E}(D \mid Z=0,X)]}.$$

In the absence of covariates, the sample analog of the previous expression is the Wald estimator [18], which is the ratio of the coefficients of Z in the reduced form regression of Y on Z to the first stage regression of D on Z,

$$\hat{\delta}^c_{wald} = \frac{\mathbb{E}_n(Y \mid Z=1) - \mathbb{E}_n(Y \mid Z=0)}{\mathbb{E}_n(D \mid Z=1) - \mathbb{E}_n(D \mid Z=0)}.$$

The following result shows that the mass of compliers, averages of potential outcomes for compliers and LATE can all be expressed as functions of potential outcomes and potential treatments defined with respect to the instrument. We use this characterization to construct estimators based on the GMM approaches described in Section 2.

Theorem 5 (LATE). *Under Assumption 3, (a) the mass of compliers is the ATE of the potential assignments with respect to the instrument,*

$$\mathrm{P}(D_1 > D_0) = \mathrm{E}(D_1 - D_0).$$

(b) The averages of the potential outcomes for the compliers can be expressed as a ratio of two ATEs with respect to the instrument,

$$\mathrm{E}(Y_d \mid D_1 > D_0) = \frac{\mathrm{E}[1(D_1=d)Y_d - 1(D_0=d)Y_d]}{\mathrm{E}[1(D_1=d) - 1(D_0=d)]}, \quad d \in \{0,1\}.$$

(c) Let $Y^*_z := D_z Y_1 + (1-D_z)Y_0$, $z \in \{0,1\}$, be potential outcomes defined with respect to the instrument Z. The LATE can be expressed as a ratio of two ATEs with respect to the instrument,

$$\delta^c = \frac{\mathrm{E}(Y^*_1 - Y^*_0)}{\mathrm{E}(D_1 - D_0)}.$$

(d) Hence, the mass of compliers, the average of the potential outcomes for the compliers, and the LATE are identified.

Part (a) follows directly from the monotonicity assumption. For part (b), if $d = 1$

$$\mathrm{E}[1(D_1=1)Y_1 - 1(D_0=1)Y_1] = \mathrm{E}(Y_1 \mid D_1 > D_0)\mathrm{E}[1(D_1=1) - 1(D_0=1)],$$

by monotonicity, where $\mathrm{E}[1(D_1=1) - 1(D_0=1)] \neq 0$ by the first stage condition. The case $d = 0$ follows similarly. Part (c) follows from part (b) since

$$\delta^c = \mathrm{E}(Y_1 \mid D_1 > D_0) - \mathrm{E}(Y_0 \mid D_1 > D_0),$$

and the definition of Y_z^*. Part (d) then follows by

$$[1(D_0 = d), 1(D_1 = d), 1(D_0 = d)Y_d, 1(D_1 = d)Y_d, Y_d^*]_{d\in\{0,1\}} \perp\!\!\!\perp Z \mid X,$$

from the conditional ignorability of the instrument, and by the support condition.

Theorem 5 shows that the LATE is identified by a ratio of ATEs with respect to the instrument. We can therefore use any of the GMM approaches described in Section 2 to estimate and make inference on the LATE. We briefly discuss the approach based on doubly-robust moment functions. Define

$p_z(X) := \mathrm{P}(Z = z \mid X), \mu_z(X) := \mathrm{E}[Y \mid Z = z, X], q_z(X) := \mathrm{E}(D \mid Z = z, X), z \in \{0, 1\}.$

Let $\alpha_z = \mathrm{E}[Y_z^*]$, and $\nu_z = \mathrm{E}[D_z]$. With this notation

$$\delta^c = \frac{\alpha_1 - \alpha_0}{\nu_1 - \nu_0},$$

where α_z and ν_z are averages of potential outcomes defined in terms of the instrument. The moment conditions of the doubly robust approach for these averages are

$$\begin{aligned}\mathrm{E}[1(Z = z)(Y - \mu_z(X))/p_z(X) + \mu_z(X)] - \alpha_z = 0,\\ \mathrm{E}[1(Z = z)(D - q_z(X))/p_z(X) + q_z(X)] - \nu_z = 0.\end{aligned}$$

