Selection on Observables, Propensity Score and Nonparametrics

Part I - Selection on observables and introduction to propensity score

In this first part, we are interested in evaluating the effect of the program on fertilizer use after the treatment using parametric estimation methods.

1. Provide OLS estimates of the treatment effects using the following samples

   (a) All farms;

   (b) Only farms from regions where the pilot was implemented ($eligible = 1$).

   Discuss your results. How can you explain the difference in treatment effects between the two groups?
   Knowing that the true ATE is -10, how can you explain the remaining difference between your results with sample (b) and the true effect?
   According to the first estimate, treated farms use 35 Kg/Ha less fertilizer in 2020. This estimate falls to -15.9 when considering only farms from eligible regions. The difference with the true effect ($\beta = -10$) is still somewhat large. Both samples lead to overestimate (in magnitude) the effect of the policy.

   The difference between estimates from a) and b) can be explained by a difference in baseline fertilizer use across regions (eligible regions using on average less fertilizer). The fact that the estimate in b) is still biased upwards could be explained by a selection in applications to treatment: farms who applied early enough to be treated may tend to be different from others.

2. Compare the characteristics of the treated farms with non-treated eligible farms. Which differences are the most striking?
Comparing treatment and control group, participating farmers are on average younger (37 vs 53 years old), have bigger farms (57 vs 48 Ha), more likely to be registered to the regional farm bureau (34% vs 28%). They are also more likely to be located in protected areas and be labeled. Participants were also already using less fertilizer in 2018 (67 vs 72 Kg/Ha).

3. Estimate the treatment effect controlling for relevant observables.
Controlling for farmer's age, past fertilizer usage, the size of the farm, the number of machines used, labeled and protected location farms and whether the individual is registered to the farm bureau, the effect is $\hat{\beta}_C = -10.85$. Comparing our estimates to the true effect, the difference is not very large.

4. (a) What is the propensity score? Discuss the advantages of using propensity score instead of observables to control for selection.
The propensity score allows to gather all the relevant information contained in the observable characteristics into a one dimensional variable. When controlling for all relevant observables, the number of cells (strata) increases exponentially with the dimension of the characteristics. When the dimension is large, there might not be enough variation in the treatment variable within each strata such that some only have one type of individuals.

   (b) Estimate the propensity score using a logit model of program participation.
Here, we estimate the following probability model

$$D_i = \mathbf{1}(\alpha_0 + \alpha_1 age_i + \alpha_2 size_i + \alpha_3 machines_i + \alpha_4 chamber_i + \alpha_5 protected_i + \alpha_6 label_i + \alpha_7 fertilizer_i^{2018} + \varepsilon_i) \quad (1)$$

Assuming logistic distribution of the error terms, we can estimate the propensity score by maximum of likelihood.

   (c) Check the common support assumption with summary statistics and plots. Interpret the results.
The common support assumption is satisfied if the probability to be treated conditional on observables is **strictly** higher than 0 and lower than 1 for all observations. Loosely speaking, it ensures that there is enough overlap between the characteristics of the treated and the control groups to find comparable subgroups.
Denoting $\hat{p}_t$ the estimated propensity score, the treatment effect is computed by averaging $(Y_1 - Y_0|\hat{p}_t)$ over the distribution of $\hat{p}_t$. Detailed statistics show that more than 99% of observations in the treatment group are included in the support for the propensity score of the control group. As expected, the distributions are very different with a vast majority of individuals in the control group having a small propensity score.

   (d) Estimate the treatment effect using the (parametric) propensity score as a control.
$\hat{\beta}_{\hat{p}_t} = -11.31$ The estimates is not statistically different from the true parameter value, $\beta = -10$.

5. Is propensity score better than controlling for observables in our situation? Why?
In our example, using all our explanatory variables as controls yields an estimate which is both closer to the true value, and more precise than replacing controls with a parametric propensity score. Note that standard errors of the regression using the PS are underestimated, since they do not take into account the fact that the PS used is itself an estimate.

This lower performance of the PS regression is due to the fact that the number of our variables and the quantity of stratas that they generate do not lead dimensionality issues. If it did, we would not have enough variation in treatment status in each strata when regressing our outcome variable on treatment and control variables. In this case, replacing controls by a PS would yield an estimate of ATE which is likely to be closer to the true value than a regression on all controls.
Using the propensity score presents several drawbacks. First, it forgoes information about the detail of each observation's covariate values - by summarizing the information into a scalar it replaces a multi-dimensional set of information by a scalar. Moreover, note that the use of the propensity score as a control relies on (i) its specification being correct and (ii) the effect of the propensity score on outcome being linear. Finally, as mentioned before, the PS is estimated and therefore when failing to adjust standard errors accordingly in the main regression we necessarily yield wrong standard errors.

