Notes on causal effects*

Johan A. Elkink

March 4, 2013

1 Decomposing bias terms

Deriving Eq. 2.12 in Morgan and Winship (2007: 46):

$$\text{Potential outcome} = \begin{cases} Y_{1i} & \text{if } T_i = 1 \\ Y_{0i} & \text{if } T_i = 0 \end{cases}$$

Using shortcut $E_{11} = E[Y_{1i}|T_i = 1]$, $E_{01} = E[Y_{0i}|T_i = 1]$, etcetera:

$$\begin{aligned}
E[\delta] =& \{\pi E_{11} + (1-\pi)E_{10}\} - \{\pi E_{01} + (1-\pi)E_{00}\} \\
=& \pi\{E_{11} - E_{01}\} + (1-\pi)\{E_{10} - E_{00}\} \\
=& \pi E_{11} - \pi E_{01} + (1-\pi)E_{10} - (1-\pi)E_{00} \\
=& E_{11} - (1-\pi)E_{11} - E_{01} + (1-\pi)E_{01} + (1-\pi)E_{10} - E_{00} + \pi E_{00} \\
E_{11} - E_{00} =& E[\delta] + (1-\pi)E_{11} + E_{01} - (1-\pi)E_{01} - (1-\pi)E_{10} - \pi E_{00} \\
=& E[\delta] + (1-\pi)E_{11} + E_{01} - (1-\pi)E_{01} - (1-\pi)E_{10} - E_{00} + (1-\pi)E_{00} \\
=& E[\delta] + (E_{01} - E_{00}) + (1-\pi)\{(E_{11} - E_{01}) - (E_{10} - E_{00})\} \\
=& E[\delta] + (E_{01} - E_{00}) + (1-\pi)\{E[\delta|T_i = 1] - E[\delta|T_i = 0]\},
\end{aligned}$$

where π is the proportion of the population where $T = 1$ and δ is the causal effect (or treatment effect). $(E_{11} - E_{00})$ is the observed difference in the data, or the "naive estimator" of δ.

$(E_{01} - E_{00})$ can then be called the baseline bias (Morgan and Winship 2007: 46) or selection bias (Angrist and Pischke 2009: 14-15), and $(1-\pi)\{E[\delta|T_i = 1] - E[\delta|T_i = 0]\}$ the differential treatment effect bias (Morgan and Winship 2007: 45). Note that Angrist and Pischke (2009: 14-15) derive a slightly different equation, focusing only on the average treatment effect on the treated (ATT) instead of the overall average treatment effect (ATE), which allows them to ignore the differential treatment effect.

*References include those from the accompanying slides.

2 Implicit weighting in matching and regression

Regression

In regression analysis, we estimate the causal effect using

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \delta\mathbf{t} + \boldsymbol{\varepsilon}. \tag{1}$$

In the derivation of the OLS estimator, we saw that this can be seen as the solution of a set of simultaneous equations, i.e. if $\mathbf{X}$ includes $\mathbf{t}$, $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}^{OLS} = \mathbf{X}'\mathbf{y}$ could be written as:

$$\begin{aligned}
\hat{\beta}_1 n + \hat{\beta}_2 \sum x_{i2} \quad &+ \cdots + \hat{\beta}_k \sum x_{ik} \quad &= \sum y_i \\
\hat{\beta}_1 \sum x_{i2} + \hat{\beta}_2 \sum (x_{i2})^2 \quad &+ \cdots + \hat{\beta}_k \sum x_{i2}x_{ik} \quad &= \sum x_{i2}y_i \\
\hat{\beta}_1 \sum x_{i3} + \hat{\beta}_2 \sum x_{i3}x_{i2} \quad &+ \cdots + \hat{\beta}_k \sum x_{i3}x_{ik} \quad &= \sum x_{i3}y_i \\
&\vdots & \\
\hat{\beta}_1 \sum x_{ik} + \hat{\beta}_2 \sum x_{ik}x_{i2} \quad &+ \cdots + \hat{\beta}_k \sum (x_{ik})^2 \quad &= \sum x_{ik}y_i
\end{aligned}$$

In a simple regression, with only one independent variable, e.g. $\mathbf{y} = \alpha + \delta\mathbf{t} + \boldsymbol{\varepsilon}$, this leads to

$$\hat{\delta} = \frac{\sum t_i y_i - n\bar{t}\bar{y}}{\sum t_i^2 - n\bar{t}^2} = \frac{cov(y_i, t_i)}{var(t_i)}.$$