Given the parametrizations

$$p_z(X) = p_z(X, \ell_d), \quad \mu_z = \mu_z(X, \beta_z), \quad q_z(X) = q_z(X, \lambda_z),$$

the moment function for the GMM approach is

$$g(Z, \theta, \eta) = \left[\begin{array}{c} 1(Z = z)(Y - \mu_z(X, \beta_z))/p_z(X, \ell_z) + \mu_z(X, \beta_z) - \alpha_z \\ 1(Z = z)(D - q_z(X, \lambda_z))/p_z(X.\ell_z) + q_z(X, \lambda_z) - \nu_z \end{array} \right]_{z\in\{0,1\}}$$

where

$$\theta = [\alpha_z, \nu_z]_{z\in\{0,1\}}, \quad \eta = [\beta_z, \lambda_z, \ell_z]_{z\in\{0,1\}}.$$

We can develop a similar GMM approach to estimate and make inference on the averages of the potential outcomes for compliers, $\mathrm{E}(Y_d \mid Y_1 > Y_0)$. Define

$$\begin{aligned} p_z(X) &:= \mathrm{P}(Z = z \mid X), \mu_{z,d}(X) := \mathrm{E}[1(D = d)Y \mid Z = z, X], \\ q_{z,d}(X) &:= \mathrm{P}(D = d \mid Z = z, X), \ \ d \in \{0, 1\}, z \in \{0, 1\}. \end{aligned}$$

Let $\alpha_{z,d} = \mathrm{E}[1(D_z = d)Y_d]$, and $\nu_{z,d} = \mathrm{E}[1(D_z = d)]$. With this notation

$$\mathrm{E}(Y_d \mid D_1 > D_0) = \frac{\alpha_{1,d} - \alpha_{0,d}}{\nu_{1,d} - \nu_{0,d}},$$

where $\alpha_{d,z}$ and $\nu_{z,d}$ are averages of potential outcomes defined in terms of the instrument and assignment. The moment conditions of the doubly robust approach for these averages are

$$\begin{aligned}\mathrm{E}[1(Z = z)(1(D = d)Y - \mu_{z,d}(X))/p_z(X) + \mu_{z,d}(X)] - \alpha_{z,d} = 0,\\ \mathrm{E}[1(Z = z)(1(D = d) - q_{z,d}(X))/p_z(X) + q_{z,d}(X)] - \nu_{z,d} = 0.\end{aligned}$$

Given the parametrizations

$$p_z(X) = p_z(X, \ell_z), \quad \mu_{z,d} = \mu_{z,d}(X, \beta_{z,d}), \quad q_{z,d}(X) = q_{z,d}(X, \lambda_{z,d}),$$

the moment function for the GMM approach is

$$g(Z, \theta, \eta) = \left[\begin{array}{c} 1(Z = z)(1(D = d)Y - \mu_{z,d}(X, \beta_{z,d}))/p_z(X, \ell_z) + \mu_{z,d}(X, \beta_{z,d}) - \alpha_{z,d} \\ 1(Z = z)(1(D = d) - q_{z,d}(X, \lambda_{z,d}))/p_z(X, \ell_z) + q_{z,d}(X, \lambda_{z,d}) - \nu_{z,d} \end{array} \right]_{z,d \in \{0,1\}}$$

where

$$\theta = [\alpha_{z,d}, \nu_{z,d}]_{z,d \in \{0,1\}}, \quad \eta = [\beta_{z,d}, \lambda_{z,d}, \ell_z]_{z,d \in \{0,1\}}.$$