Part II - Non-parametric regressions

In this part, we look into the effect of the quantity of fertilizer used on carrot production.

1. Regress carrot production on fertilizer use using a linear parametrization.

2. (a) Graph fitted values on fertilizer use.
   We run the following regression:

$$carrots_i = \alpha + \beta fertilizer_i + \varepsilon_i$$

   (b) Let's write:

$$carrots_i = g(fertilizer_i) + \varepsilon_i$$

   What's the implicit assumption about the shape $g(.)$ in the regression we ran in question 1?
   $g(.)$ is assumed to be linear.

3. Consider the command *npregress kernel*. What is it computing?
   This command computes a non-parametric local-linear regression using kernel smoothing. Observations are put in bins according to their value of the controls X. A weighted linear regression is run within each bin where weights are computed based on an observation's distance to the center of the bin - the point such that $X = x$.
   Practical implementation of non-parametric estimation requires making a tradeoff between bias and variance when choosing the size of the bandwidth. A small bandwidth will provide imprecise estimates because a given bin will not contain many observations. A large bandwidth will increase the size of the bias because a given bin will include observations which are quite different from one another, and treat them as comparable. Note that the Stata command *npregress* is using the bandwidth which minimizes the mean squared error of the prediction.

4. Regress carrot production on fertilizer use using this command.

5. Graph the relationship between carrot production and fertilizer use. What do you conclude about the assumptions mentioned in question 2?
   $g(.)$ is not linear, nor quadratic etc. When using a non-parametric regression, $g(.)$ is allowed a lot of flexibility.

6. (a) Regress linearly carrot production on all available covariates.

   (b) Do the same using a local-linear kernel estimation.

   (c) Which method has the highest R^2? What does this mean?
   The R^2 measure how well our model fits the observations. In other terms, it measures how much of the variation of *carrots_2018* we can explain using our model. If the shock explains most of the variation, the R^2 will be low. The more our model can explain of the variation of the dependent variable, the higher the R^2.
   Here, the non-parametric model seems to have a higher R^2. That was to be expected, since the non-parametric model allows for more flexibility, and therefore removes the linearity constraint from the initial model.

7. What are the pros and cons of using nonparametric regression methods?

   - Pros:
     - Nonparametric methods are more flexible, as they allow the relationship between explanatory variables and the dependent one to take any shape. This reduces the risk of misspecification.
     - As a result of this flexibility, predictions from a nonparametric model tend to be better than parametric ones. The R^2 from a non-parametric regression is generally higher than for a parametric one.

   - Cons:
     - A non-parametric regression needs a bigger sample since it requires enough observations in each bin.
     - A model that fits the data too well can fall into a trap called *overfitting*. It means that while estimating the shape of the correlation between an explanatory variable and the dependent variable, it captures patterns which are specific to the sample you estimate the model on. If you then try to predict values of the dependent variable on another sample (say, you want to predict the impact of fertilizer use on carrot production in another country), you could have a sizable bias.

Part III - Propensity Score Matching

In this part, we switch the focus back to the impact of pilot on fertilizer use, using insights from non-parametric regressions.

1. (a) Estimate again the propensity score using logistic regression and age, $fertilizer\_2018$, $size$, $chamber$, $protected$ and $label$ as regressors.

   (b) Estimate the propensity score non-parametrically using kernel estimation and the same set of explanatory variables. Check the common support assumption for this specification.

   85% of observations in the treatment group are within the support of the control group. The propensity score distribution is still very different between the two groups.

2. Perform nearest neighbor propensity score matching on the common support with replacement using each of the propensity scores computed in (1).
   Hint: To perform propensity score matching, you can install the **psmatch2** package from stata. Ignore the "sort order' warnings.
   The common support drops members of the treatment group that are not on the same support as the control group. Using parametric propensity score, we drop 5 observations. The estimated effect of the treatment $\hat{\beta}^{PSM}_{\hat{p}} = -13.75$ is biased with respect to the true effect of the treatment but precisely estimated.
   With non parametric propensity score, we drop 140 observations. The estimate is close to the true value $\hat{\beta}^{PSM}_{\hat{p}_{NP}} = -9.92$ and precisely estimated.

