ECON2390, Applied Econometrics

Problem set 2

1. This question builds on [Angrist and Evans, 1996]. To answer this question you need the replication data that can be downloaded here.

   [Angrist and Evans, 1996] estimates the effect of family size on parental labour supply (outcome) using the 'same-sex' instrument for family size (treatment). Their data is drawn from the US census public use microdata samples (PUMS). The sample used here is from the 1980 census, and consists of married women aged 21-35 with at least two children. The same-sex instrument takes a value of 1 if a parent's two oldest children are of the same sex, and a value of 0 otherwise. [Angrist and Evans, 1996] argue that parent's have a preference for mixed sex children. Hence, parents whose first two children are of the same sex are more likely to have more children. The key identifying assumption for the same-sex instrument to be a valid instrumental variable is that, conditional on observable control variables, it does not directly affect parental labor supply.

   Using a subset of this data, we want to estimate the effect of family size on family income, allowing for heterogeneity of the effects at the family level. In what follows, I list the definitions of variables available in the sample.

   - Y log family income for 1980.
   - D is 'at least three children', i.e., $D = 1$ the family has three or more children, and $D = 0$ otherwise.
   - Z is the same-sex instrument i.e. $Z = 1$ if the mother's oldest two children are of the same sex, and $Z = 0$ otherwise.
   - X consists of a categorical variable that controls for race, the mother's current age, the age of the mother at the time of the birth of her first child, and a dummy for the sex of the oldest child.

   To perform marginal treatment effect (MTE) analysis, consider implementing the following estimation procedure.

   1. Estimate the propensity score by a probit regression

      $$p_i = \Phi(X_i'\gamma + \alpha Z_i), \tag{1}$$

      and let $\hat{p}_i$ be an estimate (fitted value) of the propensity score of unit i.
   2. Next, estimate the following regression equation by OLS:

      $$Y_i = X_i'\beta_0 + \hat{p}_i X_i'\beta_1 + \kappa_1 \hat{p}_i + \kappa_2 \hat{p}_i^2 + \kappa_3 \hat{p}_i^3 + \varepsilon. \tag{2}$$

   Suppose we specify the potential outcome equations as

   $$Y_i(1) = X_i'\theta_1 + \epsilon_{1i}, \tag{3}$$
   $$Y_i(0) = X_i'\theta_0 + \epsilon_{0i}, \tag{4}$$

and the structural selection equation as

$$D_i = 1\{u(X_i, Z_i) \geq V_i\},$$

where $u(X_i, Z_i) - V_i$ is the latent utility parents derive from additional children.

(a) How do you relate the structural selection equation and the probit regression of (1)? Explain your answer.

(b) Clarify a set of assumptions that the marginal treatment effect $M(x, u)$, $u \in (0, 1)$, can be identified by the derivative of the regression equation of (2) with respect to the propensity score:

$$MTE(x, u) = \frac{\partial}{\partial p}\bigg|_{p=u} \left(x'\beta_0 + px'\beta_1 + \kappa_1 p + \kappa_2 p^2 + \kappa_3 p^3\right) \tag{5}$$

(c) Estimate the propensity score as instructed above, and assess whether the coefficient of the instrumental variable in the probit regression is significantly different from zero or not.

(d) Plot a histogram of the estimated propensity scores $\hat{p}_i$.

(e) Report the OLS estimates of the coefficients in the regression equation of (2) and their bootstrap standard errors with 1000 bootstrap replications.

(f) Plot the MTE estimates (as a function of $u \in [0, 1]$) and pointwise confidence intervals, holding X constant at its sample mean values. The confidence intervals are constructed using the boostrap results.

(g) Based on the presented estimates, interpret heterogeneity of the marginal treatment effects. What type of parent has a larger causal effect than others? In your answer, be explicit about how to interpret u.

(h) Based on the estimation results of part (b), test heterogeneity of MTE in u.

(i) Under the specifications of (2) and MTE identification by (5), explain how you identify and estimate the average treatment effect. In your answer, address any concerns about credibility on the ATE estimate.

(j) To assess robustness of the results relying on the parametric form in 2, consider estimating MTE using partial linear regression:

$$Y_i = X_i'\beta_0 + \hat{p}_i X_i'\beta_1 + K(p), \tag{6}$$

where $K(p)$ is a nonparametric function of the propensity score. Maintain the propensity score estimates as before, estimate MTE by running a partial linear regression on . In your answer, explain how you implement the partial linear regression and how you choose the smoothing parameters (bandwidth).

