ESTIMATION OF TREATMENT EFFECTS VIA PROPENSITY SCORES

AAEC 5126
INSTRUCTOR: KLAUS MOELTNER

Textbooks: Wooldridge (2010), Ch.21; Greene (2012), Ch.19;
Angrist and Pischke (2010), Ch. 3
`R` scripts: `mod5s2`

Estimating the propensity score (PS)

The first step in any PS-based approach is the estimation of the PS itself, i.e. to derive $\hat{p}(\mathbf{x}_i)$, $i = 1 \dots N$. Note that this needs to be done for *all* observations in the sample, treated or not.

In this step the binary treatment indicator is de facto the "dependent variable". The explanatory variables are the original $\mathbf{x}$ (the same we would use in a regression-based approach), plus, optionally, additional squared and interaction terms, similar to the "auxiliary regression" we used for the White's heteroskedasticity test. The addition of these extra terms was suggested by Rosenbaum and Rubin (1983) to add flexibility to the PS model.

The PS step is usually based on some *Generalized Linear Model (GLM)* that combines a continuous density function defined over $\{0, 1\}$ with a linearly parameterized mean function. Perhaps the most common choice is a *logit* model (e.g. Ch. 17 in Greene, 2012). It is given as

$$p(\mathbf{x}_i) = p(y_i = 1|\mathbf{x}_i) = \frac{exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + exp(\mathbf{x}_i'\boldsymbol{\beta})} \tag{1}$$

where $\boldsymbol{\beta}$ are the model coefficients to be estimated. This model is straightforward to estimate via MLE with known gradient and Hessian (see section 17.3 in Greene, 2012). Predicted probabilities are obtained by replacing $\boldsymbol{\beta}$ with the estimated coefficients $\hat{\boldsymbol{\beta}}$ in (1). As illustrated in script `mod5s2`, we can use `R` 's built-in `glm(.)` function to estimate the logit model, followed by `predict(.)` to obtain the predicted PS's for each observation in the sample.

As described in Wooldridge (2010), section 21.3.3, nonparametric methods are also available to estimate $p(\mathbf{x}_i)$.

Using the PS to check on overlap

As in the regression adjustment case, we are concerned about overlap of regressors between the sub-sample of treated and controls. Once you have $\hat{p}(\mathbf{x}_i)$ for each observation, you can use it to

check on overlap. Simply examine the sample statistics of $\hat{p}(\mathbf{x}_i)$, especially the lower and upper percentiles.

As mentioned in Wooldridge (2010), PS's below 0.1 and above 0.9 may be indicative of poor overlap with respect to the estimation of ATE. For ATT, only extremely high score, say 0.9 or above, pose reasons for concern, as discussed in the previous lecture notes. In that case, the analyst may again redefine the population of interest.

Other uses of $\hat{p}(\mathbf{x}_i)$ to examine overlap are via its normalized difference (ND, see regression adjustment notes) or the ND of its log-odds ratio, given as

$$ln\left(\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)}\right) = \mathbf{x}_i'\boldsymbol{\beta} \tag{2}$$

Using ND, values above 0.25 would raise a red flag, as suggested by Imbens and Rubin (forthcoming).

Estimating treatment effects

With $\hat{p}(\mathbf{x}_i)$ in hand, we can directly apply the expressions from the previous lecture notes to obtain consistent estimates of ATE and ATT, specifically:

$$\begin{aligned}\hat{\tau}_{ate} &= n^{-1}\sum_{i=1}^{n}\frac{(w_i-\hat{p}(\mathbf{x}_i))\,y_i}{\hat{p}(\mathbf{x}_i)\,(1-\hat{p}(\mathbf{x}_i))}\\ \hat{\tau}_{att} &= n^{-1}\sum_{i=1}^{n}\frac{(w_i-\hat{p}(\mathbf{x}_i))\,y_i}{\hat{\rho}\,(1-\hat{p}(\mathbf{x}_i))}\end{aligned} \tag{3}$$

where $\hat{\rho}$ is estimated as n_1/n, i.e. the proportion of treated in the sample. Note again that the summation is taken over the *entire sample* for *both estimators*.

Wooldridge (2010) discussed several approaches to obtain standard errors for these estimates. The *bootstrap* is again an attractive strategy, as it controls for sampling error in *both* stages of the analysis. This is the approach taken in script `mod5s2`.

As you can easily see from the expressions in (3), very large or very small values for $\hat{p}(\mathbf{x}_i)$ can lead to very large entries in the summation for $\hat{\tau}_{ate}$. For $\hat{\tau}_{att}$, only very large values of $\hat{p}(\mathbf{x}_i)$ are of concern.