From the Frish-Waugh-Lovell theorem (Frisch and Waugh 1933) we know that regressing Y on X_1 and X_2 gives the same coefficient estimate for X_1 as regressing Y on the residuals of regressing X_1 on X_2. Therefore, in Eq. (1),

$$\hat{\delta}_R = \frac{cov(y_i, \tilde{t}_i)}{var(\tilde{t}_i)},$$

where $\tilde{t}_i$ is a residual from regression $t_i = \mathbf{x}_i'\boldsymbol{\beta}^* + \tilde{t}_i$. This is known as the regression anatomy formula (Angrist and Pischke 2009: 74). Assuming that the regression is *saturated*, i.e. we have a dummy variable for each possible value of x (Angrist 1998: 256, fn

11):

$$
\begin{aligned}
y_i &= E[y_i|t_i, \mathbf{x}_i] + \varepsilon_i && \text{saturated model} \\
&= E[y_{0i}|\mathbf{x}_i] + E[y_{1i} - y_{0i}|\mathbf{x}_i]t_i + \varepsilon_i && \text{assumption of unconfoundedness} \\
&= E[y_{0i}|\mathbf{x}_i] + \delta_x t_i + \varepsilon_i \\
cov(y_i, \tilde{t}_i) &= E[y_i\tilde{t}_i] - E[y_i]E[\tilde{t}_i] && \text{definition of covariance} \\
&= E[y_i\tilde{t}_i] && \text{in OLS, } E[\varepsilon] = 0 \\
&= E[y_i(t_i - E[t_i|\mathbf{x}_i])] && \text{definition of residuals} \\
&= E[(E[y_{0i}|\mathbf{x}_i] + \delta_x t_i + \varepsilon_i)(t_i - E[t_i|\mathbf{x}_i])] \\
&= E[(E[y_{0i}|\mathbf{x}_i] + \delta_x t_i)(t_i - E[t_i|\mathbf{x}_i])] && \text{assuming } cov(\varepsilon_i, t_i) = 0 \\
&= E[E[y_{0i}|\mathbf{x}_i](t_i - E[t_i|\mathbf{x}_i])] + E[\delta_x t_i(t_i - E[t_i|\mathbf{x}_i])] \\
&= E[\delta_x t_i(t_i - E[t_i|\mathbf{x}_i])] && \text{assuming no baseline bias} \\
&= E[\delta_x(t_i - E[t_i|\mathbf{x}_i])^2] && T \text{ is binary} \\
var(\tilde{t}_i) &= E[(\tilde{t}_i - E[\tilde{t}_i])^2] && \text{definition of variance} \\
&= E[\tilde{t}_i^2] && \text{in OLS, } E[\varepsilon] = 0 \\
&= E[(t_i - E[t_i|\mathbf{x}_i])^2] && \text{definition of residuals} \\
\hat{\delta}_R &= \frac{cov(y_i, \tilde{t}_i)}{var(\tilde{t}_i)} = \frac{E[(t_i - E[t_i|\mathbf{x}_i])^2\delta_x]}{E[(t_i - E[t_i|\mathbf{x}_i])^2]} \\
&= \frac{E[E[(t_i - E[t_i|\mathbf{x}_i])^2|\mathbf{x}_i]\delta_x]}{E[E[(t_i - E[t_i|\mathbf{x}_i])^2|\mathbf{x}_i]]}.
\end{aligned}
$$

Using the fact that the variance of T is the variance of a Bernoulli distribution and thus $p(1-p)$:

$$\hat{\delta}_R = \frac{\sum_{\mathbf{x}} \delta_x P(T_i = 1|\mathbf{X}_i = \mathbf{x})(1 - P(T_i = 1|\mathbf{X}_i = \mathbf{x}))P(\mathbf{X}_i = \mathbf{x})}{\sum_{\mathbf{x}} P(T_i = 1|\mathbf{X}_i = \mathbf{x})(1 - P(T_i = 1|\mathbf{X}_i = \mathbf{x}))P(\mathbf{X}_i = \mathbf{x})}.$$

The regression estimator $\hat{\delta}_R$ thus provides a weighted estimator of the average treatment effect δ, weighted by the variance of the treatment, conditional on $\mathbf{X}$ (Angrist 1998; Angrist and Pischke 2009; Morgan and Winship 2007). The variance of a binary variable such as T is $p(1-p)$, thus in this particular case, $\pi(\mathbf{X})(1-\pi(\mathbf{X}))$, with $\pi(\mathbf{X})$ the propensity score. This variance is highest where $\pi(\mathbf{X})$ is close to 0.5, thus for values of $\mathbf{X}$ where there are the same number of treated as untreated cases; the variance is lowest where $\pi(\mathbf{X})$ is close to 0 or 1. If $\pi(\mathbf{X}) = 0$ or $\pi(\mathbf{X}) = 1$, the case gets a zero weight and is excluded from the estimation of $\hat{\delta}$ – so where there is lack of overlap, no contribution is made to the estimation.