The marginal distributions of the potential outcomes for the compliers, $F^c_{Y_d} := F_{Y_d|D_1 > D_0}$ $d \in \{0, 1\}$, are also identified under Assumption 3. This can be seen directly from Theorem 5(b) replacing Y by the indicators $1(Y \leq y)$ with $y \in \mathcal{Y}$ for a finite set $\mathcal{Y} \subset \mathbb{R}$. As in Section 3, we can invert these distributions to obtain quantiles of the potential outcomes and QTEs for the compliers:

$$Q^c_{Y_d}(\tau) := F^{c\leftarrow}_{Y_d}(\tau), \;\; d \in \{0, 1\}, \tau \in (0, 1).$$

By analogy with the LATE, we will call the τ-QTE for the compliers as *local τ-quantile treatment effects* or *τ-LQTE*:

$$\delta^c_\tau := Q^c_{Y_1}(\tau) - Q^c_{Y_0}(\tau), \;\; \tau \in (0, 1).$$

The estimation of the distribution $F^c_{Y_d}$ is analogous to the estimation of $\mathrm{E}(Y_d \mid D_1 > D_0)$. Thus, define

$$\begin{array}{rcl} p_z(X) & := & \mathrm{P}(Z = z \mid X), \mu_{z,d,y}(X) := \mathrm{E}[1(D = d)1(Y \leq y\} \mid Z = z, X], \\ q_{z,d}(X) & := & \mathrm{P}(D = d \mid Z = z, X), \;\; d \in \{0, 1\}, z \in \{0, 1\}, y \in \mathcal{Y}. \end{array}$$

Let $\alpha_{z,d,y} = \mathrm{E}[1(D_z = d)1(Y_d \leq y)]$, and $\nu_{z,d} = \mathrm{E}[1(D_z = d)]$. With this notation

$$F^c_{Y_d}(y) = \frac{\alpha_{1,d,y} - \alpha_{0,d,y}}{\nu_{1,d} - \nu_{0,d}},$$

where $\theta_{d,z,y}$ and $\nu_{z,d}$ are averages of potential outcomes defined in terms of the instrument. The moment conditions of the doubly robust approach for these averages are

$$\begin{array}{l} \mathrm{E}[1(Z = z)(1(D = d)1(Y \leq y) - \mu_{z,d}(X))/p_z(X) + \mu_{z,d,y}(X)] - \alpha_{z,d,y} = 0, \\ \mathrm{E}[1(Z = z)(1(D = d) - q_{z,d}(X))/p_z(X) + q_{z,d}(X)] - \nu_{z,d} = 0. \end{array}$$

Given the parametrizations

$$p_z(X) = p_z(X, \ell_d), \quad \mu_{z,d,y} = \mu_{z,d,y}(X, \beta_{z,d,y}), \quad q_{z,d}(X) = q_{z,d}(X, \lambda_{z,d}),$$

the moment function for the GMM approach is

$$g(Z, \theta, \eta) = \left[\begin{array}{c} 1(Z = z)(1(D = d)1(Y \leq y) - \mu_{z,d,y}(X, \beta_{z,d}))/p_z(X, \ell_z) + \mu_{z,d,y}(X, \beta_{z,d,y}) - \alpha_{z,d,y} \\ 1(Z = z)(1(D = d) - q_{z,d}(X, \lambda_{z,d}))/p_z(X.\ell_z) + q_{z,d}(X, \lambda_{z,d}) - \nu_{z,d} \end{array} \right]_{z,d \in \{0,1\}, y \in \mathcal{Y}}$$

where

$$\theta = [\alpha_{z,d,y}, \nu_{z,d}]_{z,d \in \{0,1\}, y \in \mathcal{Y}}, \quad \eta = [\beta_{z,d,y}, \lambda_{z,d}, \ell_z]_{z,d \in \{0,1\}, y \in \mathcal{Y}}.$$

5. Impact of 401(k) on Financial Wealth

As a practical illustration of the methods developed in this lecture, we consider estimation of the effect of 401(k) eligibility and participation on accumulated assets as in [1] and [4]. The key problem in determining the effect of participation in 401(k) plans on accumulated assets is saver heterogeneity coupled with the fact that the decision to enroll in a 401(k) is non-random. It is generally recognized that some people have a higher preference for saving than others. It also seems likely that those individuals with high unobserved preference for saving would be most likely to choose to participate in tax-advantaged retirement savings plans and would tend to have otherwise high amounts of accumulated assets. The presence of unobserved savings preferences with these properties then implies that conventional estimates that do not account for saver heterogeneity and endogeneity of participation will be biased upward, tending to overstate the savings effects of 401(k) participation.