3. Test the balancing of the matching.
   Hint: Use **pstest** command after matching.
   Pstest is used to evaluate the balancing of a set of variables (that you specify in varlist) between the treatment and the control observations. This comparison refers to the latest propensity score matching model that you estimated using psmatch2.
   Specification 1 (using logit propensity score) lead to an acceptable balance (this statement is subjective) with an average bias of 6.7% with the largest bias being in terms of machines (9.7%). Specification 2 (using non parametric propensity score) also lead to an acceptable balance. The largest difference occurs for age (25.7%).

4. Compare your estimates of ATT in parts I and III, what would be our preferred specification?
   The different specifications and estimation procedures provided the following estimates: $\hat{\beta}_C = -10.8$, $\hat{\beta}_{\hat{p}} = -11.31$, $\hat{\beta}^{PSM}_{\hat{p}} = -13.75$ and $\hat{\beta}^{PSM}_{\hat{p}_{NP}} = -9.92$. Knowing the true effect is $\beta = -10$, the less bias estimate uses nearest neighborhood PS matching with a non-parametric estimation of the propensity score. Note that in this example, nearest neighborhood PS matching provide precise estimates. In other context, it may be worth considering multiple matching (1 treatment observation matched with several control ones), which may increase the bias (since less similar observations would be compared) but would reduce the noise, providing more precise estimates.

5. What are the motivations behind the use of matching methods? How do they interact with the choice of PS specification?
   Matching methods are non-parametric, and as such they allow for a lot of flexibility in the relationship between each covariate and the dependent variable. Matching could be performed on all covariates (a bin being close to the notion of strata defined in Part I). Unfortunately, this soon hits the curse of dimensionality, where there aren't enough observations within each strata. This is what motivates the use of PS as the variable to perform the matching on. However, the PS too has to be estimated, and there is no reason it should not hit the curse of dimensionality as well if estimated non-parametrically. This is why parametric methods are often used to estimate the PS.

Appendix

A List of variables

The following variables are available to the researcher:

- **ID**: Farm identifier;
- **carrots**: Carrots production (in tons);
- **region**: region in which the farm is located;
- **age**: age of the main farmer;
- **age_cat**: group farmer's age in 6 categories;
- **size**: farm's land size (in ha);
- **machines**: number of agricultural equipment available to the farmer;
- **seed**: seed quality (from 1 to 5);
- **protected**: indicator, 1 if located in protected area;
- **label**: indicator, 1 if the farm received a care label;
- **chamber**: indicator, 1 if registered at the regional farm bureau;
- **fertilizer_2018**: pollutant fertilizer use in 2018 (in kg/ha), prior to the pilot;
- **fertilizer_2018_cat**: group farms from low to intensive pollutant fertilizer users (7 categories), in 2018;
- **fertilizer_2020**: pollutant fertilizer use in 2020 (in kg/ha), during to the pilot;
- **fertilizer_2020_cat**: group farms from low to intensive pollutant fertilizer users (7 categories), in 2020;
- **eligible**: indicator, 1 if eligible to the treatment;
- **D**: indicator, 1 if received treatment.

B Data generating process (DGP)

The implicit variable determining treatment status X_i is generated as

$$X_i = \begin{cases} 4 * age_i & -0.1 * age_i^2 + 0.2. * size + 15 * chamber_i + 0.02 * fertilizer\_2018_i \\ & -0.002 * fertilizer\_2018_i^2 + 3 * protected + 1 * label + \nu_i \quad \text{if } eligible_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

where $\nu_i \sim \mathcal{N}(0, 50)$.

Therefore, treatment status is determined by the following rule:

$$D_i = \begin{cases} 1 & \text{if } X_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

The use of fertilizer in 2020 is generated according to the following process:

$$fertilizer\_2020_i = 45 + fertilizer\_2018_i - 10 * D_i - 30 * eligible_i + \eta_i$$

where $\eta_i \sim \mathcal{N}(0, 15)$.

The production of carrots is generated according to the following process:

$$\begin{aligned} carrots\_2018_i =& 2 * fertilizer\_2018_i - 0.005 * fertilizer\_2018_i^2 + 1 * machines_i \\ &+ \sum_{k \in \text{Seed Qualities}} \delta_k \mathbb{1}\{seed_i = k\} + \sum_{r \in \text{Regions}} \gamma_r \mathbb{1}\{region_i = r\} + 0.5 * size_i \\ &- 0.01 * size_i^2 - 0.00002 * size_i^3 + \varepsilon_i \end{aligned}$$

where $\varepsilon_i \sim \mathcal{N}(0, 500)$, $\delta = (0, 0, 18, 36, 3)$, and $\gamma = (0, 21, 7, 0, 0, 7, 21, 7, 7, 35, 7, 35, 35)$.