Matching

We can apply a similar derivation for the matching estimator, using Bayes formula:

$$P(\mathbf{X}_i = \mathbf{x}|T_i = 1) = \frac{P(T_i = 1|\mathbf{X}_i = \mathbf{x})P(\mathbf{X}_i = \mathbf{x})}{P(T_i = 1)},$$

so that we get:

$$\begin{aligned}\hat{\delta}^{M}_{ATT} &= \sum_{\mathbf{x}} \delta_x P(\mathbf{X}_i = \mathbf{x}|T_i = 1) \\ &= \sum_{\mathbf{x}} \delta_x \frac{P(T_i = 1|\mathbf{X}_i = \mathbf{x})P(\mathbf{X}_i = \mathbf{x})}{P(T_i = 1)} \\ &= \frac{\sum_{\mathbf{x}} \delta_x P(T_i = 1|\mathbf{X}_i = \mathbf{x})P(\mathbf{X}_i = \mathbf{x})}{\sum_{\mathbf{x}} P(T_i = 1|\mathbf{X}_i = \mathbf{x})P(\mathbf{X}_i = \mathbf{x})},\end{aligned}$$

so while $\hat{\delta}_R$ is weighted by the variance of the treatment, conditional on $\mathbf{X}$, $\hat{\delta}^{M}_{ATT}$ is weighted by the probability of treatment, conditional on $\mathbf{X}$, or the propensity score $\pi(\mathbf{X})$. The matching estimator for ATE, $\hat{\delta}^{M}_{ATE}$ can then be seen as the unweighted estimator.

Conclusion

A regression estimator is a weighted estimator of the average treatment effect, which in cases where the treatment effect differs strongly by different values of the control variable(s), leads to a "biased" result. Otherwise, matching is quite similar to a fully saturated regression. Often, we of course do not use a fully saturated specification in our regression, so that our controls become a lot weaker, and the difference with a matching estimator larger.

References

Achen, Christopher H. 2005. "Let's put garbage-can regressions and garbage-can probits where they belong." *Conflict Management and Peace Science* 22:327–339.

Angrist, Johua D. 1998. "Estimating the labor market impact of voluntary military service using social security data on military applicants." *Econometrica* 66(2):249–288.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press.

Frisch, Ragnar and Frederick V. Waugh. 1933. "Partial time regressions as compared with individual trends." *Econometrica* 1(4):387–401.

Gelman, Andrew and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Analytical Methods for Social Research Cambridge: Cambridge University Press.

Gerring, John. 2001. *Social science methodology: a critical framework*. Cambridge University Press.

Gerring, John. 2012. *Social science methodology: A unified framework*. Cambridge: Cambridge University Press.

Imai, Kosuke, Gary King and Elizabeth A. Stuart. 2008. "Misunderstandings between experimentalists and observationalists about causal inference." *Journal of the Royal Statistical Society A* 171:481–502.

Imbens, Guido W. 2004. "Nonparametric estimation of average treatment effects under exogeneity: A review." *Review of Economics and Statistics* 86(1):4–29.

Kennedy, Peter. 2008. *A guide to econometrics*. 6th ed. Malden, MA: Blackwell.

Lee, David S. 2008. "Randomized experiments from non-random selection in U.S. House elections." *Journal of Econometrics* 142(2).

Lee, Myoung Jae. 2005. *Micro-econometrics for policy, program, and treatment effects*. Oxford: Oxford University Press.

Morgan, Stephen L. and Christopher Winship. 2007. *Counterfactuals and causal inference. Methods and principles for social research*. New York: Cambridge University Press.

Morgan, Stephen L. and David J. Harding. 2006. "Matching estimators of causal effects: Prospects and pitfalls in theory and practice." *Sociological Methods & Research* 35(1):3–60.

Pearl, Judea. 2000. *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Rubin, Donald B. 1986. "Which ifs have causal answers (Comment on 'Statistics and causal inference' by Paul W. Holland)." *Journal of the American Statistical Association* 81:961–2.

Smith, Jeffrey A. and Petra E. Todd. 2005. "Does matching overcome Lalonde's critique of nonexperimental estimators?" *Journal of Econometrics* 125:305–353.