We use the same data as [1] and [4]. The data consist of 9,915 observations at the household level drawn from the 1991 SIPP. All the variables are referred to 1990. We use net financial assets (net_tfa) as the outcome variable, Y, in our analysis.[4] Our treatment variable, D, is an indicator for having positive 401(k) balances; and our instrument, Z, is an indicator for being eligible to enroll in a 401(k) plan (e401). Among the $3,682$ individuals that are eligible, $2,594$ decided to participate in the program. The vector of covariates, X, consists of age, income, family size (fsize), years of education (educ), a married indicator, a two-earner status indicator, a defined benefit pension status indicator (db), an IRA participation indicator (pira), and a home ownership indicator (hown). Further details can be found in [4]. [11, 12, 13, 14] and [3] argued that eligibility for enrolling in a 401(k) plan in this data can be taken as exogenous after conditioning on a few observables of which the most important for their argument is income. Table 1 provides descriptive statistics for the variables used in the analysis. The unconditional APE of e401 is $19,559$ and the unconditional APE of p401 is $27,372$.

We first look at the treatment effect of e401 on net total financial assets, i.e. setting $D = Z$. This treatment is usually referred to as the *intention to treat*. Tables 2 and 3 report estimates of the ATE and ATT. The first row in each panel corresponds to a linear model

$$Y_d = d\alpha + f(X)'\beta + \epsilon_z,$$

where in panel A $f(X)$ includes indicators of marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status, and orthogonal polynomials of degrees 2, 4, 6 and 8 in family size, education, age and income, respectively. In panel B we add to $f(X)$ all the two-way interactions. The dimensions of $f(X)$

[4]Net financial assets are computed as the sum of IRA balances, 401(k) balances, checking accounts, saving bonds, other interest-earning accounts, other interest-earning assets, stocks, and mutual funds less non mortgage debts.

TABLE 1. Descriptive Statistics

	Mean	Std. Dev.	e401=1	e401=0	p401=1	p401=0
net_tfa	18,052	63,523	30,347	10,788	38,262	10,890
e401	0.37	0.48	1.00	0.00	1.00	0.15
p401	0.26	0.44	0.70	0.00	1.00	0.00
age	41.06	10.34	41.48	40.81	41.51	40.90
income	37,200	24,774	46,861	31,494	49,367	32,890
fsize	2.87	1.54	2.90	2.84	2.92	2.85
educ	13.21	2.81	13.76	12.88	13.81	12.99
db	0.27	0.44	0.42	0.19	0.39	0.23
married	0.60	0.49	0.67	0.56	0.69	0.57
two-earner	0.38	0.49	0.48	0.32	0.50	0.34
pira	0.24	0.43	0.32	0.20	0.36	0.20
hown	0.64	0.48	0.74	0.57	0.77	0.59

Source: 1991 SIPP.

in panels A and B are 25 and 275. The second and third row in each panel use the doubly robust approach of Section 2 with

$$\mu_d(X, \beta_d) = f(X)'\beta_d, \text{ and } p_d(X, \ell_d) = \Lambda(f(X)'\ell_d),$$

where Λ is the logistic link function and $f(X)$ are the same specifications in panels A and B as for the linear model. In table 2, β, β_0 and β_1 are estimated by least squares, and ℓ_1 is estimated by logit regression.[5] In table 3, β is estimated by double selection by partialing out $f(X)$ from Y and D using Lasso least squares, β_0 and β_1 are estimated by Lasso least squares, and ℓ_1 is estimated by Lasso logit regression.[6]

TABLE 2. Average Treatment Effects of e401 on net_tfa

	Est.	Std. Error	95% LCI	95% UCI
A - Without interactions (25 controls)				
Linear Model	9,003	1,238	6,577	11,429
ATE	6,367	2,012	2,424	10,310
ATT	11,237	1,520	8,258	14,216
B - With two-way interactions (275 controls)				
Linear Model	8,968	1,134	6,745	11,191
ATE	94,608	371,574	-633,663	822,880
ATT	928,407	744,720	-531,218	2,388,032

[5]Note that we do not need to estimate ℓ_0 because $p_0(X, \ell_0) = 1 - p_1(X, \ell_1)$.