The PS in regression adjustments

We can also combine regression adjustment (discussed in the previous lecture notes) and PS-based estimation. The logic of this approach is based on Rosenbaum and Rubin (1983), who show that, under Assumption [1] (conditional ignorability of treatment with respect to outcome), treatment is ignorable conditional only on the propensity score $p(\mathbf{x})$. Specifically, under Assumption [1], we

have:

$$\begin{aligned} E\left(y|w=0,p\left(\mathbf{x}\right)\right) &= E\left(y_0|p\left(\mathbf{x}\right)\right), \quad \text{and} \\ E\left(y|w=1,p\left(\mathbf{x}\right)\right) &= E\left(y_1|p\left(\mathbf{x}\right)\right) \end{aligned} \tag{4}$$

The proof proceeds in 3 steps:

Step 1: by Assumption [1], we can write:

$$E\left(w|y_0,y_1,\mathbf{x}\right) = E\left(w|\mathbf{x}\right) \tag{5}$$

Step 2: Using the law of iterated expectation:

$$\begin{aligned} E\left(w|y_0,y_1,p\left(\mathbf{x}\right)\right) =& E_{\mathbf{x}}\left(E\left(w|y_0,y_1,p\left(\mathbf{x}\right),\mathbf{x}\right)\right) = \\ & E_{\mathbf{x}}\left(E\left(w|y_0,y_1,\mathbf{x}\right)\right) = E_{\mathbf{x}}\left(E\left(w|\mathbf{x}\right)\right) = \\ & E_{\mathbf{x}}\left(p\left(\mathbf{x}\right)\right) = p\left(\mathbf{x}\right) \end{aligned} \tag{6}$$

where $p\left(\mathbf{x}\right)$ can be dropped from the inner expectation in the second line since it is deterministic (thus redundant), given $\mathbf{x}$, and we use the result from step 1 in line 3. The last line uses the alternative way of writing the propensity score, that is as $E\left(w|\mathbf{x}\right)$. The outer expectation then becomes the expectation of a constant, which of course is just that constant. Step 2 thus implies that, conditional on the PS, w is independent of y. Naturally, this also holds the other way round, i.e. conditional on the PS, y is independent of w.

Taken together, steps 1 and 2 show that ignorability of w holds conditioning only on the propensity score. The last step uses $y = (1-w)\,y_0 + wy_1$ to establish the final result:

$$\begin{aligned} E\left(y|w,p\left(\mathbf{x}\right)\right) =& (1-w) * E\left(y_0|w,p\left(\mathbf{x}\right)\right) + w * E\left(y_1|w,p\left(\mathbf{x}\right)\right) = \\ & (1-w) * E\left(y_0|p\left(\mathbf{x}\right)\right) + w * E\left(y_1|p\left(\mathbf{x}\right)\right) \end{aligned} \tag{7}$$

where we use the result from step 2 in the second line. The results in (4) then follow immediately by setting w to either 0 or 1.

Given these results, we can identify τ_{ate} via $E_{\mathbf{x}}\left(E\left(y_1|p\left(\mathbf{x}\right)\right) - E\left(y_0|p\left(\mathbf{x}\right)\right)\right)$. Letting $r_g\left(p\left(\mathbf{x}_i\right)\right) = E\left(y_g|p\left(\mathbf{x}_i\right)\right), g = 0, 1$ be some estimable regression function, we obtain

$$\begin{aligned} \hat{\tau}_{ate} &= n^{-1}\sum_{i=1}^{n} \hat{r}_1\left(\hat{p}\left(\mathbf{x}_i\right)\right) - \hat{r}_0\left(\hat{p}\left(\mathbf{x}_i\right)\right), \quad \text{and} \\ \hat{\tau}_{att} &= n_1^{-1}\sum_{i=1}^{n} w_i * \left(\hat{r}_1\left(\hat{p}\left(\mathbf{x}_i\right)\right) - \hat{r}_0\left(\hat{p}\left(\mathbf{x}_i\right)\right)\right), \end{aligned} \tag{8}$$

As before, if we believe that the effect of $p(\mathbf{x})$ on the outcome is identical for treated and controls, we can run a single pooled regression of the outcome y_i on the treatment w_i and $\hat{p}(\mathbf{x}_i)$ instead. Essentially, this corresponds to the *pooled* regression approach discussed previously, replacing $\mathbf{x}_i'\boldsymbol{\beta}$ with the estimates propensity score. In that case, we have again $\hat{\tau}_{ate} = \hat{\tau}_{att}$.

The (seeming) attraction of regression on the propensity score is that it preempts the need to specify exactly how the potentially high-dimensional set of covariates $\mathbf{x}$ should enter the regression function. However, the same dilemma surfaces for the specification of the PS model, as discussed above. Thus, little is gained in that respect.

Furthermore, as discussed in Wooldridge (2010), regression on the propensity score is generally *inefficient* compared to the standard regression adjustment approach.

References

Angrist, J. D. and Pischke, J. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics, *Journal of Economic Perspectives* **24**: 3–30.

Greene, W. (2012). *Econometric Analysis*, 7th edn, Pearson / Prentice Hall.

Imbens, I. and Rubin, D. (forthcoming). *Causal Inference in Statistics and the Social Sciences*, Cambridge University Press.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**: 41–55.

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*, MIT Press.