2. For this exercise you have to simulate the following DGP for $n = 500$ observations[1]:

$$X = (X_1, X_2, ..., X_{20})' \quad X \sim \mathcal{N}(0, \Sigma) \tag{7}$$
$$\Sigma_{ii} = 1 \quad \Sigma_{ij} = 0.8^{|j-i|}$$

[1] If you're using R, please *setseed*(1234) at the beginning of your code so everyone is using the same DGP

$$Y = g_0(X, D) + \varepsilon \quad \varepsilon \sim \mathcal{N}(0, 1) \tag{8}$$

$$V = m_0(X) + \zeta \quad \zeta \sim \mathcal{N}(0, 1) \tag{9}$$

$$D = \mathbb{1}[V > 0] \tag{10}$$

$$m_0(X) = a_0 X_1 + a_1 \frac{\exp(X_3)}{(1 + \exp(X_3))}; \quad \text{where } a_0 = 1 \text{ and } a_1 = 0.25$$

$$g_0(X, D) = \tau D + b_0 \frac{\exp(X_1)}{(1 + \exp(X_1))} + b_1 X_3; \quad \text{where } b_0 = 1 \text{ , } b_1 = 0.25 \text{ and } \tau = 0.5$$

(a) Estimate $\theta_{\text{ATE}_{\text{probit}}} = \text{E}\left[\frac{DY}{p(X)} - \frac{(1-D)Y}{(1-p(X))}\right]$ using a probit model to estimate the propensity score $p(X) = \text{E}[D|X]$ in the first step[2].

(b) Estimate $\theta_{\text{ATE}_{\text{naive}}} = g(X, 1) - g(X, 0)$ using random forest to estimate the two nuisance functions in the first step as $g(X, D) = \mathbb{E}[Y|Z, D]$.

(c) Estimate the ATE using the doubly robust estimator:

$$\theta_{ATE_{DR}} = E\left[\frac{(Y - g(X, 1))D}{p(X)} - \frac{(Y - g(X, 0))(1 - D)}{1 - p(X)} + (g(X, 1) - g(X, 0))\right]$$

where the nuisance functions $p(X) = E(D|X)$[3] and $g(X, D) = \mathbb{E}[Y|Z, D]$ are estimated using random forest as well.

(d) Estimate the ATE using the doubly robust estimator as before but compute the first step using cross-fitting for $I = 2$ sample split in the following way.

i. Divide your sample in two folds $j = 1, 2$

ii. Estimate the nuisance functions for each fold j as $\hat{\eta}_j = (\hat{g}(X, 0)_j, \hat{g}(X, 1)_j, \hat{p}(X)_j)$

iii. Compute $\hat{\theta}_{ATE_j}$ for observations in fold j using the First Step functions estimated from the other sample $\hat{\eta}_{-j}$:

iv. Compute $\hat{\theta}_{ATE_{DD}} = \frac{1}{2}\sum_{j=1}^{2} \hat{\theta}_{ATE_j}$.

(e) Repeat the process for $M = 1, ..., 1000$ Montecarlo Simulations and compute the four estimators for each sample.[4]

(f) Compute the empirical variance of each of your estimators $\hat{\sigma}^2_{\hat{\theta}_k}$ $k = \{probit, naive, DR, DD\}$

(g) Create four plots, one for each estimator, containing the histogram of the $\hat{\theta}_{ATE_k}$ and a Normal(0.5,$\hat{\sigma}^2_{\hat{\theta}_k}$) pdf.

(h) Present a table comparing the bias and mean square error of each estimator and comment on which estimator performs better and why.

References

[Angrist and Evans, 1996] Angrist, J. and Evans, W. N. (1996). Children and their parents' labor supply: Evidence from exogenous variation in family size.

[2] For this and the following points include all the covariates X in the model

[3] To ensure overlapping exclude observations with propensity score outside the range $[\alpha, 1 - \alpha]$ for $\alpha = 0.01$

[4] Given that the random forest algorithm takes sequential random samples for each simulation, this process can be time-consuming. In order to save computational time you can try parallel computing. Check library(parallel) for mac or library(doParallel) for general OS