[6]We run the Lasso least squares and logit regressions using the `R` package `glmnet` [6].

Controlling for covariates reduces the estimates of the ATE by more than half with respect to the unconditional APE of e401, indicating the presence of selection bias. The linear model produces estimates in between the ATE and ATT. A comparison between the ATE and ATT suggests the presence of heterogeneity in the average effects for treated and non treated, but this evidence is not statistically significant at the 5% level. Panel B of Table 2 shows that the doubly robust approach produces very noisy estimates in the specification with interaction of the controls due to overfitting. This overfitting is reflected in estimates of the propensity score close to 0 or 1. The selection of controls using Lasso regularizes the estimates and produces estimates that are more stable across specifications. Based on the estimates with selection of controls, the average effect of 401 eligibility is roughly $8,000$, increasing to $11,100 - 11,200$ for the treated.

TABLE 3. Average Treatment Effects of e401 on net_tfa with Selection of Controls

	Est.	Std. Error	95% LCI	95% UCI
A - Without interactions (25 controls)				
Linear Model	8,617	1,331	6,008	11,227
ATE	8,070	1,098	5,919	10,221
ATT	11,235	1,540	8,217	14,253
B - With two-way interactions (275 controls)				
Linear Model	8,365	1,326	5,767	10,963
ATE	7,922	1,113	5,741	10,104
ATT	11,116	1,528	8,122	14,111

Figures 1 and 2 report estimates and 95% confidence bands for the QTE and QTT of e401. They are constructed from estimates and confidence bands of distributions of potential outcomes using the method described in L7. The distributions are estimated using the doubly robust approach of Section 3 with the parametrizations

$$\mu_{d,y}(X, \beta_{d,y}) = \Lambda(f(X)'\beta_{d,y}) \quad \text{and} \quad p_d(X, \ell_d) = \Lambda(f(X)'\ell_d),$$

where we consider the same specifications for $f(X)$ as above. We estimate the parameters by logit regressions with and without Lasso selection of controls. The confidence bands are obtained by multiplier bootstrap with 200 and Mammen multipliers.[7]

Looking across the figures, we see a similar pattern to that seen for the estimates of the average effects in that the selection-based estimates are stable across all specifications and are very similar to the estimates obtained without selection from the specification without interactions. If we focus on the QTE and QTT estimated from variable selection methods, we find that 401(k) eligibility has a small impact on accumulated net total financial assets at low quantiles while appearing to have a larger impact at high quantiles. Looking at the uniform confidence intervals, we can see that this pattern is statistically significant at the

[7]The multipliers are drawn as $\omega = 1 + Z_1/\sqrt{2} + (Z_2^2 - 1)/2$, where Z_1 and Z_2 are independent standard normal variables. This multipliers satisfy $\mathrm{E}(\omega) = 0$ and $\mathrm{E}(\omega^2) = \mathrm{E}(\omega^3) = 1$.

