ESTIMATION OF TREATMENT EFFECTS VIA REGRESSION

AAEC 5126
INSTRUCTOR: KLAUS MOELTNER

Textbooks: Wooldridge (2010), Ch.21; Greene (2012), Ch.19;
Angrist and Pischke (2010), Ch. 3
R scripts: `mod5s1`

General Approach

The regression functions $m_0(\mathbf{x}) = E(y|\mathbf{x}, w = 0)$ and $m_1(\mathbf{x}) = E(y|\mathbf{x}, w = 1)$ can be directly used to estimate ATE and ATT.

Given a random sample of size $n = n_0 + n_1$, where n_0 is the size of the sub-sample of untreated observations, and n_1 denotes the size of the sub-sample of treated observations, consistent *regression adjustment* estimators of ATE and ATT can be derived via

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

where $\hat{m}_0(\mathbf{x})$ and $\hat{m}_1(\mathbf{x})$ are themselves consistent estimators of $m_0(\mathbf{x})$ and $m_1(\mathbf{x})$, respectively.

In general, the implementation steps are as follows:

(1) Estimate $\hat{m}_0(\mathbf{x})$ from the control sample via parametric or nonparametric regression.
(2) Estimate $\hat{m}_1(\mathbf{x})$ from the treated sample in similar fashion.
(3) For *all i in the sample*, compute $\hat{y}_{i0}$ and $\hat{y}_{i1}$ as fitted values from the two regression models, and take the difference.
(4) Average the difference over the entire sample for ATE, and over the sub-sample of the treated for ATT.
(5) Derive standard errors for these effects via analytical methods or bootstrapping.

Lets' call the model coefficients for the "control" model $\boldsymbol{\theta}_0$, and those for the the treated model $\boldsymbol{\theta}_1$. Running two separate regressions for the two sub-samples implies that we ex ante allow the effect of $\mathbf{x}$ to vary across the two groups, i.e $m_0(\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_0)$ and $m_1(\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_1)$.

If the coefficients are assumed to be the same in both models, a single "pooled" regression can be estimated instead. In that case, the ATE (which will be equal to the ATT) can be estimated as the coefficient of a treatment "dummy" variable that is added to the set of explanatory variables. The standard error of the effect is simply the standard error of that coefficient. Robust methods guarding e.g. against heteroskedasticity can be applied as usual.

Finally, if $\mathbf{x}$ are of secondary importance for identification (for example in randomized experiments), an estimate for the ATE (equal to ATT) can be obtained by simply taking the difference of the sample means of outcomes for the treated and controls.

Script `mod5s1` illustrates these various approaches.

Notice that for these regression approaches to be feasible we need to observe *the same set of covariates* for the entire sample. That is, we cannot have missing values for $\mathbf{x}$ (and, of course w) in the sample at large.

Checking for overlap

As noted in Wooldridge (2010), p. 917, Imbens and Rubin (forthcoming) suggest to use *normalized differences* (NDs) for each explanatory variable to verify the overlap assumption. For a given regressor k the ND value can be computed as

$$ND_k = \frac{(\bar{x}_{1k} - \bar{x}_{0k})}{\left(s_{1k}^2 + s_{0k}^2\right)^{1/2}} \tag{2}$$

where $\bar{x}$ is the sample mean, and s denotes the sample standard deviation. Imbens and Rubin (forthcoming) suggest any value for ND_k exceeding 0.25 a cause for concern.

If overlap is blatantly violated for some $\mathbf{x}$ it may be necessary to re-define the population of interest, as discussed in Wooldridge (2010), pp. 916-917. In fact, if we're primarily interested in the ATE for a specific subpopulation $\mathcal{R}$, overlap problems can be avoided by simply estimating the regression function(s) using only the restricted sample.

Linear regression

If the regression adjustment takes a (standard) linear form, we have

$$\begin{aligned} m_0(\mathbf{x}, \boldsymbol{\theta}_0) &= \alpha_0 + \mathbf{x}'\boldsymbol{\beta}_0, \\ m_1(\mathbf{x}, \boldsymbol{\theta}_1) &= \alpha_1 + \mathbf{x}'\boldsymbol{\beta}_1, \end{aligned} \tag{3}$$

Expressions for ATE and ATT then simplify to

$$\begin{aligned}\hat{\tau}_{ate} =& n^{-1}\sum_{i=1}^{n}\left(\hat{m}_1\left(\mathbf{x}_i,\hat{\boldsymbol{\theta}}_1\right)-\hat{m}_0\left(\mathbf{x}_i,\hat{\boldsymbol{\theta}}_0\right)\right)= \\ & n^{-1}\sum_{i=1}^{n}\left(\hat{\alpha}_1-\hat{\alpha}_0+\mathbf{x}_i'\left(\hat{\boldsymbol{\beta}}_1-\hat{\boldsymbol{\beta}}_0\right)\right)= \\ & \hat{\alpha}_1-\hat{\alpha}_0+n^{-1}\sum_{i=1}^{n}\left(\mathbf{x}_i'\left(\hat{\boldsymbol{\beta}}_1-\hat{\boldsymbol{\beta}}_0\right)\right)= \\ & \hat{\alpha}_1-\hat{\alpha}_0+\bar{\mathbf{x}}'\left(\hat{\boldsymbol{\beta}}_1-\hat{\boldsymbol{\beta}}_0\right)\end{aligned} \tag{4}$$

where $\bar{\mathbf{x}}$ is the vector of sample means for the regressors. For ATT, simply substitute $\bar{\mathbf{x}_1}$, the sample means for the treated only, in lieu of $\bar{\mathbf{x}}$.

As mentioned above, standard errors for the estimated treatment effects can be obtained via bootstrapping - as illustrated in script `mod5s1`. This procedures automatically controls for uncertainty in both $\hat{\boldsymbol{\theta}}_g$, $g = 1, 0$, and the distribution of $\mathbf{x}$ in the population.

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.

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