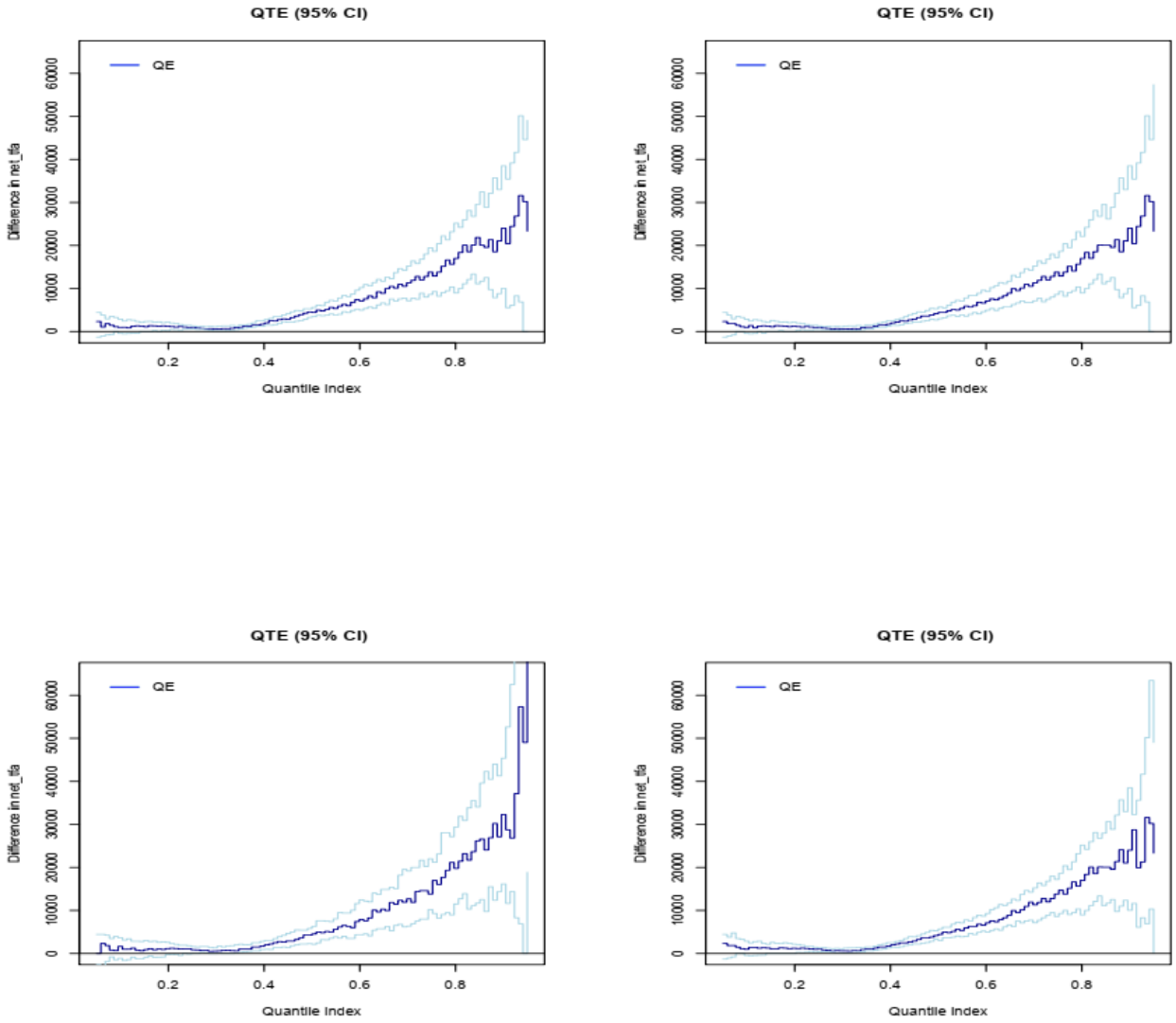

FIGURE 1. Quantile treatment effects of e401 on net_tfa. Panels differ in the specification of $f(X)$ and the estimation method. Upper-left: specification without interactions and no selection of controls. Upper-right: specification without interactions and selection of controls by Lasso. Lower-left: specification with two-way interactions and no selection of controls. Lower-right: specification with two-way interactions and selection of controls by Lasso. Conditional distribution and propensity estimated by logit regression. 95% confidence bands obtained by inversion of 95% joint confidence bands for distributions.

10% level and that we would reject the hypothesis that 401(k) eligibility has no effect and reject the hypothesis of a constant treatment effect more generally.